UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Human Development and Health

Volume 1

Genetic dissection of early-onset breast cancer and other genetic diseases

by

Rosanna Jane Upstill-Goddard

Thesis for the degree of Doctor of Philosophy

March 2015

Supervisory team: Prof. A. Collins, Prof. J. Fliege, Prof. D. Eccles, Prof. S. Ennis

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Human Genetics

Doctor of Philosophy

GENETIC DISSECTION OF EARLY-ONSET BREAST CANCER AND OTHER GENETIC DISEASES

Rosanna Jane Upstill-Goddard

Genetic variation in the genome of an individual plays a key role in susceptibility to many human diseases. Analysis of the genetic variants harboured by individuals presenting with disease phenotypes is crucial for unravelling the genetic landscape of human disease. The methods that are now available for the characterisation of genetic variants, including single nucleotide polymorphism (SNP) microarrays and next generation sequencing, make it possible to explore all genetic variants harboured within an individual with a specific disease phenotype, allowing for tailoring of treatments. This thesis focuses on the genetic dissection of early-onset breast cancer, syndromic and nonsyndromic forms of cleft lip with or without palate (CLP), and an oculopharyngeal muscular dystrophy-like (OPMD-like) phenotype through the analysis of SNP and exome data. Novel analysis approaches were used to explore the breast cancer genome-wide SNP data; a variety of machine learning algorithms were used to identify potential interactions and pathways influencing disease that cannot be uncovered using conventional analysis techniques. Such approaches are necessary because in many cases disease aetiology is likely to be complex with many genetic factors and interactions influencing disease susceptibility. Further characterisation of the genetic landscape of early-onset breast cancer, as well as the genetics of CLP and OPMD-like disease phenotypes, was possible through the use of whole exome sequencing technology. Exome sequencing identified many potentially important variants in the breast cancer samples and nonsyndromic CLP cases. Particular success was observed in the disease that were Mendelian in nature, namely syndromic CLP and the OPMD-like family; in all cases the likely causative mutation was successfully identified. Genetic studies of human disease using sequencing technologies and novel methods to analyse data are vital as personalised medicine becomes a real possibility in healthcare.

Table of Contents

List of tables

List of figures

List of published papers

Pengelly R, Upstill-Goddard R, Arias L, Martinez J, Gibson J, Knut M, Collins A, Ennis S, Collins A, Briceno I (2015). *Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing*. Clinical Genetics 88: 441-449

Smyth C., Špakulová I., Cotton-Barratt O., Rafiq S., Tapper W., Upstill-Goddard R., Hopper J. L., Makalic E., Schmidt D. F., Kapuscinski M., Fliege J., Collins A., Brodzki J., Eccles D. M., MacArthur, B. D. (2015). *Quantifying the cumulative effect of low-penetrance genetic variants on breast cancer risk*. Molecular Genetics & Genomic Medicine. doi: 10.1002/mgg3.129

Rafiq S, Khan S, Tapper W, Collins A, Upstill-Goddard R, Gerty S, Blomqvist C, Aittomäki K, Couch F. J, Liu J, Nevanlinna H, Eccles D (2014). *A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis*. PLoS ONE 9(12): e101488

Christodoulou K, Wiskin A.E, Gibson J, Tapper W, Willis C, Afzal N. A, Upstill-Goddard R, Holloway J. W, Simpson M. A, Beattie R. M, Collins A, Ennis S (2013). *Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes.* Gut 62: 977-984

Upstill-Goddard R, Eccles D, Fliege J, Collins A (2013). *Machine learning approaches for the discovery of gene–gene interactions in disease data.* Briefings in Bioinformatics 14(2): 251-260

Upstill-Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, Fliege J, Collins A (2013). *Support Vector Machine Classifier for Estrogen Receptor Positive and Negative Early-Onset Breast Cancer.* PLoS ONE 8(7): e68606

DECLARATION OF AUTHORSHIP

I, ROSANNA JANE UPSTILL-GODDARD, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

GENETIC DISSECTION OF EARLY-ONSET BREAST CANCER AND OTHER GENETIC DISEASES

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Upstill-Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, Fliege J, Collins A (2013). *Support Vector Machine Classifier for Estrogen Receptor Positive and Negative Early-Onset Breast Cancer.* PLoS ONE 8(7): e68606

   Pengelly R, Upstill-Goddard R, Arias L, Martinez J, Gibson J, Knut M, Collins A, Ennis S, Collins A, Briceno I (2015). *Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing*. Clinical Genetics 88: 441-449

Signed:.....

Date: .....

Acknowledgements

I would not have been able to complete my PhD research without the help and guidance of my supervisors and colleagues, or the support from my family and friends.

Firstly, I would like to acknowledge the Breast Cancer Campaign for providing funding, making this PhD work possible. Without their support I would not have had the opportunity to undertake this research.

My thanks go to my principle supervisor Prof. Andy Collins for his guidance and support throughout the three years of my research. His patience and advice have been invaluable.

I would also like to thank the other members of my supervisory team: Professor Joerg Fliege, Professor Diana Eccles and Professor Sarah Ennis for all their comments and advice over the past three years.

I thank all members of the Genetic Epidemiology and Genomic Informatics research group whose help has been much appreciated: Jane Gibson, Will Tapper, Sajjad Rafiq, Faisal Rezwan, Latha Kadalayil, Gaia Andreoletti, Marcin Knut, and Reuben Pengelly. It was a pleasure to work with you all.

Finally I would like to thank my parents, brother and sister for all their support. I couldn't have done any of this without you.

Definitions and Abbreviations

1000G – 1000 Genomes project database

AUC – Area under ROC curve

BCLP – Bilateral cleft lip with or without palate

CLP – Cleft lip with or without cleft palate

CPO – Cleft palate only

DAVID – Database for Annotation, Visualization and Integrated Discovery

ER – Estrogen receptor

FNR – False negative rate

FPR – False positive rate

GWAS – Genome-wide association study

HER2 – Human epidermal growth factor receptor 2

HWE – Hardy-Weinberg equilibrium

IGV – Integrative genomics viewer

LD – Linkage disequilibrium

MAF – Minor allele frequency

MBMDR – Model based multifactor dimensionality reduction

MDR – Multifactor dimensionality reduction

MDS – Multi-dimensional scaling

ML – Machine learning

NGS – Next-generation sequencing

NHLBI ESP – National Heart, Lung and Blood Institute Exome Sequencing Project Variant Server

NQP – Normalized quadratic polynomial

NSCLP – Nonsyndromic cleft lip with or without cleft palate

OPMD – Oculopharyngeal muscular dystrophy

OR – Odds ratio

POSH – Prospective study of Outcome in Sporadic versus Hereditary breast cancer

PPS – Popliteal pterygium syndrome

PR – Progesterone receptor

RBF – Radial basis function

ROC – Receiver operating characteristic

SKAT – Sequence Kernel Association Test

SKAT-O – Optimal Sequence Kernel Association Test

SNP – Single nucleotide polymorphism

SVM – Support vector machine

TNR – True negative rate

TPR – True positive rate

UCLP – Unilateral cleft lip with or without palate

VWS – Van de Woude syndrome

WTCCC – Wellcome Trust Case Control Consortium

Genetic dissection of early-onset breast cancer and other genetic diseases

Chapter 1: Introduction and Aims

The following chapter will introduce concepts relating to the genetics of human disease and the analysis methods used within this thesis to explore genetic data. The first three sections are a discussion of genetic variation in human disease (Section 1.1), methods for analysing genetic data (Section 1.2), and machine learning algorithms and their potential applications in genetic data for uncovering potentially hidden structure (Section 1.3). The next three sections are a general discussion of the genetic background of three diseases that are analysed in subsequent chapters of the thesis: breast cancer (Section 1.4), cleft lip and/or cleft palate (Section 1.5), and oculopharyngeal muscular dystrophy (Section 1.6). The final section (Section 1.7) outlines the structure of the following 5 chapters, which contain the data analysis.

1.1 Genetic Variation in Human Disease

1.1.1 Types of Genetic Variation

Genetic variants harboured within the genome influence all human traits and many human diseases. Without genetic variation evolution would not occur, nor would the wide range of traits present in the human population. Most of the heterozygosity observed in the human population results from a relatively small number of common polymorphisms (Cargill et al., 1999).

Genetic changes that occur within the DNA sequence of the genome are frequently referred to as 'variants' because the functional consequence of the DNA alteration is often unknown (Bull, 2000). Variants can occur at any nucleic acid position within the genome; both protein coding gene regions and non-coding regions between genes can be affected. The effect of each variant on the phenotype will vary, some variants will be disease-causal others will have small effects on phenotype, the majority, however, will be neutral with no phenotypic effect. Variants can be either polymorphisms or mutations. The distinction between a polymorphism and a mutation is often made based the frequency of the variant within the population. Mutations are considered to be rare in the population (typically at a frequency of less than 1%) while polymorphisms occur more frequently (typically in at least 1% of the

population) (Twyman, 2003). Mutations are considered to be DNA changes that alter the genome and therefore might be considered as more likely to cause or influence the phenotype. In contrast, polymorphisms do not usually cause a disease or trait by themselves, although they may contribute to the resultant phenotype.

Broadly categorising genetic variants results in four groups (Bull, 2000): single nucleic acid substitutions, nucleic acid insertions, deletion of nucleotides, and chromosomal abnormalities. These categories can be further refined to describe particular types of substitution, insertion, or deletion (Table 1.1). Single nucleotide variants constitute the largest class of variants.

Substitutions located within genes are classified as synonymous, nonsynonymous, stop gain, stop loss, or start loss. In all cases one nucleic acid is replaced by a different nucleic acid, the effect on the protein depends on the type and location of the substitution.

Due to the degenerate nature of the genetic code, in which several different codons represent the same amino acid, some nucleic acid substitutions will not change the amino acid sequence; such substitutions are termed synonymous, or silent, variants. There are no obvious consequences of synonymous substitutions in terms of the resultant protein, but the secondary structure of mRNA molecules can be affected (Chamary and Hurst, 2005) and synonymous variants have been implicated in human disease (Sauna and Kimchi-Sarfaty, 2011).

Nucleotide substitutions resulting in an amino acid change are termed nonsynonymous variants. Such variants change only one amino acid in the protein sequence. A single amino acid change within a protein consisting of many hundreds of amino acids is unlikely to influence a disease phenotype, particularly if the variant is common in the population (Bodmer and Bonilla, 2008).

Commonly occurring synonymous and nonsynonymous polymorphisms will, however, influence normal human variation. For example, human height is influenced by hundreds of common single nucleotide polymorphisms (SNPs) (Lango Allen et al., 2010, Wood et al., 2014) that act in combination to

Table 1.1 Types of variant detectable in the human genome

Variant category	Variant type	Explanation of variant type
Single nucleotide substitution	Synonymous	Changes one nucleotide Does not change the amino acid sequence
	Non-synonymous	Changes the amino acid that is coded for by the codon in which it occurs
	Stop gain	Introduces a premature stop codon
	Stop loss	Removes a stop codon
	Start loss	Removes the start codon
	Splicing	Alters splicing by introducing exon skipping, activation of cryptic splice sites, or intron retention
Insertion	Frameshift	Changes the reading frame of the gene, completely altering all amino acids downstream of the variant site
	Non-frameshift	Introduces amino acids into the gene sequence but does not change the reading frame
	Splicing	Alters splicing by introducing exon skipping, activation of cryptic splice sites, or intron retention
Deletion	Frameshift	Changes the reading frame of the gene, completely altering all amino acids downstream of the variant site
	Non-frameshift	Removes amino acids from the gene sequence but does not change the reading frame
	Splicing	Alters splicing by introducing exon skipping, activation of cryptic splice sites, or intron retention
Chromosomal abnormalities	Chromosome number	An increase in the number of copies of a particular chromosome
	Translocations	Exchange of parts of non-homologous chromosomes
	Insertions	Large-scale insertion of genetic sequence
	Deletions	Large-scale deletion of genetic sequence
	Inversions	Breakage and re-joining of one chromosome, inverting a portion of the chromosome
	Duplications	Genes or portions of a chromosome are duplicated

produce the phenotype. Polymorphisms are critical for natural selection and evolution, and are responsible for the diversity observable in the human population (McClellan and King, 2010).

Of the 64 possible codons, 3 code for the end of the protein and are termed stop codons. Stop codons are introduced into the protein sequence by specific nucleotide substitutions, usually termed stop-gain or nonsense variants, that can result in truncation of the protein or exon skipping (Mort et al., 2008). The introduction of premature stop codons into the DNA sequence leads to loss of function of the affected allele, most likely as a result of translation of a shortened peptide (Haraksingh and Snyder, 2013) or through nonsense-mediated decay of the resultant mRNA transcript (Lappalainen et al., 2013). Stop codons can also be removed from the amino acid sequence by nucleotide substitutions that change a stop codon into an amino acid codon, most likely resulting in elongation of the resultant protein.

The codon ATG is recognised as a start codon as well as coding for the amino acid methionine. Substitutions occurring within this codon result in a 'start-loss' variant. The result is that the efficiency of translation of the encoded protein is severely compromised (Hinnebusch, 2011, Haraksingh and Snyder, 2013).

The insertion or deletion of small numbers of nucleic acids (usually between 1 and 50 bases) is referred to as an indel (Haraksingh and Snyder, 2013). Indels occur at a much lower frequency than substitution variants (Zhang and Gerstein, 2003). Indels that are not a multiple of three base-pairs are classified as 'frameshift' variants because all codons downstream of the indel are transformed completely, changing the amino acid sequence. Indels that are multiples of three bases result in 'non-frameshift' variants in which the downstream codons are not affected. Instead, whole codons are inserted or deleted but the majority of the protein sequence remains unchanged. Analysis of genetic variants has suggested that on average, indels are 4 nucleotides in length and the majority are 3 nucleotides or fewer in length (Zhang and Gerstein, 2003).

Nucleic acid substitutions, insertions or deletions that occur close to an exon-intron boundary within a gene have the potential to affect protein splicing. Every intron within every human gene contains splice sites at both the 5' and

3' ends, both of which contain a specific dinucleotide motif – GT at the 5' end and AG at the 3' end that are essential for correct splicing and definition of the exon-intron boundary (Ward and Cooper, 2010). Mutation of these dinucleotides, or indeed variants in any other position within the splice sites, can cause errors in splicing, typically exon skipping, activation of cryptic splice sites, or intron retention (Ward and Cooper, 2010). Furthermore, mutations within the exons can alter splicing by modifying the exonic splicing enhancers (ESE) or exonic splicing silencers (ESS) (Ward and Cooper, 2010).

Genetic variation that occurs in non-genic regions of the genome can be more difficult to characterise because the relationship between variation and function can be complex. Nevertheless, functions have been identified for many non-coding regions of the genome (Birney et al., 2007, ENCODE Project Consortium, 2012) with estimates suggesting that over 80% of the genome is involved in some form of RNA- or chromatin-associated function (ENCODE Project Consortium, 2012). Furthermore, a large proportion of the genome is capable of being transcribed into RNA transcripts (Djebali et al., 2012) indicating that there are few non-functional regions of the genome. Regions that are potentially non-functional are relatively small: as many as 99% of nucleotides within the genome are actually less than 2kb away from an ENCODE functional element (ENCODE Project Consortium, 2012). Therefore, variants that occur in non-genic regions are likely to be important in disease pathology, particularly in complex diseases where multiple variants contribute to the phenotype. Results from genome-wide association studies (GWAS) (see section 1.2.2 for explanation of GWAS) support this theory: many significant associations of a SNP with a phenotype do not implicate genes but non-genic regions with regulatory functions (Schaub et al., 2012). Moreover, GWAS significant SNPs are enriched in these non-genic elements (ENCODE Project Consortium, 2012). Better characterisation of these regulatory regions is required to advance knowledge of the potential impact on phenotype of variants in non-genic regions.

1.1.2 Penetrance and Variable Expressivity

The penetrance and expressivity of a phenotype are related concepts but there is a subtle difference in what the terms describe (Cooper et al., 2013). The

penetrance of a disease variant is determined from the proportion of individuals present with both the disease variant and associated phenotype (Cooper et al., 2013). A variant or gene is completely penetrant if every individual carrying the variant presents with the phenotype. Conversely, variants that do not necessarily cause disease manifestation in all carriers are termed incompletely penetrant. While penetrance describes the proportion of individuals that present with the disease, expressivity measures the phenotypic variations. Therefore, multiple individuals can harbour the same pathogenic variant and present with the disease but the severity of the phenotype may be different due to variable expressivity.

The penetrance of a disorder can be affected by many genetic and environmental factors that act as modifiers, increasing or decreasing phenotype expression. Genetic modifiers can be variants in unrelated genes that influence whether the phenotype is expressed or not (Cooper et al., 2013). Environmental factors such as diet, smoking, and alcohol intake are major factors in many, particularly complex, diseases. It is therefore likely that such factors can act as modifiers of specific genetic variants and thereby influence phenotypic expression. Penetrance can also be age-dependent, meaning that the disease will be more likely to be expressed as the at-risk individual ages (Cooper et al., 2013).

Genes and genetic variants that are involved in disease pathogenesis can be categorised as high-penetrance, moderate-penetrance, or low-penetrance, depending on their individual influence on disease (Stratton and Rahman, 2008). Mutations classified as highly penetrant will almost always lead to expression of the corresponding phenotype. Similarly, variants in highly penetrant genes are often sufficient for phenotype presentation. At the other end of the spectrum, individual low-penetrance variants will not be sufficient to cause detectable phenotype manifestation, but may have an associated elevated risk for the phenotype (Stratton and Rahman, 2008). High penetrance variants tend to be rare in the general population due to strong selection pressures causing elimination of the variant from the gene pool (Figure 1.1) (Botstein and Risch, 2003, Gilissen et al., 2011). Conversely, low penetrance variants are weakly selected against and can have relatively high frequencies in the population because they only contribute to disease phenotypes in the presence of other risk variants (Figure 1.1).

1.1.3 Genetic Variants in Mendelian Disorders

Mendelian disorders follow principle that genes are units of inheritance that influence phenotype. Genes are inherited from one generation to the next, through both parents and can be genetically identical or different. The combination of units inherited determines the resultant phenotype.

Mendelian disorders can be either dominant (autosomal or sex-linked), in which a variant in only one allele of the gene is sufficient for presentation, or recessive (autosomal or sex-linked), in which both alleles of the gene need to be compromised for the phenotype to present. Individuals with a dominantly inherited Mendelian disease will usually present with a strong family history, demonstrating a clear parent-to-child inheritance pattern with multiple family members likely to be affected. Dominant diseases that are incompletely penetrant will not necessarily present with clear parent-to-child inheritance but the presence of affected individuals in multiple generations is suggestive of a dominant variant. In contrast, families carrying recessively inherited diseases will have few affected individuals because only individuals homozygous for the causal variant or with rare compound heterozygous variants will present with

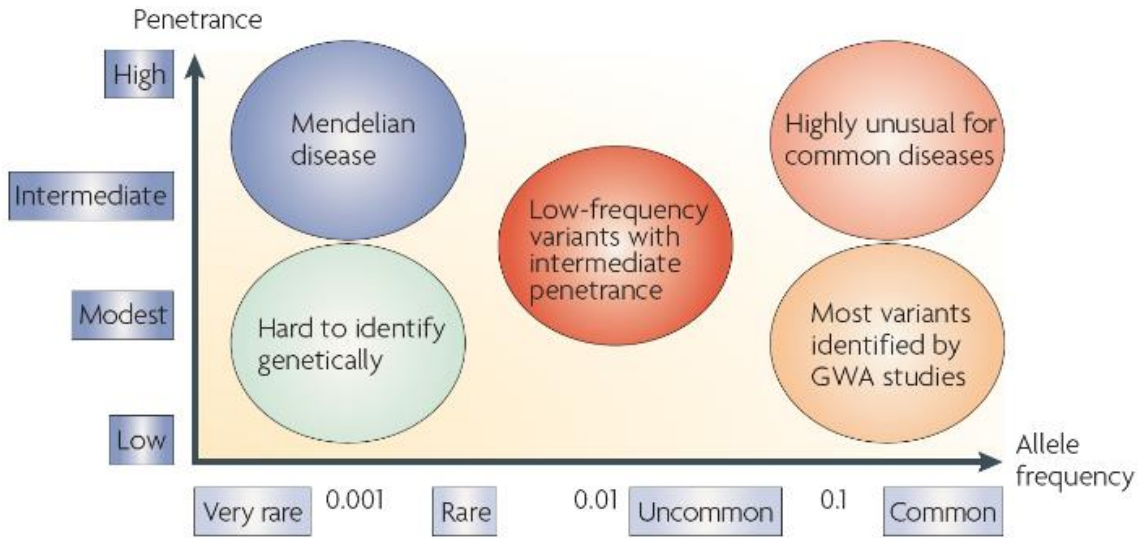


Figure 1.1. Relationship between allele frequency and penetrance. Rare variants are usually high penetrance and implicated in Mendelian diseases while common variants are typically low penetrance and implicated in complex diseases. (Figure originally published in McCarthy et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5), 356-369. Permission to reproduce this figure has been granted by Nature Publishing Group.)

the phenotype. Heterozygous individuals are carriers of the disease but will not present with the phenotype. Furthermore, the causal variant will (usually) be very rare in the general population; it is uncommon for genes in the general population to contain rare homozygous alleles or rare compound heterozygous mutations (Gilissen et al., 2011).

Causal genes have been identified for many Mendelian disorders but allelic heterogeneity is a common feature; the causal variant is rarely the same across multiple families. This is particularly true of dominantly inherited Mendelian disorders because dominant alleles are subject to negative selection pressures, causing elimination from the population over time (Botstein and Risch, 2003). Some dominant diseases present in the human population do not demonstrate this heterogeneity and are all caused by the same mutation; such diseases have a later age of onset or a low penetrance (Botstein and Risch, 2003) allowing the causal mutation to persist in the population.

Pathogenic recessive variants or compound heterozygous variants will tend to be rare in the general population (Gilissen et al., 2011), but, since they are not subjected to strong negative selection pressures, such variants can increase in frequency (Botstein and Risch, 2003) in certain isolated populations. A founder or bottleneck effect can introduce a rare pathogenic variant into the population and individuals from such populations are more likely to carry the same rare pathogenic variant than individuals from the general population. Therefore, many recessively inherited Mendelian disorders arise in particular isolated ethnic populations or in families with evidence of consanguinity, where both parents carry the same rare causal variant (Botstein and Risch, 2003) that was inherited from a recent common ancestor.

Mendelian disorders do not always arise from inherited variants shared with one or both parents; they can develop from *de novo* mutations, which arise in either one of the gametes prior to fertilisation. These types of Mendelian traits are not shared with any paternal or maternal relative but could be inherited by offspring of the affected individual.

One might expect Mendelian diseases, particularly autosomal dominant ones, to be completely penetrant but this is not necessarily the case. In fact, there is suggestion that the role of modifiers will be commonplace in Mendelian disease (Cooper et al., 2013).

1.1.4 Genetic Variants in Complex Disorders

The majority of the most commonly occurring human disease and traits are polygenic or multifactorial in origin. No single highly penetrant variant exists to explain the disease in the affected individual. Instead, the phenotype is controlled by multiple variants, usually in different genes. In many diseases many hundreds of variants could be contributing to the final phenotype.

The heritability of complex diseases is low compared to Mendelian disorders, indicative of a scenario where no single variant is sufficient to cause disease. There is, however, often evidence of common diseases clustering in families but the pattern of inheritance is not clear as it is in a Mendelian disease.

There is a genetic heterogeneity to human disease (McClellan and King, 2010) evidenced by the fact that many genes are implicated in complex diseases with different mutations giving rise to very similar phenotypes. As discussed by McClellan and King, inherited predisposition to breast cancer highlights exactly this; a number of genes with high or moderate penetrance harbour rare mutations that lead to breast cancer, however these mutations are varied and rarely the same in any two individuals.

There are two schools of thought concerning the type of genetic variants that cause common complex disease: the common disease-common variant (CD-CV) hypothesis (Reich and Lander, 2001, Pritchard and Cox, 2002) and the common disease-rare variant (CD-RV) hypothesis (Pritchard, 2001). The CD-CV hypothesis proposes that many frequently occurring variants (frequency > 5%) act additively or multiplicatively within an individual to cause presentation of disease. Each variant alone only confers a very low risk of disease and will not cause disease manifestation by itself. When considered in combination with a range of other variants with equally low associated risk however, the disease does manifest. In contrast, CD-RV implicates multiple variants that are rare in the human population (frequency < 5%) as the risk variants for common diseases. As in CD-CV, each variant will be associated with only a small risk of disease but in combination with multiple other variants, all with marginal associated disease risk, the disease will develop.

It is important to note that the pathogenesis of many complex diseases will rely on the presence of environmental factors as well as genetic factors.

Exposure to certain environmental factors will increase an individual's risk of developing a particular disease. The effects of environmental factors on disease manifestation, through their interactions with risk variants, are relevant for understanding and treating complex diseases (Hunter, 2005).

The majority of variants present in the human population are rare (Sachidanandam et al., 2001) with suggestions that ≥ 85% of all nonsynonymous, stop-gain and splice variants exist at frequencies of < 0.5% (Abecasis et al., 2012). Evidence does exist, however, to suggest that both common variants and rare variants contribute to complex human diseases (Schork et al., 2009, Manolio et al., 2009). Analysis of single nucleotide substitutions associated with complex diseases found that these variants tend to occur in weakly conserved genomic regions (Thomas and Kejariwal, 2004), as is observed for variants in 'healthy' individuals. Variants occurring in locations with weak conservation are generally expected to be only weakly deleterious or neutral, supporting the view that complex disease variants will have low associated risk when considered in isolation.

1.1.5 Genetic Variants in Cancers

For many Mendelian and non-cancer complex diseases it is variation in the germline DNA that predisposes the individual to the disease. The genetic basis of cancer however, includes mutations occurring within the genome of individual tumour cells (Stratton et al., 2009). Pathogenic mutations in the germline DNA can predispose individuals to a specific type of cancer (Balmain et al., 2003). Further mutations can then arise in specific cells of the tumour, influencing cancer progression (Balmain et al., 2003). In some cases there is no known mutation influencing predisposition, instead pathogenic somatic mutations arise through errors in DNA replication during the cell cycle or from mutagens that mutate the DNA sequence (Balmain et al., 2003).

The main model of cancer development and progression suggests that cancer cells need to acquire 'capabilities' that allow them to form tumours (Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011). The capabilities include; self-sufficiency in growth signals, insensitivity to anti-growth signals, evading apoptosis, sustained angiogenesis, limitless replicative potential, and tissue invasion and metastasis. Hanahan and Weinberg postulate that all these

capabilities need to be acquired by cells for tumourigenesis to occur but the order in which these characteristics are acquired is trivial and is likely to vary from cell to cell (Figure 1.2). Somatic mutations are acquired by cells throughout the lifetime of a patient (Stratton et al., 2009) and often lead to the development of certain characteristics that eventually cause the cell to become cancerous (Figure 1.3). Somatic variants can be classified as 'driver' or 'passenger' variants depending on their influence on cancer progression. Driver variants provide the cell with a growth advantage and are positively selected for (Stratton et al., 2009), allowing it to persist in the cell population. Passenger mutations on the other hand do not confer any selective advantage on the cell and are not positively selected for (Stratton et al., 2009), instead they arise during normal cell division and are passed on to daughter cells. The majority of variants detected within a cancer cell will actually be passengers rather than driver mutations (Burgess, 2013).

Cancers that result from somatic mutations are, on the whole, diseases of later life since the timespan required for acquisition of all the hallmarks of cancer may be decades. Cancer in the young, generally before individuals reach mid-life, is rare and will be predominantly due to germline mutations that predispose to cancer, simply because individuals have not had enough time to acquire the necessary somatic lesions.

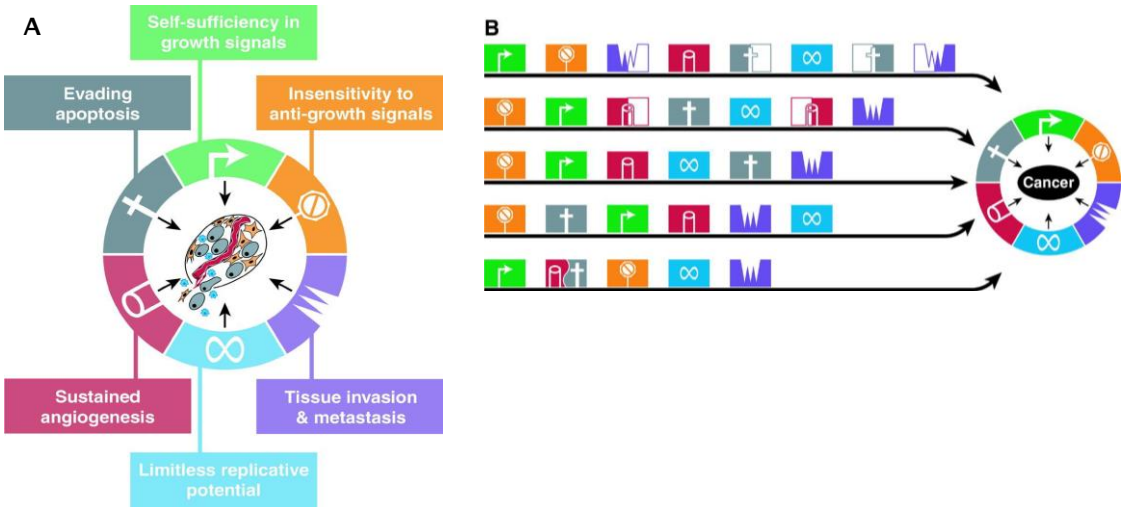


Figure 1.2. The hallmarks of cancer. (A) The six capabilities that need to be acquired by every cell before it can become cancerous. (B) The hallmarks of cancer can be acquired in any order and the order will most likely vary from cancer to cancer. (Figure originally published in Hanahan and Weinberg, (2000). The Hallmarks of Cancer. *Cell*, 100(1), 57-70. Permission to reproduce this figure has been granted by Elsevier.)

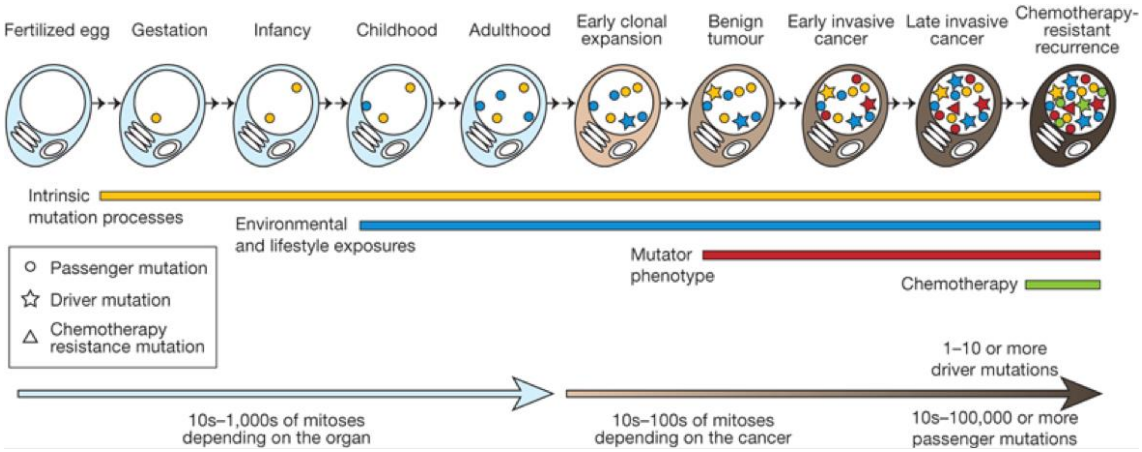


Figure 1.3. Acquisition of somatic mutations throughout life. Somatic passenger and driver mutations are acquired throughout the life of an individual, causing some cells to become cancerous. (Figure originally published in Stratton et al. (2009). The cancer genome. *Nature*, 458 (7239), 719-724. Permission to reproduce this figure has been granted by Nature Publishing Group.)

Cancers are generally separated into hereditary, familial or sporadic cases (Berliner and Fay, 2007). Hereditary cancers often have a strong family history and demonstrate autosomal dominant inheritance of a single mutation in a highly penetrant gene. If the cancer type is linked to gender, such as breast or ovarian cancer, then the inheritance pattern will not necessarily appear to be dominant if transmission is via male relatives and there is a lack of female relatives. Affected individuals tend to develop tumours at an earlier age than is observed in the general population. Such cancer syndromes tend to be rare in the population (Nagy et al., 2004). Familial cancers are similar to hereditary cancers in that a certain cancer type (or similar cancers) are observed in a family more often than would be expected by chance but the inheritance pattern is not well defined. Most cases of cancer are sporadic – they develop in individuals with no strong family history of cancer. Sporadic cancers can arise from inherited germline risk variants, *de novo* mutations, somatically acquired mutations, environmental factors, or a combination of these. Multiple variants with moderate- to low-penetrance are likely to contribute to sporadic cancer cases (Bodmer and Tomlinson, 2010), resembling complex diseases. To enrich for disease-associated rare variation, cancer cases with a relatively early age of disease onset should be selected for genetic analysis (Bodmer and Tomlinson, 2010).

1.2 Genetic Dissection of Human Disease

There are multiple methods available for characterising the genetic landscape of Mendelian and complex human diseases. Linkage studies were widely used in early genetic studies but, since sequencing of the human genome was completed in the early 2000s (Lander et al., 2001), association studies and next-generation sequencing studies have become more popular.

1.2.1 Linkage Studies

Linkage studies aim to explain the inheritance pattern of disease within families by identifying chromosomal regions that are co-inherited with the phenotype (Lander and Schork, 1994). By examining whether known genetic markers (DNA sequences that are polymorphic) segregate with the phenotype of interest, one can conclude that the disease locus is located close to the genetic marker and thus map a gene to a specific chromosome.

The genetic material contained within each human chromosome pair is transmitted from generation to generation, with one chromosome inherited from the father and the other inherited from the mother. The arrangement of genes on the chromosomes of a homologous pair is the same but the DNA sequence is not necessarily identical, giving rise to polymorphic genetic markers. Recombination between homologous chromosomes, in which regions of DNA are exchanged between the maternal and paternal chromosomes, is a common feature during meiosis. Using known genetic markers that are likely to be polymorphic in the maternal and paternal chromosomes (termed maternal and paternal alleles), one can test to see if recombination has occurred between them. If recombination has occurred then the alleles that were located on the maternal chromosome, for example, may be separated so that one allele remains on the maternal chromosome while the other is now located on the paternal chromosome (Figure 1.4A) For genes located far apart from one another on a chromosome, it will be highly likely that recombination will have occurred between them over generations. Conversely, genes located in close proximity on a chromosome will be very unlikely to be separated by recombination and specific pairs of alleles will be commonly co-inherited.

Disease genes can be identified by measuring recombination against genetic markers from across the genome (Nowak et al., 2012); frequent recombination indicates the marker and gene are located far apart while little or no recombination indicates that they are in close proximity. The linkage distance between markers can be calculated by establishing how many recombinant gametes there are in total because the percentage of recombinants relates directly to the genetic distance in centiMorgans (cM): 1% recombinants = 1cM. Genetic distance is not directly equivalent to physical distance, although 1cM is generally considered to be approximately equivalent to 1Mb. Therefore, the chromosomal region harbouring the disease gene can be mapped by identifying genetic markers that segregate with the phenotype and are likely to be flanking the disease gene (Figure 1.4B) (Nowak et al., 2012). Furthermore, if the genetic distance between two loci is found to be less than would be expected by chance, one can conclude that they are likely to be in close proximity on the same chromosome.

The lod score (Morton, 1955) calculates linkage distances and is used to determine whether the marker locus and proposed disease locus are indeed linked. The lod score is calculated by determining the frequency of recombination between two loci and then calculating the logarithm of the odds ratio (Morton, 1955). A lod score ≥ 3 indicates significant evidence of linkage between the marker and a nearby disease locus.

For Mendelian disorders with a clear inheritance pattern, linkage studies were the method of choice (Lander and Schork, 1994) to identify disease genes. However, while linkage mapping can map a gene to a region of a specific chromosome, the resolution of the method is fairly low; typically the maximum resolution achieved is 1Mb in humans (Dear, 2001). This occurs because a linkage study will usually only include meioses from a small number of generations (between 1 and 3). Complex diseases have a much more complicated inheritance pattern, involving many genetic variants that are likely to have moderate- to low-penetrance, for which linkage analysis would not be successful (Tabor et al., 2002).

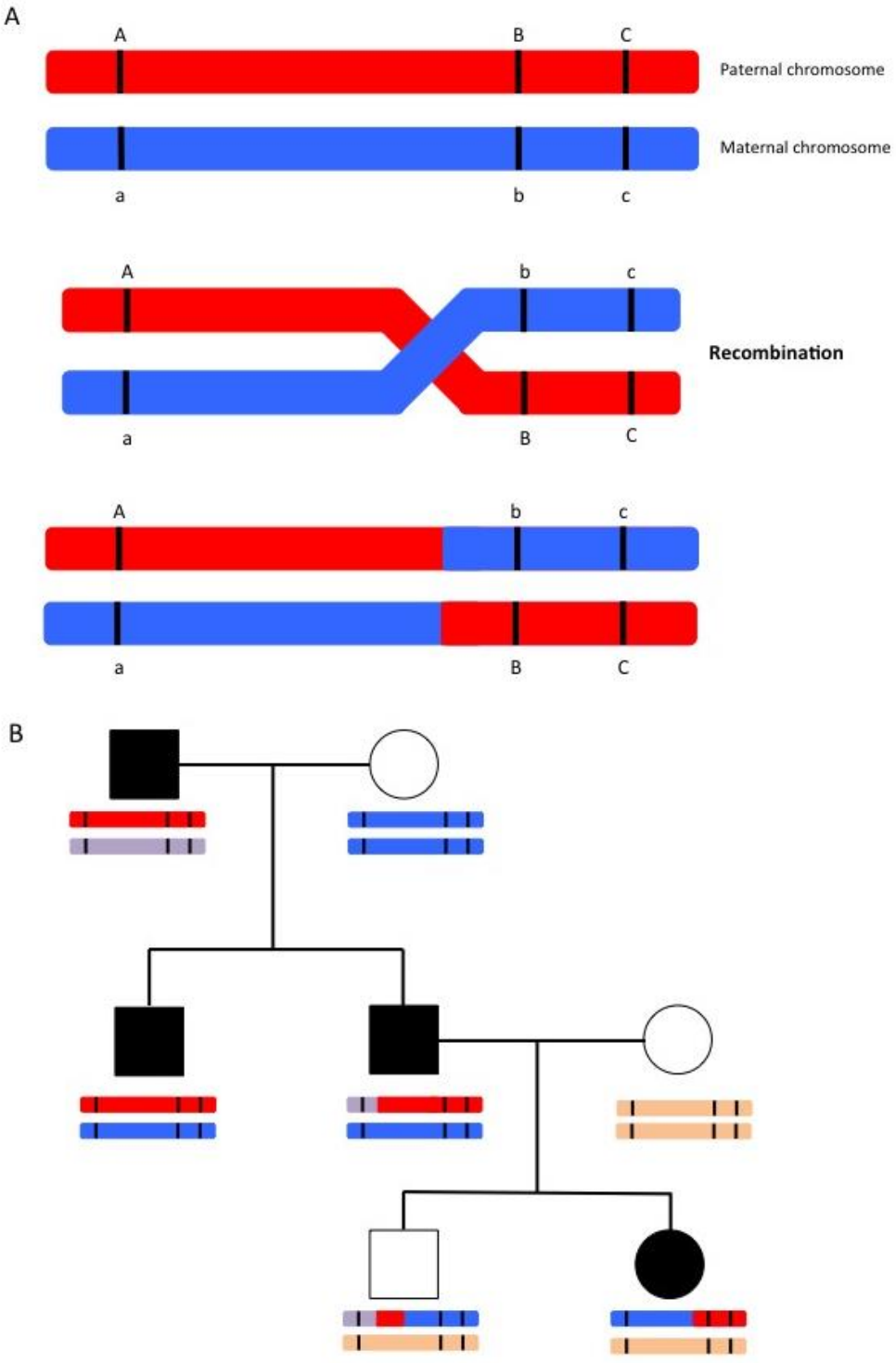


Figure 1.4. Recombination and linkage analysis to identify disease locus. (A) The process of recombination during meiosis separates alleles that were on the same chromosome in the previous generation. (B) Recombination in a single family can be used to identify a locus that is co-inherited with the phenotype. In this example the right-hand end of the red chromosome co-segregates with the disease.

1.2.2 Genome-wide Association Studies

A motivation behind genome-wide association studies (GWAS) was the common disease-common variant hypothesis. GWAS aim to identify genomic regions with an association with a particular disease using common SNP variants (frequency > 5%). Unlike linkage studies, which test for an association between a chromosomal region and a disease in a small number of individuals, GWAS search for co-occurrence of a disease and a SNP marker at the population level (Baron, 2001). By analysing the differences in frequency of common SNP variants between disease cases and controls (Kruglyak, 1999) SNPs with a disease association can be identified.

GWAS use the basic principle of linkage disequilibrium (LD) to test the entire genome for disease association signals without the need to genotype all possible SNP locations (Bush and Moore, 2012). The principle is similar to that of linkage analysis in that genetic markers that are in close proximity will not be separated by recombination and will be inherited together. Markers that are very close together (usually less than 50kb apart) are highly unlikely to be separated by recombination over many generations, and are therefore usually co-inherited (Christensen and Murray, 2007). SNPs for which this is true are in linkage disequilibrium. In a chromosomal region multiple SNPs will exist that are inherited together, so it is possible to select only one of these SNPs that can then act as a proxy for all the SNPs that it is LD with. This idea is utilised in GWAS to select SNPs that represent all non-genotyped SNPs; tag SNPs are used as markers for other variants so an entire chromosome or chromosomal region can be fully screened for disease-associated variants without the need to genotype all possible variants (Bush and Moore, 2012). The vast majority of common human SNPs are in LD with at least one other nearby SNP, so > 90% of common SNPs across the genome can be tested for disease association by using a subset of ~500,000 SNPs (The International HapMap, 2005).

In a GWAS, common SNPs are tested for a statistical association with a given trait; a higher frequency of the SNP in cases is often taken as evidence of an association with disease risk (Hirschhorn et al., 2002). A positive disease-association signal will often not be due to the tag SNP but variation located somewhere within the tagged region (often in regulatory or intergenic regions) that is responsible for the association (Hirschhorn et al., 2002, Frazer et al.,

2009). The majority of SNPs included in a GWAS are not located within any gene and are often thousands of nucleotides away from any gene. The SNP may regulate or influence the expression of a gene that it is located close to, or it may act on a distant gene, or it may have no influence whatsoever. Therefore, identifying the actual variants with disease association and inferring the biological relevance of these variants can be complex (Frazer et al., 2009). Unlike linkage analysis, the regions implicated by GWAS in the association are fairly small: 10 to 100kb opposed to regions of up to 10Mb that are identified in linkage analysis (Altshuler et al., 2008). These small regions can be sequenced or subjected to fine-mapping in order to detect the causal variant (Altshuler et al., 2008).

While GWAS works well for identifying common SNPs with small effects, it fails to identify rare variants which are often not in linkage disequilibrium with common variants, along with any potential gene-gene or gene-environment interactions (Frazer et al., 2009). Establishing links between variants and disease is further complicated by the irreproducibility of the majority of positive associations that have already been identified (Hirschhorn et al., 2002). This replication problem is likely to be partly a result of studies being underpowered; for variants with very small effect sizes many thousands of cases and controls will be necessary to detect the association signal. Often it is necessary to conduct a meta-analysis, in which multiple datasets are combined and analysed together, to achieve suitable power for detecting the variants of low effect (Zeggini and Ioannidis, 2009).

1.2.3 DNA Sequencing

Sequencing of DNA is necessary to characterise all genetic variants. For many years Sanger sequencing (Sanger et al., 1977) was the state-of-the-art method for DNA sequencing (Metzker, 2010), and is still considered to be the 'gold standard'. The Sanger method uses single stranded DNA as a template from which the DNA sequence is established by incorporating dideoxynucleotidetriphosphates (ddNTPs) that cause chain termination. Each of the four nucleotides contained within the genetic code has a corresponding ddNTP. Sequencing is separated into four separate reactions, each of which contains multiple copies of the DNA fragment to be sequenced, all four

nucleotides and one of the ddNTPs. DNA sequencing of the fragments incorporates the nucleotides and extends until one of the ddNTPs is incorporated. Sequence termination will occur at a different point for each fragment allowing for size separation of the fragments by gel electrophoresis. The nucleotide sequence can then be inferred from the results.

Since the completion of the human genome sequencing project next generation sequencing (NGS) methods have become widely used in genetic studies to characterise variants with potential importance in disease. Whole-genome and whole-exome sequencing methods are designed to sequence the entire genome or exome of an individual through the use of massively parallel DNA sequencing (Majewski et al., 2011). The basic chemistry of NGS resembles that of Sanger sequencing but hundreds of megabases to gigabases of sequence are produced through repeated cycles of polymerase-mediated nucleotide extensions (Majewski et al., 2011). The nucleotides are fluorescently or radioactively labelled in NGS so that the DNA sequence can be established by detecting the labelled nucleotides. While the sequencing method employed by NGS technologies is very similar to Sanger sequencing, the quality of reads produced during NGS tends to be inferior with more sequencing errors (Shendure and Ji, 2008, Altmann et al., 2012). Therefore, there is often a requirement to confirm candidate disease variants identified from NGS studies using Sanger sequencing.

Over recent years NGS methods have largely replaced GWAS in the search for causal variation. The advantage of the NGS approach is that all variation within an individual genome can be characterised and interrogated to identify causal variation. Whole-genome sequencing is preferable because it identifies all genetic variation, however, the quantity of data produced is vast and because current understanding of non-coding regions of the genome is poor, characterising the functional role of variation in these regions will be difficult. In contrast, the exome consists of only the sequences of exons so produces a much smaller quantity of data (~1% of the entire genome) yet still captures ~85% of coding variation per individual (Majewski et al., 2011). While this greatly expands the potential for genetic studies, the interpretation of sequencing results is challenging; in most cases it is not immediately obvious which variants are pathogenic (Gilissen et al., 2012). Sequencing approaches are likely to be more successful at identifying causal variants in individuals

presenting with monogenic Mendelian disorders (Gilissen et al., 2010, Ng et al., 2010). Establishing the effect of substitution variants on protein function is difficult (Mooney, 2005, Karchin, 2009), demonstrating the importance of correctly identifying functional variation and disregarding irrelevant variants in genetic studies (Karchin, 2009) particularly when considering complex disorders. In order to maximise the potential for identifying pathogenic variants in complex diseases, it may be more appropriate to sequence familial patients or patients with an extreme phenotype (Bamshad et al., 2011, Majewski et al., 2011).

Exome sequencing uses a targeted capture approach with the aim of sequencing only the exons. Inevitably, targeted capture is not perfect meaning some exons will not be covered or will only be sequenced at a very low depth, leading to the possibility of important variants being missed (Majewski et al., 2011). Additionally, exome sequencing is not particularly successful at identifying large structural variants such as insertions, deletions, block substitutions, inversions, and copy number variants (Majewski et al., 2011). Structural variants may account for as much as 20% of variation per individual (Frazer et al., 2009) so much variation is missed using this approach. Although most variation is contained in the exons some important variation is contained within the non-coding regions of the genome (Manolio et al., 2009). This variation is not identified through exome sequencing but will be identified in whole genome sequencing. It is currently extremely difficult to assess the consequence of variants in non-coding DNA on gene expression (Mooney, 2005) but as sequencing technologies move towards whole genome sequencing there will be a requirement for greater characterisation of non-exonic variants.

1.2.4 Sequencing of the Human Genome

The Human Genome Project (HGP) began in 1990 as a worldwide collaborative effort to sequence the euchromatic part of the human genome (Lander et al., 2001). The sequencing of the human genome has revolutionised biomedical research and increased overall understanding of the genome structure (Lander et al., 2001).

Genetic dissection of early-onset breast cancer and other genetic diseases

Sequencing of the entire human genome was incredibly important in the identification of human genes. Without this sequence many of the genetic discoveries made over the past decade would not have been possible. The human reference sequence is a necessary component in any sequencing study; as discussed, the method of NGS produces millions of short fragments of DNA that need to be mapped to their original location in the genome. The human genome is thus used as a reference sequence to which DNA fragments are mapped. Potential disease variants can thus be identified: any nucleotides in the DNA sequence of these fragments that do not match the reference sequence can be identified.

1.2.5 Heritability and Missing Heritability

Complex traits develop from a combination of multiple genetic and environmental factors. The genetic component of a trait can be measured using an estimate of heritability – a measure of how much of the variance of a particular phenotype is due to genetic factors (Visscher et al., 2008).

All genes and loci identified as associated with the majority of complex human traits and diseases only account for a small proportion of the estimated heritability (Manolio et al., 2009). Therefore, there is a large 'missing heritability' component to most complex diseases (Manolio et al., 2009). GWAS have not resolved as much of the missing heritability as was perhaps expected. As a consequence, there is perception of only partial success for GWAS in genetic variant detection despite the many new variants that have been identified and improved understanding of mechanisms and pathways that underlie many diseases.

There are several explanations as to why GWAS have been perceived as perhaps not as successful as was originally hoped. One explanation for this is that disease-associated variation is overlooked in GWA studies because it does not reach genome-wide significance or causal SNPs are not in complete LD with common SNPs and so are not detected (Yang et al., 2010, Lee et al., 2011). Small samples sizes in GWAS could produce this problem; few variants will have effect sizes that reach the significance threshold in a small sample size. To maximise the chances of detecting disease-associated SNPs in a population, sample sizes in the tens of thousands will probably be necessary (Park et al.,

2010). Further potential sources of missing heritability include the presence of epistatic interactions (Clarke and Cooper, 2010), rare variants with low-penetrance (Clarke and Cooper, 2010, Bodmer and Tomlinson, 2010), locus heterogeneity (Clarke and Cooper, 2010), *de novo* CNVs (Clarke and Cooper, 2010), and overestimates of heritability.

Missing heritability was in part responsible for the recent drive towards NGS studies and the identification of rare variants in diseased individuals. While it is likely that some common low-penetrance variants will be involved in the polygenic model of complex disease, it is unlikely that such variants will be responsible for the missing heritability. Instead, it is highly likely that much more of the heritability will be explained by low-frequency variants with moderate penetrance and greater effects on risk (McCarthy et al., 2008, Li et al., 2010).

1.2.6 Functional Characterisation of Genetic Variants

Exome sequencing studies identify in excess of 20,000 genetic variants per individual, this number will be greatly increased when considering the entire genome. Assessing the functional relevance of variants identified through NGS is crucial for understanding and targeting the mechanisms underlying genetic disease.

As a general rule, analysis of sequencing results focuses on variants that are considered more likely to be damaging. However, determining which variants are likely to be damaging is not an easy task. There has been suggestion that many nonsynonymous variants SNVs affect protein structure and are thus likely to affect phenotype (Sunyaev et al., 2000), implying that all nonsynonymous SNVs identified in an individual are potentially disease-related. More recent analysis of nonsynonymous variants has, however, suggested that the functional consequences of the majority of common nonsynonymous SNVs (with MAF > 5%) are essentially neutral, suggesting that focus should be on rare variation (Boyko et al., 2008).

Comprehensive *in vivo* and *in vitro* analysis is required to confirm that identified variants disrupt protein function. However, due to the large numbers of identified variants, even after filtering for those that are rare in the human

population, prioritisation of candidate causal variants is often required. There are a number of scoring algorithms available that aim to predict how damaging a particular variant may be or use conservation scores to predict the consequence of a variant. Such programs should, however, be used with caution; it has been suggested that *in silico* predictions are not as effective for complex diseases (Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011)(Kumar et al., 2011) and often there is limited agreement between programs (Liu et al., 2011). As the volume of genetic data from next generation sequencing increases it is becoming increasingly important to have programs capable of accurately identifying relevant variation and excluding non-functional variation.

1.3 Machine Learning Applications and Algorithms

Machine learning (ML) is a field of computer science that uses algorithms to 'learn' from training data to make predictions and solve problems in other data based on learned patterns and rules. ML algorithms are useful for genetic data analysis for the identification patterns within the data that standard genetic analysis does not detect. There are a range of machine learning algorithms available for use with genetic data, some were specifically designed with biological applications in mind while others were designed for other purposes but are applicable to genetic datasets. Some of these methods are designed for or are capable of identifying gene-gene interactions while others are aimed more at sample classification.

1.3.1 Classification of Disease Samples

Many of the commonly used ML algorithms are classification models that learn patterns from training data and use these patterns to classify unseen test data. As discussed, human diseases are, on the whole, highly heterogeneous. In the case of cancer particularly, each cancer type is actually composed of multiple subtypes that can have wildly diverse disease mechanisms, characteristics and

prognoses. Therefore, over recent years there has been increased interest in using ML approaches to classify cancers.

ML approaches can be applied to many classification problems including disease detection and diagnosis, subtype classification, and risk/prognosis prediction (Cruz and Wishart, 2006). There are many classification algorithms available but the support vector machine (SVM) is widely regarded as state-of-the-art (Chen et al., 2008). The SVM is a supervised learning classification technique (Cortes and Vapnik, 1995) that aims to produce a classification model capable of distinguishing between samples from two classes. Each sample has associated data for a number of 'features'; the class separation is based on these features. A model is developed using a training set in which the class of each sample is known. Each data point is mapped in high dimensional feature space and a separator – termed a hyperplane – used to divide the samples from the two classes (Figure 1.5). Maximising the distance between the hyperplane and each data point, where possible, produces the optimal model. In its simplest form a linear hyperplane is identified. Such a classifier works well for simple datasets where an almost linear separation exists between the sample classes.

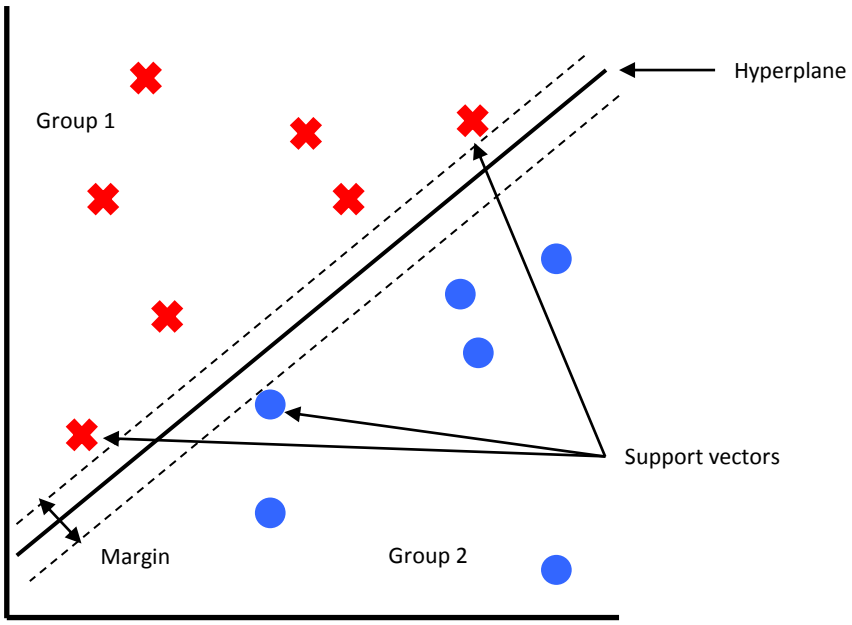


Figure 1.5. SVM classifier to separate two groups of samples. The hyperplane is placed to separate the two groups of samples with the maximum margin between the closest samples (support vectors) and the hyperplane.

For more complex datasets it is often necessary to use a kernel function to map the data points into a higher- dimensional feature space in which a more appropriate separation may be possible. In situations where it is not possible to construct a hyperplane that classifies every data point correctly, a 'soft margin hyperplane' (Burges, 1998), which is plotted based on the smallest sum of all classification errors, is used instead. From the resulting model new samples can be classified into one of the two groups dependent upon their location in feature space with respect to the hyperplane.

A simple linear kernel will produce the optimal model in many cases (Ben-Hur and Weston, 2010, Scheubert et al., 2012, Waddell et al., 2005), regardless of the type of data being considered. All data sets contain hidden complexities, meaning a suitable choice of kernel cannot be determined prior to model development. Therefore, multiple kernel models should be tested to identify the most appropriate model for the data in question (Ben-Hur and Weston, 2010).

An advantage of the SVM method is the relative interpretability of the output (Chen et al., 2008), something that cannot be said of all ML algorithms. Furthermore, beyond simply classifying samples, SVMs can report which features contribute most to the separation between classes through the assignation of weight values. Features with the largest absolute weight values are those with the most discriminatory power, while those with weights close to 0 add little to the model.

One limitation of the SVM approach is the poor performance that is often associated with missing data (Chen et al., 2008). This is a major issue, and is particularly problematic in genetic studies involving SNP data because SNP genotyping is very often incomplete. Pre-processing of SNP data is normally necessary to either impute any missing data or remove features that do not have a complete set of feature values (Wall and Elser). Sparse SVMs have been specifically designed to include a pre-processing feature selection step to counteract any model instabilities (Bi et al., 2003).

A number of genetic studies have implemented the SVM approach in classification problems using SNP data (Listgarten et al., 2004, Chen et al., 2008, Waddell et al., 2005) and achieved high classification accuracy. These studies all implicate a large number of features in the final model, indicating

the importance of multiple genes and loci in disease (Chen et al., 2008, Listgarten et al., 2004, Waddell et al., 2005), with multiple SNP variants in different genomic regions better at distinguishing between sample groups than single SNPs (Listgarten et al., 2004).

1.3.2 Gene-gene Interaction Detection

One proposed source of the missing heritability component of many diseases is gene-gene or gene-environment interactions (Clarke and Cooper, 2010, Moore and Williams, 2005). Epsitatic interactions may account for a large proportion of the missing heritability (Zuk et al., 2012) so analysis of potential epistasis within disease is an important area of research.

The term epistasis usually refers to several distinct phenomena: biological epistasis, statistical epistasis and compositional epistasis (Phillips, 2008). Biological epistasis is used to represent physical interactions between biomolecules, so that the impact of a gene on the phenotype depends on additional genes. Such interactions occur within specific genetic pathways or networks and ultimately have an impact on the phenotype (Moore and Williams, 2005). Statistical epistasis is based on differences in biological epistasis on a population level (Moore and Williams, 2005, Moore and Williams, 2009) while compositional epistasis represents the traditional definition of epistasis: the phenotypic effect of an allele is masked by an allele at a separate locus (Phillips, 2008). The relationship between the different types of epistasis is complicated, with statistical epistasis not necessarily suggestive of a biological interaction (Cordell, 2002). Therefore detecting interactions from genetic data that have biological relevance is a complex problem.

SNPs are commonly used as markers for genes in interaction detection in genetic studies. As the number of SNPs considered increases the number of potential interactions also increases but at an exponential rate (Moore and Ritchie, 2004), known as the ‘curse of dimensionality’. This adds another level of complexity to the problem.

There are many difficulties associated with the detection of epistasis in genome-wide SNP data related to both the data to be analysed and the capabilities of the ML methods being used. Firstly, there is the complexity of

the disease data, often including allelic/locus heterogeneity, phenocopies, trait heterogeneity, phenotypic variability (Thornton-Wells et al., 2004) and incomplete penetrance (Cordell, 2002). Some ML algorithms are capable of dealing with such data complexities but in the majority of situations the ability of the algorithms to detect gene-gene interactions will be impaired.

The computational burden associated with the search for gene-gene interactions is potentially huge (Wang et al., 2011), particularly when searching for interactions between more than two SNPs (Cordell, 2009). Furthermore, the power to detect interactions is often reduced when high order interactions are considered. Aside from the computational burden the outputs may present serious challenges for biological interpretation.

Exhaustive search of pairwise interactions in genome-wide SNP data may be computationally feasible but extensive validation of candidate interactions in independent samples is essential to confirm or refute discoveries. More sophisticated approaches capable of modelling higher order interactions need to be developed but may require the use of expert knowledge of biological and biochemical pathways to choose SNPs likely to be associated with a particular disease (Moore et al., 2010). It may also be necessary, and more powerful, to employ a two-stage model, in which filter algorithms select a subset of SNPs and a ML method exhaustively searches for interactions (Thornton-Wells et al., 2004, Marchini et al., 2005, Cordell, 2009). This approach may be less time consuming and produce easier-to-interpret models (Marchini et al., 2005). However, some argue it is likely that SNPs with strong epistasis but weak main effects will be filtered out (Wan et al., 2010a), so these methods will not necessarily find the optimal solution. Moreover, it is often the case that individual SNPs are assessed for disease association based on an importance score that does not take into account interactions with other SNPs. A SNP with a high importance score but no involvement in SNP-SNP interactions is clearly not useful in this context.

There are numerous algorithms available for gene-gene interaction detection. The rationale and methodology behind each algorithm is different but they all have the capacity to identify potential interactions. Several algorithms have been developed specifically for application to genetic data to directly identify gene-gene interactions. These include: multifactor dimensionality reduction

(MDR) (Ritchie et al., 2001), which considers the genotypes of a group of SNPs and identifies any genotype combinations with distribution differences in cases compared to controls; SNPHarvester (Yang et al., 2009), which aims to identify small subsets of SNPs with a significant association with the phenotype; and SNPRuler (Wan et al., 2010b), which aims to recognise predictive rules that describe the relationship between features of the data and phenotype. Many other ML algorithms that have been successfully applied in other scientific fields can also be applied to gene-gene interaction detection studies: neural networks (NNs) and tree-based methods are two examples. NNs resemble directed graphs in their structure, composed of nodes that represent features of the data (such as SNPs) and arcs that link nodes and represent associations (or interactions) between the features (Motsinger-Reif et al., 2008). The potentially most important interactions can be identified from weights that are assigned to each arc during model training and optimisation (Lucek and Ott, 1997). Tree-based methods, such as random forests (RF), are predictive models comprising multiple classification or regression trees (Bureau et al., 2005) generated from random vectors (Breiman, 2001). Each tree of the forest takes a slightly different set of features as input and produces a decision tree. Potential interactions can be identified from routes through the decision tree that produce a good classification of the data.

1.4 Breast Cancer

Breast cancer is one of the most commonly occurring cancers in Western society, with the majority of cases being post-menopausal women. In the UK ~50,000 new cases are diagnosed every year, yielding an incidence rate of 155 cases per 100,000 women (Cancer Research UK). In 2012 the 5-year prevalence of breast cancer, i.e. those women still alive 5 years after diagnosis, was estimated at ~200,000 (755 women per 100,000) (Bray et al., 2013).

Breast cancer in younger women, usually occurring before 40 years of age, is rare: ~2,000 cases were diagnosed in the UK in 2011 (Cancer Research UK). Yet the survival rates in this subgroup of women are poor, despite improvements in detection and treatment of breast cancer in recent years leading to much increased survival rates overall.

1.4.1 Early-Onset Breast Cancer

Approximately 5-7% of all breast cancer cases occur in women under the age of 40. Breast tumours that develop in young women are generally more aggressive and associated with poor prognosis and higher mortality rates (Walker et al., 1996, Kollias et al., 1997, Gonzalez-Angulo et al., 2005, Bharat et al., 2009, Kheirelseid et al., 2011). Early-onset cases are more likely to have an underlying genetic basis, and a number of gene expression patterns and pathways have been identified as unique to early-onset breast carcinomas, including gene sets related to immune function and multiple oncogenic signalling pathways (Anders et al., 2008).

The 'hallmarks of cancer' hypothesis (Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011) does not apply to early-onset cancer because patients have not had the time to acquire the necessary hallmarks to cause oncogenesis. Genetic predisposition to breast cancer is likely to play a much greater role in early-onset disease and explains why patients can develop tumours at such a young age; germline mutations (both inherited and *de novo*) with an associated breast cancer risk will be harboured by such individuals. A small proportion of early-onset breast cancer cases will arise from variation in a highly penetrant gene, the vast majority of cases are considered to be sporadic and likely to follow a polygenic model of inheritance (Venkitaraman, 2002).

1.4.2 Estrogen and Progesterone in Breast Cancer

A number of female hormones have been linked to increased risk of breast cancer. The majority of breast cancers are estrogen receptor (ER) positive and grow in response to estrogen. In general these types of tumour are less aggressive and more treatable, using therapies aimed at targeting and blocking the hormone receptors on the cancer cell (Fabian and Kimler, 2005, Deroo and Korach, 2006).

There are two proposed mechanisms of estrogen influence on breast cancer (Deroo and Korach, 2006). The first hypothesises that the binding of estrogen to the receptor stimulates mammary cell proliferation. The subsequent increase in DNA replication increases the potential for acquisition of mutations

due to replication errors (Deroo and Korach, 2006). The second hypothesis is that DNA damage could arise from genotoxic by-products that are produced during estrogen metabolism (Deroo and Korach, 2006).

Treatment of estrogen receptor-positive breast cancer is often through the use of selective estrogen receptor modulators (SERMs) such as tamoxifen and raloxifene (Deroo and Korach, 2006, Fabian and Kimler, 2005). Both drugs exhibit tissue-specific activity with antagonistic activity in the breast (Deroo and Korach, 2006). Most (although not all) ER-positive breast tumours respond well to these anti-hormone therapies. Estrogen receptor-negative cancers on the other hand, do not respond to such treatments and instead require other treatment regimens to combat the disease.

The mechanisms underlying progesterone influence on breast cancer development and progression is less well characterised than for estrogen. The role of progesterone in breast cancer development is not well characterised but there is evidence that increased progesterone levels stimulate rapid cell proliferation in progesterone receptor (PR) positive cells (Lange and Yee, 2008). Furthermore, progesterone has a potential pro-survival effect (Moore et al., 2006) and can drive tumour cell differentiation (Sartorius et al., 2005). Understanding how estrogen and progesterone influence oncogenesis is especially important because these hormones are generally found in the presence of each other (Sartorius et al., 2005).

1.4.3 Human Epidermal Growth Factor Receptor 2 in Breast Cancer

Human epidermal growth factor receptor 2 (HER2) overexpression is common in breast tumours (Kraus et al., 1987) affecting more than 1 in 5 breast cancer cases (Mitri et al., 2012). The *ERBB2* gene encoding HER2 is a proto-oncogene that becomes oncogenic when HER2 is overexpressed. Due to its role as a growth factor receptor, overexpression of HER2 on the cell surface increases the number of growth signals received by the cell leading to an increase in tumour cell proliferation and growth, contributing to oncogenesis in the breast. The clinical prognosis for tumours overexpressing HER2 is relatively poor; the tumours are generally very aggressive and survival rates are low (Mitri et al., 2012).

Genetic dissection of early-onset breast cancer and other genetic diseases

Treatment for cancers overexpressing HER2 is the antibody trastuzumab. Trastuzumab binds HER2 on the surface of tumour cells, compromising their ability to interact with growth factors. Blocking this interaction reduces the growth signals received by the tumour cells, slowing down and potentially halting cell proliferation.

1.4.4 Subtypes

Breast cancer can be divided into a number of subtypes each with distinct gene expression profiles and clinical characteristics. Many of the features of breast carcinomas, including hormone receptor status, nodal status, tumour size, tumour grade, and menopausal status, significantly influence the behaviour, aggressiveness and prognosis of breast cancer and thus the required treatment regime (Sotiriou et al., 2003).

A number of breast cancer subgroups have been described, largely based on gene expression profiles. As many as 10 distinct subtypes have been characterised based on copy number variation and gene expression profiles in almost 2000 individuals (Curtis et al., 2012). It is generally accepted that there are at least four core subtypes constructed depending on the presence or absence of the estrogen receptor (ER+/ER-), progesterone receptor (PR+/PR-), and over- or under-expression of human epidermal growth factor receptor 2 (HER2+/HER2-); Luminal A (ER+ and/or PR+, HER2-), Luminal B (ER+ and/or PR+, HER2+), triple-negative/basal-like (ER-, PR-, HER2-), HER2-enriched (ER-, PR-, HER2+) (Sørlie et al., 2001, 2012). Much progress has been made into understanding the genetic basis of these subtypes over the past couple of decades, however recently it has become apparent that the genetic component of each subtype is highly complex and unique, and effective sub-type specific treatment is still lacking in many cases.

The estrogen receptor plays a major role in both the gene expression profile of breast tumours and the final phenotype of the cancer (Sotiriou et al., 2003); using gene expression patterns, the clearest distinction between breast cancer samples is often observed between ER-positive and ER-negative tumours (Sorlie et al., 2006), indicating that these tumour types are discrete. It has been suggested that ER-positive and ER-negative tumours represent different stages of tumour evolution (Allred et al., 2004) although it is now thought that they in

fact constitute distinct forms of disease (Garcia-Closas et al., 2008) that arise through different mechanisms. There are two distinct epithelial cell types present in the mammary gland: basal (myoepithelial) cells and luminal epithelial cells (Perou et al., 2000). The estrogen receptor phenotype and subsequent gene expression may be related to the cellular make-up of the breast tissue; tumour samples classified as ER-positive have been shown to express relatively high levels of genes normally expressed by luminal cells while ER-negative tumours often express genes associated with the basal cell type (Perou et al., 2000).

1.4.5 Breast Cancer Susceptibility Genes

An individual's risk of developing a disease depends upon variation within susceptibility genes. A range of genes and alleles with high-, moderate-, and low-penetrance has been linked to an increase in breast cancer risk (Figure 1.6) (de Jong et al., 2002, Antoniou and Easton, 2006, Stratton and Rahman, 2008).

Mutations in highly penetrant breast cancer susceptibility genes greatly increase disease risk and tend to be associated with hereditary cancer syndromes. *BRCA1* and *BRCA2* are the most well defined breast cancer susceptibility genes, with multiple breast cancer causing mutations identified in both genes (Turnbull and Rahman, 2008). The risk associated with carrying a deleterious variant in either of these genes is increased by between 10- and 30-fold (Antoniou et al., 2003) and combined, they account for as much as 20% of familial cases (Lalloo and Evans, 2012). Both are highly penetrant autosomal dominant predisposition genes that function as tumour suppressors with roles in DNA repair mechanisms and genome instability suppression (Venkitaraman, 2002, Turnbull and Rahman, 2008, Gudmundsdottir and Ashworth, 2006). *BRCA1* was identified through linkage mapping of early-onset breast cancer patients who harboured familial forms of the disease (Hall et al., 1990, Hall et al., 1992, Miki et al., 1994). Linkage between the 17q21 chromosomal region and breast cancer cases was only observed for these early-onset cases; no linkage was observed in families presenting with late-onset disease (Hall et al., 1990). *BRCA2* was linked to the 13q12-13 region in 1994 (Wooster et al., 1994) and the exact location was more clearly defined a year later (Wooster et al., 1995). Variants in *BRCA1* and *BRCA2* confer a high risk of early-onset

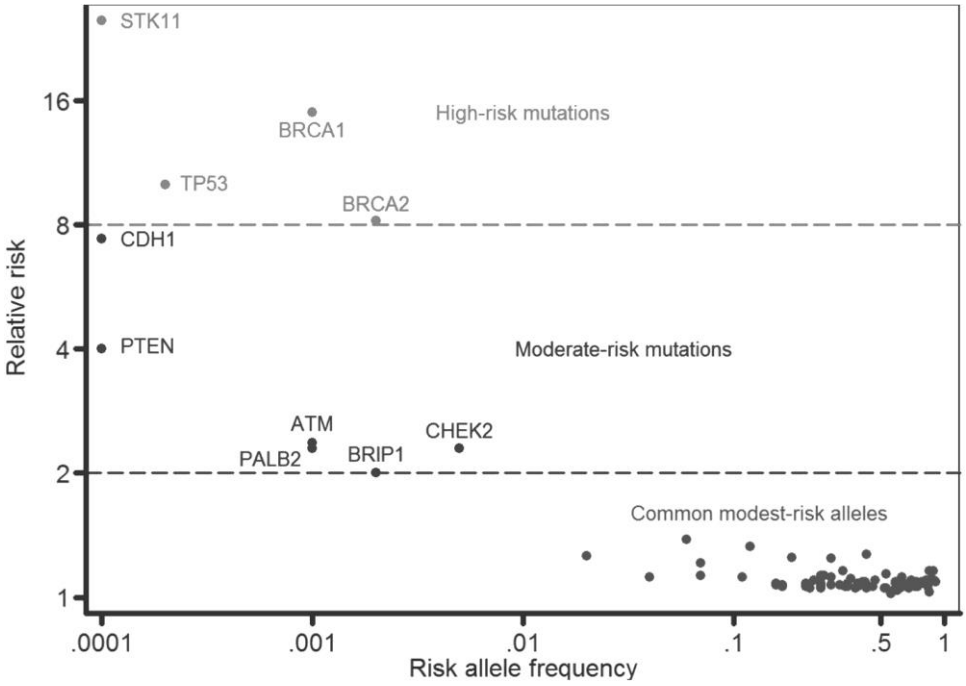


Figure 1.6. Breast cancer risk loci. A number of high-risk, moderate-risk and modest-risk variants have been associated with breast cancer. (Figure originally published in Ghoussaini et al. (2013). Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning?. *The American Journal of Pathology*, 183(4), 1038-1051. Permission to reproduce this figure has been granted by Elsevier.)

breast cancer (Wooster et al., 1994), but mutations in these genes are rare in the population (Antoniou and Easton, 2006) and only account for less than 20% of familial breast cancer cases (Peto et al., 1999). A very small proportion of early-onset patients harbour any variants in either *BRCA1* or *BRCA2*, although individuals with at least two affected relatives have a much-increased likelihood of harbouring such a variant (Peto et al., 1999) and early-onset triple-negative cancers have a greater than 10% chance of carrying a *BRCA1* mutation (Robertson et al., 2012). Many variants have been identified in both *BRCA1* and *BRCA2*. While thousands of truncating mutations in these genes have been associated with the disease there are many more, mainly missense mutations, whose functional significance is unclear (Easton et al., 2007a). Further rare high-penetrance susceptibility genes that increase the risk of developing breast cancer include *TP53* (Malkin et al., 1990), *PTEN* (Nelen et al., 1996) and *STK11* (Boardman et al., 1998). Many of these genes were identified through strong association to familial cancer syndromes, of which breast cancer was a component. Mutations in these genes occur rarely in breast cancer cases where

the cancer syndrome is not present (Hilbers et al., 2013, Ginsburg et al., 2009, Rapakko et al., 2001, FitzGerald et al., 1998, Guenard et al., 2010).

Disease-related mutations in moderate-penetrance susceptibility genes tend to have a frequency of less than 5% in the population and confer a lower increase in disease risk than mutations in highly penetrant genes. Often the mutations do not segregate with the disease (Stratton and Rahman, 2008). A number of moderate-penetrance susceptibility genes have been identified in breast cancer based on interrogation of candidate genes in pathways that include *BRCA1* and *BRCA2*: *ATM*, *CHEK2*, *BRIP1* and *PALB2* (Antoniou and Easton, 2006, Stratton and Rahman, 2008). Mutations are found at a much lower rate in these genes when compared to *BRCA1* and *BRCA2* (Stratton, 1997). Although many genes have been identified using the candidate gene approach it is often difficult to replicate the results in subsequent association studies. There are also difficulties relating to the selection of candidate genes; there is not enough understanding of the underlying disease mechanisms to inform which genes should be interrogated (Tabor et al., 2002). This makes identifying real moderate-penetrance susceptibility genes difficult.

GWAS using breast cancer populations have identified at least 10 moderate-to-high penetrance risk genes and over 70 low-penetrance susceptibility alleles (Easton et al., 2007b, Hunter et al., 2007, Stacey et al., 2007, Ahmed et al., 2009, Zheng et al., 2009, Thomas et al., 2009, Turnbull et al., 2010, Fletcher et al., 2011, Haiman et al., 2011, Ghoussaini et al., 2012, Milne et al., 2014a, Ahsan et al., 2014, Purrington et al., 2014), as well as increasing understanding of the molecular pathways underlying human diseases (Frazer et al., 2009). The majority of positive associations have been identified in cohorts of late-onset breast cancer. One might expect early-onset and late-onset breast cancers to have different underlying aetiology but evidence from early-onset cohorts suggests that many of the risk regions are the same between early-onset and late-onset disease (Ahsan et al., 2014) despite vast differences between age-of-onset. Each identified common low-penetrance susceptibility allele confers only a very small risk of breast cancer (Stratton and Rahman, 2008, Michailidou et al., 2013); considering all identified breast cancer-associated SNPs, risk is typically only increased by less than 1.2-fold per SNP (Milne et al., 2014a) and there are no examples of SNPs that increase risk by more than 1.5-fold (Ghoussaini et al., 2013). Much of the inherited genetic

component of breast cancer is unexplained by all susceptibility genes and loci identified to date: only ~14% is explained by all low-penetrance susceptibility loci (Ghoussaini et al., 2013).

A family history of breast cancer is a major risk factor (Antoniou and Easton, 2006) but only about 10% of breast cancer cases actually cluster in families. Early-onset breast cancer in individuals who do not have a clear family history cannot be attributed to mutations in high- or moderate-penetrance genes because of the lack of family history, but these cases may still arise from inherited susceptibility variants. These types of cases are more difficult to explain but are likely to result from multiple risk variants as well as some *de novo* variants and environmental factors that interact to cause disease development. Indeed, as the number of variants associated with breast cancer continues to increase it is becoming clear that in most cases the disease is polygenic, resulting from multiple variants in multiple susceptibility loci (Pharoah et al., 2002, Stratton and Rahman, 2008).

1.5 Cleft Lip with or without Cleft Palate

Cleft lip with or without cleft palate (CLP) is a birth defect arising from abnormal foetal development. Patients can be affected with cleft lip only (CLO), cleft lip and cleft palate (CLP), or cleft palate only (CPO). Facial clefting occurs in approximately 1 in 700 live births (Dixon et al., 2011), making it a fairly common birth defect. Population incidence rates of CLP vary worldwide; the highest incidence is observed in Asian and Native American populations (1 in 500 births), while populations of African descent show the lowest incidence (1 in 2500 births) (Vanderas, 1987, Dixon et al., 2011). Cleft lip with or without cleft palate generally occurs at twice the rate of cleft palate only (Grosen et al., 2010). Clefting phenotypes also demonstrate sex-specific incidence, with CLP occurring more commonly in males and CPO more common in females (Grosen et al., 2010). Recurrence of CLP, CLO and CPO phenotypes are common in affected families, with sibling recurrence and parent-to-child recurrence rates comparable (Sivertsen et al., 2008). The recurrence risk differs for CLP, CLO and CPO; in particular, cleft palate only has very different recurrence rates when compared to any type of cleft lip phenotype, indicative of the recognised difference in causes of cleft lip and cleft palate phenotypes (Sivertsen et al.,

2008). Furthermore, it has been reported that there is no significant difference in occurrence of either CLO or CLP in first degree relatives of individuals with either of these phenotypes (Sivertsen et al., 2008). Sibling recurrence rates of oral clefts have been reported for a large study of the Danish population: CLO has an overall sibling recurrence risk of 2.5%, CPO a risk of 3.3%, and a risk of 3.9% for CLP (Grosen et al., 2010).

Clinical management of facial clefting usually takes the form of facial surgery although many other kinds of care may be necessary, such as speech therapy and counselling (Mossey et al., 2009). Often individuals born with clefting phenotypes have increased morbidity and mortality throughout their lives (Mossey et al., 2009). Better characterisation of the genetic variants that contribute to cleft lip and/or palate phenotypes is necessary for eventual disease prevention.

1.5.1 Syndromic Cleft Lip and Cleft Palate

Cleft lip and palate or cleft palate are common features of genetic syndromes: CLP features in over 200 syndromes while more than 400 syndromes include CPO in the phenotype (Mossey et al., 2009). These syndromes are usually developmental disorders that include a range of birth and developmental defects. Such disorders are often Mendelian in nature and the causal genes for many of these disorders are usually highly-penetrant and have already been characterised (Leslie and Marazita, 2013).

1.5.2 Nonsyndromic Cleft Lip with or without Cleft Palate

Isolated clefting phenotypes account for approximately 70% of CLP cases and are generally classified as nonsyndromic cleft lip with or without cleft palate (NSCLP). NSCLP is a complex and multifactorial disorder influenced by both genetic and environmental factors (Mossey et al., 2009, Mangold et al., 2011) with gene-environment and gene-gene interactions also likely to play a role in disease susceptibility.

The genetic basis of NSCLP has been investigated through a number of linkage studies, genome scans and genome-wide association studies. Estimates of the number of loci underlying the disorder suggest that a small number of genes

influence the disease: a few major genes and a small number of minor genes (Schliekelman and Slatkin, 2002). Many studies have implicated a handful of genes and genomic regions in the aetiology of NSCLP in many different populations (Prescott et al., 2000, Marazita et al., 2004, Zucchero et al., 2004, Blanton et al., 2005, Ghassibe et al., 2005, Scapoli et al., 2005, Park et al., 2007, Vieira et al., 2007, Rahimov et al., 2008, Jugessur et al., 2008, Birnbaum et al., 2009, Beaty et al., 2010, Rojas-Martinez et al., 2010, Ludwig et al., 2012, Lennon et al., 2012, Fontoura et al., 2012, Grant et al., 2009, Mangold et al., 2010, Letra et al., 2012, Camargo et al., 2012, Otero et al., 2007, Nikopensius et al., 2009, Blanton et al., 2010, Younkin et al., 2014). Four major GWA studies implicated 12 genetic loci in NSCLP (Birnbaum et al., 2009, Grant et al., 2009, Mangold et al., 2010, Beaty et al., 2010), all of which were confirmed in a meta-analysis study (Ludwig et al., 2012). The results of many of these genome scans and GWAS show little concordance, rarely implicating the same genes or genetic regions in NSCLP (Leslie and Marazita, 2013), this is likely the result of the highly heterogeneous nature of NSCLP (Rahimov et al., 2011)

As with most complex diseases, it is likely that common and rare variation influences the phenotype of NSCLP, however, much of the genetic analysis of NSCLP to date has focussed on common variants. There is evidence that over 70% of rare variants identified in patients with NSCLP are not found in non-NSCLP individuals (Leslie and Murray, 2013) suggesting a role for rare variation in NSCLP.

Many of the genes responsible for syndromic clefting disorders are also important in NSCLP, perhaps due to variable penetrance or action of modifiers (Stanier and Moore, 2004). A major example is *IRF6*, the gene responsible for Van de Woude syndrome (VWS) and popliteal pterygium syndrome (PPS) (Kondo et al., 2002). VWS is a good model for NSCLP since most patients present with only minor additional phenotypes; in 15% of cases isolated cleft lip is the only phenotype and patients are phenotypically no different to NSCLP patients (Stanier and Moore, 2004). *IRF6* has been strongly implicated as a susceptibility factor for NSCLP in a range of populations (Zucchero et al., 2004, Ghassibe et al., 2005, Park et al., 2007, Vieira et al., 2007, Jugessur et al., 2008).

1.6 Oculopharyngeal Muscular Dystrophy

Oculopharyngeal muscular dystrophy (OPMD) is a Mendelian neuromuscular disorder primarily affecting craniofacial muscles. The disorder can be either autosomal dominant or autosomal recessive, with dominant forms being most common. The disease affects both sexes equally and is completely penetrant (Calado et al., 2000). OPMD prevalence rates differ between populations, with the highest prevalence observed in the Bukharan Jewish population at 1 in 600 (Abu-Baker and Rouleau, 2007). The estimated prevalence in Europe is 1 in 100,000 (Abu-Baker and Rouleau, 2007). The largest cluster of cases are observed in the French-Canadian population, where the prevalence is 1 in 1000 (Abu-Baker and Rouleau, 2007).

Disease onset begins in later life, usually the fifth decade, and is characterised by progressive ptosis (drooping of the eyelids) and weakness of the extraocular muscles, dysphagia (difficulty swallowing) and is later characterised by progressive weakness of proximal limbs. As with many degenerative diseases, protein aggregates in tissue cells are recognised as a hallmark of OPMD.

Muscle biopsies from OPMD patients identified rimmed vacuoles (RV) and intranuclear inclusions (INI) as the major morphological features within muscle fibre nuclei. Rimmed vacuoles have been described in a number of muscle disorders and do not necessarily occur in all cases of OPMD (Tomé et al., 1997). The type of inclusion bodies observed in OPMD patients however, are unique to the disorder; there are no reports of these type of INIs in other disorders and they are different from other types of inclusion bodies previously described (Tomé et al., 1997).

1.6.1 OPMD Causal Mutations

The genomic locus 14q11.2-q13 was identified as the OPMD disease locus through linkage mapping (Brais et al., 1995). The gene responsible was subsequently identified as *PABPN1* (Brais et al., 1998). The most common mutation in *PABPN1* responsible for the disorder is an expansion of a $(GCG)_6$ repeat in the polyalanine tract at the N terminus of the protein (Figure 1.7A). Expansions of between 2 and 7 extra GCG codons were originally observed in French-Canadian families (Brais et al., 1998). Further studies have identified

similar expansions in other ethnic groups (Blumen et al., 2000, Mirabella et al., 2000, Nagashima et al., 2000, Hill et al., 2001, Müller et al., 2001, Rodríguez et al., 2005) as well as GCG repeats interspersed with GCA codons causing the alanine repeat (Scacheri et al., 1999, van der Sluijs et al., 2003, Robinson et al., 2005). An expansion by only one codon does not confer the phenotype; $(GCG)_7$ has been observed in 2% of controls (Brais et al., 1998). However, individuals homozygous for this mutation presented with an autosomal recessive form of OPMD and individuals found to be compound heterozygous for this mutation and the $(GCG)_9$ mutation had more severe phenotypes than individuals with only one mutated copy of the gene (Brais et al., 1998) suggesting that the $(GCG)_7$ allele is a modifier of the autosomal dominant phenotype (Banerjee et al., 2013).

A second *PABPN1* mutation has been identified in three patients with OPMD who did not have the GCG codon expansion (Robinson et al., 2006, Robinson et al., 2011a): a G > C mutation at nucleotide 35 in exon 1 of the gene. This mutation has the same effect as the GCG expansion because it changes the glycine immediately after the 10 alanine tract to an alanine which is then followed by a further two alanine residues (Figure 1.7B). As a result, this mutation generates a sequence of 13 alanine residues.

1.6.2 Pathogenicity of *PABPN1* Mutations

The mechanism by which *PABPN1* mutations cause the craniofacial muscular phenotype of OPMD has not been fully characterised. *PABPN1* codes for an ubiquitously expressed nuclear protein that binds to pre-mRNA molecules in the nucleus, specifically the poly(A) tails at the 3' ends of the molecules. The protein is important for stimulating and maintaining polyadenylation of the mRNA molecule by poly(A) polymerase (PAP) (Calado et al., 2000, Kühn and Wahle, 2004, Corbeil-Girard et al., 2005).

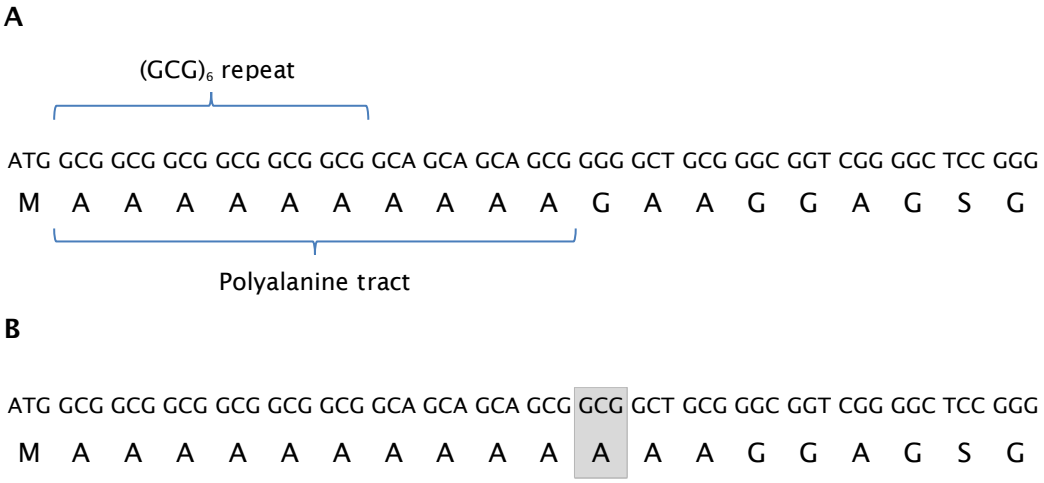


Figure 1.7. Summary of N-terminal region of PABPN1 protein. (A) Coding sequence of the first 20 amino acids at the N-terminal end of the PABPN1 protein. The repeat of 6 GCG codons coding for the first 6 alanine residues is highlighted along with the polyalanine tract of 10 alanine residues. The most common expansion mutation observed is insertion of between 2 and 7 extra GCG codons. Insertions of GCG and GCA codons have also been observed. (B) G > C point mutation observed in 3 English patients that has a similar effect as the GCN codon expansion by increasing the polyalanine tract from 10 to 13 alanine residues.

Although specifically a nuclear protein, PABPN1 shuttles across the nuclear membrane to the cytoplasm; it is likely that PABPN1 enters the cytoplasm while still bound to the poly(A) tail of mRNA where it then dissociates and returns tothe nucleus (Kühn and Wahle, 2004). Depletion of PABPN1 in cell studies detected an increase in nuclear accumulation of polyadenylated pre-mRNA molecules (Apponi et al., 2010), suggestive of a direct role for PABPN1 in mRNA export. Another possible explanation is that polyadenylation is adversely affected by reduced PABPN1 levels, impacting on further RNA processing steps necessary for efficient export (Banerjee et al., 2013). It is therefore likely that PABPN1 has a role in mRNA export from the nucleus (Kühn and Wahle, 2004).

The pathological features of OPMD are mostly confined to skeletal muscle tissue despite PABPN1 protein expression being ubiquitous, suggestive of a specific role for PABPN1 in skeletal muscle cells (Fan and Rouleau, 2003). Wild-type PABPN1 is anti-apoptotic but this function is partially lost in mutated protein, which may lead to the disease phenotype through loss of muscle cells (Davies et al., 2008). There is also evidence that PABPN1 binds the Ski-

interacting protein (SKIP) and together they control the expression of muscle-specific genes (Kim et al., 2001), which may explain why OPMD is restricted to muscle tissue. Evidence from cellular studies suggests that a reduction of PABPN1 levels in muscle cells impacts on cell differentiation and proliferation (Apponi et al., 2010), while overexpression of expanded PABPN1 might decrease the cellular levels of muscle-specific proteins (Abu-Baker et al., 2003) and increase the rate of aggregate formation in muscle cell lines (Raz et al., 2011). Aggregates sequester a range of proteins including a number of myogenic transcription factors critical for differentiation of muscle cells (Banerjee et al., 2013). This could lead to a functional depletion of muscle specific proteins, causing a compensatory response through up-regulation of corresponding genes (Banerjee et al., 2013). It may also be the case that inherent properties of muscle tissue cause the OPMD phenotype to manifest specifically in muscle (Banerjee et al., 2013).

The OPMD nuclear inclusions are considered toxic and associated with skeletal muscle cell death (Fan et al., 2001). It was originally thought that the expansion of the polyalanine tract in PABPN1 introduces a toxic gain-of-function causing protein aggregation (Brais et al., 1998, Calado et al., 2000, Robinson et al., 2006, Davies et al., 2008), however, the PABPN1 expansion is relatively small making it unlikely that this alone is disease-causing (Chartier et al., 2006). Furthermore, wild-type PABPN1 also forms aggregates with similar properties to mutated PABPN1 aggregates (Tavanez et al., 2005) and aggregate formation is actually independent of the alanine tract (Winter et al., 2012). Therefore, the aggregates themselves are unlikely to be toxic or the cause of muscle degeneration in OPMD cases. Indeed, both wild-type and mutated PABPN1 can move freely in and out of the inclusion bodies (Tavanez et al., 2005) suggesting that the inclusions themselves are not pathogenic. Instead, extension of the polyalanine tract maybe lead to stabilisation of the PABPN1 protein, causing it to accumulate over time and become toxic (Chartier et al., 2006). Further evidence suggests that the aggregates may actually form during the polyadenylation process; disruption of PABPN1's ability to bind RNA or recruit poly(A) polymerase stops the formation of inclusion bodies (Tavanez et al., 2005). The inclusion bodies also contain mRNA molecules and it is thus thought that they constitute 'mRNA traps' that prevent mRNA export from the nucleus, which may be detrimental to the cells, contributing to cell death (Fan

and Rouleau, 2003). A small proportion of cells containing inclusion bodies undergo apoptosis so expression of mutated PABPN1 in the nucleus could weaken the nuclear structure, leading to cell death (Sasseville et al., 2006).

PABPN1 protein expression has been found to decrease with age and this process is accelerated in individuals with OPMD (Anvar et al., 2013). This decrease in expression was mainly restricted to skeletal muscle cells, suggestive of a role for PABPN1 in regulation of muscle cell aging (Anvar et al., 2013), which may explain why OPMD-affected individuals present with late-onset muscle weakness.

1.7 Research Project Aims

The overall aim of this research project was to use genetic studies to explore the aetiology of human diseases. The underlying genetic aetiology of three diseases (breast cancer, cleft lip/palate and OPMD) was explored using both SNP array-derived data and whole-exome sequence data (Figure 1.8).

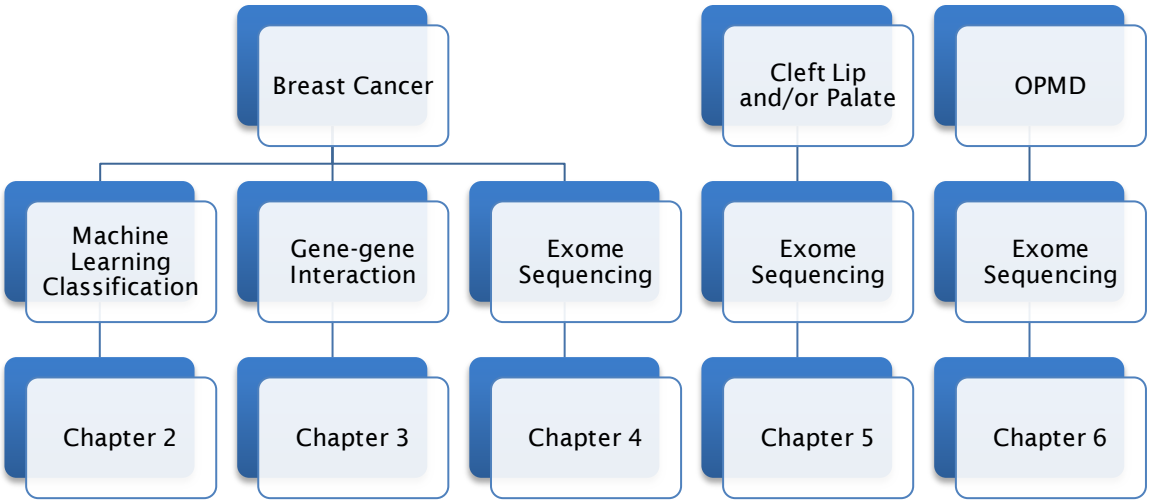


Figure 1.8. Flow-chart showing the content of each chapter in terms of disease analysed and method used.

1.7.1 Aim 1 – Analysis of Breast Cancer Genetics

The first aim was to explore the genetic landscape of early-onset breast cancer through the analysis of genome-wide SNP array data and whole-exome sequencing. Genome-wide SNP data were analysed using machine learning algorithms, rather than the standard association test, as a novel method for exploring SNP data. Machine learning methods have been little used for genome-wide association data so this represents a novel area of research that has the potential to uncover complex aspects of the data that will be missed by association testing.

Chapter 2 explores SNPs underlying the ER subtype distinction using a support vector machine. The SVM model was chosen because it is a state-of-the-art classification technique. Using a small subset of genome-wide SNPs, the resultant model could distinguish between ER-positive and ER-negative cancer subtypes successfully. The results implicate over 100 genes in the distinction between ER cancer subtypes.

Chapter 3 further explores the importance of common polymorphisms in breast cancer genes genotyped in early-onset breast cancer through the analysis of potential interactions between SNPs. Both breast cancer cases and non-cancer controls were selected for this analysis to allow for identification of SNPs underlying breast cancer as a whole or either of the ER subtypes. Three machine learning methods were used to analyse the data alongside the widely accepted logistic regression method for detecting interaction, to identify all possible interactions. Analysis identified many potential interactions underlying the ER-positive and ER-negative subtypes.

The focus of chapter 4 is on the more recently developed next-generation whole-exome sequencing. A small number of early-onset breast cancer cases, presenting with an extreme phenotype or a family history of early-onset breast cancer, were selected for exome sequencing. Two tiers of rare variant analysis were implemented to fully explore the genetic profile of each patient. Many rare variants were identified with potential roles in breast cancer pathology.

1.7.2 Aim 2 – Analysis of Cleft Lip and Palate Genetics

Chapter 5 focuses on the genetics of Colombian patients presenting with various forms of syndromic or non-syndromic cleft lip and/or palate. A total of 18 patients were selected for whole-exome sequencing and all rare variants in candidate genes were identified. Likely causative mutations were identified in all syndromic families due to the Mendelian inheritance patterns of these syndromes. Many variants were detected in non-syndromic cases but the mutational spectrum is more complicated due to the likely complex and polygenic genetic model for this phenotype.

1.7.3 Aim 3 – Analysis of an OPMD-like Phenotype

Chapter 6 focuses on identifying the causative mutation in a dominant OPMD-like phenotype in a single affected family. Six family members were selected for whole-exome sequencing. A potential causative mutation was identified in all affected family members.

Chapter 2: Classification of Estrogen Receptor-Positive and Estrogen Receptor-Negative Early-Onset Breast Cancer Cases Using the Support Vector Machine

2.1 Background

Breast tumours can express receptors for a number of growth hormones, including estrogen. Expression of estrogen receptors produces two distinct tumour types; tumours expressing many estrogen receptors are labelled estrogen receptor-positive while tumours expressing few or no receptors are estrogen receptor-negative. The cellular make-up of these tumour types differ, with ER-positive tumours tending to be more differentiated (Putti et al., 2004) while ER-negative tumours tend to have rapid cell proliferation and cells have large irregular nuclei and a high mitotic index. It has long been recognised that ER-positive and ER-negative breast cancers display different prognoses and patterns of recurrence (Knight et al., 1977, Anderson et al., 2002) with ER-negative tumours tending to have poor prognoses and reduced survival rates.

ER-positive tumours, which proliferate in response to estrogen, can be treated fairly effectively using estrogen suppression treatments, which might block estrogen receptors or remove estrogen from circulation. There are, however, a proportion of ER-positive tumours that are resistant to anti-hormone treatment but identifying which patients will not respond is difficult. ER-negative cancer cannot be treated with hormonal therapies; instead surgery, radiation therapy and chemotherapy are required.

Estrogen is crucial for the development of breast cancer regardless of the receptor status of the cells; estrogen can act on cells of the tumour microenvironment, influencing angiogenesis and stromal cell recruitment (Gupta et al., 2007, Péqueux et al., 2012). The role of estrogen in breast cancer development may act down two hypothesised routes. In the first case, estrogen binds to estrogen receptors, stimulating mammary cell proliferation, thus increasing DNA replication which has an associated risk of replication errors

and acquisition of deleterious mutations (Deroo and Korach, 2006). The second hypothesis considers DNA damage to be a result of genotoxic by-products of estrogen metabolism (Deroo and Korach, 2006). There is evidence to support both hypotheses.

Linkage studies including individuals with familial forms of early-onset breast cancer identified the genomic regions of the *BRCA1* (Miki et al., 1994, Hall et al., 1990, Hall et al., 1992) and *BRCA2* (Wooster et al., 1994, Wooster et al., 1995) genes as associated with this phenotype. These two genes were identified as highly-penetrant breast cancer genes involved in familial forms of the disease. Further high-to-moderate penetrance genes have also been identified, including *TP53* (Malkin et al., 1990), *STK11* (Boardman et al., 1998), and *PTEN* (Nelen et al., 1996). These genes are associated with cancer syndromes rather than specifically with early-onset breast cancer and they only occur in families presenting with the associated syndrome (FitzGerald et al., 1998, Rapakko et al., 2001, Ginsburg et al., 2009, Guenard et al., 2010). Mutations in all these highly- and moderately-penetrant genes only account for a small proportion of non-familial forms of disease, therefore, many cases of non-familial early-onset breast cancer are likely to be more complex and polygenic in nature, arising from multiple risk variants, all of which contribute to disease progression.

The polygenic model of complex disease suggests that many variants, perhaps in the region of thousands (Michailidou et al., 2013) with low-to-moderate penetrance will act together to cause disease (Pharoah et al., 2002). Evidence for this type of model has been collected from GWA studies (Easton et al., 2007b, Hunter et al., 2007, Stacey et al., 2007, Ahmed et al., 2009, Zheng et al., 2009, Thomas et al., 2009, Turnbull et al., 2010, Fletcher et al., 2011, Haiman et al., 2011, Ghoussaini et al., 2012, Garcia-Closas et al., 2013), which have implicated over 70 SNP variants and loci (Ghoussaini et al., 2013) as risk variants for breast cancer development. These variants are generally common and low-penetrance so their individual contribution to disease risk is small but the combined risks of multiple variants can lead to disease manifestation. There is also evidence that women with various forms of early-onset breast cancer have much more ‘disordered’ genomes than non-breast cancer controls (Smyth et al., 2015), i.e. breast cancer cases were associated with common SNPs at thousands of locations across the genome. Although association

studies have identified many breast cancer risk variants, combined they only explain a small proportion of disease cases, suggesting that many more common low-penetrance breast cancer risk variants are yet to be identified.

2.2 Aim

To explore the contribution of inherited SNP variants to the broadly classified estrogen receptor subtypes (ER-positive and ER-negative) of breast cancer within a SVM classification model.

2.3 Materials and Methods

2.3.1 Data and Data Processing

Patients with early-onset estrogen receptor-positive or –negative breast cancer were selected from the 'Prospective study of Outcome in Sporadic versus Hereditary breast cancer' (POSH) cohort (Eccles et al., 2007) of ~3000 patients with disease onset before the age of 40 years. A total of 542 samples that were selected for an earlier GWAS genotyping study (Rafiq et al., 2013) were included in the analysis. These samples were selected based on early or no relapse at the time of observation. Characterisation of estrogen receptor status classified 170 cases as ER-positive and 372 cases as ER-negative.

Genotyping of the breast cancer samples was conducted using the Illumina 660-Quad SNP array. Genotyping was conducted at the Mayo Clinic, Rochester, Minnesota, USA, (261 samples) and the Genome Institute of Singapore, National University of Singapore (281 samples) (Rafiq et al., 2013). To ensure complete harmonisation of genotype calling, the intensity data available from both locations, in form of .idat files, were combined and used to generate genotypes using the algorithm in the genotyping module of Illumina's Genome Studio software. A GenCall threshold of 0.15 was selected and the HumanHap660 annotation file was used. All 542 patients were genotyped for 502,330 genome-wide single nucleotide polymorphisms (SNPs). SNPs were located on all chromosomes, including the X chromosome and in mitochondrial DNA regions.

SNPs were excluded from further analysis if they had a minor allele frequency (MAF) below 0.01, genotyping call rate < 95% or showed significant deviation from Hardy-Weinberg equilibrium (HWE, p-value <0.0001). We used the pairwise Identity-By-State (IBS) and multidimensional scaling, implemented in PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/)(Purcell et al., 2007), to confirm that patients were ethnically homogeneous.

Genotyping was not complete, resulting in a number of missing genotype values for 84,381 of the total 502,330 SNPs. The MACH 1.0 program (http://www.sph.umich.edu/csg/abecasis/MACH/index.html) was used along with genotype and haplotype phase data specific for CEU (Caucasian residents of Utah with northern and western European ancestry) population (available from HapMap phase 2 project) to impute missing genotypes where possible. Imputation was done for the 490,732 SNPs located on the autosomes.

All SNP genotypes were transformed into numerical coding. The major (reference) and minor (alternate) alleles for each SNP were determined from all genotypes within the sample. SNP genotypes were labelled 1 for reference homozygous SNPs, 0 for heterozygous, and -1 for alternate homozygous SNPs. ER-negative cases were given a phenotype label of 1 and ER-positive cases were labelled 0.

2.3.2 SNP Feature Selection

SNPs showing significant association with ER-negative cases were identified from the additive chi-squared association test implemented in PLINK. All SNPs were ranked based on the results of the chi-squared test: those SNPs showing strongest association with the ER-negative phenotype were ranked highest. Subsets of SNPs showing strongest disease association were selected as features for SVM models from the ranked list of SNPs and models were produced from subsets of 50, 100 and 200 SNPs. Additional subsets of 200 SNPs, without strong association with the phenotype, were selected as model input: (i) the final 200 SNPs in the chi-squared test ranked list; (ii) a randomly selected set of 200 SNPs with a range of chi-square values.

2.3.3 SVM Classification Model

An SVM model is a supervised machine learning algorithm that produces a classification model based on information contained within a 'training' set and uses the model to classify samples from a 'test' set (Cortes and Vapnik, 1995). Samples are separated into a minimum of two 'classes' and a number of 'features' are selected. All samples within the training set have class labels and feature values that are used to produce the classification model. All samples within the training set are plotted in high-dimensional space and a 'hyperplane' is used to separate the two classes of samples. There are many possible hyperplanes; the selected hyperplane is the one that maximises the distance between all samples from each class and the hyperplane. The resultant SVM model is applied to the test set in which the class labels are hidden from the algorithm. Therefore, samples are classified based on the position of the data point relative to the hyperplane.

The robustness and reliability of the SVM classifier can be tested using cross-validation, where the data is split into *n* equally sized sets testing *n* models. Ten-fold cross-validation was selected for this study: data were divided into 10 approximately equal-sized sets and a classifier built based on the data in 9/10 of these sets. The remaining 10% of data were used as a test set to determine the accuracy of the classifier. The procedure is repeated 10 times with each set representing the test data once and average classification accuracy is determined. Cross-validation was repeated 10 times for the tier 1 analysis (see section 2.3.4) and 100 times for tier 2 analysis (see section 2.3.5). Therefore, the classification accuracies presented are the mean values from 100 and 1000 resultant models for tier 1 and tier 2 analyses respectively.

In all cases classifier performance was assessed using measures of classification accuracy, true positive/negative rates (TPR/TNR), false positive/negative rates (FPR/FNR), and area under receiver operating characteristic (ROC) curve values (AUC). In this study, classification accuracy provides a measure of the samples that are classified into the correct ER subtype while the TPR/TNR indicates the proportion of ER-positive/ER-negative samples that are correctly classified as ER-positive/ER-negative, and the FPR/FNR indicates the proportion of ER-positive/ER-negative samples that are incorrectly classified as ER-negative/ER-positive. ROC curves can be

constructed by plotting TPR values against FPR values, the area under such a curve (AUC) can be used as a measure of classifier performance, with values closer to 1 indicative of a good classifier and values closer to 0.5 indicative of a classifier that cannot discriminate between samples of the two classes.

The SVM models were produced from the Weka data mining software (http://www.cs.waikato.ac.nz/ml/weka/) (Hall et al., 2009) through implementation of the Sequential Minimal Optimization (SMO) algorithm for training a SVM classifier. Tier 1 analysis evaluated 6 kernel models; linear, normalized quadratic polynomial (NQP), quadratic polynomial, cubic polynomial, Pearson VII function-based universal kernel (Puk), and radial basis function (RBF). The Puk kernel was not evaluated in tier 2 analysis and models were produced from the remaining five kernels.

2.3.4 Tier 1 Analysis

Classification models were produced using 6 kernels and 3 SNP subsets. SNP subsets were selected from the chi-squared association test ranked SNP list: the 50, 100 or 200 highest-ranked SNPs. Missing genotype data were excluded from the SVM models by either (i) removing any samples without a full set of SNP genotypes, or (ii) removing any SNPs without genotypes for all 542 samples. Therefore, for each subset of SNPs, 3 classification models were produced; one model containing all SNPs and all samples, one model containing only samples with a complete set of SNP genotypes, and one model containing only SNPs with a complete set of genotypes.

2.3.5 Tier 2 Analysis

Classification models were produced using 5 kernels and 3 SNP subsets. SNP subsets were selected from the chi-squared association test ranked SNP list: the 50, 100 or 200 highest-ranked SNPs. Imputed genotype data was used to resolve missing data where possible. Any SNPs with missing genotypes after imputation were removed from the subset and replaced with the next most disease-associated SNPs, from the ranked list that contained a full set of genotypes. Only SNPs located on the autosomes were included in this analysis.

Additional subsets of SNPs, not showing strong disease association, were selected and SVM classification models produced. Both subsets contained 200 fully genotyped SNPs (including imputed genotypes). One subset contained the 200 SNPs ranked at the bottom of the chi-squared test ranked list, the other containing 200 SNPs randomly selected from the whole ranked list.

2.3.6 Gene Annotation

SNPs included as features in the classification models were annotated using ANNOVAR (http://www.openbioinformatics.org/annovar/annovar_gene.html) (Wang et al., 2010). Gene-based annotation was carried out using the UCSC 'Known Gene' database. For SNPs situated outside genes the SNP was annotated as associated with the closest gene. Official gene names were obtained from the HUGO Gene Nomenclature Committee database (http://www.genenames.org/). A total of 139 unique gene names were linked to the set of 200 SNPs used in tier 2 analysis.

2.3.7 Functional Gene Classification

Functional gene annotation clusters were identified using the 'Gene Functional Classification' tool in DAVID (Database for Annotation, Visualization and Integrated Discovery; http://david.abcc.ncifcrf.gov/home.jsp) (Huang et al., 2008, Huang et al., 2009). DAVID determines significant enrichment of function within a submitted gene list by contrasting with a 'whole genome' background. Annotation clusters were identified from the 139 genes using the 'Functional Annotation Clustering' tool and five annotation categories: disease, functional categories, gene ontology, pathways and protein domains. Enriched pathways were identified using only the 'Pathways' annotation category with BBID, BIOCARTA, and KEGG selected.

2.3.8 SNP Feature Weights

Feature weights for the 200 SNPs included in tier 2 analysis were obtained by running a linear model in the Weka Explorer environment. The linear kernel model output in the Weka Explorer provides feature weights rather than support vectors.

2.4 Results

Overall classification accuracy is a measure of how successful a model is at assigning a sample to the correct class. For all subsets of SNPs considered classification accuracy exceeds 70%.

Missing genotypes are a common feature of SNP genotype data but SVM models do not work well in the presence of missing data. Therefore, the aim of tier 1 analysis was to investigate classification accuracy using different approaches to the missing SNP genotype data. Based on the results of tier 1 analysis, tier 2 analysis explored the classification accuracy of a small subset of genome-wide SNPs that included imputed genotypes.

2.4.1 Tier 1 Analysis

2.4.1.1 Classification Accuracy using Subsets with Samples Removed

To remove missing genotypes from the data, any breast cancer samples that did not have a full set of SNP genotypes were removed from the data set. Considering the subset of 200 most disease associated SNPs, 73 samples had at least one missing SNP genotype. Removal of these samples reduced the dataset to 469 samples, all of which were fully genotyped for the 200 SNPs. Classification accuracy exceeded 93% (Table 2.1) for five of the six kernels. The cubic polynomial kernel model performs best in this case with 95.3% accuracy. In this model 92% of ER-positive cases and 97% of ER-negative cases are classified correctly. In all models, except the Puk kernel model, at least 88% of ER-positive cases are correctly classified (TPR Table 2.1) and at least 96% of ER-negative cases are correctly classified (TNR Table 2.1). Therefore, the majority of models can distinguish between the two sample classes successfully, with very few samples incorrectly classified.

Smaller subsets of the most associated SNPs were also assessed for their ability to distinguish between the sample classes. Considering only the 100 most strongly associated SNPs, 485 samples were fully genotyped. In the case of 50 SNPs, 512 samples were fully genotyped. In both cases, a decrease in the number of SNPs used as features in the SVM model led to a reduction in

Table 2.1. Classification results for 200 SNPs with strongest disease association genotyped in 469 samples

Kernel type	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	93.77	0.90	0.04	0.96	0.10	0.93
NQP	94.90	0.88	0.02	0.98	0.12	0.93
Quadratic polynomial	94.78	0.91	0.03	0.97	0.09	0.94
Cubic polynomial	95.33	0.92	0.03	0.97	0.08	0.94
Puk	66.31	0.00	0.00	1.00	1.00	0.50
RBF	94.90	0.88	0.02	0.98	0.12	0.93

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

Table 2.2. Classification results for subsets of 100 SNPs and 50 SNPs with strongest disease association

Kernel type	Number of SNPs	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	100	90.08	0.85	0.07	0.93	0.15	0.89
	50	84.40	0.73	0.10	0.90	0.27	0.81
NQP	100	91.40	0.81	0.04	0.96	0.19	0.89
	50	84.57	0.65	0.06	0.94	0.35	0.79
Quadratic polynomial	100	89.56	0.84	0.08	0.92	0.16	0.88
	50	78.58	0.67	0.16	0.84	0.33	0.76
Cubic polynomial	100	91.13	0.85	0.06	0.94	0.15	0.90
	50	78.77	0.67	0.16	0.84	0.33	0.76
Puk	100	67.46	0.00	0.00	1.00	1.00	0.50
	50	73.13	0.18	0.01	0.99	0.82	0.58
RBF	100	89.66	0.73	0.02	0.98	0.27	0.85
	50	80.33	0.45	0.03	0.97	0.55	0.71

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

classification accuracy; accuracy was < 92% for the subset of 100 SNPs (Table 2.2) and < 85% for the subset of 50 SNPs (Table 2.2).

Six kernel models were produced for each subset of SNPs to compare classification accuracy from different of SVM models. All except the Puk kernel model performed well; the Puk models behaved as majority class classifiers (classifying all samples as ER-negative) for the subsets of 100 and 200 SNPs, while it classified few ER-positive cases correctly for the subset of 50 SNPs.

2.4.1.2 Classification Accuracy using Subsets with All Samples Included

Classification accuracy was assessed using all genotyped samples but any SNPs that had a genotyping rate of less than 100% were removed from consideration. After removal of poorly genotyped SNPs, a subset of the 200 most strongly disease-associated SNPs was produced. Classification accuracy using this dataset exceeded 94% for all kernel models except the Puk kernel. True positive rates and true negative rates for the five well performing kernels are consistently high; at least 90% of ER-positive cases and 97% of ER-negative cases were classified correctly. The best performing models, in terms of classification accuracy, were those produced using the NQP kernel (96.2%) and RBF kernel (96.3%) (Table 2.3). Both models classified 99% of ER-negative patients and 90% of ER-positive cases correctly (Table 2.3). Sample classification from smaller subsets of SNPs also performed well, with > 80% accuracy achieved in most cases (Table 2.4).

Table 2.3. Classification results for 200 SNPs genotyped in all samples after removal of incompletely genotyped SNPs

Kernel type	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	94.93	0.90	0.03	0.97	0.10	0.94
NQP	96.22	0.90	0.01	0.99	0.10	0.95
Quadratic polynomial	95.31	0.91	0.03	0.97	0.09	0.94
Cubic polynomial	95.87	0.92	0.02	0.98	0.08	0.95
Puk	68.63	0.00	0.00	1.00	1.00	0.50
RBF	96.31	0.90	0.01	0.99	0.10	0.95

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

Table 2.4. Classification results for subsets of 100 SNPs and 50 SNPs genotyped in all samples after removal of incompletely genotyped SNPs

Kernel type	Number of SNPs	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	100	90.40	0.84	0.06	0.94	0.16	0.89
	50	84.76	0.74	0.10	0.90	0.26	0.82
NQP	100	91.96	0.83	0.04	0.96	0.17	0.89
	50	85.48	0.67	0.06	0.94	0.33	0.80
Quadratic polynomial	100	89.19	0.83	0.08	0.92	0.17	0.87
	50	80.37	0.67	0.14	0.86	0.33	0.77
Cubic polynomial	100	89.83	0.83	0.07	0.93	0.17	0.88
	50	80.83	0.67	0.13	0.87	0.33	0.77
Puk	100	68.71	0.00	0.00	1.00	1.00	0.50
	50	74.02	0.19	0.01	0.99	0.81	0.59
RBF	100	91.16	0.79	0.03	0.97	0.21	0.88
	50	80.94	0.43	0.02	0.98	0.57	0.71

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

2.4.2 Tier 2 Analysis

2.4.2.1 Classification Accuracy using Subsets with all Samples Included and Imputed Genotype Data

The second tier of analysis investigated classification accuracy using all samples and fully genotyped SNPs. Unlike tier 1 analysis where all samples and fully genotyped SNPs were considered; this dataset included imputed genotypes which resolved some of the missing genotypes. A subset of the 200 most strongly disease associated SNPs were selected as features for the models.

Five kernels were used to produce SVM models. Classification accuracy exceeded 93% in all cases (Table 2.5

Table 2.5). The highest classification accuracy was achieved using the RBF kernel and NQP kernel; 95.95% and 95.69% respectively. In both cases 99% of the ER-negative cases and 89% of the ER-positive cases were classified correctly. The true positive rate is equal in all five models, demonstrating that

they are equally successful at recognising and classifying any ER-positive cases in the test data. The true negative rate always exceeded 0.95, indicating that at least 95% of ER-negative cases are classified correctly in each model. All models are superior at classifying ER-negative cases compared to ER-positive cases. Classification accuracy was reduced when only 50 (<86%) or 100 (<93%) SNPs were considered (Table 2.6).

Table 2.5. Classification results for 200 SNPs genotyped in all samples including imputed SNP genotype data

Kernel type	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	93.28	0.88	0.04	0.96	0.12	0.92
NQP	95.69	0.89	0.01	0.99	0.11	0.94
Quadratic polynomial	93.89	0.89	0.04	0.96	0.11	0.93
Cubic polynomial	94.54	0.89	0.03	0.97	0.11	0.93
RBF	95.95	0.89	0.01	0.99	0.11	0.94

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

Table 2.6. Classification results for subsets of 100 SNPs and 50 SNPs genotyped in all samples including imputed SNP genotype data

Kernel type	Number of SNPs	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	100	90.00	0.83	0.07	0.93	0.17	0.88
	50	85.22	0.73	0.09	0.91	0.27	0.82
NQP	100	92.42	0.82	0.03	0.97	0.18	0.90
	50	85.09	0.64	0.05	0.95	0.36	0.79
Quadratic polynomial	100	88.30	0.81	0.08	0.92	0.19	0.86
	50	78.15	0.66	0.16	0.84	0.34	0.75
Cubic polynomial	100	89.42	0.82	0.07	0.93	0.18	0.87
	50	78.40	0.66	0.16	0.84	0.34	0.75
RBF	100	91.25	0.77	0.02	0.98	0.23	0.87
	50	79.81	0.38	0.01	0.99	0.62	0.69

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

2.4.2.2 Classification Accuracy using Subsets without Strong Disease Association

Subsets of 200 SNPs not showing strong association with the ER-negative phenotype SVM classifiers were selected to test whether the high classification accuracy, observed using strongly associated SNPs, can be replicated using SNPs without strong phenotype association. From the ranked list of SNPs the lowest ranked 200 that were fully genotyped were selected as one subset and 200 other SNPs were randomly selected to produce the other subset. SVM models were built from both subsets and classification accuracy was poor in both cases at less than 69% (Table 2.7 and Table 2.8). Less than one third of ER-positive cases were correctly classified using either subset of SNPs and the area under ROC curve values were ~0.5, which is equivalent to the placing of the hyperplane being completely random. Therefore, the high accuracy achieved using the top ranked SNPs is not a coincidence, suggesting that the top ranked SNPs do contain information that may be relevant for distinguishing between ER-positive and ER-negative breast cancer cases.

Table 2.7. Classification results for 200 SNPs with no disease association, genotyped in all samples

Kernel type	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	58.17	0.32	0.30	0.70	0.68	0.51
NQP	68.63	0.00	0.00	1.00	1.00	0.50
Quadratic polynomial	57.47	0.31	0.31	0.69	0.67	0.51
Cubic polynomial	58.42	0.29	0.29	0.71	0.69	0.51
RBF	68.63	0.00	0.00	1.00	1.00	0.50

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

Table 2.8. Classification results for 200 SNPs with varying disease association, genotyped in all samples

Kernel type	Classification accuracy (%)	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC curve
Linear	66.75	0.00	0.03	0.97	1.00	0.49
NQP	68.63	0.00	0.00	1.00	1.00	0.50
Quadratic polynomial	37.37	0.00	0.46	0.54	1.00	0.27
Cubic polynomial	40.46	0.00	0.41	0.59	1.00	0.30
RBF	68.63	0.00	0.00	1.00	1.00	0.50

Classification accuracy – percentage of cases classified correctly; True positive rate – proportion of ER-positive case classified correctly; False positive rate – proportion of ER-positive case classified incorrectly; True negative rate – proportion of ER-negative case classified correctly; False negative rate – proportion of ER-negative case classified incorrectly

2.4.2.3 SNP Feature Weights

SNP weights were obtained from a linear SVM model for the final set of 200 SNPs (tier 2 analysis) to assess the importance of each SNP feature in the ER subtype classification (Appendix I). The linear model assigned a weight value to each of the 200 SNPs, implying all SNPs were involved in producing a linear classification model and were also likely to all be involved in producing the other non-linear kernel models. Feature weights can be either positive or negative depending upon which class the feature is more associated with. Taking the absolute values of the weights, those SNPs with the largest weights were those that were most discriminating in the linear classifier.

The chi-squared association test was used to rank all genotyped SNPs based on association with the ER-negative phenotype. Comparison of the chi-squared values for each individual SNP and the weights derived from the linear classification model revealed that there was no correlation between the two values; Pearson's correlation coefficient of $r = -0.026$ (Figure 2.1). Therefore, the SNPs identified as most strongly associated with the ER-negative phenotype were not the same SNPs that were most important in distinguishing between the ER subtypes in a linear classification model.

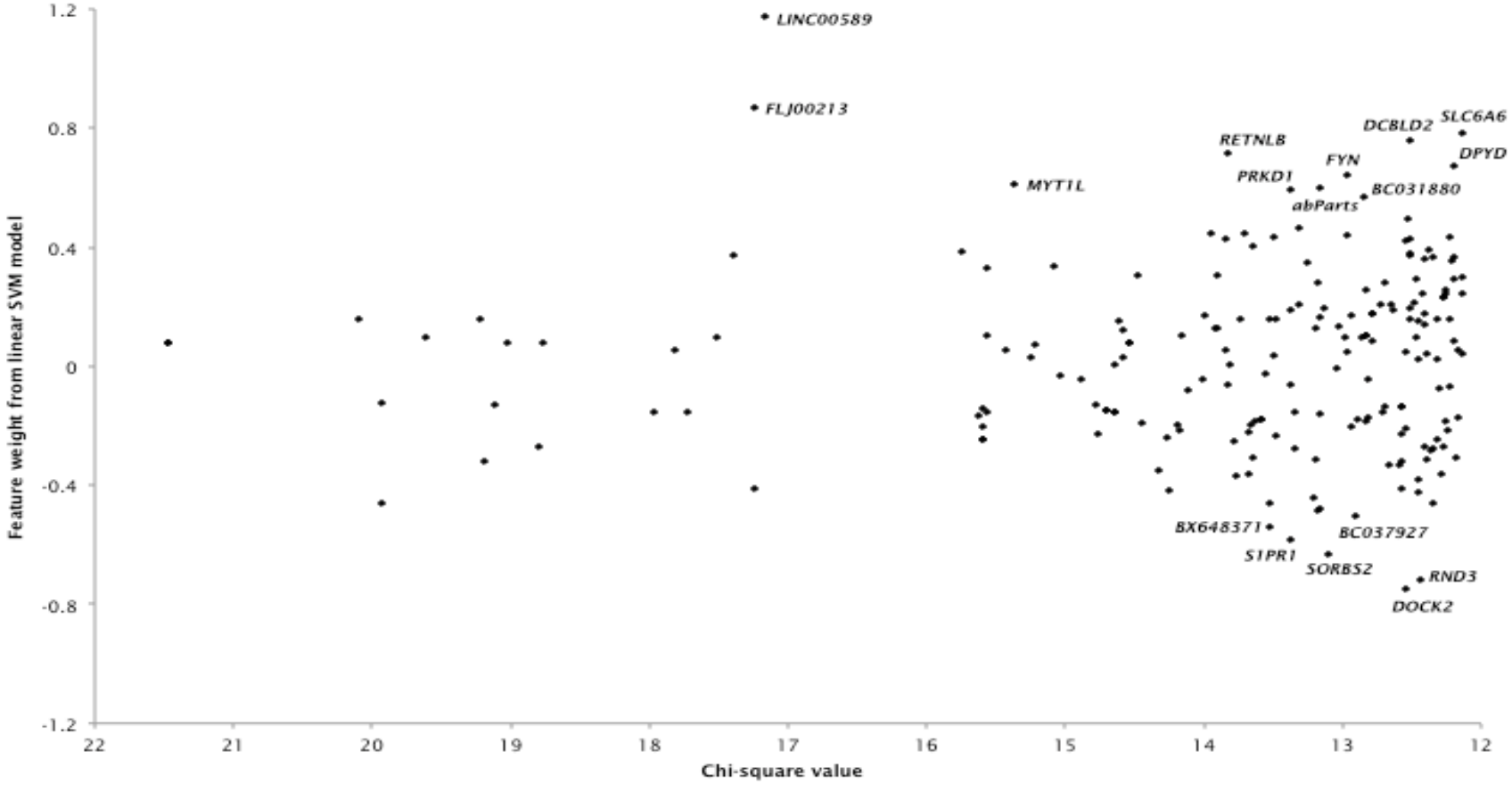


Figure 2.1. Relationship between chi-square value and SNP feature weight. Feature weights were obtained from a linear classification model. SNPs with an absolute feature weight > 0.5 are labelled with the gene in which they are located/closest to.

2.4.2.4 DAVID Functional Annotation

To identify biological terms and pathways that are particularly enriched for the 139 genes represented by the set of 200 SNPs used in the tier 2 classification model (Appendix I), the DAVID tool was used. DAVID identified four gene annotation clusters, three enriched pathways and 36 term annotation clusters from the 139 genes. Of these, two gene annotation clusters and 9 term annotation clusters were particularly enriched (enrichment score > 1.00) relative to the whole genome background. The cluster with the highest enrichment score contains genes related to the inflammatory response (Table 2.9). Pathways identified as enriched for genes in the gene set included axon guidance and signalling (Table 2.10).

Table 2.9. DAVID annotation cluster (enrichment score 1.97) showing enrichment for the inflammatory response

Annotation term	Number of genes identified from gene set	Percentage of total gene set	P-value	Fold enrichment
calcium ion transport	7	6.03	0.00018	8.64
T cell proliferation	4	3.45	0.00051	25.05
di-, tri-valent inorganic cation transport	7	6.03	0.00056	6.73
T cell activation	6	5.17	0.00084	8.05
lymphocyte proliferation	4	3.45	0.00187	16.10
leukocyte proliferation	4	3.45	0.00214	15.37
mononuclear cell proliferation	4	3.45	0.00214	15.37
lymphocyte activation	6	5.17	0.00613	5.09
positive regulation of immune system process	6	5.17	0.01269	4.26
leukocyte activation	6	5.17	0.01356	4.19
cell proliferation	8	6.90	0.01360	3.10
response to abiotic stimulus	7	6.03	0.02048	3.22
cell activation	6	5.17	0.02619	3.54

Table 2.10. Enriched KEGG pathways identified by DAVID

KEGG pathway	Genes implicated in pathway	p-value
Axon guidance	*EPHA4*, *FYN*, *NRP1*, *NTN4*, *PPP3CA*	0.007
T cell receptor signalling pathway	*FYN*, *IL5*, *PPP3CA*, *PTPRC*	0.027
Fc epsilon RI signalling pathway	*FYN*, *IL5*, *MAP2K4*	0.081

Table 2.11. DAVID annotation cluster (enrichment score 1.49) showing enrichment for the inflammatory response from 100 SNPs with largest absolute weight values

Annotation term	Number of genes identified from gene set	Percentage of total gene set	P-value	Fold enrichment
T cell activation	4	5.80	0.00842	9.34
lymphocyte activation	4	5.80	0.02830	5.91
positive regulation of immune system process	4	5.80	0.04442	4.94
leukocyte activation	4	5.80	0.04629	4.86
cell proliferation	4	5.80	0.06987	4.10

Reducing the gene set to include only those genes represented by the 100 SNPs with the highest absolute weight values identified similar annotation clusters. The gene ontology (GO) term cluster with highest enrichment score (1.49) related to immune cell activation (Table 2.11), and other annotation clusters resembled those implicated from the full gene set. These 100 SNPs were most important in the classification model and are, therefore, potentially the SNPs that are most important in discriminating between the ER cancer subtypes in this sample. A smaller gene set was represented by these SNPs yet very similar annotation terms were identified, suggesting that these annotation terms are likely to be relevant in the development of one of the two ER cancer subtypes.

2.5 Discussion

Common SNP variants can be disease risk variants, with breast cancer risk SNPs identified from GWA studies (Ghoussaini et al., 2013). Detecting risk SNPs

through association testing within a GWAS has had limited success in resolving the missing heritability issue. Novel approaches to analysing SNP data, such as ML techniques, are necessary for the discovery of underlying disease mechanisms, which may include gene-gene interactions or pathway/gene enrichment. Support vector machines have become recognised as state-of-the-art classification techniques with demonstrated success in genetic studies investigating genetic differences between populations based on SNP data (Listgarten et al., 2004, Waddell et al., 2005, Ban et al., 2010). SVMs are designed for binary classification problems and are thus useful for distinguishing between samples with two distinct disease phenotypes, such as ER-positive and ER-negative breast cancer, which have individual biologies.

Missing data values is a common problem in genetic studies. This is true of genome-wide SNP studies because the SNP genotyping process is imperfect, meaning that genotype data is often not available for all SNPs in all samples. The SVM algorithm is not suited to deal with missing feature values so it is necessary to deal with the missing data prior to model building. Quality control measures were applied to exclude any SNPs with a genotyping rate of less than 95%. Further pre-processing was done to handle missing data values by using three common approaches: 1) remove any features that have missing values for any sample; 2) remove any samples that do not have values for all features; 3) use imputation to replace the missing values (Wall and Elser). Each method will produce slightly different results.

Models were built from subsets of the SNP data using the three approaches outlined above to investigate how the results compared. Discounting any SNP features containing missing values meant that 36 of the 200 highest ranked SNPs were removed from consideration; this included the SNP identified as most disease-associated by the chi-square association test. It is highly likely that important information is contained within these features and this information will be lost from the model, producing an inferior classifier. However, this approach is preferable to removing samples. The amount of genetic variation available to the classifier will be decreased for every sample that is removed, ultimately reducing the model's ability to classify unseen samples correctly. Quality control measures ensured that all samples had a very high genotyping rate, however, there were 73 samples that were incompletely genotyped for the 200 top ranked SNPs. Removal of these

samples reduced the sample size by over 13%. The final approach, using imputation to replace the missing values is perhaps the most preferable approach. As previously stated, SNPs were only included in the full dataset if they were genotyped in at least 95% of the samples, therefore the number of missing values per SNP is likely to be low. Consequently, it is preferable to impute these small numbers of missing values rather than discounting samples or features, which ultimately means removing potentially relevant data from the model, possibly resulting in a less accurate classifier. Missing genotype values were replaced using genotype imputation. Not all missing values could be resolved by this method so SNPs that were still incomplete after imputation were removed and replaced with the next most disease-associated, fully genotyped SNPs.

The results produced from classifiers built using the three different approaches are remarkably similar (Table 2.1, Table 2.3 and Table 2.5), with over 93% classification accuracy for all kernel models (except Puk) in all cases. Of the three approaches used the most successful classifier is produced using 200 SNPs genotyped in all 542 cases but using no genotype imputation (Table 2.3). The least accurate models are produced from the 200 highest ranked SNPs genotyped in 469 samples (Table 2.1). However, the difference between the models produced using the three different approaches is marginal.

For each approach to the missing data, a slightly different set of 200 SNPs was selected, with 161 SNPs common to all 3 SNP subsets. Yet the choice of SNPs seems to make little difference to the overall classification accuracy; accuracy is extremely high and almost identical in all cases. This suggests that a large number of SNPs are involved in the phenotype distinction and it is likely that not all important variation has been captured using any of these SNP sets. Information obtained on SNP weight values from the linear SVM model supports this suggestion; comparing weight values to the chi-square values for each SNP indicates that there is no correlation between the ranking provided by the chi-square association test and the importance of the SNP in the subsequent model (Figure 2.1). Therefore, it is unlikely that the feature selection method used here will actually identify all relevant variation; indeed it appears from Figure 2.1 that there will be more SNPs with lower chi-square values but high absolute weight values that may been important for classification. The choice of 200 SNPs was an arbitrary cut-off point to select a subset of SNPs

though further exploration of larger SNP subsets of SNPs would be valuable, in order to assess at which point more SNP data does not improve classification accuracy.

Feature selection is an important component of building a machine learning classifier. Many of the SNPs genotyped in these samples will have little use for discriminating between the ER cancer subtypes so it was necessary to select a subset of potentially important SNP features from which to build a classifier. Feature selection prior to SVM implementation is essential to avoid the 'curse of dimensionality', which tends to arise from training of too few examples with too many variables (Cruz and Wishart, 2006). A major problem of many feature selection methods for SVMs is that they assess feature importance on an individual basis, which will exclude some important features because they do not satisfy the selection criteria. In this study the chi-squared association test was used to rank SNPs based on their individual association with the ER-negative phenotype, with the top ranked SNPs those with strongest association. Based on results of association testing, in which few SNPs with strong individual disease association have been identified or replicated, it is highly likely that some particularly important SNPs will not show individual association in the phenotype, however they are very important in the presence of other SNP variants. In fact, by comparing the chi-square values for 200 highly ranked SNPs with the weights produced by the linear model, the SNPs predicted to be most disease associated are not the SNPs that appear to have the most discriminatory power in the model (Figure 2.1). It is likely that the feature selection method used here is not the most appropriate approach, but knowing which method is most suitable is virtually impossible.

Another potential issue relating to the feature selection method used here is that all 542 samples were used to test the association of each SNP with the phenotype. Therefore, samples that are being used to test the model were involved in the original selection of the SNPs, which may introduce a bias. To overcome this issue it would be more appropriate to split the samples into a training set and a test set and use only the training set to select the SNP features, thus the samples in the test set have no influence on feature selection and are completely independent.

The success of ML classifiers is assessed through the classification accuracy measure of the number of test set samples correctly classified. This measure can, however, be misleading, particularly if there are unequal numbers of samples in the two classes (Ben-Hur and Weston, 2010). This is the case in this study; the ER-negative group is over twice as large as the ER-positive group. Other measures, such as the number of cases correctly classified into each group and the area under ROC curve (AUC) values should also be considered. When presented with an unbalanced dataset it is often the case that a majority class classification model is produced in which all cases are classified as members of the majority class (Ben-Hur and Weston, 2010), making the classifier appear more accurate than in reality. For example, using all 542 samples, an accuracy of 68.6% can be achieved by simply classifying all 542 samples into the larger ER-negative group, giving the impression that the classifier is correctly identifying a reasonable proportion of the samples. True positive and true negative rates indicate the proportion of samples from each class that are correctly classified. If a majority class classifier is produced, the true positive and negative rates of 0.00 and 1.00 respectively indicate that this is not a valid classification model. For all classification models, TPR and TNR values were obtained to ensure that the models were valid (Table 2.1, Table 2.3 and Table 2.5). Majority class classifiers were produced when the Puk kernel was used in tier 1 analysis (Table 2.1, Table 2.2) indicating that the Puk kernel produces poor classification models for these data. In all other cases majority class classifiers were not produced, therefore, it is reasonable to conclude that these SVM models are successful as ER-positive/-negative classifiers and are suitable as classifiers for unseen data. It is evident in all cases that the ER-negative samples are classified more accurately than ER-positive samples, which is likely to reflect the unbalanced data. The greater difficulty in classifying ER-positive cases arises from the more limited variation in the SNP profile given the smaller number of cases available to the classifier.

Classifier performance can be further evaluated using receiver operator characteristic (ROC) curves that are based on the true positive and true negative rates at several different thresholds. One of the major advantages of the ROC curve is that it is unaffected by unbalanced datasets (Fawcett, 2006). The area under ROC curve (AUC) measure (Bradley, 1997) takes values between 0.00 and 1.00 with values closer to 1.00 indicating good performance. A

random classification would produce an AUC of 0.5. AUC values were obtained for all ER-positive/negative classifiers (Table 2.1, Table 2.3 and Table 2.5) and for all models except those produced using the Puk kernel, AUC values were in the range of 0.92-0.95, suggesting excellent classification ability.

The major aim of this work is to identify the underlying causal factors of the disease subtypes and to link this to reasonable biological assumptions. Ideally the model would be based on only a few features whose interaction or joint presence are causal of disease, however it is often the case that SVM models are based on a large number of the input features (Waddell et al., 2005), making it challenging to uncover any relevant or biologically meaningful results. Since weight values imply that all 200 SNPs are important in the classification models, there may be particular enrichment in certain pathways that influence which ER subtype develops in an individual. Exploration of gene enrichment identified a number of annotation clusters and pathways that may be important in the ER-positive/negative distinction. The annotation cluster with the highest enrichment score contained genes with roles in immune/inflammatory cell activation, differentiation and proliferation, while there also seems to be enrichment for genes with a role in the T-cell receptor (TCR) signalling pathway. This suggests that one of the distinctions between ER-positive and ER-negative tumours relates to genetic variation in immune system pathways.

The role of the immune/inflammatory response in influencing tumourigenesis and tumour progression, through the formation of an inflammatory microenvironment at the tumour site, is well characterised (Ménard et al., 1997, Chen et al., 2007, Mantovani et al., 2008, Grivennikov et al., 2010). Immune system cells such as tumour-associated macrophages (TAMs) and tumour-infiltrating lymphocytes (TILs), which establish the tumour microenvironment, could account for almost 50% of breast tumour volume (Reed and Purohit, 1997). Immune cells that infiltrate the tumour have the capacity to release signalling molecules and bioactive mediators, making them a likely major source of pro-tumourigenic factors at the tumour site. Furthermore, infiltrating immune cells regulate a number of processes, including enhanced cell survival, angiogenesis and suppression of antitumour immune responses (DeNardo and Coussens, 2007) suggesting a role in both tumour development and progression. Leukocytes infiltrating breast tumours

may be a major source of estrogen expression in breast tumours (Mor et al., 1998) which could contribute to disease development and progression, while TAMs have been implicated as a source of mitogenic signals for tumour cells through cytokine secretion (Ch'ng et al., 2013) potentially enhancing cell division and tumour growth.

Infiltrating lymphocytes, which are often T cells (Whitford et al., 1990, Mahmoud et al., 2011), appear to have an important role in breast cancer patient survival. Evidence from several studies suggests that high levels of lymphocyte infiltration are associated with an improved prognosis in ER-negative disease while the opposite appears to be true for ER-positive tumours (Mahmoud et al., 2011, Teschendorff et al., 2007, Calabrò et al., 2009). Conflicting evidence was presented in (Curtis et al., 2012) in which tumour subgroups were created based on gene expression and copy number variant profiles; one particular subgroup with good prognosis contained many samples, both ER-positive and ER-negative, that exhibited extensive lymphocyte infiltration. Patients presenting with disease before the age of 40 also appear to have better survival rates if there are high levels of lymphocyte infiltration (Ménard et al., 1997).

In contrast, TAM levels in breast tumours appear to positively correlate with aggressiveness of disease and poor prognosis (Leek et al., 1996, Solinas et al., 2009). TAMs secrete mitogenic cytokines that stimulate tumour growth, and angiogenic cytokines that stimulate angiogenesis, which is vital for tumour survival (Leek and Harris, 2002, Ch'ng et al., 2013). In order for tumours to become malignant an ‘angiogenic switch’, which sets up a vascular network, is required (Lin et al., 2006). TAMs promote the development of this vascular network and an increased infiltration of macrophages into the tumour stroma is observed shortly before the development of vasculature (Lin et al., 2006).

Gene enrichment analysis also identified five genes implicated in the ‘axon guidance’ pathway (Table 2.10). Axon guidance molecules are important in the mammary gland for maintaining normal cell proliferation and adhesion during tissue development (Harburg and Hinck, 2011) and the proximity of nerves and blood vessels in a number of tissues suggests that there may be molecular cross-talk and common cues between these structures (Klagsbrun and Eichmann, 2005). Dysregulation of these guidance molecules in the mammary

gland has been linked to breast cancer initiation and progression (Harburg and Hinck, 2011). One of the genes involved in this pathway, *NTN4*, has been identified as a predictor of overall breast cancer patient survival and high levels of *NTN4* are often found in ER-positive tumours (Esseghir et al., 2007).

Genome wide association studies have identified risk-related SNPs for many diseases. Thirty-five SNPs, which lie in or near to 36 genes, are identified as breast cancer risk SNPs in the Catalog of Published Genome-Wide Association Studies (www.genome.gov/gwastudies) (Hindorff). The final set of 200 SNPs identified in this study as related to the ER-positive/ER-negative tumour sub-type distinction was compared to the list of SNPs; none of the 35 risk SNPs is present in this list. Nor are any of the 200 SNPs in or near the 36 catalogued genes. Therefore, the SNPs identified in this study represent a set of genes that have not yet been conclusively linked to breast cancer risk.

2.6 Conclusion

ER-positive and ER-negative breast tumour subtypes are likely to have distinct genetic profiles, with common and rare variants contributing to disease risk. A small subset of 200 common SNP variants, representing 139 genes, was used to produce SVM classification models. Classification accuracy of over 90% was achieved, indicating that the selected SNPs can be used to discriminate between ER tumour subtypes in early-onset breast cancer patients with a high degree of accuracy. Due to the polygenic nature of complex phenotypes many more genes will be involved in this distinction. Indeed, it is clear from the results presented that the different SNP subsets produce similar classification accuracy, suggesting that the 139 genes identified are not a definitive list and further important variation has likely been overlooked. The success achieved using the SVM approach with distinct breast cancer subtypes is encouraging and opens up the possibility of extending this work to look at the underlying genetic profile of other breast cancer subtypes as well as investigating differences between breast cancer and non-breast cancer individuals.

Chapter 3: Detecting SNP-SNP Interactions Underlying Early-onset Breast Cancer

3.1 Background

Genome-wide association studies of common SNP variants have been unable to explain a high proportion of the missing heritability in the majority of diseases (Eichler et al., 2010, Manolio et al., 2009). Epistatic interactions between genes are suggested to be a potential explanation for some of this missing heritability and it is likely that gene-gene interactions are actually ubiquitous in human disease (Moore, 2003).

GWA studies provide a wealth of data for further analysis beyond the classical association testing. It is likely that common variants with small individual effect sizes interact additively or multiplicatively, contributing to disease phenotypes (Pritchard and Cox, 2002, Reich and Lander, 2001). Logistic regression analysis is a potential approach to the detection of statistical interaction (Cordell, 2009). Through inclusion of terms for each predictor variable (e.g. SNP) individually and a term for interaction of the variables, the model can test whether the interaction term is equal to zero. Logistic regression assumes that the relationship between each predictor variable and the outcome variable (e.g. phenotype) is linear, which will not necessarily be true for genetic disorders since the underlying structure of interactions is usually unknown (Cordell, 2009).

Epistasis is a complex phenomenon that may involve SNPs with detectable main effects, SNPs with no detectable independent effects, or a combination of SNPs with and without main effects. To assess all possible situations, including SNPs with and without main effects, exhaustive search is often considered (Cordell, 2009). On a genome-wide scale however, exhaustive search of all SNP-SNP pairs rapidly becomes computationally demanding as the number of tests required is in the order of thousands of millions. Furthermore, when higher-order interactions are also considered the problem is further exacerbated. Based on this, a form of feature selection may be necessary to reduce the initial search space. Another approach is to utilise knowledge of pathways and mechanisms with involvement in the disease phenotype as a means to select

potentially relevant variables (Moore et al., 2010) since certain pathways are often perturbed in certain diseases. For example, the highly penetrant breast cancer genes *BRCA1* and *BRCA2* both have roles in DNA repair (Yoshida and Miki, 2004) and disruption of gene function can lead to disease. Therefore, it is possible that multiple common variants in certain pathways, when present in specific combinations, will cause disease. Thus, a pathway or gene-based approach to SNP selection is a valid method for feature selection to reduce the search space in epistasis detection studies.

3.2 Aim

To investigate potential SNP-SNP interactions between breast cancer susceptibility loci in a cohort of early-onset breast cancer patients and non-disease control samples, using a variety of machine learning algorithms and statistical tests.

3.3 Materials and Methods

3.3.1 Data and Data Processing

SNP genotypes were available for all breast cancer case and non-disease control samples. Breast cancer cases comprised of 574 patients from the 'Prospective study of Outcome in Sporadic versus Hereditary breast cancer' (POSH) cohort of ~3000 women diagnosed with breast cancer before age 40 (Eccles et al., 2007). Non-disease controls comprised of 5200 Wellcome Trust Case Control Consortium (WTCCC) samples from the 1958 British Birth Cohort (1958BC) and the UK National Blood Service Collection (NBS).

Breast cancer cases were genotyped using the Illumina 660-Quad SNP array in two batches at separate locations; Mayo Clinic, Rochester, Minnesota, USA (274 samples) and Genome Institute of Singapore, National University of Singapore (300 samples) (Rafiq et al., 2013). Only 536 samples passing quality control metrics were considered in the analysis. All WTCCC controls were genotyped using the Illumina 1.2M SNP array.

Only SNP data for the autosomes was considered and all SNPs were subjected to quality control filters. SNPs were excluded if they had a minor allele

frequency < 1%, a genotyping call rate < 95%, or showed significant deviation from Hardy-Weinberg equilibrium ($p < 0.0001$). All samples had a SNP genotyping rate > 95%. All quality control checks on the data were implemented using the PLINK package (Purcell et al., 2007). Post-QC there were ~491,000 SNPs genotyped in the POSH cohort, ~898,000 SNPs in the WTCCC NBS cohort, and ~896,000 SNPs in the WTCCC 1958BC cohort. The set of ~475,000 SNPs common to all three populations was considered in the analysis. All SNPs were annotated with respect to genes using ANNOVAR (Wang et al., 2010).

To exclude potential effects from linkage disequilibrium (LD) among SNPs, LD based pruning was applied to the SNP set using the PLINK package. A sliding window approach was used; all SNPs within a 50 SNP window were compared with one another to obtain an r^2 value of correlation. If the r^2 value for a pair of SNPs exceeded 0.5 one SNP from the pair was removed. LD based pruning reduced the SNP set to 216,015 SNPs.

To ensure all samples were ethnically uniform (Caucasian) multidimensional scaling (MDS) was implemented in PLINK. Breast cancer case and non-disease control samples were compared to reference populations from the HapMap Project (Gibbs et al., 2003); African (Yoruba in Ibadan, Nigeria; YRI), Asian (Han Chinese in Beijing; CHB and Japanese in Tokyo; JPT) and Caucasian (Utah residents with northern and western European ancestry; CEU). MDS was applied to a set of ~133,000 SNPs common to all five populations. Eight POSH samples were outliers and thus showed evidence of ethnic admixture (Appendix II). These samples were removed from further consideration, leaving 528 POSH samples all showing evidence of Caucasian ethnicity (Appendix III).

Estrogen receptor (ER) status was determined for all 528 remaining breast cancer samples. Samples with high expression of ER were designated ER-positive (161 samples) while samples with low or no expression of ER were designated ER-negative (366 samples). ER status could not be determined for one sample, therefore this sample was removed from consideration in ER-specific analysis.

All SNP genotypes for all case and control samples were transformed into numerical coding based on the number of alternate alleles: 0 for reference homozygous SNPs, 1 for heterozygous, and 2 for alternate homozygous SNPs.

All breast cancer cases were given a phenotype label of 1 and all WTCCC controls a label of 0.

3.3.2 SNP Marker selection

To reduce the search space of potential SNP-SNP interactions a subset of SNPs located within or close to 87 breast cancer genes (Appendix IV) was selected. All 87 genes had shown evidence of association with breast cancer through linkage mapping or GWA studies (Stratton and Rahman, 2008, Ghoussaini et al., 2013). Official gene names were obtained from the HUGO Gene Nomenclature Committee database (http://www.genenames.org/). A total of 1980 SNPs from 78 genes were selected. Using previous gene names and/or synonyms obtained from HUGO a further 5 SNPs were identified in 5 genes.

Seventy-two loci have been shown to be associated with breast cancer (Ghoussaini et al., 2013), including 80 specific SNPs. Fourteen of these SNPs are included in the 1985 selected SNPs. The pre-LD pruned set of ~500,000 SNPs was interrogated to select all of these SNPs that were genotyped in the samples but were filtered out during LD pruning. A further 30 of these SNPs were identified and were included in the analysis, giving a final total of 2015 SNPs in 91 genes, of which 49 have been previously linked to breast cancer through GWAS (Appendix V).

3.3.3 Single SNP Analysis

Genome-wide association p-values were calculated for each of the 2015 SNPs to identify any SNPs with potential genome-wide effects. SNP association with the breast cancer phenotype was assessed using the basic chi-squared association test in PLINK.

3.3.4 Interaction Analysis

Potential SNP-SNP interactions were detected by testing the genotypes of SNP pairs in cases versus controls using five methods: i) PLINK's 'fast epistasis' test; ii) PLINK's 'case-only' test; iii) Model-Based Multifactor Dimensionality Reduction; iv) SNPRuler; and v) SNPHarvester.

Each model was applied to three partitions of the data: all controls and all cases (5728 samples), all controls and ER-positive only cases (5361), all controls and ER-negative only cases (5566).

3.3.4.1 'Fast-epistasis' and 'Case-only' Tests in PLINK

The 'fast-epistasis' test in PLINK is based on a Z-score for the difference in SNP-SNP association (odds ratio) between cases and controls. In the same way as an allelic test in a single locus is constructed, the three genotype categories of each SNP are collapsed into two allele categories. For example, consider the following genotype distribution of two SNPs:

		SNP1		
		AA	*Aa*	*aa*
SNP2	*BB*	a	b	c
	Bb	d	e	f
	bb	g	h	i

First count the number of *A* and *a* alleles at SNP1 conditional on the genotypes at SNP2:

		SNP1	
		A	*a*
SNP2	*BB*	2a+b	2c+b
	Bb	2d+e	2f+e
	Bb	2g+h	2i+h

Then the number of *B* and *b* alleles at SNP2:

		SNP1	
		A	*a*
SNP2	*B*	4a+2b+2d+e	4c+2b+2f+e
	b	4g+2h+2d+e	4i+2h+2f+e

The odds ratio between SNP1 and SNP2 can be calculated. The procedure outlined above is applied to both cases and controls separately. The test for epistasis is then the difference between the odds ratios for the two SNPs:

$$Z = (\log(X) - \log(Y)) / \sqrt{(se(X) + se(Y))}$$

Where X and Y represent the odds ratios for the cases and controls respectively and se(X) and se(Y) represent the standard error of the corresponding odds ratio. In practice, this test will produce very similar results to the standard logistic regression test but is much more efficient.

The 'case-only' test is based on the fact that, under some conditions, an interaction term in logistic regression depends on relevant predictor variables within the cases. The test detects correlation between the genotypes or alleles of two SNPs within the cases. Implementation of the test is through a χ^2 test of independence between the genotypes or alleles (Cordell, 2009). Although this method is considered to be a more powerful approach than testing cases versus controls, there are certain assumptions that the test makes that, if they do not hold, will lead to the identification of false-positives. PLINK's case-only test assumes that the SNP pairs being considered are in linkage equilibrium in the general population, and, as a means to ensure that LD effects are not detected, only SNPs that are > 1 Mb apart are considered in the analysis.

3.3.4.2 Model-Based Multifactor Dimensionality Reduction

MB-MDR is an extension of the multifactor dimensionality reduction (MDR) algorithm. MDR aims to reduce dimensionality by pooling genotypes for multiple SNPs into two groups based on the genotype distribution in cases and controls. For each pair of SNPs there are 9 possible genotypes (Figure 3.1). MDR assigns a case-control ratio value to each genotype and designates the genotype as either high- or low-risk based on this ratio value, thus reducing the problem to only one dimension – i.e. one variable with two classes. Each SNP pair is used as a predictor of phenotype and the SNP pair with the highest predictive accuracy is selected.

MB-MDR instead tests for association of each of the 9 genotypes of a SNP pair with the phenotype, through a χ^2 test with 1 d.f. In theory, each association test statistic T_i can be positive or negative, dependent upon the direction of the effect. However, the χ^2 test is always positive so it is assumed that T_i is equal to the square root of the χ^2 test, with $T_i > 0$ if $OR_i > 1$ and $T_i < 0$ if $OR_i < 1$, where OR_i is the derived odds ratio. The p-values p_i for the association tests T_i for each genotype are compared to a reference p-value, p_r, usually $p_r = 0.1$.

Genotypes with (i) $p_i < p_r$ and $T_i > 0$ are labelled 'high-risk', (ii) $p_i < p_r$ and $T_i < 0$ are labelled 'low-risk', (iii) $p_i > p_r$ are labelled 'no evidence'. Thus, a single variable with three possible classes is established, which can be further tested for association with the phenotype.

MB-MDR v4.0.1 was implemented in Unix using open source code (available from http://www.statgen.ulg.ac.be/software.html). Within MB-MDR there is the option to apply a model of interaction to the data: no model assumed, additive model assumed, or co-dominant model assumed. Since the underlying structure of potential SNP-SNP interactions within this cohort is unknown MB-MDR was implemented using each of the three models. Collapsing the genotypes for a SNP pair into one variable leads to a simplified model with inflated false positive rates (Cattaert et al., 2011). The p-values obtained from MB-MDR are therefore adjusted using the Westfall and Young step-down maxT method, which uses permutations to adjust the p-values.

3.3.4.3 SNPRuler

SNPRuler (Wan et al., 2010b) uses a predictive learning rule to identify potential epistatic interactions. The basis of the method is that each real interaction implicitly contains predictive rules that describe the relationship between a feature and the class label; finding and evaluating these predictive rules is quicker and easier than evaluating all possible interactions. For example, consider SNPs S_1 and S_2, with a genotype distribution described in Figure 3.1. A rule for these SNPs could be "if S_1 is *AA* and S_2 is *BB* (top-left cell in Figure 3.1) then the probability that the sample is unaffected is 0.7". There are a further three rules that could be detected in this particular example, but detecting only one of these rules is sufficient for identification of the interaction.

Predictive rule learning is utilised by the algorithm to select rules with high confidence and use these rules to identify potential interactions. SNPRuler begins by selecting one SNP and building a search tree by adding new SNPs. Each node of the tree represents a SNP with each edge representing a potential interaction. An upper bound is applied to each node to prune out any unnecessary expansion. Once the search tree is built, the possible interactions are evaluated through the chi-squared statistic and significant interactions are reported.

SNPRuler was implemented in Unix using the open-source code (available from http://bioinformatics.ust.hk/Software.html). Parameters implemented in the model were: the depth of the interaction, *depth* = 2; the step size of updating a rule, *updateRatio* = 0.1 (default is 0.2); the expected number of interactions, *listSize* = 10,000.

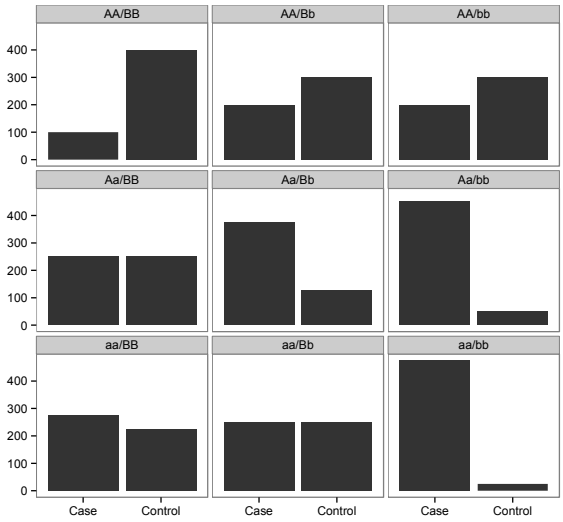


Figure 3.1. Illustration of the joint genotype distribution of two SNPs in cases and controls. Four genotypes show a strong contrast between cases and controls, suggestive of an interaction effect

3.3.4.4 SNPHarvester

SNPHarvester (Yang et al., 2009) is a stochastic search method that aims to identify SNP groups with a significant association with the phenotype from hundreds of thousands of genome-wide SNP markers. SNPHarvester works on the basis that SNPs are either (i) not involved in disease susceptibility; (ii) have an independent association with the disease; (iii) have a marginal/no independent association with disease but are involved in disease susceptibility jointly with other SNPs. SNPs in the second category are identified through GWAS while SNPHarvester aims to identify SNPs in the third category. Therefore, any SNPs with strong main effects, assessed using a χ^2 test with 2 d.f, are removed from the SNP set and not considered in the interaction analysis.

SNPHarvester employs a two-step process. In the first step, a group of k SNPs, $A = [S_1,\ldots,S_k]$, is randomly selected from the full set of SNPs and tested for

phenotype association using a χ^2 test with 3^k-1 d.f, giving χ^2_A. One SNP, S_1, is swapped out of the group and replaced by another SNP, $S_i \notin A$. The new group of SNPs, B, is tested for association with the phenotype. If $\chi^2_B > \chi^2_A$ then group B is retained, else group A is retained and SNP S_2 is replaced by S_i, and so forth until no further swaps are found that increase the χ^2 value. This process is repeated multiple times with different starting groups of SNPs, allowing for all local optima to be identified. Once a local optimum has been identified the corresponding SNPs are removed from consideration to ensure that the algorithm identifies all local optima. The second step then assesses all identified SNP groups using L_2 penalized regression to remove spurious results. SNP groups that pass an internally applied significance threshold are reported as potential interactions.

SNPHarvester was implemented in Eclipse Java EE IDE (https://www.eclipse.org/downloads/) using the open-source code (available from http://bioinformatics.ust.hk/SNPHarvester.html). Parameters included in the model were: the number of successive runs of the algorithm, *SuccessiveRun* = 20; the level of significance, $p = 0.05$; the number of SNPs per group, $k = 2$. SNPHarvester applies an internal Bonferroni corrected significance threshold, p_{corr}, based on p and only SNP groups passing this threshold are reported. In this case $p_{corr} = 2.46 \times 10^{-8}$. The SNP group size was set to 2 to identify only 2-way interactions.

3.3.4.5 Logistic Regression Analysis

Logistic regression analysis of all interactions identified by any of the five methods described above was implicated in PLINK using the standard 'epistasis' option. This fits an allelic-by-allelic model to the SNP pairs, with terms for the individual SNPs, a term for the interaction and a constant term:

$$Y \sim \alpha_0 + \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_1 S_2$$

where S_1 and S_2 are SNPs and α_0 is the constant term. The model tests whether the interaction term, α_3, is equal to zero.

3.4 Results

3.4.1 Single SNP Association Testing

A total of 2015 SNPs, identified as residing in or in close proximity to a total of 91 genes were selected for analysis. Single SNP association with the breast cancer phenotype was assessed using the association capability in PLINK. A basic chi-squared test was applied to each SNP to assess association with the disease phenotype. Using a Bonferroni correction, any SNPs with p-value < 2.48×10^{-5} are considered to be significantly associated with the phenotype. Of the 2015 SNPs, none reached significance (Figure 3.2A) indicating that none of these SNPs has a genome-wide main effect large enough to be detected in a sample of this size.

The most significant SNP was rs3757318, which lies in an intronic region of the *CCDC170* gene ($p = 8.56 \times 10^{-4}$; OR = 1.43). This particular SNP has a reported association with breast cancer with a p-value of 2.9×10^{-6} and an odds ratio of 1.30 in a GWA study including ~9000 samples (Turnbull et al., 2010).

To establish whether there were any SNPs strongly associated with estrogen receptor subtype within this cohort, the association test was applied to ER-positive cases versus disease-free controls and ER-negative cases versus disease-free controls. No SNPs were significantly associated with either subtype (Figure 3.2 B and C). The SNP with the lowest p-value for ER-positive disease was rs11865267 ($p = 2.51 \times 10^{-4}$; OR = 0.66). This SNP is more strongly significant than for breast cancer overall, for which it has $p = 8.51 \times 10^{-3}$ and OR = 0.84.

For ER-negative breast cancer the most significant SNP is rs3757318 ($p = 4.43 \times 10^{-4}$; OR = 1.54), the same SNP that is most significant for breast cancer overall. This SNP is in fact more strongly significant in ER-negative cases than for ER-positive and ER-negative cases considered together, suggesting that the association picked up for breast cancer is in fact influenced by the ER-negative cases. There are more than twice as many ER-negative cases in the cohort, which is likely to bias the results in favour of ER-negative specific associations when considering all cases as a single phenotype.

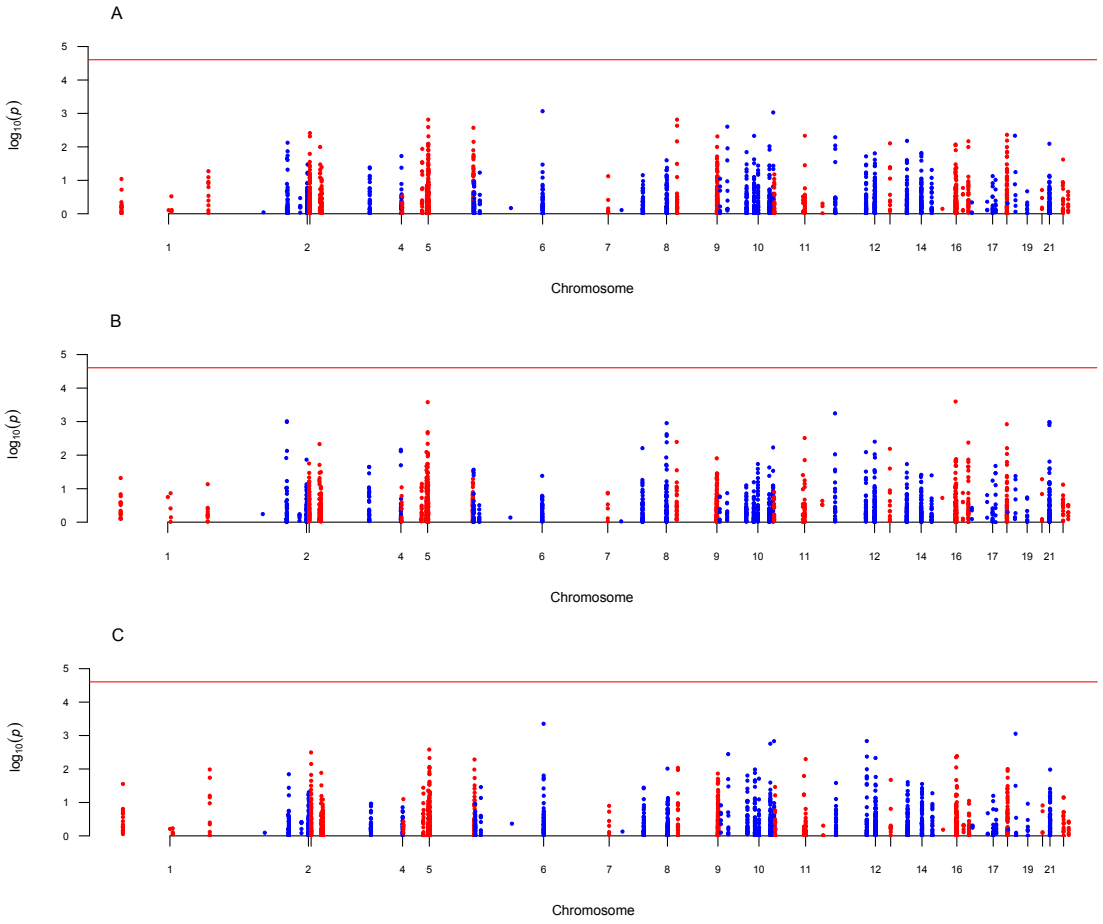


Figure 3.2. Manhattan plots for association of all 2015 SNPs with breast cancer phenotypes. All 2015 SNPs were tested for association with each of three phenotypes using a chi-squared test: (A) overall breast cancer; (B) estrogen receptor-positive breast cancer; (C) estrogen receptor-negative breast cancer. Horizontal red lines indicate the genome-wide significance threshold.

Considering ER subtype as distinct disease phenotypes demonstrates a difference between the associations of each SNP with phenotype (Figure 3.2 B and C), with strong signals in one phenotype not necessarily being present in the other. For example, the SNP identified as most strongly associated with ER-positive disease has $p = 0.37$ (OR = 0.9) in ER-negative cases and the SNP most strongly associated with ER-negative samples (and overall breast cancer) is not associated with ER-positive disease: $p = 0.37$ (OR = 1.2).

Considering only the 49 SNPs with reported genome-wide association with breast cancer from GWA studies (Appendix V), the significance threshold for association with breast cancer in this study is increased to 1.02×10^{-3}. At this significance level one SNP, rs3757318 in *CCDC170*, shows significant association with breast cancer ($p = 8.56 \times 10^{-4}$; OR = 1.43) (Figure 3.3A).

Considering ER-positive and ER-negative breast cancer subtypes as separate phenotypes identifies no SNPs with significant association to ER-positive disease (Figure 3.3B) but identifies two SNPs significantly associated with ER-negative disease: rs3757318 in *CCDC170* ($p = 4.43 \times 10^{-4}$; OR = 1.54) and rs8170 in *BABAM1* ($p = 8.81 \times 10^{-4}$, OR = 1.4) (Figure 3.3C). SNP rs8170 was reported associated with breast cancer risk in carriers of *BRCA1* gene mutations with $p = 2.3 \times 10^{-9}$ (OR = 1.26) (Antoniou et al., 2010). Analysis of risk of breast cancer in the general population did not find an association for rs8170 with breast cancer overall but did identify a significant association with ER-negative breast cancer; $p = 0.0029$ (Antoniou et al., 2010).

As was evident in the larger set of SNPs, the SNP associations across the two ER subtypes are distinct. SNPs more strongly associated with one subtype are generally not strongly associated with the other subtype (Figure 3.2 B and C). This result supports reports in the literature that certain SNPs are more strongly associated with one ER subtype than the other (Haiman et al., 2011, Purrington et al., 2014, Stacey et al., 2007).

3.4.2 SNP-SNP Interaction Detection

3.4.2.1 Full Set of SNPs

To detect potential SNP-SNP interactions five methods were applied to the data, each using unique search strategies. A Bonferroni significance threshold of $p < 2.46 \times 10^{-8}$ was applied to all results to identify any potentially real interactions.

Two epistasis tests were implemented in the PLINK package: 'fast-epistasis' test and 'case-only' test. The fast-epistasis test produces very similar results to a standard allelic-by-allelic logistic regression but is a much more efficient test. Using a Z score based statistic the odds ratio for a SNP pair is calculated in both cases and controls and the difference between these ratios is calculated.

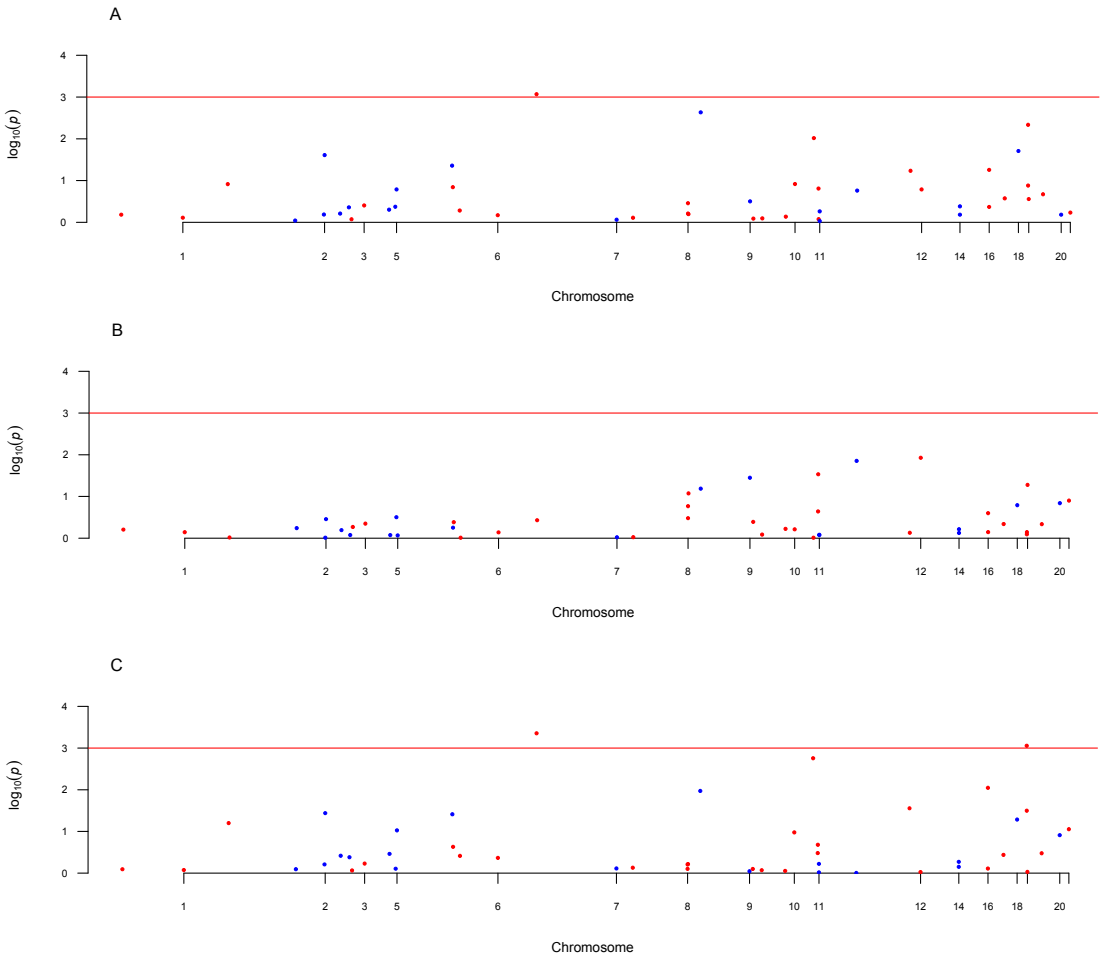


Figure 3.3. Manhattan plots for association of 49 SNPs associated with breast cancer from GWAS. All SNPs were tested for association with each of three phenotypes using a chi-squared test: (A) overall breast cancer; (B) estrogen receptor-positive breast cancer; (C) estrogen receptor-negative breast cancer. Horizontal red lines indicate the significance threshold.

Analysis of all SNP pairs using fast-epistasis failed to identify any interactions that were significant at $p < 2.46 \times 10^{-8}$. The most significant p-value was between two SNPs on chromosomes 6 and 16; $p = 1.83 \times 10^{-7}$.

Case-only analysis identified one significant SNP pair (Table 3.1). The interaction was, however, between two SNPs on chromosome 5 so could result from a genotype correlation due to LD. The SNPs are ~ 1.12 Mb apart and the r^2 value for LD is low ($r^2 = 0.076$), suggesting that the test is not simply identifying an LD effect between these SNPs.

Further analysis of potential SNP-SNP interactions was assessed using three algorithms specifically designed for the detection of epistasis in genome-wide SNP data; SNPHarvester, SNPRuler, and MB-MDR. All three algorithms use a chi-

square test to assess SNP pairs for potential interaction effects but all employ unique strategies to search for potential interactions. Each method identified many potential epistatic interactions, however, few were significant at $p < 2.46 \times 10^{-8}$ (Table 3.1). SNPRuler identified two such interactions; SNPHarvester identified only one, whilst MBMDR identified none.

Table 3.1. Significant interactions identified in breast cancer

SNP	Chr	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corr. p value	Met.
rs2330572	5	44740989	*MRPS30*	Intergenic	0.497	0.076	6.94×10^{-10}	1.41×10^{-3}	PCO
rs4289567	5	45869015	*HCN1*	Intergenic	0.172				
rs10073340	5	1321873	*CLPTM1L*	Intronic	0.993	7.48×10^{-7}	1.17×10^{-8}	0.024	SR
rs7815815	8	129265338	*MIR1208*	Intergenic	0.944				
rs9924319	16	52725857	*CASC16*	Intergenic	0.033	3.92×10^{-5}	1.53×10^{-8}	0.031	SH
rs17274750	21	16353809	*NRIP1*	Intronic	0.111				
rs6877743	5	58454191	*PDE4D*	Intronic	0.437	1.63×10^{-5}	1.83×10^{-8}	0.037	SR
rs2808375	9	109987123	*RAD23B*	Intergenic	0.023				

Chr – the chromosome on which the SNP is located; Base pair – the base pair location of the SNP in hg19 coordinates; Gene – the gene that the SNP is located within or in closest proximity to; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corr. p-value – Bonferroni corrected p-value; χ^2 p value – p-value for individual SNP association with phenotype, obtained from χ^2 test; Met. – the algorithm that identified the interaction: PCO – PLINK case-only; SR – SNPRuler; SH – SNPHarvester

Table 3.2. P-values for all significant interactions across all five methods and results from logistic regression analysis

SNP 1	SNP 2	PLINK 'fast-epistasis' p-value	PLINK case-only p-value	SNPHarvester p-value	SNPRuler p-value	MBMDR p-value	Logistic regression p-value
rs2330572	rs4289567	0.709	**6.94×10^{-10}**	.	.	.	0.700
rs10073340	rs7815815	0.830	0.822	7.55×10^{-8}	**1.17×10^{-8}**	1.00	0.824
rs9924319	rs17274750	0.275	0.231	**1.53×10^{-8}**	.	.	0.251
rs6877743	rs2808375	0.004	0.008	1.03×10^{-7}	**1.83×10^{-8}**	.	0.002

Significant p-values are indicated in bold

All SNP pairs identified as significant were further assessed using logistic regression, and p-values from all other methods were identified where possible (Table 3.2). Interestingly, SNPHarvester also detected both interactions identified by SNPRuler and although the associated p-values did not reach significance they were close to the significance threshold (Table 3.2).

The interaction between rs10073340 and rs7815815 was the third most significant interaction detected by SNPHarvester, while the interaction between rs6877743 and rs2808375 was the seventh most significant.

The interaction identified by the PLINK case-only test was tested in the fast-epistasis test but was not detected by any of the other algorithms. The p-value associated with this interaction from fast-epistasis was not significant, nor was the p-value when the SNP pair was assessed using logistic regression. The case-only test examines only genotypes within the cases, which provides a more powerful test, so when considering cases and controls together it is likely that the effect will not be detected.

Logistic regression analysis of the four significant interactions was significant for the interaction between SNPs rs6877743 and rs28018375 identified as significant by SNPRuler. This SNP pair was detected by all tests except MBMDR and the p-value assigned by SNPHarvester was close to the significance threshold. Both epistasis tests in PLINK assigned a p-value < 0.01 to this SNP pair. Therefore, this SNP pair may be a good candidate for a SNP-SNP interaction in breast cancer.

To identify any potential interactions specific to the ER-positive or ER-negative breast cancer subtypes the breast cancer cases were separated and the subtypes were individually analysed in comparison with the controls. In both cases, many more significant interactions were identified than in breast cancer overall.

For ER-positive breast cancer in excess of 400 interactions were identified across the five methods, with SNPRuler and SNPHarvester identifying the majority.

One significant interaction was identified through case-only analysis (Table 3.3) between SNPs on separate chromosomes. The significant result cannot be due to an LD effect and neither SNP has a strong main effect in ER-positive disease,

so it does not appear that a single SNP is driving the association effect observed.

In breast cancer overall this SNP pair has $p = 0.002$ from the case-only test while in ER-negative samples the result is $p = 0.972$. Therefore, it appears that there is a strong correlation between the genotypes for these two SNPs in ER-positive cases that is not present in ER-negative cases.

The distribution of the genotypes of this SNP pair in ER-negative cases versus ER-positive cases is illustrated in Figure 3.4. In ER-positive disease certain alleles or genotypes tend to appear together, for example, if a sample has at least one A allele for SNP rs7581219 then they tend to also have at least one G allele for SNP rs1346907 (top two rows of Figure 3.4). In ER-negative disease there is not such a clear correlation between certain alleles.

Although fast-epistasis does not identify any significant interactions the interaction identified by the case-only test was the most significant interaction tested by fast-epistasis, with $p = 4.53 \times 10^{-8}$, just above the significance threshold. Furthermore, logistic regression analysis of this interaction is highly significant with $p = 3.33 \times 10^{-8}$ and the odds ratio for interaction is 0.37.

Of the hundreds of interactions identified by SNPRuler and SNPHarvester, 47 were identified by both algorithms. Logistic regression analysis identified 20 of these interactions as significant at $p < 0.05$.

Table 3.3. Significant interactions identified in ER-positive breast cancer by the 'case-only' test

SNP	Chr	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corr. p value	Met.
rs7581219	2	174363574	*CDCA7*	Intergenic	0.118	7.42×10^{-4}	1.21×10^{-8}	0.025	PCO
rs1346907	3	30723470	*TGFBR2*	Intronic	0.584				

Chr – the chromosome on which the SNP is located; Base pair – the base pair location of the SNP in hg19 coordinates; Gene – the gene that the SNP is located within or in closest proximity to; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corr. p-value – Bonferroni corrected p-value; $\mathbf{X}^2$ p value – p-value for individual SNP association with phenotype, obtained from $\mathbf{X}^2$ test; Met. – the algorithm that identified the interaction: PCO – PLINK case-only; SR – SNPRuler; SH – SNPHarvester

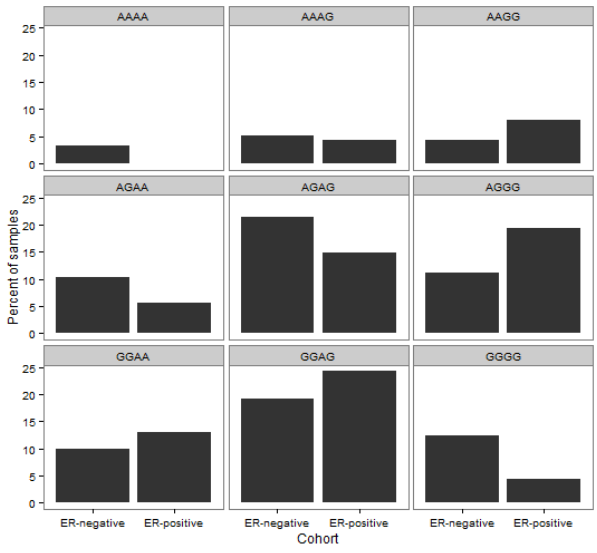


Figure 3.4. Genotype distribution of SNPs rs7581219 and rs1346907 in ER-negative and ER-positive disease. The percentage of samples from each cohort presenting with each of nine possible genotypes is depicted.

Thirty-one significant interactions were identified in ER-negative disease by only two of the five methods: SNPRuler and SNPHarvester (Table 3.4). One interaction was identified as significant by both algorithms; $p = 2.09 \times 10^{-8}$ and $p = 2.21 \times 10^{-8}$ for SNPHarvester and SNPRuler respectively (Table 3.4). This interaction is between SNPs on chromosomes 5 and 9, neither of which has a strong main effect that is likely to be influencing the result. One further significant interaction that was identified by SNPRuler, between SNPs rs11819509 and rs3731217, was also detected by SNPHarvester but with $p = 3.21 \times 10^{-5}$, which does not reach the significance threshold.

The most significant interaction identified in ER-negative samples was between rs1525608 and rs17601696 ($p = 3.89 \times 10^{-11}$) close to the genes *IGFBP2* and *FGFR2* respectively. Neither of these SNPs is strongly associated with the phenotype when considered individually; $p = 0.559$ and $p = 0.707$ respectively. Logistic regression analysis found that this interaction was significant; $p = 4.52 \times 10^{-3}$.

Table 3.4. Significant interactions in ER-negative breast cancer

SNP	Chromosome	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corrected p value	Method	Logistic regression p-value
rs1525608	2	217493755	*IGFBP2*	Intergenic	0.559	2.66×10^{-4}	3.89×10^{-11}	7.89×10^{-5}	SH	4.52×10^{-3}
rs17601696	10	123120036	*FGFR2*	Intergenic	0.707					
rs10073340	5	1321873	*CLPTM1L*	Intronic	0.567	1.20×10^{-5}	4.47×10^{-11}	9.07×10^{-5}	SH	0.324
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841					
rs10822036	10	64374360	*ZNF365*	Intronic	0.857	0.002	4.91×10^{-11}	9.96×10^{-5}	SR	6.74×10^{-3}
rs12413946	10	64431206	*ZNF365*	UTR3	0.090					
rs7852896	9	110361166	*KLF4*	Intergenic	0.175	3.74×10^{-5}	1.61×10^{-10}	3.27×10^{-4}	SR	0.102
rs606555	11	69384851	*CCND1*	Intergenic	0.742					
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841	2.55×10^{-5}	3.16×10^{-10}	6.41×10^{-4}	SH	0.012
rs2244814	14	69025870	*RAD51B*	Intronic	0.217					
rs610118	3	27336213	*NEK10*	Intronic	0.420	4.91×10^{-6}	5.43×10^{-10}	1.10×10^{-3}	SR	8.05×10^{-3}
rs17035162	4	105890551	*TET2*	Intergenic	0.713					
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841	2.19×10^{-5}	7.11×10^{-10}	1.44×10^{-3}	SH	0.179
rs9325542	10	114977369	*TCF7L2*	Intergenic	0.979					
rs3803662	16	52586341	*CASC16*	ncRNA exonic	0.009	1.37×10^{-5}	2.21×10^{-9}	4.49×10^{-3}	SH	0.011
rs8170	19	17389704	*BABAM1*	Exonic	0.001					
rs1423369	5	58115159	*RAB3C*	Intronic	0.605	1.79×10^{-5}	2.32×10^{-9}	4.72×10^{-3}	SH	0.017
rs2302674	16	54060594	*FTO*	Intronic	0.980					
rs6697258	1	120485335	*NOTCH2*	Intronic	0.595	6.82×10^{-5}	2.38×10^{-9}	4.83×10^{-3}	SR	0.178
rs10093823	8	76864260	*HNF4G*	Intergenic	0.687					
rs4685829	3	4872198	*ITPR1*	Intronic	0.113	7.47×10^{-6}	3.15×10^{-9}	6.39×10^{-3}	SH	0.015
rs9340941	6	152313146	*ESR1*	Intronic	0.246					
rs2888479	2	218139904	*DIRC3*	Intergenic	0.795	4.26×10^{-7}	3.22×10^{-9}	6.53×10^{-3}	SR	3.18×10^{-4}
rs1884219	14	37178991	*SLC25A21*	Intronic	0.701					
rs10979136	9	110725754	*KLF4*	Intergenic	0.094	4.99×10^{-4}	3.65×10^{-9}	7.41×10^{-3}	SR	0.573
rs16945643	17	59893990	*BRIP1*	Intronic	0.165					
rs1548942	2	217619036	*IGFBP5*	Intergenic	0.982	6.27×10^{-5}	6.12×10^{-9}	0.01	SH	1.80×10^{-3}
rs10896050	11	65577516	*OVOL1*	Intergenic	0.016					
rs7906302	10	115056097	*TCF7L2*	Intergenic	0.758	0.001	6.53×10^{-9}	0.01	SR	3.24×10^{-4}
rs17105965	14	69069594	*RAD51B*	Intergenic	0.592					
rs16861690	2	174306537	*CDCA7*	Intergenic	0.659	1.18×10^{-4}	9.08×10^{-9}	0.02	SR	0.063
rs13190249	5	58732661	*PDE4D*	Intronic	0.559					
rs4234019	2	217457187	*IGFBP2*	Intergenic	0.299	9.84×10^{-5}	9.57×10^{-9}	0.02	SH	0.042
rs12509636	4	106010433	*TET2*	Intergenic	0.180					
rs10816531	9	110321771	*KLF4*	Intergenic	0.099	2.03×10^{-4}	9.61×10^{-9}	0.02	SH	2.97×10^{-5}
rs405394	12	28240440	*PTHLH*	Intergenic	0.325					

rs1548942	2	217619036	*IGFBP5*	Intergenic	0.982	8.69×10^{-5}	1.00×10^{-8}	0.02	SH	4.00×10^{-3}
rs1105237	18	24994052	*CHST9*	Intergenic	0.018					
rs4645956	8	128750212	*MYC*	Intronic	0.443	2.69×10^{-5}	1.23×10^{-8}	0.03	SH	0.698
rs4262880	14	69071936	*RAD51B*	Intergenic	0.498					
rs6435979	2	218231649	*DIRC3*	ncRNA intronic	0.817	8.06×10^{-4}	1.34×10^{-8}	0.03	SH	7.80×10^{-3}
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841					
rs1105237	18	24994052	*CHST9*	Intergenic	0.018	1.72×10^{-4}	1.50×10^{-8}	0.03	SH	0.066
rs8170	19	17389704	*BABAM1*	Exonic	0.001					
rs11612530	12	96159659	*NTN4*	Intronic	0.004	1.31×10^{-5}	1.50×10^{-8}	0.03	SH	0.013
rs8170	19	17389704	*BABAM1*	Exonic	0.001					
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841	3.77×10^{-4}	1.58×10^{-8}	0.03	SH	0.517
rs2253168	14	69050465	*RAD51B*	Intronic	0.224					
rs12338922	9	110482025	*KLF4*	Intergenic	0.026	1.86×10^{-4}	1.76×10^{-8}	0.04	SH	0.039
rs7160841	14	37414154	*SLC25A21*	Intronic	0.469					
rs3731217	9	21984661	*CDKN2A*	Intronic	0.486	8.94×10^{-4}	1.95×10^{-8}	0.04	SR	0.592
rs11819509	10	115001714	*TCF7L2*	Intergenic	0.280		3.21×10^{-5}	1.00	SH	
rs726501	5	56127866	*MAP3K1*	Intronic	0.228	7.51×10^{-5}	2.03×10^{-8}	0.04	SR	0.104
rs12894505	14	91808042	*CCDC88C*	Intronic	0.053					
rs1391651	5	58376455	*PDE4D*	Intronic	0.782	4.87×10^{-6}	2.09×10^{-8}	0.04	SH	0.650
rs1981047	9	22173499	*CDKN2B-AS1*	Intergenic	0.161		2.21×10^{-8}		SR	
rs12493413	3	4877199	*ITPR1*	Intronic	0.301	4.78×10^{-5}	2.12×10^{-8}	0.04	SH	0.235
rs4262880	14	69071936	*RAD51B*	Intergenic	0.498					
rs7815815	8	129265338	*MIR1208*	Intergenic	0.841	1.95×10^{-4}	2.20×10^{-8}	0.04	SH	6.65×10^{-4}
rs1628089	10	80872270	*ZMIZ1*	Intronic	0.997					
rs10978761	9	109986827	*RAD23B*	Intergenic	0.068	6.64×10^{-6}	2.23×10^{-8}	0.05	SR	0.025
rs2244814	14	69025870	*RAD51B*	Intronic	0.217					

Base pair – the base pair location of the SNP in hg19 coordinates; Gene – the gene that the SNP is located within or in closest proximity to; ncRNA intronic/exonic refers to RNA without coding annotation – may or may not code for protein; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corrected p-value – Bonferroni corrected p-value; χ^2 p value – p-value for individual SNP association with phenotype, obtained from χ^2 test
Method – the algorithm that identified the interaction: SR – SNPRuler; SH – SNPHarvester

3.4.2.2 Forty-nine Breast Cancer-related SNPs

Due to the large number of tests required when considering all 2015 SNPs (> 2 million), the smaller subset of 49 SNPs with reported association with breast cancer, was also analysed for potential epistasis. Any SNP pair with an interaction p-value < 4.25×10^{-5}, detected by any method, was reported as significant.

Analysis of these 49 SNPs using all five methods yielded only two significant results, one from SNPHarvester and one from SNPRuler (Table 3.5). SNP rs3803662 in *CASC16* is implicated in both significant interactions. Results from logistic regression identify the interaction detected by SNPHarvester as significant.

Five potential ER subtype-specific interactions were identified; two ER-positive specific interactions (Table 3.6) and three ER-negative interactions (Table 3.7). Both interactions in ER-positive disease were detected by SNPRuler; the most highly significant result was between rs2363956 in *ANKLE1* and rs8170 in *BABAM1*; $p = 3.48 \times 10^{-10}$. Both *ANKLE1* and *BABAM1* lie on chromosome 19 and the two SNPs are located 4.42kb apart so it is likely that an LD effect may be influencing this result. Furthermore, the r^2 value for these two SNPs is only 0.239, implying that the SNPs may be in weak LD and this may be responsible for the result.

The three significant interactions identified in ER-negative disease were all identified by SNPHarvester and have $p < 4 \times 10^{-5}$ (Table 3.7). Four unique SNPs were involved, with rs3803662 in *CASC16* and rs8170 in *BABAM1* each

Table 3.5. Significant interactions between known susceptibility loci in breast cancer

SNP	Chromosome	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corrected p value	Method	Logistic regression p value
rs3803662	16	52586341	*CASC16*	ncRNA exonic	0.055	2.34×10^{-5}	4.89×10^{-6}	5.75×10^{-3}	SH	0.037
rs8170	19	17389704	*BABAM1*	Exonic	0.005					
rs614367	11	69328764	*CCND1*	Intergenic	0.174					
rs3803662	16	52586341	*CASC16*	ncRNA exonic	0.055	7.17×10^{-5}	7.73×10^{-6}	9.09×10^{-3}	SR	0.729

Base pair – the base pair location of the SNP in hg19 coordinates; Gene – the gene that the SNP is located within or in closest proximity to; ncRNA exonic refers to RNA without coding annotation – may or may not code for protein; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corr. p-value – Bonferroni corrected p-value; χ^2 p value – p-value for individual SNP association with phenotype, obtained from $\mathbf{X}^2$ test; Met. – the algorithm that identified the interaction: SR – SNPRuler; SH – SNPHarvester

Table 3.6. Significant interactions identified between known susceptibility loci in ER-positive breast cancer

SNP	Chromosome	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corrected p value	Method	Logistic regression p value
rs8170	19	17389704	*BABAM1*	Exonic	0.806	0.239	$3.48x10^{-10}$	$4.09x10^{-7}$	SR	0.339
rs2363956	19	17394124	*ANKLE1*	Exonic	0.714					
rs3757318	6	151914113	*CCDC170*	Intronic	0.370	$1.60x10^{-5}$	$1.40x10^{-6}$	$1.64x10^{-3}$	SR	$1.11x10^{-4}$
rs4808801	19	18571141	*ELL*	Intronic	0.053					

Base pair – the base pair location of the SNP in hg19 coordinates ; Gene – the gene that the SNP is located within or in closest proximity to; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corr. p-value – Bonferroni corrected p-value; χ^2 p value – p-value for individual SNP association with phenotype, obtained from $\mathbf{X}^2$ test; Met. – the algorithm that identified the interaction: SR – SNPRuler; SH – SNPHarvester

Table 3.7. Significant interactions identified in known susceptibility loci in ER-negative breast cancer

SNP	Chromosome	Base pair	Gene	Type	χ^2 p value	LD r^2 value	Interaction p value	Corrected p value	Method	Logistic regression p value
rs3803662	16	52586341	*CASC16*	ncRNA exonic	0.009	$1.37x10^{-5}$	$2.21x10^{-9}$	$2.60x10^{-6}$	SH	0.011
rs8170	19	17389704	*BABAM1*	Exonic	0.806					
rs614367	11	69328764	*CCND1*	Intergenic	0.991	$5.75x10^{-5}$	$7.33x10^{-6}$	$3.06x10^{-3}$	SH	0.750
rs3803662	16	52586341	*CASC16*	ncRNA exonic	0.009					
rs1011970	9	22062134	*CDKN2B-AS1*	ncRNA intronic	0.011	$4.48x10^{-6}$	$3.86x10^{-5}$	0.05	SH	0.126
rs8170	19	17389704	*BABAM1*	Exonic	0.806					

Base pair – the base pair location of the SNP in hg19 coordinates ; Gene – the gene that the SNP is located within or in closest proximity to; LD r^2 – the linkage disequilibrium r^2 value for the pair of SNPs; Interaction p-value – raw p-value for the interaction from the corresponding algorithm; Corr. p-value – Bonferroni corrected p-value; χ^2 p value – p-value for individual SNP association with phenotype, obtained from $\mathbf{X}^2$ test; Met. – the algorithm that identified the interaction: SR – SNPRuler; SH – SNPHarvester

involved in two interactions. None of these interactions occur between SNPs on the same chromosome so there is no evidence that LD between the SNPs might be responsible for the interactions. Two of these ER-negative related interactions are the same two identified for breast cancer overall, suggestive of a bias towards detecting ER-negative specific interactions when considering all samples as one phenotype.

3.5 Discussion

Epistasis – the interaction between genes – is likely to be an important aspect in many diseases, particularly those in which there is still a large missing heritability component (Zuk et al., 2012). The impact of such gene-gene interactions on complex disease risk is likely to be small (Aschard et al., 2012), therefore, detecting real interactions that contribute to disease risk will be challenging and will require large case-control cohorts.

Through the analysis of a small cohort of early-onset breast cancer cases and non-disease control samples, a number of potential interactions between known susceptibility loci have been identified. Analysis involved using five different statistical tests and algorithms to assess SNP pairs for potential interaction effects. The underlying nature of epistasis within disease is unknown and is likely to occur in many different forms, for example some interactions may arise from additive effects of SNPs while some will be due to multiplicative effects. Rather than using a single model in analysis it is more appropriate to use multiple different methods because no single method will identify real interactions in all possible situations.

A relatively small subset of SNPs was analysed in this study, focusing on SNPs that a located within or in close proximity to established breast cancer genes. The majority of sporadic cases of breast cancer are unexplained by known disease variants and are likely to arise from multiple variants in multiple genes that only cause disease manifestation in the presence of one another. Therefore, it is plausible that multiple common, low-penetrance variants within breast cancer genes will lead to tumour formation and may explain disease in some sporadic cases.

Considering all breast cancer cases as a single phenotype, four significant interactions were detected across the five methods. Two of these interactions were identified by SNPHarveter, SNPRuler, fast-epistasis, and case-only. In both cases the interaction was identified as significant by SNPRuler and was only just shy of the significance threshold when analysed in SNPHarvester. One of the two SNP pairs was also found to be significant when analysed by logistic regression, providing compelling evidence that there may indeed be some interaction or correlation between the genotypes of these two SNPs in the breast cancer patients, which is not observed in control samples. The SNPs involved in this interaction lie on chromosomes 5 and 9, associated with the genes *PDE4D* and *RAD23B* respectively.

Limited significant results were detected when all cases were considered as one single disease entity. While these interactions may reflect true interaction effects it is likely that some interactions, specific to certain subtypes of the disease, are being masked. Breast cancer consists of numerous recognised subtypes, the characteristics of which are unique. Therefore, by considering breast cancer as one single entity will potentially mask some subtype specific interactions. By separating ER-positive and ER-negative cases and running ER subtype specific analysis, many more unique interactions were identified. Comparison of the subtype-specific interactions indicates that the genes and chromosomes involved are different (Figure 3.5). Therefore, it is likely that distinct breast cancer subtypes are characterised by different gene-gene interactions, leading to diverse mechanisms of oncogenesis.

Many significant interactions were detected in ER-positive and –negative disease subtypes despite the relatively small sample size considered; 5200 control samples and only 528 cases, of which 161 were ER-positive and 366 were ER-negative. It is thus likely that many of these significant results are in fact false positives, particularly in the case of ER-positive disease where in excess of 400 interactions were significant. With only 161 cases compared to 5200 controls it is likely that some SNP pairs will appear significantly different in cases and controls simply due to a lack of data in the case cohort. Replication in a larger cohort is required to provide further supporting evidence that some of these interactions are real and not artefacts of the data.

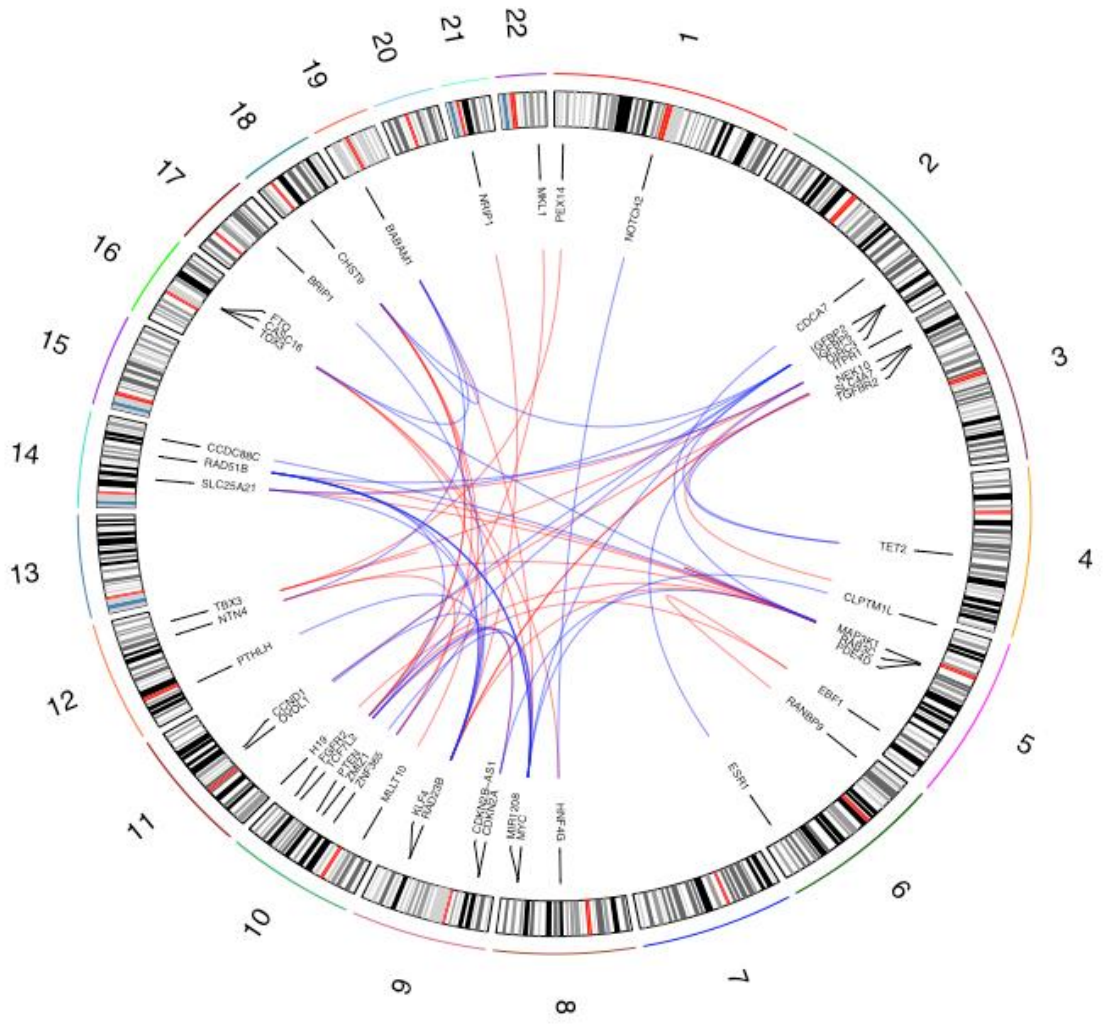


Figure 3.5. Circos plot depicting the most significant interactions identified in ER-positive and ER-negative breast cancer. The thirty most highly significant interactions detected in ER-positive breast cancer, by any method, are depicted in red. The 31 significant interactions identified in ER-negative breast cancer are depicted in blue.

The aim of this study was to exploring potential gene-gene interactions between known breast cancer susceptibility loci. As such, SNPs in the major early-onset breast cancer genes *BRCA1* and *BRCA2* were included in the analysis. No significant interactions were detected that included a SNP in or close to either of these genes. Previous studies have identified interactions between mutations in *BRCA1* and *BRCA2* with mutations in *ATM* and *CHEK2* (Turnbull et al., 2012), both of which are also breast cancer susceptibility loci and were included in this analysis. No interactions between these 4 genes were replicated here but the study in which these interactions were detected

included over 7000 samples. Further SNP-SNP interaction studies carried out in breast cancer also did not identify any interactions involving either *BRCA1* or *BRCA2* despite including these genes in the analysis (Sapkota et al., 2013, Milne et al., 2014b)

Systematic errors can be incorporated into the genotype call data due to the technical difficulties of the genotyping technology. Such errors can lead to spurious associations if the populations being considered were genotyped at different location or different times. There is the potential for SNP-SNP interactions detected in this study to be the result of systematic differences between the case and controls populations. Cases were genotyped in two batches at separate locations (Rafiq et al., 2013); all ER-positive samples were genotyped at one location while the ER-negative cases were split into two groups, one sequenced at each location. There is, therefore, the potential that some of the interactions identified when comparing the breast cancer subtypes to one another are due to batch effects rather than reflecting true interactions effects. The same problem arises when the breast cancer cases were compared to controls samples because the control samples were not genotyped as part of the same study and different SNP arrays were used for the genotyping. A second problem of using genotype data from controls not chosen specifically for this study is that the methodology used for sampling is not the same between the two populations. This could also lead to systematic differences between the two populations that may be detected by the interaction detection methods, and thus reported as interactions.

Low statistical power is a major limitation for epistasis detection studies and is likely to have been a limiting factor here due to the relatively small sample size. A recent large-scale analysis of 2-way epistasis in breast cancer analysed in excess of 40,000 samples in both the case and control groups but was unable to identify any strongly significant interactions (Milne et al., 2014b). Due to the large sample size the statistical power was high; > 90% power to detect odds ratios of 1.29 or more for SNPs with MAF > 10%. Therefore, the reported lack of success implies that many tens of thousands of case and control samples will be required to identify real interactions effectively.

None of the tests implemented in this study is optimal for all possible situations in which an interaction may occur. For example, SNPHarvester has

been shown to perform moderately well when simulated interactions do not involve main effects but it performs poorly when main effect SNPs are involved in interactions (Ibrahim et al., 2013) because a pre-processing step of the algorithm prunes out all SNPs with detectable main effects (Yang et al., 2009). SNPRuler on the other hand, tends to perform better when main effect SNPs are involved in an interaction (Ibrahim et al., 2013). This may be due to the rule-based approach of SNPRuler, which identifies predictive rules describing the relationship between variables and the phenotype (Wan et al., 2010b); a main effect SNP will have an individual association with the phenotype which may drive the detection of a SNP pair including the main effect SNP. In this study no interactions were identified that involved SNPs with particularly strong association with the phenotype; this may be because no such interactions exist or because the algorithms are not well suited to analysing real genetic data. Furthermore, evaluation of methods inevitably depends upon testing simulated data, which is highly unlikely to reflect the true underlying structure of real genetic data, as such it is extremely difficult to assess whether interactions detected in real data are in fact true epistasis. Therefore, one must be cautious when using such approaches and when drawing conclusions from the results.

All tests used in this study to identify interactions tended to identify different SNP pairs rather than consistently implicating the same SNPs. This is likely to be due to the different search strategies and the individual strengths and weaknesses of each test. Determining whether the results represent real interactions or are mere artefacts in the data requires replication in other, preferably larger, datasets. One way to prioritise potentially significant interactions is to select those that are found to be significant by multiple methods; this greatly reduces the number of significant interactions detected in this study.

Only a small subset of all genotyped SNPs was considered in this analysis (~5%). Therefore, it is highly likely that any true interactions contained within the data are going to be overlooked because they simply were not assessed. It is necessary however, to reduce the search space investigated because of the vast number of tests that would otherwise be carried out. Assessing all possible SNP-SNP pairs would not only be time consuming but would have very low power to detect interactions in data sets that are currently available.

Therefore it is likely that a large number of false-positive interactions will be detected in such studies.

3.6 Conclusion

Different machine learning algorithms applied to genome-wide SNP genotype data successfully identified many highly significant SNP-SNP interactions in a relatively small subset of SNP markers. All detected interactions have the potential to be real and contribute to early-onset breast cancer risk. However, it is unlikely that all detected interactions are true interactions. To enhance understanding of SNP-SNP interactions in early-onset breast cancer will require many more case-control studies, with many thousands of individuals genotyped. Only through the use of large-scale genotyping studies can sufficient power to detect interactions be achieved.

Chapter 4: Next Generation Whole-Exome Sequencing of Eight Early-Onset Breast Cancer Patients with Extreme Phenotypes

4.1 Background

Early-onset breast cancer is likely to have a stronger genetic component than late-onset breast cancer because the cellular hallmarks of cancer (Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011) are amassed over many years; young people with breast cancer simply do not have enough time to accrue all the necessary hallmarks. Few cases of breast cancer arise from a single variant in a highly penetrant gene, instead they are likely polygenic, resulting from multiple variants at multiple susceptibility loci (Pharoah et al., 2002, Stratton and Rahman, 2008). This is further supported by the results of numerous GWA studies; many common variants with an association with breast cancer have been identified (Easton et al., 2007b, Hunter et al., 2007, Stacey et al., 2007, Ahmed et al., 2009, Zheng et al., 2009, Thomas et al., 2009, Turnbull et al., 2010, Fletcher et al., 2011, Haiman et al., 2011, Ghoussaini et al., 2012).

In general, GWA studies only identify common variation, with MAF > 5%, that has an association with disease (Frazer et al., 2009, Bodmer and Bonilla, 2008). It has become increasingly apparent however, that the majority of common variants have only modest effects on disease susceptibility (Schork et al., 2009). Therefore, it may be necessary to search for rare variants that are moderately penetrant and have greater effects on disease risk (McCarthy et al., 2008), through the use of NGS technology.

Next generation exome sequencing is capable of identifying all coding variation, whether common or rare, in a single patient (Kiezun et al., 2012). Identifying variation underlying complex traits is inherently more difficult than detecting causal variation in Mendelian disorders (Bamshad et al., 2011). To maximise the potential for identifying relevant disease variants and genes, focus should be on sequencing only those individuals with extreme disease phenotypes since such patients are more likely to harbour pathogenic variation in the same genes or pathways (Bamshad et al., 2011).

Genetic dissection of breast cancer and other genetic diseases

The genetic profile of individual early-onset breast cancer patients is likely to be unique, with multiple variants contributing to disease manifestation. Mutations within particular pathways and genomic regions are likely to influence disease risk (Gilissen et al., 2012), indicating that it is necessary to focus on specific pathways to identify causal variants.

The *BRCA1*, *BRCA2* and *TP53* genes are tumour suppressors with DNA repair and genome stability functions. A single germline mutation in one of these genes is often sufficient for breast cancer development, although such mutations explain disease in only a small proportion of breast cancer cases (Antoniou and Easton, 2006, Peto et al., 1999). These genes are members of important gene pathways and interact with many other genes. *TP53* is a particularly important tumour suppressor gene that inhibits the proliferation of abnormal cells by initiating cell cycle arrest and apoptosis (Lacroix et al., 2006). *TP53* mutations have been identified in almost all common cancers, however, germline *TP53* mutations occur rarely in breast cancer. Instead, germline mutations within genes of the *TP53* gene network may impair *TP53* function, leading to breast cancer development and progression (Gasco et al., 2002). A number of well-characterised breast cancer susceptibility genes, including *BRCA1*, *BRCA2*, *ATM* and *CHEK2*, interact with the *TP53* gene (Gasco et al., 2002), thus variation in other genes that interact with *TP53* has the potential to disrupt the function of this gene and cause oncogenesis.

Germline mutations are often implicated in early-onset breast cancer but mutations that occur early in embryo development are also possible disease candidates. Postzygotic mosaic variants may be involved in a wide range of diseases, particularly those that are sporadic (Forsberg et al., 2012). Postzygotic mutations in the *PP1MD* have been identified as involved in predisposition to breast cancer (Ruark et al., 2013). Therefore, it is possible that a number of breast cancer cases, for which the causative mutation has not been identified, may be the result of postzygotic variation.

Early-onset breast cancer patients with extreme phenotypes (triple-positive or triple-negative breast cancer) or strong family history were selected for exome sequencing (Table 4.1). Both patients with a family history have a close relative also affected by early-onset breast cancer (Figure 4.1 and Figure 4.2) but no identified causative mutation.

Table 4.1. Summary of patient phenotypes and family history

Patient POSH ID	Sample Label	Age at diagnosis (yrs)	Receptor phenotype	Ethnicity	Family history	Subtype group
2008063006	DE1	22	ER+, PR+, HER2+	White Caucasian	No	Triple positive
2003120374	DE2	25	ER+, PR+, HER2+	Asian	No	Triple positive
2007012360	DE3	25	ER+, PR+, HER2+	White Caucasian	No	Triple positive
2007122818	DE4	25	ER+, PR+, HER2+	White Caucasian	No	Triple positive
2004010436	DE5	23	ER-, PR-, HER2-	White Caucasian	No	Triple negative
2004120932	DE6	24	ER-, PR-, HER2-	White Caucasian	No	Triple negative
2004070685	DE7	24	ER-, PR-, HER2+	White Caucasian	Yes	Familial
2007022382	DE8	30	ER+, PR+, HER2-	Unknown	Yes	Familial

ER – Estrogen receptor; PR – Progesterone receptor; HER2 – Human epidermal growth factor receptor 2
ER+/PR+/HER2+ - over-expression of corresponding receptor
ER-/PR-/HER2- - under-expression of corresponding receptor

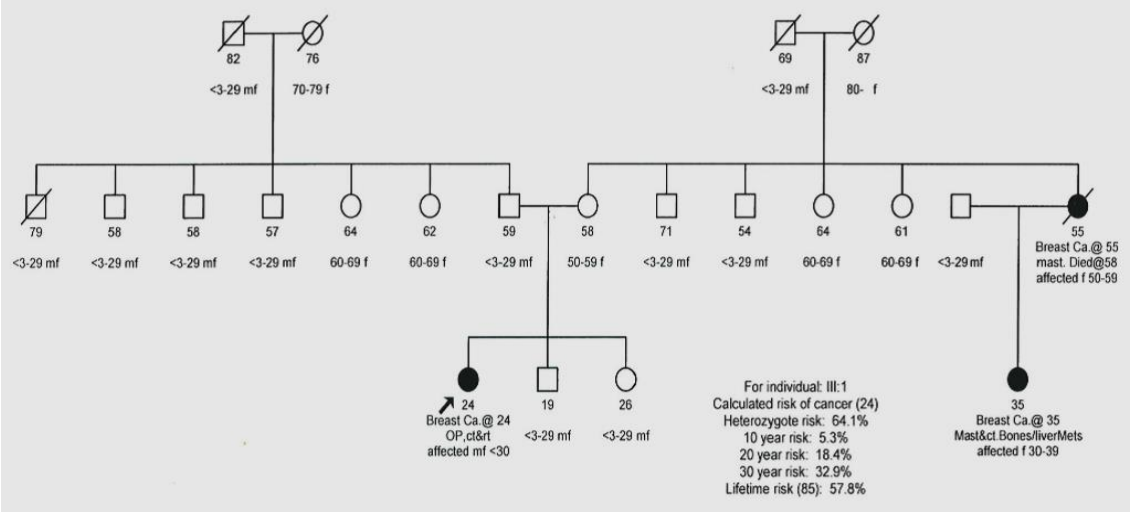


Figure 4.1. Pedigree of sample DE7 showing family history of breast cancer. Sample DE7 was diagnosed with breast cancer aged 24 and has a cousin who was diagnosed with early-onset breast cancer aged 35.

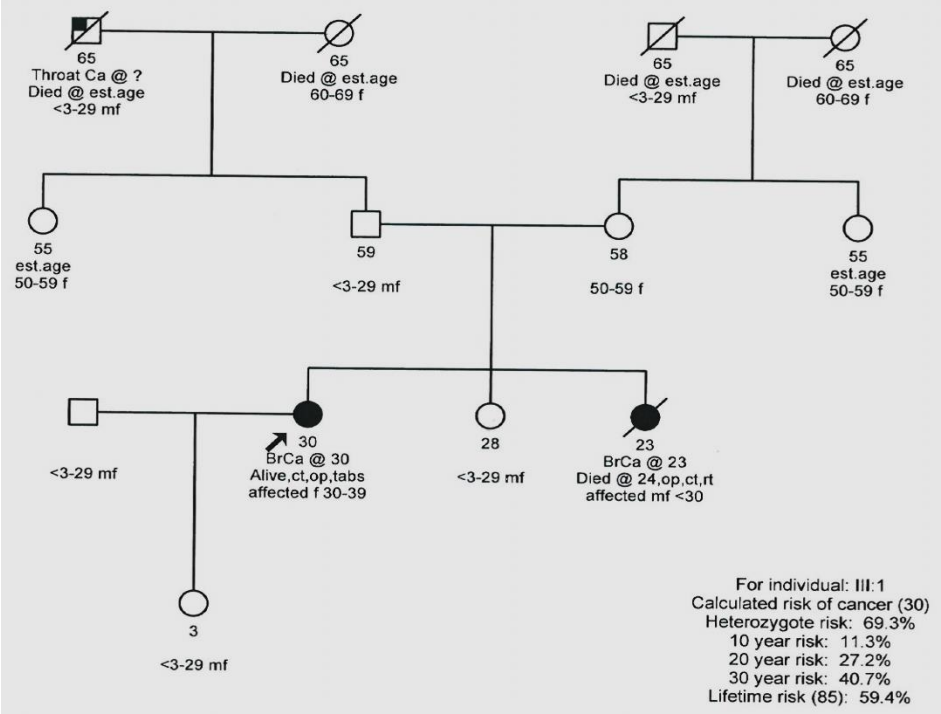


Figure 4.2. Pedigree of sample DE8 showing family history of breast cancer. Sample DE8 was diagnosed with early-onset breast cancer aged 30 and has a sister who was diagnosed with early-onset breast cancer aged 23.

4.2 Aim

To characterise the spectrum of recessive and compound heterozygous rare and novel genetic variation harboured in the coding regions of the genome, in 8 individuals with particularly early-onset breast cancer.

4.3 Materials and Methods

4.3.1 Exome Sequencing

DNA samples from 8 breast cancer patients were exome sequenced at the Peter MacCallum Cancer Centre, Melbourne, using the Agilent SureSelect Human All Exon 50Mb Kit. Patients with extreme phenotypes or a strong family history were selected from the Prospective study of Outcome in Sporadic versus Hereditary breast cancer (POSH) cohort of ~3000 patients with early-onset breast cancer. All patients within this cohort were diagnosed before the age of 40 (Eccles et al., 2007).

4.3.2 *TP53* Pathway Candidate Genes

A comprehensive list of 327 candidate genes that directly interact with *TP53* or are members of the *TP53* pathway, was constructed based on information from the following websites: Wikipedia, KEGG, The TP53 Website, and NCBI (see Appendix VI for more detail). One further gene (*WDR*) was included following communication with Magali Olivier (Database Manager, IARC. Contacted September 2011). Correct gene names were obtained from the 'HUGO Gene Nomenclature Committee' website (http://www.genenames.org/). For the full list of 327 TP53 pathway genes see Appendix VII.

4.3.3 Rare Variation in the *TP53* Gene Pathway

4.3.3.1 Sequence Alignment, Variant Annotation and Variant Filtering

Paired-end sequence data was aligned against the human genome reference sequence 19 (hg19) and all variants were identified and annotated at the Peter MacCallum Cancer Centre, Melbourne.

Non-reference allele frequencies were obtained for all variants from the full phase 1 1000 Genomes Project database using ANNOVAR (http://www.openbioinformatics.org/annovar/) (Wang et al., 2010). Only variants with a non-reference allele frequency of 5% or less were retained. Variants were annotated with respect to genes and transcripts using ANNOVAR and only variants identified as exonic were retained. All exonic variants were thoroughly filtered to remove synonymous SNVs, nonframeshift indels, and variants identified as mapping to segmental duplications (likely to represent sequence alignment errors). Variants located in *TP53* pathway candidate genes were extracted to produce a subset of 372 variants in 95 genes.

In silico functional prediction scores were obtained, where possible, from SIFT (Ng and Henikoff, 2001, Ng and Henikoff, 2002, Ng and Henikoff, 2003) (http://sift.jcvi.org/) and PolyPhen2 (Adzhubei et al., 2010) (http://genetics.bwh.harvard.edu/pph2/index.shtml) for all variants.

4.3.3.2 Quality Control Measures

Variants with low read depth (<10 reads) or strand bias (<1 read on either the forward or reverse strand) were considered to be poor quality and were removed from consideration.

Any variants failing a number of recommended GATK filtering criteria (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3) were considered to be poor quality and were filtered out of the subset. Variants were removed from consideration if they satisfied at least one of the following criteria (description of criteria in Appendix VIII): QD<2, MQ<40, HaplotypeScore >13, MQRankSum<-12.5, ReadPosRankSum<-8. Insertion or deletion variants with QD<2 or ReadPosRankSum<-20 were also removed.

Presence of all identified variants was confirmed through the manual examination of BAM files containing raw sequence reads using the Integrative Genomics Viewer (IGV; http://www.broadinstitute.org/igv/) (Robinson et al., 2011b, Thorvaldsdóttir et al., 2013) software. Three frameshift variants present in all 8 exomes, that are likely to represent alignment artefacts, were removed from further analysis along with 4 variants which were all located within a region of 10 nucleotides in a single gene, present in multiple individuals.

4.3.3.3 Potentially Damaging Variation

The identified subset of rare variants was further filtered to retain only those variants that were either potentially compound heterozygous (two or more variants present in the same gene within one individual) or potentially recessive. Further variants were retained if they were classified as deleterious by the SIFT scoring algorithm.

4.3.4 Rare Variation in all Genes of the Exome

4.3.4.1 Sequence Alignment and Variant Calling

Paired-end sequence data was aligned against the human genome reference sequence 18 (hg18) using the Novoalign software (Novocraft Technologies:

http://www.novocraft.com/main/index.php). Gene exon coverage and read depth statistics were calculated using the BEDTools software (http://bedtools.readthedocs.org/en/latest/) and custom scripts. Single nucleotide polymorphisms and small insertion or deletion variants were identified post alignment with the SAMtools software (http://samtools.sourceforge.net/) using custom scripts. All variants were annotated with respect to genes and transcripts using the ANNOVAR software tool and custom scripts. Functional predictions were obtained from SIFT, PolyPhen2, and Grantham score (Grantham, 1974, Li et al., 1984) for each variant where possible. Only coding variation was retained following annotation.

4.3.4.2 Variant Filtering

All rare and novel variation was identified in each sample by applying a thorough filtering regime. To retain only rare variants, all variants with a non-reference allele frequency of > 1% in the full phase 1 1000 Genomes Project database and the National Heart, Lung and Blood Institute Exome Sequencing Project Exome Variant Server (NHLBI EPS ESV) were removed. Synonymous and nonframeshift insertions or deletions were removed from further consideration along with any SNV with an rsID in either the dbSNP129 (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=129) or dbSNP135 (https://www.ncbi.nlm.nih.gov/SNP/) variation databases. Variants annotated as located within a homopolymer or repeat region were removed because they are likely to represent sequence alignment errors. A number of genes have been identified as highly mutable (Fuentes Fajardo et al., 2012) and by in-house custom scripts; any variants identified in these genes were not considered further. All variants were cross-referenced against all variation observed in 40 previously sequenced exomes from unrelated individuals. Each of these 40 exomes is from a patient presenting with a distinct clinical diagnosis but not presenting with breast cancer; 26 had imprinting disorders, 11 were paediatric inflammatory bowel disease patients and 3 had Mendelian disorders. Any variant also present in any one of these exomes was unlikely to be disease-causing and was thus removed from consideration. Further variant cross-referencing was done using the COSMIC gene database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) to identify any variants in genes implicated in cancer, and the Human Gene

Mutation Database (HGMD) Trial version (http://www.biobase-international.com/product/hgmd) was used to identify any variants previously linked to disease.

A final set of variants was identified by isolating any potentially compound heterozygote variants, potentially recessive variants, variants in cancer genes catalogued in COSMIC, or variants identified as disease-causing in HGMD.

4.3.4.3 Identifying Potential Splice Sites

Some exonic missense, nonsense or synonymous variants modulate splicing by causing exon skipping or activating ectopic splice sites. To detect variants with the potential to influence splicing the SKIPPY software (http://research.nhgri.nih.gov/skippy/index.shtml) (Woolfe et al., 2010), which generates quantitative scores for exonic variation, was used. Positive scores represent splice-affecting variants while variants given a negative score are not associated with splicing. The final set of variants identified after thorough filtering of all exonic variation described above (rare variation in all genes of the exome) was submitted to the SKIPPY program to obtain splice scores.

4.3.5 *BRCA1* and *BRCA2* Variants

Variation in the highly penetrant *BRCA1* and *BRCA2* genes is strongly implicated in breast cancer susceptibility. Any variants called in the *BRCA1* and *BRCA2* genes were identified in each exome and the clinical impact of each variant was assessed by querying a number of databases: Breast Cancer Information Core (BIC) (http://research.nhgri.nih.gov/bic/), Align GVGD (http://agvgd.iarc.fr/agvgd_input.php), and the Breast Cancer genes IARC (http://brca.iarc.fr/PRIORS/). Functional predictions (SIFT, PolyPhen2, Grantham score and prediction) were provided for each variant in the annotation step. To identify any known clinical or disease-associated roles for any of the variants the Human Genome Mutation Database (HGMD) trial software was queried.

4.3.6 Post-zygotic Mosaic Variants

Potential post-zygotic mosaic variants were identified in *BRCA1*, *BRCA2* or *TP53* in each sample exome. Paired-end sequence data was aligned against the human genome reference sequence 19 (hg19) using the Novoalign software. To identify any variation present at low levels within these genes in any of the eight exomes, multi-sample calling was implemented within the SAMtools software using custom scripts. All variants were annotated with respect to genes and transcripts using the ANNOVAR software tool and custom scripts. SAMtools calling reported the genotypes at all base-pair locations in each gene region; all reference homozygous, heterozygous and alternate homozygous genotypes were called.

Reference and alternate allele read counts were obtained for each base-pair location in each individual sample by implementing single sample variant calling in SAMtools.

Potential post-zygotic mosaic variants are identified through the presence of a low number of alternate allele reads in variants that are classified as reference homozygote. Potential variants were identified from the output from the multi-sample calling step. Reference and alternate allele read counts on the forward and reverse DNA strands are reported in the output from the single sample variant calling step. All reference homozygote positions with a minimum of three alternate allele reads in any one of the eight samples were identified and classified as potential post-zygotic variants. To confirm that all identified potential post-zygotic variants were likely to be real and not alignment artefacts, the Integrative Genomics Viewer (IGV) software was used to examine the raw sequence reads associated with each identified variant.

4.4 Results

4.4.1 Sequencing Coverage

Each breast cancer exome had a minimum of 69% of mappable bases covered by at least 20 reads (Appendix IX). Approximately 73% of reads were mapped to target reads and there was a mean read depth of at least 74 for each sample (Appendix IX).

4.4.2 Rare Variation in Genes of the *TP53* Gene Pathway

Thorough filtering of all called variants in each exome sequenced sample identified 41 variants in 20 candidate *TP53* pathway genes (Table 4.2).

Functional prediction scores obtained from the SIFT algorithm classified the majority of the identified variants as potentially deleterious. No potentially recessive variants were identified that passed all filtering criteria. Potential compound heterozygous variants were identified in 5 genes.

Variants common to individuals presenting with the same disease subtype may possibly be disease-associated although this is more likely to be the case if the variants are highly penetrant or the individuals are related. Only 2 of the 41 identified variants are present in more than one sample; a variant in the *THBS1* gene is present in samples DE4 (triple positive) and DE7 (familial) while a variant in *TP53BP2* is found in two triple positive cases and in both the familial cases. All other rare variants identified are present in only a single individual.

Most cases of breast cancer are likely to be polygenic and it is unlikely that the same variant will be identified in multiple individuals, even though they are all characterised by extremely early disease onset, which is unusual. When considering only 8 samples that represent at least 4 disease subtypes, it is particularly unlikely that the same variants will be implicated in disease. Instead, it is likely that variation within a particular gene or region may confer risk of developing a certain disease subtype. Three genes identified in this study have variation in more than one individual; *CUL9*, *HTT* and *MDC1*. Variants identified as deleterious by SIFT were identified in *CUL9* in both DE7 and DE8 who have familial forms of the disease, although these samples present with different subtypes. Three variants are identified in *HTT* present in sample DE3, DE4 and DE6. A large amount of variation is observed in the *MDC1* gene in both DE6 and DE7.

Potential compound heterozygous variants were identified in samples DE4, DE6, DE7 and DE8. Of particular interest are the variants identified in *BRCA1* and *BRCA2* present in samples DE4 and DE6 respectively. Both *BRCA1* variants in DE4 are predicted deleterious by SIFT. One of the *BRCA2* variants present in sample DE6 is a stopgain variant. Both DE6 and DE7 harbour compound heterozygous variants in the *MDC1* gene. The result for sample DE7 is

particularly peculiar; 13 variants are identified in *MDC1*. Finally, there are two genes harbouring more than one variant in sample DE8; *LAMA4* and *TEP1*. One variant in *LAMA4* has a SIFT score of 0 indicating it is deleterious while the other variant has a SIFT score of 0.95. Neither of the *TEP1* variants appears to be deleterious by SIFT.

4.4.3 Rare Variation in All Genes of the Exome

A total of 172,989 variants were identified in the exomes of all 8 patients, of which 83 were novel. Each exome harbours numerous compound heterozygous and recessive variants, many of which are novel or predicted to be damaging by SIFT. There are also a large number of variants identified in disease-related genes, both cancer causing and disease-related more generally. However, none of the cancer causing genes identified is a breast cancer related gene and the majority of the disease-related variants are unlikely to be relevant in these 8 patients because they are often not disease causing on their own.

4.4.3.1 Compound Heterozygous and Recessive Variation

Following the thorough filtering regime outlined in Materials and Methods, only very rare and novel variation remained (only unrecorded variants or variants with MAF ≤ 1% in the 1000 Genomes Project data or the Exome Variant Server database). Of particular interest were any potential compound heterozygous variation and potential recessive homozygous variation in any of the 8 breast cancer samples. A total of 104 such variants were identified across the 8 exome samples.

Variants were considered to be potentially pathogenic if they had a SIFT score < 0.2 and/or were classified as possibly/probably damaging (P or D) by the PolyPhen2 algorithm. Potentially pathogenic compound heterozygous variants were identified in 20 genes (Table 4.3) and 15 recessive variants were classified as potentially pathogenic (Table 4.4). Compound heterozygous and recessive variants not considered pathogenic by this classification are listed in Appendices X and XI respectively.

Annotation of all called variants identified a number of rare or novel variants that were intronic splice sites, i.e. occurring within 10 base pairs of the beginning or end of an exon.

Table 4.2. Variants identified in genes of the *TP53* gene pathway

Gene	Exon	Chromosome	Variant type	Base pair location in hg19	Nucleotide change	Protein change	Frequency in 1000G	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
ABL1	2	9	ns	133729511	G197A	R66H	.	0.00	D	29	C		+						
ATR	26	3	ns	142232392	A4592G	H1531R	.	0.03	D	29	C	+							
AURKA	11	20	ns	54945309	A1117G	M373V	0.006	0.03	B	21	C		+						
BRCA1	22	17	ns	41197708	A5438C	H1813P	.	0.00	B	77	MC				◇				
BRCA1	10	17	ns	41246481	A1067G	Q356R	0.018	0.02	D	43	C				◇				
BRCA2	11	13	ns	32914236	C5744T	T1915M	0.006	0.13	B	81	MC						◇		
BRCA2	27	13	sg	32972626	A9976T	K3326*	.	-	.	.	.						◇		
CDC25C	7	5	ns	137625228	G670C	G224R	0.007	0.00	D	125	MR			+					
CHEK2	4	22	ns	29121060	A497G	N166S	.	0.00	D	46	C		+						
CUL9	2	6	ns	43155718	G223A	E75K	0.003	0.00	B	56	MC							+	
CUL9	21	6	ns	43172498	G1808T	S603I	.	0.01	B	142	MC								+
ERCC3	6	2	ns	128046416	C112T	R38C	.	0.00	P	180	R						+		
HTT	8	4	ns	3117122	A839G	H280R	.	0.01	D	29	C				+				
HTT	34	4	ns	3176780	G4353T	L1451F	.	0.00	D	22	C						+		
HTT	39	4	ns	3188417	C4960T	R1654W	.	0.00	D	101	MR			+					
LAMA4	35	6	ns	112439072	T4830G	C1610W	.	0.00	D	215	R								◇
LAMA4	17	6	ns	112471715	G2150A	R717K	.	0.95	B	26	C								◇
MDC1	11	6	ns	30671312	T5565G	D1855E	0.019	1.00	B	45	C							◇	
MDC1	10	6	ns	30671588	T5372A	V1791E	0.018	1.00	B	121	MR							◇	
MDC1	10	6	ns	30671726	C5234G	P1745R	0.015	0.30	D	103	MR							◇	
MDC1	10	6	ns	30672326	A4634G	Q1545R	0.040	0.28	B	43	C						◇		
MDC1	10	6	ns	30673064	G3896A	R1299Q	0.017	0.52	B	43	C							◇	
MDC1	10	6	ns	30673113	C3847A	P1283T	0.028	0.34	B	38	C							◇	
MDC1	10	6	ns	30673163	A3797C	Y1266S	0.048	0.80	B	144	MR							◇	
MDC1	10	6	ns	30673185	C3775T	P1259S	.	0.67	B	74	MC							◇	
MDC1	10	6	ns	30673340	T3620C	L1207P	.	0.46	B	98	MC							◇	
MDC1	10	6	ns	30673359	A3601C	T1201P	0.009	1.00	B	38	C							◇	
MDC1	10	6	ns	30673403	T3557C	V1186A	0.019	0.75	B	64	MC							◇	

MDC1	10	6	ns	30673625	C3335T	S1112F	0.014	0.02	D	155	R						◇	
MDC1	10	6	ns	30673662	C3298G	P1100A	0.015	0.03	B	27	C						◇	
MDC1	6	6	ns	30680124	G1211A	G404E	.	0.20	B	98	MC					◇		
MDC1	5	6	ns	30681477	C151T	R51C	0.015	0.09	B	180	R						◇	
PRKDC	64	8	ns	48739305	C8692T	R2898C	0.018	0.02	P	180	R			+				
SMARCA4	16	19	ns	11123647	T2297G	V766G	.	0.00	D	109	MR				+			
SSTR3	2	22	ns	37603155	G688A	V230M	.	0.00	D	21	C							+
TAF1A	10	1	ns	222732027	A986G	Y329C	0.014	0.00	B	194	R		+					
TEP1	40	14	ns	20844384	C5804T	T1935M	.	0.17	B	81	MC							◇
TEP1	24	14	ns	20851756	A3434G	E1145G	0.011	0.44	B	98	MC							◇
THBS1	3	15	ns	39882178	A515G	N172S	0.043	0.01	D	46	C			+			+	
TP53BP2	7	1	ns	223991119	C685A	Q229K	0.040	0.01	D	53	MC		+	+			+	+
TSC2	16	16	ns	2120487	G1600A	A534T	.	0.02	D	58	MC	+						

Where a specific variant is present in a sample this is indicated by +. Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined

Where a specific variant is a potential compound heterozygous variant this is indicated by ◇

B, benign; C, conservative; Cl, clinical; D, probably damaging; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; ns/exsp nonsynonymous or potential exonic splice site; P, possibly damaging; R, radical; sg, stopgain

Table 4.3. Potential compound heterozygous variants identified from all genes of the exome

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
C2orf71	1	2	ns	29147573	A3059G	Q1020R	.	.	0.000	0.17	B	43	C	.			+					
C2orf71	1	2	ns	29147574	C3058A	Q1020K	N	.	.	0.27	P	53	MC	.			+					
CDH23	24	10	ns	73134795	A2855G	E952G	N	.	.	0.12	B	98	MC	1.360					+			
CDH23	26	10	ns	73138932	C3178T	R1060W	.	.	0.000	0.00	D	101	MR	0.768					+			
CES5A	10	16	ns	54444305	G1262C	R421P	.	.	0.001	0.03	P	103	MR	-2.246							+	
CES5A	11	16	ns	54441128	C1332A	D444E	.	.	0.001	0.23	P	45	C	-2.962							+	
EPS8L1	5	19	ns	60282887	C135G	C45W	N	.	.	0.00	D	215	R	-2.962		+						
EPS8L1	10	19	ns	60285310	G940T	A314S	N	.	.	0.16	B	99	MC	1.730		+						
GOLGA3	9	12	ns	131885028	C1909T	R637C	.	.	0.000	0.03	D	180	R	1.290								+
GOLGA3	5	12	ns	131894889	C839T	A280V	.	.	0.001	0.12	B	64	MC	-0.568								+
LILRA1	6	19	sg	59799035	G781T	G261X	.	.	.	0.11	.	.	.	-2.962						+		
LILRA1	6	19	ns	59799039	A785G	E262G	.	.	.	0.44	B	98	MC	0.052						+		
MAGI1	23	3	ns	65317473	T4009G	L1337V	.	.	0.005	0.33	D	32	C	.				+				
MAGI1	12	3	ns	65390332	T2070G	N690K	N	.	.	0.02	D	94	MC	-5.884				+				
MEFV	3	16	ns	3239587	C1105T	P369S	Cl	.	0.006	0.02	D	74	MC	-1.284								+
MEFV	3	16	ns	3239469	G1223A	R408Q	Cl	.	0.006	0.10	B	43	C	-3.894								+
MRVI1	19	11	ns	10560065	C2012T	A671V	.	.	0.002	.	B	64	MC	-5.884								+
MRVI1	9	11	ns/sp	10604705	A479G	N160S	.	.	0.002	.	P	46	C	-1.284								+
PCNXL2	5	1	ns	231461401	G830C	G277A	.	.	0.000	.	P	60	MC	-1.284						+		
PCNXL2	33	1	ns	231188795	C5906T	T1969M	.	.	0.007	0.16	D	81	MC	-0.568						+		
PER3	11	1	ns/sp	7792540	C1240G	P414A	.	.	0.004	0.00	D	27	C	-1.284	+							
PER3	11	1	ns/sp	7792547	A1247G	H416R	.	.	0.005	0.00	B	29	C	0.768	+							
RIN2	7	20	ns	19903714	C1045T	P349S	.	.	0.002	0.47	B	74	MC	0.749						+		
RIN2	7	20	ns	19903673	C1004G	P335R	.	.	0.002	0.21	D	103	MR	-1.284						+		
RTN3	2	11	ns	63243586	C979T	P327S	N	.	.	0.00	B	74	MC	1.474							+	
RTN3	2	11	ns	63243880	G1273T	D425Y	N	.	.	.	D	160	R	1.360							+	
SERPINB2	8	18	ns	59721158	G887A	S296N	N	.	.	0.75	B	46	C		+							
SERPINB2	3	18	ns	59713534	C225A	N75K	N	.	.	0.12	B	94	MC	2.190	+							
SHROOM3	10	4	ns	77910858	G5405A	G1802E	N	.	.	0.33	B	98	MC	-5.168					+			

SHROOM3	5	4	ns	77879527	G1177A	A393T	N	.	.	0.01	P	58	MC	-1.284			+	
SLC25A41	5	19	ns	6378487	G650A	R217H	.	.	0.000	0.29	D	29	C	-1.284			+	
SLC25A41	5	19	ns	6378481	C656A	T219K	.	.	0.007	0.04	D	78	MC	-2.962			+	
SORBS1	3	10	ns	97182227	C173T	P58L	N	.	.	0.00	D	98	MC	-3.894	+			
SORBS1	7	10	ns	97164382	G669T	Q223H	N	.	.	0.00	B	24	C	-2.216	+			
TOM1	10	22	ns	34059433	G970A	D324N	N	.	.	0.01	B	23	C	-5.884				+
TOM1	11	22	ns	34060426	C1133T	A378V	.	.	0.003	0.10	P	64	MC	2.696				+
ZNF135	5	19	ns	63270613	G985A	E329K	.	.	0.001	0.88	P	56	MC	.			+	
ZNF135	5	19	ns	63270428	C800T	S267L	.	.	0.001	0.07	D	145	MR	.			+	
ZNF813	4	19	ns	58686954	G1656T	K552N	.	.	0.001	.	D	94	MC	.			+	
ZNF813	4	19	ns	58686431	C1133G	T378R	.	.	0.005	.	D	71	MC	.			+	

Where a specific variant is present in a sample this is indicated by +. Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined
B, benign; C, conservative; Cl, clinical; D, probably damaging; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; ns/sp nonsynonymous or potential exonic splice site; P, possibly damaging; R, radical; sg, stopgain

Table 4.4. Potential recessive variants identified from all genes of the exome

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
ATG2B	11	14	ns	95867620	C1576G	P526A	.	.	0.001	0.05	B	27	C	-2.246		+						
C21orf63	2	21	ns	32747566	G236A	R79Q	.	.	0.010	0.15	D	43	C	-5.884					+			
C9orf68	8	9	ns	4607977	C767T	T256I	N	.	.	0.17	P	89	MC	-0.568		+						
CDK11A	14	1	ns	1625887	A1496G	K499R	.	.	0.000	0.05	D	26	C	-2.246								+
CFHR1	5	1	ns	195066419	A774T	E258D	N	.	.	0.39	P	45	C	0.749		+						
FHOD1	20	16	ns	65821584	G3100T	V1034F	.	.	0.000	0.11	P	50	C	2.006		+						
ISCU	1	12	fd	107480605	77_78del	26_26del	.	.	.	.	.	.	.	.						+		
KIR2DS4	1	19	ns	60036036	C5T	S2L	N	.	.	0.02	B	145	MR	.	+							
KRTAP7-1	1	21	fd	31123841	47_48del	16_16del	N	.	.	.	.	.	.	.		+						
LCE3C	1	1	fd	150840054	223delA	R75fs	N	.	.	.	.	.	.	.							+	
LRRC56	10	11	ns	541741	C887T	S296F	.	.	0.004	0.01	D	155	R	0.052								+
MZF1	6	19	ns	63766228	C1228G	R410G	N	.	.	0.00	B	125	MR	.		+						
PRSS53	3	16	ns	31006501	G100A	G34S	.	.	0.000	0.19	D	56	MC	-3.894		+						
RIBC2	2	22	fd/sp	44188943	191delC	T64fs	.	.	.	.	.	.	.	.								+
UBTFL1	1	11	fd	89459537	772delT	W258fs	N	.	.	.	.	.	.	.	+			+				

Where a specific variant is present in a sample this is indicated by +. Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined
B, benign; C, conservative; D, probably damaging; fd, frameshift deletion; fd/sp, frameshift deletion or potential exonic splice site; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; P, possibly damaging; R, radical

Current understanding of the importance of intronic splice variants in disease status is limited. Therefore, any compound heterozygous or recessive variants annotated as intronic splicing were not considered as pathogenic (see Appendices X and XI for all rare and novel intronic splicing variants).

Potential compound heterozygous variants were identified in all samples however no variants or genes were consistently implicated in disease (Table 4.3). One *PCNXL2* variant (T1969M) was identified in both triple-negative cases, DE5 and DE6, however, DE6 is compound heterozygous for variants in *PCNXL2* but DE5 only harbours this one variant in this gene.

Potentially recessive variants are identified in all samples except DE3 (Table 4.4). Only one of the 15 variants is found in more than one sample; a frameshift deletion in *UBTFL1* is present in both DE1 and DE4, both of whom present with triple-positive breast cancer.

4.4.3.2 Pathogenic Variation in Cancer Genes

Variation, classified as pathogenic by associated SIFT and PolyPhen2 scores, was identified in 25 genes listed in the COSMIC database. Twenty-eight variants were classified as pathogenic by at least one of either SIFT score or PolyPhen2 score (Table 4.5), a further 24 variants were identified which are either not classified as damaging by SIFT or Polyphen2, or are annotated as intronic splice sites (Table XII in Appendix). Three genes contain potentially damaging variants in multiple individuals; *CARS*, *NCOA1*, and *TCF3*. However, these variants do not occur in patients with the same subtype. Once again, no genes are consistently implicated by disease subtype.

4.4.3.3 Pathogenic Variation in HGMD Disease Genes

A total of 28 variants have been previously linked to disease in the HGMD variant database, 21 of these variants are classified as damaging by SIFT or PolyPhen2 score (Table 4.6). (Any variants that are intronic splice variants or are not classified as damaging by the SIFT or PolyPhen2 algorithms are listed in Appendix XIII). Again, each variant is only present in one sample and only one gene is flagged more than once: *MEFV*. The *MEFV* gene contains three variants that are all listed in HGMD. Two are present in DE8 and one in DE7.

4.4.3.4 Variation in Triple Positive Samples

The variants implicated in triple positive breast cancer differ from sample to sample, with no genes consistently mutated across individuals. All four patients harbour compound heterozygous variants, with at least one novel compound heterozygous variant per patient (Table 4.3). Of particular note are the *PER3* variants in DE1 which are both in potential exonic splice sites and have SIFT scores of 0, and the novel *SORBS1* variants in DE2 which are both deleterious by SIFT. Only DE1 and DE2 harbour potentially damaging recessive variants; DE1 harbours a nonsynonymous variant with SIFT score 0.02 in the *KIR2DS4* gene while DE2 harbours 7 recessive variants (Table 4.4).

4.4.3.5 Variation in Triple Negative Samples

Both triple negative patients harbour numerous compound heterozygous variants (10 in DE5 and 6 in DE6) (Table 4.3). In particular, DE5 harbours two novel variants in *SHROOM3*, one of which is predicted to be damaging by both SIFT and PolyPhen2. Both triple negative exomes harbour one potentially recessive variant (Table 4.4); DE5 has a variant in *C21orf63* that is predicted damaging by PolyPhen2 while DE6 harbours a frameshift deletion in the *ISCU* gene.

4.4.3.6 Variation in Samples with Family History

Both patients with familial disease harbour 6 potential compound heterozygous variants (Table 4.3); DE7 has two novel variants in *RTN3*, one of which has a SIFT score of 0 while the other is damaging by PolyPhen2, as well as a further novel variant in *TOM1* with SIFT score 0.01, DE8 harbours two clinical variants in *MEFV* which are predicted to be pathogenic by both dbSNP135 (Table 4.3) and HGMD (Table 4.6). These variants are related to Familial Mediterranean Fever, however it has been suggested that these two variants are actually often observed in controls and patients with these variants are often asymptomatic (Ryan et al., 2010). DE7 harbours one potentially recessive variant; a frameshift deletion in *LCE3C* (Table 4.4). DE8 contains 3 potentially recessive variants, two of which are nonsynonymous with SIFT scores of 0.05 and 0.01 (Table 4.4).

Table 4.5. Variants identified in known cancer genes catalogued in the COSMIC database

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
ABL1	2	9	ns	132719332	G140A	R47H	.	.	0.000	0.00	D	29	C	-1.284		+						
APC	14	5	ns	112207258	G8014A	A2672T	.	.	0.001	0.01	B	58	MC	.			+					
CARS	11	11	ns	2998115	C1177A	R393S	N	.	.	0.01	B	110	MR	-2.962			+					
CARS	15	11	ns	2995694	T1696C	W566R	.	.	0.000	0.00	B	101	MR	-1.284						+		
CBLB	12	3	ns	106903846	G1741A	V581M	.	.	0.000	0.05	P	21	C	-1.284	+							
CIITA	14	16	ns	10916980	A2941G	T981A	N	.	.	0.09	B	58	MC	-1.284							+	
COL1A1	32	17	ns	45623837	C2141T	A714V	N	.	.	0.09	D	64	MC	-0.568			+					
ERCC3	7	2	ns	127762886	C847T	R283C	.	.	0.002	0.00	P	180	R	2.190						+		
ERCC4	11	16	ns	13949041	C2087T	P696L	N	.	.	0.00	D	98	MC	.					+			
ETV5	7	3	ns	187280495	C455A	P152Q	N	.	.	0.23	P	76	MC	-1.284		+						
FANCG	1	9	ns/sp	35069445	A77G	Q26R	N	.	.	0.00	D	43	C	.		+						
MSN	4	X	ns	64866086	A254G	Y85C	N	.	.	0.00	D	194	R	-5.884					+			
MUTYH	9	1	ns	45570714	G640T	V214L	.	.	0.000	0.22	P	32	C	-0.568								+
MYH9	2	22	ns	35075092	C136T	L46F	.	.	0.003	0.00	P	22	C	-2.216						+		
NCOA1	19	2	ns	24834459	A3995G	N1332S	.	.	0.001	0.00	D	46	C	1.360		+						
NCOA1	19	2	ns	24834432	G3968A	G1323D	N	.	.	0.10	P	94	MC	-3.894						+		
NDRG1	6	8	ns	134340645	C337T	P113S	N	.	.	0.01	D	74	MC	-2.962							+	
NIN	14	14	ns	50302827	T1583A	L528Q	N	.	.	0.57	D	113	MR	-1.284	+							
NOTCH2	14	1	ns	120297762	T2292A	N764K	.	.	0.000	0.01	P	94	MC	-3.894						+		
NTRK1	16	1	ns	155118006	G2321A	R774Q	Cl	.	0.005	0.13	B	43	C	.				+				
NUMA1	6	11	ns	71411815	G235A	E79K	N	.	.	0.01	B	56	MC	3.934			+					
PCM1	5	8	ns	17840653	C467T	A156V	N	.	.	0.04	B	64	MC	1.730		+						
PMS1	2	2	ns	190364809	G29A	R10Q	N	.	.	0.01	D	43	C	0.394		+						
TCF3	16	19	ns	1566716	A1555G	K519E	N	.	.	0.14	B	56	MC	-1.284							+	
TCF3	18	19	ns	1566457	G1649T	R550L	N	.	.	0.00	D	102	MR	-5.884					+			
TPR	26	1	ns	184579820	G3443A	R1148H	.	.	0.001	0.01	D	29	C	0.768			+					
WAS	10	X	ns	48432328	G1267A	G423R	N	.	.	0.04	D	125	MR	-1.284					+			
XPC	15	3	ns	14163870	A2417G	K806R	N	.	.	0.00	P	26	C	0.394		+						

Where a specific variant is present in a sample this is indicated by +. Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined
B, benign; C, conservative; D, probably damaging; fd, frameshift deletion; fd/sp, frameshift deletion or potential exonic splice site; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; P, possibly damaging; R, radical

Table 4.6. Variants identified as disease-causing and catalogued in HGMD

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	Associated disease from HGMD
AR	1	X	ns	66682359	G646A	G216R	N	.	.	0.00	D	125	MR	.		+							Androgen insensitivity syndrome
BCS1L	3	2	sg	219234120	C166T	R56X	Cl	.	.	0.11	.	.	.	-0.568			+						Complex 3 deficiency
CDH23	26	10	ns	73138932	C3178T	R1060W	.	.	0.000	0.00	D	101	MR	0.768					+				Non-syndromic autosomal recessive deafness
CFTR	8	7	ns	116967410	G890A	R297Q	.	.	0.002	0.02	B	43	C	-2.962			+						Cystic fibrosis
CNGB3	11	8	ns	87714208	G1208A	R403Q	.	.	0.002	0.13	D	43	C	0.749				+					Progressive cone dystrophy
DISC1	4	1	ns	229952430	G1253A	R418H	.	.	0.000	0.00	P	29	C	1.730							+		Schizophrenia
F11	15	4	ns	187446742	G1858C	E620Q	.	.	0.000	0.16	B	29	C	.							+		Factor XI deficiency
FECH	2	18	ns/sp	53398312	C185G	P62R	.	.	0.002	0.45	P	103	MR	-1.284						+			Protoporphyria, erythropoietic
GUSB	3	7	ns	65082276	G454A	D152N	.	.	0.001	0.00	P	23	C	-2.246					+				Mucopolysaccharidosis VII
MEFV	3	16	ns	3239587	C1105T	P369S	Cl	.	0.006	0.02	D	74	MC	-1.284								+	Mediterranean fever
MEFV	3	16	ns	3239469	G1223A	R408Q	Cl	.	0.006	0.10	B	43	C	-3.894								+	Mediterranean fever
MTFMT	4	15	ns	63100924	C626T	S209L	.	.	0.002	0.28	D	145	MR	3.152								+	Leigh syndrome and combined OXPHOS deficiency
NLRP3	5	1	ns	245653966	G598A	V200M	Cl	.	0.010	0.07	B	21	C	-0.568				+					Familial cold autoinflammatory syndrome
PKD1	23	16	ns	2093766	C8293T	R2765C	.	.	0.007	0.01	D	180	R	-1.626	+								Polycystic kidney disease 1
PNKD	2	2	ns	218844377	G97C	A33P	Cl	.	0.000	0.00	D	27	C	-5.884		+							Paroxysmal nonkinesiogenic dyskinesia
POMC	4	2	ns	25237552	C706G	R236G	Cl	.	0.005	0.00	D	125	MR	.				+					Obesity
PON3	9	7	ns	94827315	G971A	G324D	.	.	0.003	0.00	D	94	MC	.						+			Paraoxonase activity variant
SLC22A5	3	5	ns	131747881	C641T	A214V	N	.	.	0.02	D	64	MC	3.323		+							Carnitine deficiency, systemic primary
TACR3	5	4	ns	104730365	C1321T	R441C	.	.	0.000	0.00	B	180	R	.	+								Hypogonadotropic hypogonadism
TMPRSS6	18	22	ns	35792119	G2383A	V795I	.	.	0.009	0.56	D	29	C	.					+				Iron deficiency anaemia
TTC21B	23	2	ns	166455691	C3004G	L1002V	.	.	0.009	0.00	D	32	C	0.394				+					Meckel-Gruber-like syndrome

Where a specific variant is present in a sample this is indicated by +. Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined
B, benign; C, conservative; D, probably damaging; fd, frameshift deletion; fd/sp, frameshift deletion or potential exonic splice site; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; P, possibly damaging; R, radical

4.4.4 *BRCA1* and *BRCA2* Variants

Variant annotation identified 11 variants located in *BRCA1* and 10 *BRCA2* variants in at least one affected sample (Table 4.7). Almost half of the variants are synonymous so do not affect the amino acid sequence and are thus highly unlikely to have a significant effect on the gene. The remaining 11 variants have greater potential to be clinically relevant because they are either nonsynonymous, affecting the amino acid sequence, or are stop-gains which introduce a premature stop codon and result in a truncated protein.

To assess the clinical significance of the variants, *in silico* functional predictions were obtained for the 11 variants. Clinical information was obtained for each variant from the Breast Cancer Information Core (BIC) (http://research.nhgri.nih.gov/bic/), Align GVGD (http://agvgd.iarc.fr/agvgd_input.php), the Breast Cancer genes IARC (http://brca.iarc.fr/PRIORS/) and the Human Genome Mutation Database. None of the variants is predicted to have a significant effect on the phenotype or a pathogenic role in breast cancer (Table 4.7).

4.4.5 Post-zygotic Mosaic Variants

All potential post-zygotic mosaic variants were identified in the breast cancer susceptibility genes *BRCA1*, *BRCA2* and *TP53*. Candidate variants were selected if there was evidence for at least three alternative allele reads, with at least one read on both the forward and negative strands, in a reference homozygous call. A total of five candidate variants were identified, all of which were intronic (Appendix XIV). Four of these variants were flagged as 'novel' in the analysis pipeline.

Two variants were identified in the *BRCA1* gene, one in the *BRCA2* gene and two in the *TP53* gene. Further analysis of these variants in fact indicated that all except those in *TP53* were in homopolymer tract regions and thus may represent alignment artefacts, rather than bona fide variation. IGV was used to inspect each variant identified and confirmed that all variants in *BRCA1* and *BRCA2* were in homopolymer repeat regions and were, thus, likely to be the result of poor alignment to the reference genome.

Table 4.7. Characterisation and clinical significance of all *BRCA1* and *BRCA2* variants

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	Clinical significance of variant determined from HGMD
BRCA1	23	17	ns	38451234	A5579C	H1860P	.	.	0.013	0.00	D	77	MC	+			+	+		+	+	.
BRCA1	15	17	ns	38476501	G4956A	M1652I	.	.	0.017	0.25	B	10	C		+			+				Likely neutral or not clinically significant
BRCA1	15	17	ns	38476620	A4837G	S1613G	.	0.37	0.326	0.10	P	56	MC		+	+		+				Neutral or not clinically significant. In LD with E1038G.
BRCA1	12	17	sy	38487996	T4308C	S1436S	.	0.37	0.326	.	.	.	.		+	+		+				.
BRCA1	10	17	ns	38497526	A3548G	K1183R	.	0.37	0.324	1.00	B	26	C		+	+		+				K residue possibly has protective role - associated with increased breast cancer risk
BRCA1	10	17	ns	38497961	A3113G	E1038G	.	0.37	0.325	0.06	P	98	MC		+	+		+				Neutral or not clinically significant. In strong LD with S1613G.
BRCA1	10	17	ns	38498462	C2612T	P871L	.	0.38	0.335	1.00	B	98	MC		+	+		+				Neutral or not clinically significant.
BRCA1	10	17	sy	38498763	T2311C	L771L	.	0.37	0.324	.	.	.	.		+	+		+				.
BRCA1	10	17	sy	38498992	C2082T	S694S	.	0.37	0.324	.	.	.	.		+	+		+				.
BRCA1	10	17	ns	38500007	A1067G	Q356R	.	0.03	0.065	0.03	D	43	C				+					Neutral or not clinically significant. Rare R allele possibly has protective role.
BRCA1	10	17	sy	38500381	G693A	T231T	.	.	.	.	.	.	.			+						.
BRCA2	10	13	ns	31804729	A1114C	N372H	Cl	0.33	0.290	0.19	.	68	MC				+		+		•	Homozygotes for H residue have significantly increased breast cancer risk. Associated with women from high-risk families without any *BRCA1/2* mutations. No effect on BRCA2 function in any assay.
BRCA2	11	13	sy	31809888	A3396G	K1132K	.	0.25	0.305	.	.	.	.	+		•		+	+	+		.
BRCA2	11	13	sy	31810299	T3807C	V1296V	.	0.23	0.186	.	.	.	.	+	•					+		.
BRCA2	11	13	sy	31811055	A4563G	L1521L	.	1.00	1.000	.	.	.	.	•	•	•	•	•	•	•	•	.

BRCA2	11	13	ns	31812236	C5744T	T1915M	.	0.02	0.028	0.13	B	81	MC							+			More common in controls - possible protective role (except when a *CHEK2* variant is present).
BRCA2	11	13	sy	31813005	G6513C	V2171V	.	.	1.000	.	.	.	.	•	•	•	•	•	•	•	•	•	.
BRCA2	14	13	sy	31827232	A7242G	S2414S	.	0.15	0.212	.	.	.	.			•		+	+	+			.
BRCA2	14	13	ns	31827387	T7397C	V2466A	.	0.99	1.000	0.54	.	64	MC	•	•	•	•	•	•	•	•	•	No clear evidence that this variant is pathogenic or that it is not clinically significant.
BRCA2	19	13	sy	31842628	G8421A	S2807S	.	.	0	.	.	.	.		+								.
BRCA2	27	13	sg	31870626	A9976T	K3326X	.	.	0.009	0.16	.	.	.							+			93 residues removed but not associated with increased cancer risk. Frequently observed with no effect on BRCA2 function in any assay.

Where a specific variant is present in a sample this is indicated by + if heterozygous or • if homozygous.
Functional scores/predictions (from SIFT, PolyPhen2, or Grantham) considered damaging are underlined
B, benign; C, conservative; D, probably damaging; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; P, possibly damaging; R, radical; sy, synonymous

Genetic dissection of early-onset breast cancer and other genetic diseases

Protein-truncating variants (PTVs) in *PPM1D* have recently been identified as associated with a predisposition to both breast and ovarian cancer in a large-scale study of over 13,500 cases and controls (Ruark et al., 2013). A total of 26 PTVs, clustering in the final exon of the *PPM1D* gene, were identified in this study. The sequence data for all eight exomed breast cancer cases were therefore interrogated for any evidence of PTVs present at low-levels. Nineteen stopgain and insertion/deletion variants were identified in exon 6 of *PPM1D*, however the percentage of reads for the alternate allele was consistently below 0.015%; in most cases only one read had evidence of an alternate allele. This is likely to be due to errors in sequencing rather than evidence of true variants.

4.5 Discussion

Exome sequencing of eight patients with various subtypes of early-onset breast cancer demonstrated the unique genetic profile of individual patients; a wide range of genes were identified as harbouring rare variation in each sample but no genes were consistently implicated in disease.

The hypothesis of this small-scale study was that cases with 'extreme' phenotypes – very early age of onset, triple-negative disease, strong family history of early-onset breast cancer – might arise from Mendelian or near-Mendelian inheritance, rather than representing complex cases arising from polygenic inheritance. Mendelian inheritance patterns in breast cancer are usually suggestive of variation in a highly penetrant gene, particularly *BRCA1, BRCA2* or *TP53*. However, it was determined prior to exome sequencing that known variation in any one of these genes was not responsible for the cancer in any of these cases. Therefore, the cases were subjected to exome sequencing to identify any rare or novel variants that were recessive, potentially compound heterozygous, or predicted deleterious by *in silico* predictors, that could be the causal mutation(s) in any of these patients. Multiple tiers of analysis, including analysis of pathway-specific genes and high-penetrance susceptibility genes, were implemented to thoroughly explore the genetic variation present in all affected samples. Cases with particularly early-onset are likely to have a strong genetic component and it was hoped that a small number of deleterious mutations may have been detected through exome sequencing.

TP53-pathway genes were selected for interrogation because *TP53* is a tumour suppressor implicated in many cancer types that interacts with a number of high- and moderate-penetrance breast cancer susceptibility genes (Gasco et al., 2002). Moreover, germline *TP53* mutations have been associated with Li-Fraumeni syndrome, of which early-onset breast cancer is a common feature (Li and Fraumeni, 1969; Malkin et al., 1990), and early-onset cases of HER2+ breast cancer (Wilson et al., 2010; Melhem-Bertrandt et al., 2012) making *TP53* a strong candidate gene for cases of early-onset breast cancer. Due to the limited variants identified in over 300 *TP53*-related genes, all rare and potentially deleterious variants in any gene of the exome were also reported.

Potentially interesting compound heterozygous variants were identified in a number of the exome samples (Table 4.2 and Table 4.3), representing a range of genes. The majority of these variants were predicted deleterious by at least one of SIFT or PolyPhen2. Of particular note were the potential compound heterozygous variants identified in *BRCA1* and *BRCA2* (Table 4.2). The occurrence of two deleterious *BRCA1* mutations in one patient is considered to be highly unlikely, at a frequency of 1 in 10,000, based on evidence that homozygous and compound heterozygous deleterious *BRCA1* mutations are often embryonic lethal (Easton et al., 2007a). Therefore, the likelihood that both identified *BRCA1* variants are deleterious is very low. Evidence from HGMD suggests that one of the *BRCA1* variants, Q356R in sample DE4, has been previously identified but is predicted to have little or no clinical significance (Abkevich et al., 2004) and may have a protective effect against breast cancer (Wenham et al., 2003). The potential compound heterozygous *BRCA2* variants observed in sample DE6 (Table 4.3) are both recorded in HGMD (Table 4.6). The T1915M amino acid substitution is actually found more often in control samples without any evidence of disease rather than in breast cancer cases, suggesting a protective effect (Serrano-Fernández et al., 2009). The stopgain K3326X variant, which will result in a potentially non-functional truncated protein, does not appear to increase risk of breast cancer (Mazoyer et al., 1996). The variant is frequently observed but a functional effect is yet to be identified (Wu et al., 2005). In addition, the stopgain variant is usually found in control populations, so is presumed to have little associated cancer risk (Wu et al., 2005). Consequently, it is considered unlikely that these *BRCA1* and

BRCA2 variants are the cause of disease in the patients in which they are found.

Rare or novel variants identified in any one of the affected samples were cross-referenced against HGMD to identify any potentially disease-related variation. Despite 21 of these variants being identified as disease-related (Table 4.6), none appears to have any role in breast cancer development. A number of diseases and disorders are represented by these variants but none share similar development pathways to breast cancer or can explain the occurrence of disease in these 8 samples. The clinical relevance of the 21 variants in these samples is unclear; it appears that many of the variants may actually only cause mild forms of disease or require the presence of certain variants for disease to manifest. It is also possible that these variants exhibit an age-dependent penetrance so these patients do not yet exhibit the associated phenotype. Another possibility is that these variants do not always cause disease when they are present in an individual, supporting the notion that many diseases are polygenic and certain variants will only cause disease in the presence of other specific variants. While these variants are interesting for demonstrating the complexity of an individual exome, the relevance to breast cancer in these patients is unclear.

Each exome was found to harbour a number of rare variants that were recessive or potentially compound heterozygous, but implicating any of these variants as the causal mutations requires functional evidence of an adverse effect of the variant on protein function or expression. Based on the hypothesis that these samples may result from Mendelian-like inheritance, *BRCA1* and *BRCA2* were interrogated because they are highly-penetrant early-onset breast cancer genes. A handful of variants were observed in *BRCA1* and *BRCA2* (Table 4.7). *In silico* functional predictions were obtained for each variant as well as information regarding pathogenicity from HGMD. All 11 nonsynonymous and stopgain variants have been identified previously; all have a SNP rsID and all are recorded in the NHLBI Exome Variant Server database. Seven of the variants have minor allele frequencies exceeding 5%, thus they are not rare alleles and are unlikely to have an important role in breast cancer due to their prevalence in non-disease samples. Variant V2466A is observed in both the 1000 Genomes Project database and NHLBI ESP with frequencies of 99.2% and 99.96% respectively, suggesting that the reference genome used for alignment

contained the rare allele, which was therefore considered to be the reference allele. This variant is listed in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/), here the nucleotide change is identified as C>T rather than the T>C change highlighted in our results. Therefore, it is likely that the C allele is the common allele. HGMD found no conclusive evidence that this variant was pathogenic but nor did it disprove the possibility of clinical significance. The N372H variant does not appear to be damaging by any of the *in silico* predictions but it is listed as 'probably pathogenic' in dbSNP135. This variant is observed in ~30% of the population (based on MAF values from 1000G and NHLBI ESP) suggesting that it may not have an important role in disease. HGMD presents conflicting evidence regarding the pathogenicity of this variant however: it was previously suggested that homozygotes for the alternate allele have an increased risk of breast cancer (Healey et al., 2000), but, another study could find no effect of this variant on *BRCA2* function in any assay and instead suggested that this variant has little or no cancer risk (Wu et al., 2005).

Analysis of *BRCA1, BRCA2* and *TP53* for potential post-zygotic mutations identified 5 variants, of which 4 were flagged as novel, potentially warranting further investigation. All variants in *BRCA1* and *BRCA2* were, however, in homopolymer repeat regions and are therefore more likely to result from errors in the sequence alignment process rather than represent actual variation. The two variants in *TP53* however, do appear to be real variants after inspection in IGV. Both are present in sample DE6 only. The TP53 variants were cross-referenced against the IARC TP53 database (http://p53.iarc.fr/) to determine whether these variants had previously been identified. Neither variant was recorded in the germline mutation database, somatic mutation database, or gene variation database. Due to a lack of evidence relating to these variants, the clinical relevance of these mutations remains unclear.

To increase the power to detect causal variants other samples with similar extreme phenotypes or multi-case families including distantly related affected individuals, need to be included in the analysis. In particular, it would have been desirable to include affected relatives of samples DE7 and DE8, both of whom have a family history of early-onset disease. The presentation in the family of sample DE8 is particularly interesting because sample DE8 had a sister diagnosed with breast cancer at age 23 (Figure 4.2). Two cases of

particularly early-onset disease in first-degree relatives are suggestive of a recessive inheritance pattern in this family. Exome sequencing of the affected sister and some unaffected relatives may have allowed for identification of the causal mutation in this family, however, the sister was deceased at the time the family history was taken for sample DE8, so exome sequencing would not have been an option. Sample DE7 has a cousin diagnosed with breast cancer at age 35 (Figure 4.1) who could have been included in the analysis to add extra power. As cousins they are third-degree relatives who will share < 50% of exonic genetic variants, so inclusion of this sample would increase power to detect the causal variant(s) in sample DE7. It is, however, possible that the two cases in the family of DE8 actually arise from unrelated mutations and represent different subtypes of the disease rather than the same subtype.

The importance of the variants identified within this study is currently unclear. Any variants identified through exome sequencing should be confirmed with Sanger sequencing (Sanger et al., 1977) because high-throughput NGS technology produces short DNA reads that contain more sequencing errors than Sanger sequencing reads (Shendure and Ji, 2008, Altmann et al., 2012). Variants can be prioritised for Sanger sequencing using potential functional roles for the variants, inferred from *in silico* prediction algorithms: prediction scores can be used to identify potentially pathogenic variation. While these scores are useful as an indication of which variants are potentially more damaging they must be used with caution because different algorithms can give vastly different predictions for the same variant (Kumar et al., 2011, Liu et al., 2011). A number of variants identified in this study are predicted deleterious by SIFT and benign by PolyPhen2 (or vice versa), however, SIFT and PolyPhen2 tend to concur more often than not. The relatively large number of rare variants predicted to be deleterious in each patient leads to the possibility that these cases are not Mendelian-like disorders but are polygenic. Many cases of breast cancer are likely to arise from a polygenic model (Pharoah et al., 2002, Stratton and Rahman, 2008) so it is plausible that these cases are in fact polygenic.

Quantitative splice site scores were provided for all exonic variants, where possible, to identify variants with potential roles in exonic splicing. Of the 102 rare variants identified from all genes of the exome (Table 4.4- Table 4.6) splice site scores were obtained for 77 variants. 27 of these variants have positive

splice scores suggesting a role for these variants in modulating exonic splicing. Splice variants play vital roles in the production of functional proteins. Variation in the human genome that causes the production of new splice variants has the potential to affect splicing, causing exon skipping and activation of ectopic splice sites. Nonsynonymous and synonymous variants do not only alter the nucleotide sequence of the genome but can lead to exon skipping which can cause the exclusion of protein domains from the final protein (Woolfe et al., 2010). This is often more deleterious than nonsynonymous changes that result in amino acid changes. Another potential result of exonic variants is the production of a new splice site that is then used in preference to the original splice sites, resulting in a shortened exon (Woolfe et al., 2010). However, it is extremely difficult to establish the role of exonic variants in modulating splicing without functional evidence.

Analysis of the exomes of 8 breast cancer patients has identified rare, potentially pathogenic, variation in a wide range of genes. It has not been possible, however, to identify any variants or genes that appear to show a definite link to any of the breast cancer subtypes. The filtering regimes used in the different tiers of analysis were fairly stringent, making it likely that a proportion of relevant variation was removed. For example, the dbSNP database contains a small number of variants that are pathogenic (Gilissen et al., 2012, Walsh et al., 2010), so filtering against this database will remove potentially pathogenic variants that may be disease-related. Genetic variation databases, namely the 1000 Genomes Project and Exome Variant Server databases, were also used to exclude any variation with frequency >1%. It is possible however, that important variants with clinical relevance will be contained within these databases. Combined, these two databases contain genetic variation taken from over 6000 samples so it is expected that some of these individuals may harbour the same disease as samples in the study, and by filtering against these databases important variation will be excluded from consideration. In an attempt to minimise this risk a cut-off of 1% prevalence in the population was applied, thus any rare variants observed in less than 1% of these patients were retained for the analysis. This decision was based on the assumption that very few samples in these databases will have early-onset breast cancer and any variants present at higher frequencies will be too common to be responsible for this disease.

4.6 Conclusion

Exome sequencing of patients with particularly early-onset breast cancer identified many rare and novel variants per individual, which may or may not be related to disease onset and/or progression. The 8 selected individuals represent different breast cancer subtypes that are likely to arise via different mechanisms, involving distinct mutations but all had very early age-of-onset, suggestive of Mendelian-type inheritance. The genetic profile of each individual is highly unique, regardless of the breast cancer subtype they present with; there are no specific variants or consistently mutated genes implicated in any of the subtypes represented by this cohort. Exome analysis was unable to implicate any genes as causal in this study; this may be because important variants were filtered out of the results or because the cases are actually polygenic and the result of multiple variants. Further analysis of these samples alongside many more samples presenting with the same subtypes will be necessary to improve understanding of the relevance of the detected variants and improve the search for and identification of causal variants. Functional analysis of any candidate variants is also necessary to improve understanding of how the mutations influence cancer development and progression.

Chapter 5: Next Generation Exome Sequencing in Syndromic and Non-syndromic Cleft Lip and Palate Patients

5.1 Background

Cleft lip with or without cleft palate (CLP) is a common birth defect affecting approximately 1 in 700 live births (Dixon et al., 2011). More than 200 genetic syndromes include CLP in the phenotype while cleft palate only (CPO) features in over 400 disorders (Mossey et al., 2009).

The inheritance patterns underlying syndromic and non-syndromic forms of CLP/CPO differ, with Mendelian inheritance usually observed in syndromic forms while non-syndromic forms tend to be complex disorders influenced by multiple genetic and environmental factors (Mossey et al., 2009, Mangold et al., 2011). The genes underlying many of the syndromic forms of CLP/CPO have already been characterised (Leslie and Marazita, 2013), with a single variant often sufficient to explain the phenotype. In non-syndromic cases of CLP/CPO (NSCLP) fewer than 20 genes have so far been implicated in disease aetiology (Collins et al., 2014), with many cases likely to be unexplained by variants in these genes. Genes responsible for syndromic clefting disorders, such as *IRF6* which causes cases of Van de Woude syndrome and popliteal pterygium syndrome, are also important in NSCLP pathology (Mossey et al., 2009) and will likely harbour causal mutations in NSCLP cases.

The importance of rare variation in many complex phenotypes is becoming clear. Rare variants are likely to play an important role in non-syndromic CLP/CPO cases since nearly three quarters of rare variants identified in NSCLP cases are not found in control samples (Leslie and Murray, 2013).

The advent of next-generation sequencing technology has provided the opportunity for the exploration and characterisation of all genetic variation contained in the entire coding region of an individual's genome, regardless of allele frequency. Exome sequencing will be valuable for identifying the causative mutations in syndromic and non-syndromic forms of CLP/CPO,

although identification of the causal mutation is much more likely in patients with Mendelian inheritance patterns. The entire spectrum of rare variation harboured by CLP/CPO affected individuals can be revealed through exome sequencing, allowing for greater understanding of the genetics of CLP/CPO (Collins et al., 2014).

5.2 Aim

To characterise the spectrum of rare and novel variation harboured in the coding regions of the genome in Colombian patients with syndromic or non-syndromic CLP.

5.3 Materials and Methods

5.3.1 Exome Sequencing

DNA samples were taken from 18 individuals from 11 multi-case families collected at Operation Smile, Bogota, Colombia (Table 5.1). All individuals were affected with syndromic (6 cases) or nonsyndromic (12 cases) cleft lip with or without cleft palate. Exome sequencing was carried out at Oxford Gene Technology using a HiSeq 2000 sequencer. Exome capture was done on the Agilent SureSelect Human All Exon V4 capture kit for sample CL001_1 and on the Agilent SureSelect All Human Exon V5 capture kit for all other samples.

5.3.2 The Exome Pipeline

Raw exome sequence data in fastq format were analysed using the in-house Exome Pipeline. The Exome Pipeline carries out raw sequence alignment to the reference sequence, variant calling and variant annotation.

Resultant paired end reads were aligned against the human genome reference sequence 19 (hg19) using Novoalign v2.08.02 (Novocraft Technologies: http://www.novocraft.com/main/index.php).

All single nucleotide substitutions and small insertions or deletions present in each individual were identified using SAMtools v0.1.18 (http://samtools.sourceforge.net/). Variants were annotated with respect to

Table 5.1. Description of samples selected for exome sequencing

Family	Number of affected individuals	Number of exome sequenced individuals	Individual ID of exome sequenced samples	Sample ID	Age at examination (yrs)	Gender	Syndromic or nonsyndromic	Phenotype	Relationship to other sequenced samples
CL001	6	2	CL001-SC	CL001_1	.	M	Nonsyndromic	CLP	Brother of CL001-MAC
			CL001-MAC	CL001_2	.	M	Nonsyndromic	CLP	Brother of CL001-SC
CL002	3	2	CL002-3507	CL002_1	19	M	Nonsyndromic	UCLP (left)	Brother of CL002-3548
			CL002-3548	CL002_2	18	F	Nonsyndromic	UCLP (left)	Sister of CL002-3507
CL003	2	1	CL003-3246m	CL003_1	.	F	Nonsyndromic	CPO (submucous)	.
CL004	2	2	CL004-3266m	CL004_1	.	F	Nonsyndromic	UCLP (left)	Mother of CL004-3266d
			CL004-3266d	CL004_2	1.7	F	Nonsyndromic	BCLP	Daughter of CL004-3266m
CL005	5	1	CL009-2108p	CL005_1	.	M	Syndromic	Van de Woude syndrome	.
CL006	1	1	CL006-KP	CL006_1	9	M	Syndromic	Nager syndrome	.
CL007	5	1	CL007-3114jg	CL007_1	4	M	Nonsyndromic	UCLP (left)	.
CL010	5	2	CL010-2112-Pro	CL010_1	30	M	Nonsyndromic	BCLP	Son of CL010-2112-Pat
			CL010-2112-Pat	CL010_2	60	M	Nonsyndromic	BCLP	Father of CL010-2112-Pro
CL012	10	2	CL012-3753-Pro	CL012_1	34	F	Syndromic	Incontinentia pigmenti	Half niece of CL012-3753-Unc
			CL012-3753-Unc	CL012_2	.	M	Syndromic	Facial clefting	Half uncle of CL012-3753-Pro
CL014	2	2	CL014-3655-Pro	CL014_1	0.25	M	Nonsyndromic	UCLP (left)	Nephew of CL014-3655-Aunt
			CL014-3655-Aunt	CL014_2	.	F	Nonsyndromic	UCLP (left)	Aunt of CL014-3655-Pro
CL018	5	3	CL018-3868-Pro	CL018_1	.	F	Syndromic	Pierre Robin syndrome	Niece of CL018-3868-Unc, second cousin of CL018-3868-Cous
			CL018-3868-Unc	CL018_2	.	M	Syndromic	CPO	Uncle of CL018-3868-Pro
			CL018-3868-Cous	CL018_3	.	M	Syndromic	BCLP	Second cousin of CL018-3868-Pro

CLP – cleft lip and palate; UCLP – unilateral CLP; BCLP – bilateral CLP; CPO – cleft palate only

genes and transcripts using ANNOVAR (Wang et al., 2010). ANNOVAR was used to cross-reference all variants with dbSNP137, 1000 Genomes Project database, and the European American population in NHLBI ESP database. Variants were also cross-referenced against our in-house exome database; this contains ~300 exomes from unrelated individuals presenting with a range of other disorders, including autoimmune disorders, imprinting disorders, solid tumour and blood cancers.

Seven *in silico* functional prediction scores were obtained for all SNVs using KGGseq v0.4 (Li et al., 2012, Li et al., 2013) which implements dbNSFP v2 (Liu et al., 2011, Liu et al., 2013). Scores of deleteriousness of each variant were obtained from SIFT (Ng and Henikoff, 2003, Ng and Henikoff, 2001, Ng and Henikoff, 2002); PolyPhen2 HumVar (synonymous cases) and HumDiv (non-synonymous cases) (Adzhubei et al., 2010); LRT (Chun and Fay, 2009); and MutationTaster (Schwarz et al., 2010). Conservation scores were obtained from GERP++ (Davydov et al., 2010) and PhyloP 100way vertebrate (Siepel et al., 2006). The multiple annotation combination PHRED-scaled CADD score (Kircher et al., 2014) was obtained for all variants where possible. KGGseq implements a logistic regression model, trained on an internal dataset, to obtain a conditional probability that a variant is Mendelian disease-causal based on 13 prediction scores available in KGGseq. Each variant is subsequently classified as disease-causal or neutral using the logistic regression model (logit model) (Li et al., 2012, Li et al., 2013). Conditional probabilities and classification from the logit model were obtained for all variants where possible. Grantham scores were also included for all nonsynonymous variants (Grantham, 1974). (See Appendix XV for description of algorithms used to predict effect of substitutions).

5.3.3 Candidate Gene Selection

A comprehensive list of candidate genes for nonsyndromic and syndromic forms of cleft lip with or without cleft palate was constructed from information in the Human Gene Mutation Database (HGMD professional http://www.hgmd.org/), accessed in July 2014 using the following search terms: cleft lip; cleft palate; cleft; syndactyly; brachydactyly; Pierre Robin; incontinentia pigmenti; Nager syndrome; hyperpigmentation; craniofacial;

clubbing; dysmorphic; dysmorphia; micrognathia. This list comprised 363 genes. Additional genes were included from interrogation of OMIM (http://omim.org/, accessed July 2014 using the same search terms as described above). Further CLP-related genes from the review by Collins et al (2014) were also included. The complete list of 865 interrogated genes is given in Appendix XVI.

5.3.4 Potentially Damaging Variation in Candidate Genes

Exome sequence data for the 18 samples were cross-referenced against the panel of candidate genes and only variants present in these genes were retained. To understand the spectrum of potentially damaging variation all non-synonymous, stopgain and insertion or deletion variants with an alternative allele frequency of < 1% in 1000 Genomes Project database (http://www.1000genomes.org/) were identified. Variants were further excluded if they: (i) were present in homopolymer tracts or repeat regions as they are likely to reflect alignment errors of sequence reads to the reference genome; (ii) had read depth < 10; (iii) were located in genes identified as 'highly mutable', based on Fuentes Fajardo et al., (2012) and a custom program to identify genes with a higher number of mutations per base pair than expected; (iv) had a strand bias or base quality bias.

5.3.5 Rare Variant Association

To assess the potential role of rare variants in NSCLP in the cohort of Colombian patients, all genes containing a variant were tested for association with the phenotype using SKAT-O (Optimal unified Sequence Kernel Association Test, (Lee et al., 2012)). A control sample of 107 whole-exome sequenced individuals was selected from the Soton Exome Database. Control samples were not affected with cleft lip and/or palate and were selected from an in-house exome database of disease and control exomes. The breakdown of samples selected as controls was: 34 cancer samples, 3 complex disease samples, 33 imprinting disorder samples, 35 Mendelian disease samples, and 2 non-disease control samples.

5.3.5.1 SKAT-O

For each individual gene, cases were compared to controls to test for an excess of rare variants. SKAT-O was implemented in EPACTS v3.2.5 (http://genome.sph.umich.edu/wiki/EPACTS).

SKAT (Wu et al., 2011) is a non-burden test that uses a multiple regression model to regress the phenotype on genetic variants and covariates. In situations where most variants within a gene/region are causal and have effects with similar magnitudes, burden tests are more powerful. SKAT, on the other hand, is more powerful when a gene/region contains both causal and neutral variants, or protective and damaging variants. Burden tests collapse all variants within a region into a single variable and then regress the phenotype on this variable to test the cumulative effect of rare variants on the phenotype. Burden tests make the assumption that all rare variants within a region are causal and exert their effects in the same direction. Nonburden tests, such as SKAT, do not make these assumptions; instead they aggregate the individual variant-score test statistics with weights. SKAT-O (Lee et al., 2012) therefore includes both burden testing and nonburden testing to find the test that is optimal for the particular scenario under investigation. SKAT-O differs from SKAT in that it combines burden testing with the non-burden testing of SKAT to fit the optimal model.

SKAT uses a weight function because upweighting causal variants and downweighting non-causal variants makes the test more powerful. However, the causal variants are often unknown prior to testing, therefore weights a derived from the beta distribution density function:

$$\sqrt{w_i} = Beta(MAF_i; a_1, a_2)$$

with weights w_i, and beta distribution density function parameters a_1 and a_2 evaluated at the minor allele frequency (derived from all cases and controls in the sample population) of the i-th variant. By default $a_1 = 1$ and $a_2 = 25$ because this increases the weight of rare variants (MAF < 1%) while still assigning non-zero weights to moderately rare variants (MAF = 1-5%).

5.3.5.2 Identification of Variant Base Pair Locations and Genotype Calling

All genomic locations containing a variant in at least one of the 119 exome samples to be included in the SKAT-O test were identified from fully annotated variant files. The Soton Exome pipeline calls all variants from BAM files using SAMtools and all identified variants are annotated by ANNOVAR to produce the variant files. In order to carry out rare variant association testing with SKAT-O, VCF files containing genotype calls for all locations of interest were produced for each sample.

Base pair locations were extracted from the variant files for each sample. However, the base pair locations of insertion and deletion variants (indels) identified from the variant files differ from the locations stored in the BAM files because ANNOVAR annotation requires a specific input format, which is different to the VCF format produced by SAMtools. Therefore, to extract the correct genotype calls, the genomic locations of all indels had to be reverted back to the original VCF format.

The base pair location of indels in the variant files depends on the number of bases in the reference allele column of the original VCF file. If the reference allele is only one base then the location remains the same in the variant file. If the number of bases in the reference allele column > 1 then the new location in the variant file is the original location plus the number of bases in the reference column, minus one.

The change in base pair location for deletions depends on the number of bases in the alternative allele column of the original VCF file; the new location in the variant file is the original location plus the number of bases in the alternative allele column of the original VCF file.

Indel variant files were produced for each sample from the corresponding VCF file. All variants annotated ‘INDEL’ in the original VCF files were extracted. The VCF files contain many poor quality variants and variants in non-exonic regions of the genome that are filtered out prior to creation of the variant files in the Soton Exome pipeline. Therefore, many of the indels that appear in the indel variant files will not be in the original variant files, only those that do appear in the original variant files need to be extracted. This is achieved by comparing

the original base pair location in the indel variant file (taken from the VCF file) with the altered base pair locations in the variant file and if there is a match, extracting the original base pair location. These locations were added to the base pair locations for SNVs to produce a list of variant base pair locations per sample. All variant locations identified in all samples were amalgamated to produce a final list of ~136,000 locations across all chromosomes.

Samples were exome sequenced using four exome capture kits: Truseq, Agilent SureSelect Human All Exon V3, V4 or V5. To minimise the amount of missing genotype data, only variant locations covered by all four capture kits were considered in the test. A BED file of base pair locations covered by all four capture kits was produced using BEDTools multi-intersect (http://bedtools.readthedocs.org/en/latest/) (see Appendix XVII for a breakdown of the intersection of the number of locations covered by each capture kit). In total, ~88,000 of the identified variant locations were contained within the intersection of locations covered by all capture kits.

Genotype calls for all ~88,000 variant locations were made for each sample. SAMtools was used to make the calls from individual sample BAM files and produce VCF files for each sample. Resultant VCF files were merged using VCFtools to produce a single VCF file of all variant locations with genotype calls from all 119 samples.

5.3.5.3 Quality Control

Custom scripts were used to remove any variants that were not biallelic or had missing genotype calls in more than 5% of samples. Four control samples with missing calls in more than 5% of locations were removed from consideration. Hardy-Weinberg equilibrium tests were applied to all 103 remaining control samples to remove any variants that fail Hardy-Weinberg equilibrium at $p < 0.0001$, implemented in VCFtools. A final VCF file containing genotype calls for 23,053 base pair locations in 115 exome samples was produced.

5.3.5.4 Rare Variant Association Testing with SKAT-O

The VCF file containing all genotype calls in all 115 samples was annotated with respect to genes against gencodeV7 using EPACTS v3.2.5 (http://genome.sph.umich.edu/wiki/EPACTS). A ‘groupfile’ of all variants

(nonsynonymous, insertion, deletion, start loss, stop loss, stop gain) in all genes was produced from the annotated VCF file. Variants were identified in 4,895 genes.

SKAT-O was applied to all variants in all 4,895 genes contained within the groupfile using the genotype calls in the annotated VCF file. SKAT-O was implemented in EPACTS v3.2.5. Due to a sample size of < 2000 samples an adjustment for small sample size was included in the test.

The p-values for association of each gene with the phenotype differ for each run of SKAT-O because the test estimates a null model for each run. The null model is included in producing the final results, thus each run of the program will produce a slightly different null model, which will affect the final p-value. To achieve accurate estimations for the p-values SKAT-O was implemented 1000 times and the average p-value for each gene was calculated.

5.3.5.5 Correcting for Population Stratification

Two distinct ethnic groups are considered here: all affected samples have Colombian ancestry whilst the majority of control samples are of European descent. Population stratification is likely to affect the results of the SKAT-O test. The effect of population stratification can be accounted for in the SKAT-O model through the use of principal components.

All 115 samples included in the SKAT-O test were examined for population stratification using EIGENSTRAT (Price et al., 2006). EIGENSTRAT is based on principal components analysis and can model ethnic differences between case and control populations (Price et al., 2006). Ten principal components were calculated for the data and were included as covariates in SKAT-O. EIGENSTRAT was implemented in Unix using EIGENSOFT v4.0.2 (http://www.hsph.harvard.edu/alkes-price/software/).

5.3.5.6 Single Variant Analysis

Genes identified as significantly associated with the phenotype by the SKAT-O test contain multiple variants, some of which contribute to the significant result and some which do not. For all variants in significant genes genotype counts were made for all reference homozygous, heterozygous, and alternate

homozygous genotypes in the case and control populations. Genotype counts were made using custom scripts.

To assess the association of each variant with the phenotype, the single variant Logistic Score Test was implemented in EPACTS v3.2.5. Application of this test to all variants located in genes identified by SKAT-O as significant, determined whether any particular variants were driving the association. One or two variants with a strong association might produce the significant SKAT-O result, rather than all variants within the gene making a contribution.

5.3.6 Functional Analysis of Genes Significantly Associated with NSCLP

SKAT-O association testing identified 60 genes that were significantly associated with the NSCLP phenotype at a Bonferroni-corrected significance threshold of 1.02×10^{-5}. 'Functional Annotation Clustering' of these significantly associated genes was implemented in DAVID (http://david.abcc.ncifcrf.gov) (Huang et al., 2008, Huang et al., 2009) to identify any potential gene ontology terms that were associated with more than one of the significant genes.

5.4 Results

5.4.1 Sequencing Coverage

All 18 exome sequenced samples had a minimum of 83% of mappable bases covered by at least 20 reads (Appendix XVIII). At least 71% of reads were mapped to target reads and each sample had a mean read depth of at least 50 for each sample (Appendix XVIII).

5.4.2 Evaluating the Spectrum of Rare Variation in Candidate Genes in Syndromic CLP Cases

Syndromic CLP cases are often Mendelian (Leslie and Marazita, 2013), therefore, all syndromic cases of CLP sequenced in this study were considered likely to be Mendelian disorders. Each of the four families presented with a slightly different syndrome, each with a distinct causative mutation.

Application of the filtering pipeline (see Materials and Methods) identified a total of 110 variants in 99 genes. Each syndromic family presented with a known syndrome for which candidate genes existed, analysis focused on identifying rare variants within these candidate genes. All rare variants in each sample (passing all filtering criteria) are reported although in all cases a single variant in a highly penetrant gene is likely to be sufficient to explain the phenotype.

5.4.2.1 Family CL005

Family CL005 has 5 known CLP affected samples, with the mode of inheritance likely to be dominant. The candidate gene for this family was predicted to be *IRF6* due to the phenotype in this family being Van de Woude (VWS)/popliteal pterygium syndrome (PPS). One affected sample was exome sequenced.

Analysis of the exome data identified a single rare variant in *IRF6* that was unique to this individual (Table 5.2). This same variant was previously reported and is recorded in dbSNP135 as a clinical variant. The R84C variant has been previously identified in multiple families with PPS (Kondo et al., 2002). The mutation is located within the Smad-interferon regulatory factor-binding domain (SMIR), which is proposed to be critical for IRF6 protein function (Kondo et al., 2002).

The proband's unaffected grandmother was also exome sequenced in this study. Analysis of the exome data confirmed that the *IRF6* variant was not present in this individual. Therefore, based on evidence from previous PPS studies and confirmation that the variant was not carried by an unaffected family member, it is likely that the *IRF6* variant is the causative mutation in this individual.

5.4.2.2 Family CL006

The syndrome observed in the affected proband of this family was believed to be the rare Nager syndrome. Nager syndrome is one of a group of acrofacial dysostoses (AFDs) (OMIM 154400). The main phenotypic features include craniofacial, limb and muscloskeletal malformations (McDonald and Gorski, 1993). The affected male had symptoms of micrognathia, auditory canal defects and digit malformations, all of which are recognised phenotypic

features of Nager syndrome (McDonald and Gorski, 1993). As with the majority of Nager syndrome cases (Bernier et al., 2012), the proband in this family represents an isolated case of the syndrome with no known affected relatives. The gene harbouring mutations that cause Nager syndrome was identified as *SF3B4* (Bernier et al., 2012). Therefore, *SF3B4* was considered to be the candidate gene for this individual.

Analysis of the variants contained within the candidate genes identified 12 rare or novel variants in 11 genes (Table 5.2); however, no variants were identified in the candidate *SF3B4* gene. The majority of identified variants (9 variants) were considered disease-causal by the logit model.

Manual analysis of the raw sequence reads for the *SF3B4* gene was carried out using IGV to ensure that no variants were overlooked due to errors in the variant calling step of the exome pipeline. A previously characterised causal frameshift mutation was identified in *SF3B4* (Bernier et al., 2012) in the affected patient (Table 5.2). Insertion of a C allele was identified in a number of reads at this location suggestive of a heterozygous frameshift mutation. Many of the Nager syndrome-causing mutations are frameshifts, suggesting that Nager syndrome may result from haploinsufficiency of *SF3B4* (Bernier et al., 2012).

5.4.2.3 Family CL012

The phenotype in this family was suspected to be X-linked and lethal in males; all affected males died in infancy. The disorder was originally suspected to be the rare X-linked disorder incontinentia pigmenti (IP), but males in IP families generally die *in utero*, which is not a feature of this particular family. The incidence of IP is between 1 in 10,000 and 1 in 20,000, with skin abnormalities the main characteristic, although other features include abnormalities of the eyes, nails, hair, teeth and central nervous system (Fusco et al., 2012). All cases of IP are the result of mutations within the *IKBKG* gene (Fusco et al., 2012)

An affected female who presented with bilateral cleft lip and palate as well as other syndromic features (abnormal nail pigmentation and cutaneous

Table 5.2. Rare and novel variants identified in CLP candidate genes in syndromic CLP patients

Gene	Chromosome	Exon	Base pair in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	Minor allele frequency in 1000G	Minor allele frequency in NHLBI ESP	Novel or clinical	SIFT score	Polyphen2	LRT	MutationTaster	Grantham Score	PhyloP	GERP++	CADD score	Logit score	Logit classification	CL005_1	CL006_1	CL012_1	CL012_2	CL018_1	CL018_2	CL018_3
MTHFR	1	5	11856370	ns	A673G	I225V	rs200100285	0.001	0.000	.	0.21	0.152	0.000	0.968	29	1.325	1.46	9.651	0.020	No	◇						
SZT2	1	68	43913968	ns	G9611A	R3204Q	rs190363418	0.003	0.000	.	0.57	0.001	0.002	1.000	43	2.405	3.49	7.06	0.014	No		◇						
PGM1	1	5	64100551	ns	C143T	A48V	.	.	.	N	0.13	0.025	0.000	1.000	64	7.651	5.13	19.05	0.072	Y			◇					
NOTCH2	1	34	120458122	ns	T7223A	L2408H	rs35586704	0.001	0.003	.	0.00	0.969	0.006	0.739	99	2.431	5.35	15.05	0.100	Y			◇					
SF3B4	1	5	149895760	fi	1060_1061 insC	R354fs	.	.	.	.					.			.	.	.		◇						
PKLR	1	1	155271095	ns	C92T	A31V	rs150077703	.	0.001	.	0.03	0.935	0.000	0.742	64	1.654	4.74	26.10	0.962	Y			◇	◇				
IRF6	1	5	209964011	ns	G604A	V202I	.	.	.	D	.	0.916	0.000	1.000	29	7.311	6.17	27.10	0.119	Y					◇	◇	◇	
IRF6	1	4	209969822	ns	C250T	R84C	rs121434226	.	.	C	.	1.000	0.000	1.000	180	5.046	4.72	20.3	0.214	Y	◇							
COLEC11	2	3	3660961	ns	C191T	T64M	rs142369741	0.001	0.000	.	0	0.555	.	0.962	81	2.882	4.61	19.13	0.063	Y	◇							
IFT172	2	33	27676956	ns	G3604T	V1202L	.	.	.	D	.	.	.	1.000	32	4.362	5.70	28.30	0.148	Y		◇						
ECEL1	2	9	233347880	ns	A1516G	M506V	rs142971707	.	0.000	.	0.01	0.760	0.000	1.000	21	3.251	5.36	14.54	0.048	Y				◇				
XPC	3	12	14190411	ns	G2042A	R681Q	.	.	0.000	.	0.31	0.987	0.000	1.000	43	7.629	4.82	19.81	0.157	Y	◇							
IFT122	3	7	129195170	ns	C496T	R166W	rs61744448	0.001	0.001	.	0.02	0.880	0.001	1.000	101	3.254	4.80	15.93	0.041	Y		◇						
SLCO2A1	3	10	133664019	ns	G1381C	D461H	.	.	.	D	0.06	0.940	0.128	1.000	81	1.765	3.67	12.90	0.045		◇							
IDUA	4	7	995942	ns	T965A	V322E	rs76722191	0.002	0.000	.	0.00	0.999	0.000	1.000	121	5.962	5.15	23.30	0.151	Y			◇					
NKX3-2	4	2	13544126	ns	G493C	D165H	rs61795263	.	.	.	0.17	0.419	0.022	1.000	81	3.045	5.31	21.40	0.212	Y			◇	◇				
CC2D2A	4	22	15559105	ns	G2804A	R935Q	rs187003641	0.002	.	.	0.27	0.983	0.001	1.000	43	6.747	4.80	24.20	0.118	Y	◇							
FRAS1	4	40	79360063	ns	G5374A	A1792T	rs150916370	0.004	0.001	.	1.00	0.014	0.000	1.000	58	3.652	3.96	1.59	0.021	No	◇							
NIPBL	5	15	37002789	ns	A3690C	E1230D	.	.	.	D	0.36	0.239	0.000	0.997	45	0.710	2.22	15.11	0.035	No	◇							
SH3PXD2B	5	13	171765821	ns	C2288T	P763L	.	.	.	D	.	0.997	0.000	1.000	98	7.565	5.29	19.54	0.116	Y				◇				
TUBB2B	6	4	3225580	ns	C743T	A248V	rs2808001	.	.	.	0.00	0.082	0.000	1.000	64	9.506	4.18	7.09	0.098	Y		◇						
ABCB1	7	29	87133651	ns	G3751A	V1251I	rs28364274	0.004	0.000	.	0.60	0.004	0.002	0.990	29	0.428	-3.21	8.63	0.014	No	◇				◇			
COL1A2	7	48	94056566	ns	C3226T	P1076S	.	.	.	N	0.02	0.063	0.000	1.000	74	3.858	5.32	16.87	0.147	Y			◇					

TFR2	7	15	100224533	sp	1483-7A>C	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.			◇	
MCPH1	8	8	6302018	ns	A775C	K259Q	.	.	0.000	.	0.43	0.506	0.164	1.000	53	-1.592	-1.78	18.37	0.053	Y		◇		
MCPH1	8	8	6302516	ns	T1273A	Y425N	rs201261159	.	0.000	.	0.75	0.001	0.096	1.000	143	0.073	-1.86	5.58	0.006	No		◇		
ASPH	8	22	62438574	ns	A1775G	E592G	rs200474335	.	0.000	.	0.17	0.970	0.000	1.000	98	7.450	4.26	19.71	0.032		◇			
TRPS1	8	4	116616196	ns	C2000T	S667L	rs7002384	.	0.000	.	0.01	0.024	0.418	1.000	145	1.658	4.90	7.06	0.044	Y		◇		
ABO	9		136131722	un			.	.	.	D											◇			
TNNT3	11	6	1947946	sp	82+7C>T		rs188811451	0.001	0.000	.	.	.	.	.	.	.	.	.	.	.			◇	
KMT2A	11	27	118376934	ns	G10327A	A3443T	rs201345346	.	0.000	.	0.23	0.001	0.016	1.000	58	3.393	1.73	8.28	0.040	Y		◇		
COL2A1	12	24	48379557	ns	A1427G	N476S	rs145042175	.	0.001	.	0.73	0.463	0.000	0.936	46	-0.172	5.58	4.63	0.019	No	◇			
UBE3B	12	3	109921492	ns	C136T	R46W	rs149953168	.	0.000	.	0.00	1.000	0.000	1.000	101	1.832	4.43	16.02	0.050	Y		◇		
GJB6	13	3	20797144	ns	A476G	N159S	rs35277762	0.002	0.000	.	0.76	0.038	0.007	0.913	46	1.71	3.95	6.08	0.015	No			◇	
SEC23A	14	18	39509962	ns	A2116G	I706V	.	.	.	N	0.55	0.042	0.000	1.000	29	5.022	5.75	12.73	0.043	Y			◇	
IFT140	16	20	1576628	ns	G2569A	G857S	rs200876696	.	.	.	0.07	0.995	0.000	1.000	56	7.661	5.22	34.00	0.275	Y			◇	
TSC2	16	33	2133688	sp	3884-8C>G		.	.	.	D	.	.	.	.	.	.	.	.	.	.	◇			
ABCA3	16	22	2336921	ns	G3052A	G1018S	rs201700257	0.001	0.000	.	0.45	0.064	0.000	1.000	56	2.165	4.65	12.09	0.062	Y			◇	
GRIN2A	16	3	10032161	ns	A662G	K221R	rs61731464	0.001	0.000	.	.	0.027	0.000	1.000	26	6.107	5.09	12.59	0.057	Y			◇	
ABCC6	16	16	16276768	ns	C1963A	Q655K	.	.	0.000	.	1.00	0.004	0.061	1.000	53	3.499	4.97	4.84	0.017	No			◇	
PALB2	16	4	23647602	ns	G265C	D89H	.	.	.	D	0.04	0.063	0.084	1.000	81	0.699	0.03	7.76	0.066	Y				◇
SRCAP	16	23	30734036	ns	A3859G	T1287A	.	.	.	D	0.47	0.091	.	1.000	58	2.494	5.17	8.04	0.062	Y				◇
RPGRIP1L	16	6	53720397	sg	G724T	E242X	.	.	.	N	0.14	.	0.000	1.000	.	4.463	5.87	36.00	0.164	Y			◇	◇
COG1	17	4	71193355	sp	743-10C>G	.	rs185592690	0.001	.	.	.	.	.	.	.	.	.	.	.	.			◇	
ERCC2	19	20	45856006	ns	A1900G	K634E	.	.	.	D	0.00	0.925	0.000	1.000	56	5.182	5.13	28.30	0.123	Y		◇		
PROKR2	20	2	5283122	ns	C719T	T240I	.	.	.	D	0.23	0.841	0.000	1.000	89	5.246	5.16	23.10	0.096	Y		◇		
COL6A2	21	3	47532093	ns	G316A	E106K	rs141703710	0.002	0.011	.	0.30	0.437	0.088	0.991	56	1.335	4.34	12.22	0.040	No		◇		
IKBKG	X	2	153780386	ns	G169A	E57K	rs148695964	.	0.002	C	0.16	0.997	0.000	0.998	56	5.105	5.60	21.90	0.094	Y			◇	

Where a specific variant is present in a sample this is indicated by • if heterozygous or + if homozygous.
Functional scores/predictions (from SIFT, PolyPhen2, LRT, MutationTaster, or Grantham) considered damaging are underlined
C, clinical; D, unique to individuals with this disease; fi, frameshift insertion; N, novel; No,variant not classified as disease-causal; ns, nonsynonymous; sg, stop-gain; sp; splicing; sy, synonymous; Y, variant classified as disease-causal

syndactyly) was exome sequenced along with a half-uncle who presented with features of facial clefting, syndactyly and brachydactyly. Since all other IP-affected male members of this family died in early infancy, it was suspected that the phenotype in the half-uncle was not IP but a second disorder including clefting as a feature. Based on the apparent X-linked inheritance pattern the causal mutation was expected to be in a gene on the X chromosome. *IKBKG* was the candidate gene due to the suspected IP phenotype in CL012_1.

A missense variant was identified in the *IKBKG* gene in CL012_1 (Table 5.2). This was the only rare variant in a gene on the X chromosome in this individual. The mutation was not shared with the male relative so may be the variant responsible for the phenotype in CL012_1. The variant was recorded as clinical in dbSNP135 and had been previously identified in a mild form of IP but whether it was compatible with male survival into infancy was undetermined (Aradhya et al., 2001). Evidence from this particular family suggest that this variant is compatible with male survival, although sequencing of affected males within this family would be required to confirm that this variant was present in these individuals. This *IKBKG* mutation results in the substitution of a single amino acid that appears to disrupt TRAF6 binding and IL-1β signalling (Gautheron et al., 2010). The substitution has no identified impact on the NEMO protein itself (encoded by *IKBKG)* but disrupts the NEMO/TRAF6 interaction, affecting TRAF6-dependent signalling (Gautheron et al., 2010).

To determine the variant responsible for the clefting observed in sample CL012_2, all rare variants within this individual were analysed. A novel heterozygous stopgain mutation was identified in *RPGRIP1L* on chromosome 16 (Table 5.2). This variant was identified as damaging by most *in silico* prediction metrics. Conservation metrics identified the variant as located in a strongly conserved genomic region: GERP++ score of 5.87 and PhyloP score of 4.46. Moreover, the variant was assigned a CADD scaled score of 36, meaning it is ranked within the top 0.1% of over 8 million variants identified across the human genome. Homozygous mutations in *RPGRIP1L* are associated with the developmental disorders Joubert syndrome and Meckel syndrome (Delous et al., 2007) both of which present with cleft lip and palate and digit malformations as clinical features (OMIM 611561, (Gopalakrishna et al., 2014)). It is therefore possible that this individual has a mild form of either Meckel syndrome or Joubert syndrome. CL012_1 also harboured this *RPGRIP1L*

variant, which may contribute to the observed phenotype since cleft lip and palate in patients with IP has only been described a couple of times previously (Stewart et al., 1979, Yell et al., 1991). Analysis of the phenotypic features of over 1000 reported IP cases suggested that cleft palate was a feature of ~1.6% of all IP cases (Minić et al., 2013).

5.4.2.4 Family CL018

Family CL018 has a complex pattern of affected samples with varying penetrance. Multiple CLP phenotypes are present within this family, including unilateral cleft lip and palate, bilateral cleft lip and palate, and cleft palate. Three samples were selected from this family to undergo exome sequencing in order to identify the causative mutation: the proband who presented with Pierre Robin sequence at birth, one of his uncles who was affected with cleft palate, and a cousin who presented with bilateral cleft lip and palate.

Pierre Robin sequence is a craniofacial anomaly comprising of several events that occur in succession (Jakobsen et al., 2007). Micrognathia is present at birth and leads to glossoptosis, preventing fusion of the palatal shelves (Jakobsen et al., 2007, Tan et al., 2013). The affected proband in this family presented with micrognathia at birth leading to respiratory failure, requiring a ventilator and oral surgery at a later stage.

Candidate gene analysis, which included genes implicated in both Pierre Robin syndrome and CLP, identified one variant in *IRF6* that was common to the three affected individuals and unique to this family (Table 5.2). The variant caused the replacement of a valine residue with an isoleucine residue and was classified as damaging by all prediction scores except Grantham. Grantham is a score based on the properties of amino acids, such as polarity and charge, so amino acids with similar properties are given low Grantham scores. Both valine and isoleucine have very similar biochemical properties; both are non-polar with a neutral charge. Hence they are given a low Grantham score. The two measures of conservation, PhyloP and GERP++, identify the location of the mutation as a highly conserved region of the genome. The GERP++ score for this variant is 6.17, which is the highest possible GERP++ score that can be assigned. This implies that any variant at this location is likely to be damaging.

IRF6 mutations are most commonly associated with VWS or PPS (Kondo et al., 2002). Associations between this gene and non-syndromic forms of CLP have been identified through GWAS (Zucchero et al., 2004, Blanton et al., 2005, Scapoli et al., 2005, Nikopensius et al., 2010). Therefore, mutations within this gene are not limited to the VWS/PPS phenotypes and may give rise to various forms of CLP.

5.4.3 Evaluating the Spectrum of Rare Variation in Candidate Genes in Non-Syndromic CLP Cases

NSCLP is considered to be a complex trait, rather than a Mendelian trait, with multiple genetic and environmental factors contributing to the disorder (Mossey et al., 2009, Mangold et al., 2011). NSCLP cases from 6 families were exome sequenced to identify all rare and novel variation in each individual. In all NSCLP cases it was suspected that multiple rare variants, likely shared by all affected family members, would contribute to the phenotype in each family. These variants could be recessive or heterozygous in nature. Therefore, the spectrum of rare variation in each patient was of interest.

Application of the filtering pipeline (described in Materials and Methods) to the 12 exome sequenced samples identified 141 variants in 119 of the 865 candidate genes. Forty-five variants were classified as novel – present in one patient only (in heterozygous or homozygous form) and not previously reported in the public databases (dbSNP135, 1000G or NHLBI ESP) or the in-house exome database. A further 37 variants were designated 'disease-only'; these variants were identified in more than one of the 18 patients with CLP phenotypes analysed in this study, but were not reported in the public databases or the in-house database.

5.4.3.1 Family CL001

Cleft lip affects 6 individuals in family CL001, of which 5 are male. A pedigree such as this would normally be suggestive of an X-linked inheritance pattern but NSCLP is a complex trait likely to arise from multiple variants. Nevertheless, genes on the X chromosome were examined thoroughly to ensure that the phenotype in this family was not the result of X-linked

inheritance. Two affected brothers (CL001_1 and CL001_2) were selected for exome sequencing.

Consideration of heterozygous variants shared by both sequenced samples identified 6 variants in CLP candidate genes, 5 of which were unique to these samples (Table 5.3). Two variants were located in the *IGF1R* gene, which encodes a growth factor receptor. CLP is not a reported clinical feature of individuals with *IGF1R* mutations but *IGF1R* variants are associated with growth abnormalities *in utero* (Abuzzahab et al., 2003). One of the *IGF1R* variants is a non-frameshift insertion so *in silico* predictions were unavailable. The other variant was nonsynonymous variant predicted to be a well-conserved region of the genome (GERP++ = 5.67) and predicted to be causal by the logit model.

Nonsynonymous variants were identified in genes located on the X chromosome in both sequenced samples, however none of these variants was common to both brothers. Confirmation that each variant was present in only one of the affected samples was achieved through inspection of the raw sequence reads from both samples, using IGV. The identified X chromosome variants are unlikely to be relevant to the phenotype because they are found in only one of the two samples.

5.4.3.2 Family CL002

Three members of family CL002 presented with left-sided unilateral CLP: a pair of siblings, one male and one female, and their mother. The affected siblings were selected for exome sequencing.

Rare homozygous variation, shared by both samples, was not detected in any of the CLP candidate genes. Considering homozygous variation in any gene identified no variants shared by both affected members of family CL002.

The siblings shared rare heterozygous variation in 13 CLP candidate genes (Table 5.3). One potentially interesting result is the two separate heterozygous variants in the *SNAP29* gene in both samples. Homozygous variation in *SNAP29* causes CEDNIK syndrome (OMIM 609528), of which facial dysmorphism is a feature. If the two heterozygous variants are compound heterozygous then they will potentially affect both copies of the *SNAP29* gene, particularly if they are both deleterious. Some of the *in silico* metrics predict that these variants are deleterious but there is no consensus across all scores.

5.4.3.3 Family CL003

A mother presented with submucous palate while her daughter was affected with left-sided CLP. The female had 3 miscarriages prior to the birth of the CLP affected daughter. Only the mother (CL003_1) was selected for exome sequencing.

Sample CL003_1 did not harbour a single homozygous variant in a candidate gene. She did, however, harbour heterozygous variation in 10 candidate genes (Table 5.3). The logit model classified 7 of these variants as disease causal. Heterozygous variants unique to this individual were identified in 7 genes.

5.4.3.4 Family CL004

Two members of family CL004 are affected with CLP phenotypes: a mother (CL004_1) presenting with left-sided unilateral CLP and her daughter (CL004_2) presenting with bilateral CLP.

Analysis of heterozygous variation in CLP candidate genes identified 10 variants shared by both individuals (Table 5.3). One of these variants, in *SETBP1*, was shared with sample CL007_1 who also presented with left-sided unilateral CLP (Table 5.1). A variant unique to these affected individuals was identified in the *MSX1* gene. Point mutations within *MSX1* have been detected in several cases of isolated CLP or CPO (Jezewski et al., 2003).

One homozygous variant, in *BMPR1A*, was identified in the mother (CL004_1) and was found to be heterozygous in her daughter. Loss-of-function of *Bmpr1a* in a mouse model resulted in a completely penetrant bilateral CLP phenotype (Liu et al., 2005). Therefore, there is potential for a damaging variant in *BMPR1A* to be involved in the NSCLP phenotype, however, the *in silico* prediction metrics do not suggest that this particular variant is damaging.

5.4.3.5 Family CL007

A single sample was selected from family CL007 for sequencing (CL007_1). A large pedigree was obtained for this family: 5 individuals affected with CLP, bilateral CLP, or unilateral CLP. Sample CL007_1 presented with left-sided unilateral CLP (Table 5.1).

Variant filtering identified 14 rare variants, all of which were heterozygous (Table 5.3). Four variants were unique to this single patient while 2 were shared with unrelated individuals who also presented with left-sided unilateral CLP or bilateral CLP. Implication of the same variants across families with NSCLP increases the evidence in favour of those variants being disease-related. In the case of the *COLQ* variant shared with CL002_2, the same variant is not detected in her similarly affected brother. Whether this particular variant is contributing to the phenotype in CL002_2 is unclear since one might expect the same variants to be influencing the phenotype in first degree relatives.

One variant was identified in a gene previously implicated in NSCLP: *COL11A2* (Nikopensius et al., 2010). This same variant was observed at low frequency in the 1000 Genomes Project database (in < 0.09% of genomes), although it is plausible that individuals sequenced as part of the 1000 Genomes Project have NSCLP or carry NSCLP risk variants.

5.4.3.6 Family CL010

An affected child and father were selected for exome sequencing from family CL010 (CL010_1 and CL010_2 respectively). Both individuals were affected with bilateral CLP (Table 5.1) and share 5 heterozygous variants in candidate CLP genes (Table 5.3). One of these variants is also present in the child of family CL004 (CL004_2). All were predicted to be disease causal by the logit model.

One of the variants was a stop-gain variant in *FRAS1*. Stop-gain variants tend to be disruptive to gene function, particularly if they occur early in the gene sequence. This particular variant does not seem to be close to the start codon of the gene however. Three of the *in silico* metrics predict this variant to be damaging while both GERP++ and PhyloP indicate that the affected residue is highly conserved. Furthermore, mutations in *FRAS1* are linked to Fraser syndrome (OMIM 219000) and cleft lip with or without cleft palate has been described as a feature of this syndrome.

Table 5.3. Rare and novel variants identified in candidate genes and shared by relatives, in non-syndromic cleft lip and palate patients

Gene	Chromosome	Exon	Base pair in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	Minor allele frequency in 1000 Genomes Project	Minor allele frequency in Exome Variant Server	Novel or clinical	SIFT score	Polyphen2	LRT	MutationTaster	Grantham Score	PhyloP	GERP++	CADD score	Logit score	Logit classification	CL001_1	CL001_2	CL002_1	CL002_2	CL003_1	CL004_1	CL004_2	CL007_1	CL010_1	CL010_2	CL014_1	CL014_2
AHDC1	1	6	27876631	ns	C1996G	R666G	.	.	.	D	0.03	1.000	0.000	1.000	125	8.73	5.51	22.80	0.089	Y			◇	◇								
ARHGAP29	1	15	94654402	ns	A1672G	I558V	.	.	.	D	0.59	0.001	0.013	1.000	29	1.89	0.69	12.33	0.025	No			◇	◇								
NOTCH2	1	9	120509101	ns	G1465T	V489L	.	.	.	D	0.33	0.704	0.004	0.838	32	0.87	5.73	12.51	0.149	Y						◇	◇				
LYST	1	3	235993575	ns	A143G	H48R	rs200132460	.	0.000	.	0.50	0.002	0.000	0.931	29	4.65	4.58	8.65	0.031	Y								◇				
GLI2	2	13	121747113	ns	G3623A	R1208H	rs200537256	0.001	0.000	.	.	0.002	0.777	1.000	29	0.09	-0.30	0.01	0.015	No								◇				
SCN2A	2	13	166187894	ns	T2204C	M735T	.	.	.	N	1.00	.	0.000	0.968	81	0.47	2.85	2.95	0.007	No					◇						
ABCA12	2	3	215928852	ns	C254T	T85I	.	.	.	N	0.39	0.983	0.018	1.000	89	4.18	5.65	15.26	0.197	Y					◇						
SETD5	3	23	9517336	ns	C3890T	T1297M	.	.	.	D	0.18	0.883	0.212	1.000	81	2.34	4.62	10.19	0.038	Y			◇	◇								
COLQ	3	11	15507874	ns	C686T	P229L	rs146619514	.	0.000	.	0.20	0.999	0.000	1.000	98	2.62	3.66	12.53	0.040	Y				◇				◇				
LAMB2	3	32	49158863	sp	5260+3 A>G	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.						◇	◇					
IL17RD	3	4	57144258	ns	A392C	K131T	rs184758350	0.001	0.001	.	0.19	0.993	0.000	1.000	78	5.75	5.66	22.20	0.044	Y			◇	◇								
FLNB	3	2	58062842	ns	A362T	Y121F	rs147846832	0.002	0.000	.	0.03	0.999	0.000	1.000	22	9.29	5.04	28.50	0.111	Y					◇						
PIK3CA	3	21	178951879	sp	2937-3 T>C	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.			◇	◇								
MASP1	3	16	186937872	ns	G2087A	G696E	.	.	.	D	0.05	0.261	.	1.000	98	1.65	4.36	14.53	0.235	Y									◇	◇		
DLG1	3	8	196846393	nd	923_ 925del	308_ 309	.	.	.	D	.	.	.	.	.	.	.	.	.	.						◇	◇					
MSX1	4	2	4864736	ns	C778A	P260T	.	.	.	D	0.01	0.958	0.000	1.000	38	5.96	4.81	27.60	0.422	Y						◇	◇				
EVC2	4	16	5617202	ns	G2776A	E926K	.	.	.	D	0.08	0.998	0.000	1.000	56	1.14	3.47	16.13	0.033	Y			◇	◇								
FRAS1	4	51	79391228	sg	G7354T	E2452X	.	.	.	D	0.05	.	0.000	1.000	.	9.73	4.67	37.00	0.523	Y									◇	◇		
SMARCAD1	4	20	95201837	ns	G2519A	R840Q	rs200626666	.	0.000	.	0.66	0.001	0.717	0.976	43	1.97	2.76	13.52	0.030	No								◇				
FGF10	5	1	44388655	ns	A130G	M44V	rs201168313	0.000	.	.	0.50	0.002	0.000	0.915	21	1.14	4.72	9.50	0.027	Y								◇				
SPRY4	5	3	141693887	ns	C856T	R286C	.	.	.	N	0.07	0.999	0.000	1.000	180	2.45	5.05	13.49	0.059	Y					◇						
COL11A2	6	55	33135081	ns	G3878A	R1293Q	rs181895110	0.001	.	.	0.29	0.864	0.008	0.667	43	0.45	3.26	19.17	0.051	Y								◇				

COL11A2	6	8	33148880	sp	921+8 A>T	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.				◇							
ZBTB24	6	2	109802863	ns	A367G	K123E	.	.	.	D	0.07	0.801	0.000	0.730	56	1.52	4.66	14.67	0.038	Y					◇	◇					
SEMA3A	7	17	83591080	ns	G1923C	Q641H	rs115335242	0.001	0.000	.	0.08	0.153	0.046	0.543	24	1.69	5.28	12.78	0.054	Y			◇	◇							
TFR2	7	4	100238501	sp	287-6 C>T	.	.	.	0.000	.	.	.	.	.	.	.	.	.	.	.										◇	◇
WDR60	7	15	158711426	sp	1792-5 G>C	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.							◇				
WDR60	7	23	158727115	ns	A2653G	I885V	.	.	.	D	1.00	0.059	0.026	1.000	29	0.41	-5.44	0.01	0.056	Y										◇	◇
ZFHX4	8	10	77764484	ns	A5327G	E1776G	rs139920573	0.003	0.003	.	0.04	0.999	0.000	1.000	98	7.33	4.71	14.65	0.030	Y						◇		◇	◇		
PTCH1	9	8	98241314	ns	A730G	I244V	rs139498131	.	0.000	.	1.00	0.161	0.000	1.000	29	4.87	5.05	11.44	0.038	Y							◇				
BMPR1A	10	3	88635834	ns	G59A	R20H	.	.	.	D	0.68	0.000	0.146	1.000	29	0.08	0.16	0.02	0.029	Y					+	◇					
ALDH18A1	10	14	97376380	sp	1468-9 G>C	.	.	.	.	D	.	.	.	.	.	.	.	.	.	.										◇	◇
COL17A1	10	49	105795306	ns	T3434C	I1145T	.	.	.	D	0.58	0.682	0.081	1.000	89	5.46	4.83	12.18	0.045	Y							◇				
DEAF1	11	11	654023	ns	C1532G	A511G	.	.	.	N	0.10	0.978	0.000	1.000	60	9.01	3.49	17.71	0.047	Y				◇							
WT1	11	1	32456755	ns	C137T	A46V	.	.	.	D	0.22	.	.	1.000	64	0.33	0.91	12.21	0.015	No					◇	◇					
PHF21A	11	18	45955628	ns	A1796G	N599S	rs191903911	0.001	.	.	0.32	0.739	0.000	0.984	46	3.47	5.78	14.73	0.041	Y					◇	◇					
INPPL1	11	27	71949096	ns	T3563G	L1188R	.	.	.	D	.	.	.	0.991	102	0.45	2.95	10.20	0.009	No	◇	◇									
DLG2	11	2	84245750	ns	G67C	A23P	.	.	.	N	.	0.221	0.137	1.000	27	1.63	4.02	19.59	0.041	Y				◇							
CDON	11	6	125887051	ns	A860G	N287S	.	.	.	D	0.72	0.752	0.000	0.907	46	3.10	5.06	15.32	0.043	Y										◇	◇
PTHLH	12	2	28122357	ns	G71A	G24E	.	.	.	D	0.01	1.000	0.000	1.000	98	5.75	5.48	32.00	0.055	Y			◇	◇							
GRIP1	12	10	66849248	ns	C1139T	T380I	rs201326216	0.000	.	.	.	.	.	1.000	89	4.84	5.51	14.82	0.043	Y										◇	◇
GNPTAB	12	1	102224384	ns	T70G	F24V	rs141329633	.	0.000	.	0.03	0.780	0.000	1.000	50	5.14	3.76	11.54	0.047	Y								◇	◇		
FREM2	13	1	39263993	ns	A2512G	T838A	.	.	.	D	0.50	0.000	0.000	1.000	58	2.49	4.64	7.38	0.053	Y						◇	◇				
SUCLA2	13	7	48528684	ns	A811G	M271V	rs142020748	0.002	0.000	.	0.02	0.382	0.000	1.000	21	5.86	5.67	12.13	0.045	Y	◇	◇									
ERCC5	13	23	103528048	ns	C4718T	A1573V	.	0.002	0.001		0.25	0.004	0.979	1.000	64	-0.46	-3.01	10.80	0.011	No				◇							
FMN1	15	4	33446492	ns	G624C	R208S	.	.	.	D	0.56	0.225	0.629	0.988	.	0.12	4.22	5.80	0.023	No			◇	◇							
FBN1	15	39	48758053	ns	G4750A	E1584K	rs148888513	0.001	0.000	.	0.13	0.741	0.000	1.000	56	4.91	5.74	36.00	0.607	Y										◇	◇
HCN4	15	8	73615214	ns	C3220T	L1074F	rs200399862	0.001	.	.	0.06	0.697	.	1.000	22	-0.02	2.84	12.20	0.313	Y							◇				
IGF1R	15	21	99500507	ni	3940_3941ins CGTCCTCCC	L1314delins PSSL	.	.	.	D	.	.	.	.	.	.	.	.	.	.	◇	◇									
IGF1R	15	21	99500597	ns	C4030G	Q1344E	.	.	.	D	0.25	0.012	0.000	1.000	29	4.78	5.67	13.05	0.038	Y	◇	◇									
ADAMTS17	15	13	100657099	ns	G1841C	R614P	rs144928920	.	0.000	.	0.21	0.929	0.000	1.000	103	4.73	5.18	18.47	0.027	Y				◇							
ABCC6	16	10	16292021	ns	G1195A	G399S	.	.	0.000	.	1.00	0.097	0.051	0.848	56	-0.25	2.39	0.02	0.007	No			◇	◇							

CDH3	16	11	68719119	ns	T1436C	L479P	rs138190335	.	0.003	.	0.23	0.916	0.863	0.974	98	2.61	4.45	15.48	0.029	Y								◇				
CFDP1	16	5	75429103	ns	A535T	T179S	.	.	.	D	0.42	0.058	0.021	1.000	58	2.66	5.89	15.68	0.045	Y	◇	◇										
MYH3	17	29	10539158	ns	G3869A	R1290H	.	.	.	D	0.05	0.355	.	1.000	29	4.95	5.42	21.30	0.108	Y							◇	◇				
KCNJ2	17	2	68171569	ns	C389T	T130M	rs151168735	.	0.000	.	.	0.386	0.000	1.000	81	4.97	5.92	10.12	0.043	Y										◇	◇	
ASXL3	18	12	31323195	ns	G3383A	R1128Q	.	.	.	D	0.11	0.913	0.477	0.992	43	2.76	5.04	15.59	0.026	Y			◇	◇							
SETBP1	18	4	42530509	ns	A1204G	I402V	.	.	.	D	0.31	0.770	0.000	0.742	29	4.47	5.78	18.64	0.044	Y							◇	◇	◇		
CEP89	19	4	33444604	ns	G409A	G137S	.	.	.	D	0.89	0.181	0.000	1.000	56	-1.46	-10.2	3.52	0.016	No			◇	◇							
RYR1	19	34	38976754	ns	G5459T	R1820L	.	.	.	D	0.16	0.091	0.344	0.992	102	0.93	2.65	8.87	0.023	No								◇			
CLPTM1	19	9	45491357	ns	A1058G	N353S	.	.	.	D	0.13	0.867	0.000	1.000	46	6.60	3.37	17.19	0.072	Y	◇	◇									
FERMT1	20	13	6064778	ns	G1627A	V543M	rs76933768	0.001	.	.	0.00	0.937	0.000	1.000	21	3.16	4.80	20.60	0.083	Y					◇						
COL6A2	21	28	47551876	ns	G2470A	V824M	.	.	.	D	0.22	1.000	0.001	1.000	21	9.43	4.18	13.53	0.084	Y								◇			
SNAP29	22	1	21213511	ns	C113T	P38L	rs200628919	.	0.000	.	0.30	0.030	0.000	0.738	98	1.50	5.49	14.93	0.061	Y			◇	◇							
SNAP29	22	2	21224814	ns	A427G	N143D	.	.	.	D	0.77	0.772	0.010	0.754	23	8.77	4.01	11.41	0.053	Y			◇	◇							

Where a specific variant is present in a sample this is indicated by • if heterozygous or + if homozygous
D, unique to individuals with this disease; N, novel; No,variant not classified as disease-causal; ns, nonsynonymous; sg, stop-gain; sp; splicing; Y, variant classified as disease-causal
Functional scores/predictions (from SIFT, PolyPhen2, LRT, MutationTaster, or Grantham) considered damaging are underlined

5.4.3.7 Family CL014

Two second-degree relatives in family CL014 were affected with CLP: the proband (CL014_1) and his affected aunt (CL014_2). Both presented with left-sided unilateral CLP.

Six heterozygous variants, common to the two samples, were identified in candidate CLP genes (Table 5.3). Three of these variants were unique to these two samples. Two of the variants are splicing variants and the role of such variants in disease manifestation is poorly understood. Of the 4 nonsynonymous variants, 3 are considered to be in highly conserved residues and are considered to be deleterious by at least one of the *in silico* prediction metrics. Therefore, these 3 variants have the potential to be damaging although functional analysis would be necessary to confirm that they disrupt protein function.

5.4.4 Rare Variant Association testing using SKAT-O

The SKAT-O association test was applied to all genes that contained a single nucleotide or indel variant. A total of 4,895 genes were tested using SKAT-O. SKAT-O considers all variants (common and rare) in a gene and tests for an association of that gene with the phenotype. Through use of a beta weighting function, variants with a low allele frequency in the sample population are up-weighted while commonly occurring variants are down-weighted.

Applying a Bonferroni-corrected significance threshold of 1.02×10^{-5} to all results identified 60 genes as significant (Table 5.4). The most significant result was *THBS4* with $p = 4.05 \times 10^{-13}$ in which 6 variants were genotyped in all 115 samples, 4 of these variants had individual p-values < 0.05 (Table 5.5).

Of the 865 candidate NSCLP genes, 2 are included in the 60 significant genes: *WNK1* and *SETBP1*. One rare variant was identified in *SETBP1* in 3 of the NSCLP cases (Table 5.3). This same variant has a p-value of 2.72×10^{-7} from the single variant test (Table 5.5) although it is not the most significant variant in *SETBP1*.

Analysis of the variants underlying the significant results revealed that in the majority of cases, one or two of the variants were significant (p = 0.05) in the single variant test and were therefore likely to be the variants driving the gene

Table 5.4. Genes identified as significantly associated with the NSCLP phenotype

Gene	Chromosome	Genomic location	Fraction of samples carrying an alternate allele (n = 115)	Number of variants included in SKAT-O test	P-value from SKAT-O test
THBS4	5	79351735-79373994	0.69	6	4.05×10^{-13}
CCSER2	10	86130899-86273249	0.55	9	1.85×10^{-10}
KANK1	9	710818-740858	0.95	26	5.13×10^{-10}
NLRC3	16	3594296-3614029	0.30	9	7.76×10^{-10}
UNC5D	8	35583815-35647974	0.09	2	8.05×10^{-10}
ICE1	5	5447607-5489383	0.63	13	1.66×10^{-8}
MLST8	16	2258575-2258767	0.09	2	2.31×10^{-8}
NCLN	19	3198817-3207646	0.16	5	3.41×10^{-8}
GHRL	3	10328453-10328453	0.27	1	3.57×10^{-8}
RAB36	22	23503121-23503155	0.44	2	3.61×10^{-8}
CENPE	4	104044197-104102563	0.62	11	3.83×10^{-8}
BCR	22	23615946-23637312	0.78	7	4.53×10^{-8}
ZNF74	22	20759672-20761193	0.36	5	5.39×10^{-8}
WNK1	12	968489-999638	1.00	19	7.92×10^{-8}
NRIP2	12	2937150-2944024	0.10	3	7.97×10^{-8}
CEP104	1	3745924-3764123	0.81	9	8.15×10^{-8}
NCF4	22	37260123-37273742	0.51	5	8.81×10^{-8}
ZCCHC4	4	25314495-25370666	0.31	6	1.19×10^{-7}
ZWINT	10	58118381-58120988	0.64	3	1.43×10^{-7}
TSC1	9	135771709-135786904	0.43	10	1.97×10^{-7}
SETBP1	18	42281778-42643270	0.76	10	2.44×10^{-7}
BHMT	5	78417132-78422035	0.52	3	2.46×10^{-7}
VGLL4	3	11600135-11684957	0.06	2	3.71×10^{-7}
CDH23	10	73269891-73572643	0.93	35	4.10×10^{-7}
CACNA2D4	12	1910786-2024092	0.50	9	4.23×10^{-7}
ATN1	12	7044838-7047911	0.22	8	4.28×10^{-7}
LAMA3	18	21333720-21523890	0.58	24	4.35×10^{-7}
OR2C1	16	3405986-3406838	0.72	7	5.14×10^{-7}
ZNF300	5	150275312-150276288	0.16	5	5.30×10^{-7}
DDX60	4	169158471-169227787	0.30	11	6.43×10^{-7}
CLN8	8	1719273-1728648	0.09	6	7.29×10^{-7}
ASH2L	8	37985897-37985897	0.09	1	7.36×10^{-7}
FANK1	10	127668854-127697997	0.22	6	9.86×10^{-7}
FLT1	13	28880871-29008086	0.60	11	1.00×10^{-6}
FAM114A1	4	38879720-38945169	0.96	8	1.01×10^{-6}
POLR3A	10	79742011-79782047	0.19	9	1.07×10^{-6}
TLR2	4	154624656-154626402	0.78	7	1.23×10^{-6}
ASPA	17	3397702-3402376	0.57	3	1.43×10^{-6}
TSTD2	9	100364848-100388197	0.58	7	1.49×10^{-6}
GBGT1	9	136029045-136031433	0.17	7	1.50×10^{-6}
TBATA	10	72532314-72537020	0.43	3	1.58×10^{-6}
PAPD4	5	78944960-78977875	0.09	4	1.60×10^{-6}

ZFAND2A	7	1192797-1195215	0.36	2	2.29×10^{-6}
ZNF37A	10	38404188-38407745	0.08	2	2.33×10^{-6}
ZNF407	18	72343008-72347253	0.47	13	2.39×10^{-6}
CATSPERD	19	5733874-5778597	0.70	12	2.91×10^{-6}
MFSD8	4	128842761-128861133	0.16	3	3.75×10^{-6}
ATXN1	6	16306751-16328491	0.68	10	3.81×10^{-6}
LEF1	4	108969864-109010339	0.07	3	3.90×10^{-6}
PTCHD3	10	27687225-27702624	0.60	12	4.54×10^{-6}
FGD5	3	14862085-14939089	0.29	9	5.11×10^{-6}
TTC16	9	130479638-130493257	0.10	6	5.21×10^{-6}
OR4K1	14	20403971-20404614	0.69	6	6.38×10^{-6}
ANKRD1	10	92676019-92678920	0.06	3	6.73×10^{-6}
PPRC1	10	103898685-103907129	0.33	10	6.75×10^{-6}
UTS2	1	7909737-7913445	0.70	3	7.17×10^{-6}
KL	13	33628138-33638026	0.93	12	7.77×10^{-6}
FBP2	9	97325726-97349666	0.31	3	8.53×10^{-6}
SV2C	5	75427935-75594743	0.64	7	8.71×10^{-6}
FOXRED2	22	36886199-36900812	0.59	8	1.01×10^{-5}

level results (Table 5.5 for example of variants underlying the significant results for three genes). Results from the single variant test give an indication of which direction each of the variants underlying the gene level result may be acting. Variants with small p-values will be associated with either the case population or control population, the population it is associated with will indicate whether the variant is potentially protective or deleterious in NSCLP.

The SKAT-O results may reflect population specific variants rather than disease-related variants since the affected samples are of Colombian descent while the majority of the control samples are of European descent. There is the capability within SKAT-O to include covariates in the test to correct for complexities within the data. Covariates are taken into account by producing a null model of no association that regresses the phenotype on only the covariates and then produces a regression model by regressing the phenotype on genetic variants and covariates.

Ten principal components were calculated for all samples using EIGENSTRAT (Price et al., 2006), a program for detecting and correcting for population stratification. SKAT-O was implemented with all ten principal components included as covariates. Averaging the p-values from 1000 runs of the test did not identify any genes that reached significance. The lowest p-value obtained was p = 0.498 for gene *PDLIM5*.

Table 5.5. Breakdown of variants underlying SKAT-O results for 3 genes

Gene	Genomic location	Nucleotide change	Alternate allele count	Alternate allele freq	Alternate allele freq in cases	Genotype distribution in cases (n = 12)			Alternate allele freq in controls	Genotype distribution in controls (n = 103)			P value
						Reference hom	Het	Alternate hom		Reference hom	Het	Alternate hom	
THBS4	Chr5:79351735	C > T	46	0.200	0.375	0.417	0.417	0.167	0.180	0.660	0.320	0.019	0.021
THBS4	Chr5:79355606	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
THBS4	Chr5:79361265	G > C	48	0.209	0.000	1.000	0.000	0.000	0.233	0.583	0.369	0.049	7.84×10^{-3}
THBS4	Chr5:79363860	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
THBS4	Chr5:79372721	G > A	1	0.004	0.042	0.917	0.083	0.000	0.000	1.000	0.000	0.000	3.26×10^{-3}
THBS4	Chr5:79373994	G > A	10	0.043	0.417	0.250	0.667	0.083	0.000	1.000	0.000	0.000	1.61×10^{-18}
WNK1	Chr12: 968489	T > C	34	0.148	0.167	0.750	0.167	0.083	0.146	0.748	0.214	0.039	0.800
WNK1	Chr12: 970489	A > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 971291	C > T	17	0.074	0.083	0.833	0.167	0.000	0.073	0.864	0.126	0.010	0.855
WNK1	Chr12: 977170	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 977283	G > C	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 980497	A > G	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 987482	G > A	173	0.248	0.875	0.000	0.250	0.750	0.738	0.078	0.369	0.553	0.149
WNK1	Chr12: 988894	G > A	38	0.165	0.250	0.500	0.500	0.000	0.155	0.718	0.252	0.029	0.235
WNK1	Chr12: 989049	A > G	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 990067	G > C	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 990912	A > C	194	0.157	0.792	0.083	0.250	0.667	0.850	0.039	0.223	0.738	0.490
WNK1	Chr12: 990934	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 993930	C > T	100	0.435	0.417	0.333	0.500	0.167	0.437	0.330	0.466	0.204	0.853
WNK1	Chr12: 994014	C > T	34	0.148	0.208	0.667	0.250	0.083	0.141	0.757	0.204	0.039	0.415
WNK1	Chr12: 994534	G > A	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 994546	G > A	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
WNK1	Chr12: 998338	A > G	6	0.026	0.250	0.500	0.500	0.000	0.000	1.000	0.000	0.000	1.69×10^{-13}
WNK1	Chr12: 998365	G > T	81	0.352	0.125	0.750	0.250	0.000	0.379	0.408	0.427	0.165	0.019

WNK1	Chr12: 999638	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
SETBP1	Chr18:42281778	G > A	8	0.035	0.000	1.000	0.000	0.000	0.039	0.922	0.078	0.000	0.317
SETBP1	Chr18:42529996	G > C	42	0.183	0.042	0.917	0.083	0.000	0.199	0.641	0.320	0.039	0.060
SETBP1	Chr18:42530509	A > G	3	0.013	0.125	0.750	0.250	0.000	0.000	1.000	0.000	0.000	2.72×10^{-7}
SETBP1	Chr18:42531184	C > T	7	0.030	0.292	0.417	0.583	0.000	0.000	1.000	0.000	0.000	1.26×10^{-15}
SETBP1	Chr18:42531274	G > A	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
SETBP1	Chr18:42532571	C > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732
SETBP1	Chr18:42532606	G > A	29	0.126	0.167	0.667	0.333	0.000	0.121	0.767	0.223	0.010	0.513
SETBP1	Chr18:42532693	C > A	23	0.100	0.083	0.833	0.167	0.000	0.102	0.816	0.165	0.019	0.782
SETBP1	Chr18:42532923	T > C	3	0.013	0.000	1.000	0.000	0.000	0.015	0.971	0.029	0.000	0.549
SETBP1	Chr18:42643270	G > T	1	0.004	0.000	1.000	0.000	0.000	0.005	0.990	0.010	0.000	0.732

freq, frequency; hom, homozygous; Het, heterozygous

5.4.5 Functional Annotation Clustering of Genes

All 60 genes identified through SKAT-O as being significantly associated with the NSCLP phenotype were assessed for shared gene ontology terms using DAVID. Twenty clusters of gene ontology terms were identified in total. Two clusters had enrichment scores above the suggested threshold of 1.2: the cluster with highest enrichment score was related to neurodegeneration and cell death while the second was related to GTPase regulation (Table 5.6).

Table 5.6. Functional annotation clusters obtained from DAVID using 60 genes

GO annotation term	Genes included	P value
Cluster 1. Enrichment score 1.39		
Neurodegeneration	*ATXN1, ATN1, CLN8, MFSD8*	0.0015
Cell death	*ATXN1, ATN1, CLN8, MFSD8, UNC5D*	0.22
Death	*ATXN1, ATN1, CLN8, MFSD8, UNC5D*	0.22
Cluster 2. Enrichment score 1.27		
Regulation of Rho GTPase activity	*FGD5, MLST8, TSC1*	0.0039
Regulation of Rho protein signal transduction	*FGD5, MLST8, TSC1, BCR*	0.0043
Regulation of Ras protein signal transduction	*FGD5, MLST8, TSC1, BCR*	0.032
Regulation of Ras GTPase activity	*FGD5, MLST8, TSC1*	0.047
Regulation of small GTPase mediated signal transduction	*FGD5, MLST8, TSC1, BCR*	0.051
Regulation of GTPase activity	*FGD5, MLST8, TSC1*	0.063
Regulation of hydrolase activity	*FGD5, MLST8, TSC1*	0.31
GTPase regulator activity	*FGD5, TSC1, BCR*	0.38
Nucleoside-triphosphatase regulator activity	*FGD5, TSC1, BCR*	0.39

5.5 Discussion

Facial clefts can occur as a single phenotypic feature or as part of a syndrome that includes many more, often developmental, features. Understanding the aetiology of CLP/CPO phenotypes, particularly those that are not related to a syndrome, is complex with many genes implicated in the pathology.

The CLP phenotypes were syndromic in 4 families (6 cases). Syndromic cases of CLP exhibit Mendelian inheritance patterns and usually result from a mutation in a single highly penetrant gene. All the syndromic cases had strong candidate genes known to harbour causal mutations for the corresponding syndrome. The advantage of using an exome sequencing approach rather than candidate gene screening for these patients was that all coding variation could be

identified and interrogated. Although each syndrome had a candidate gene there are examples of trait heterogeneity in Mendelian diseases and some of the observed phenotypes did not completely fit with the clinical features of the suspected syndrome. By running a candidate gene screen it is possible that causal variation would not be identified because the causal variant is located in another gene. Exome sequencing removed this problem by allowing for all genes to be screened at once. Causal variants were identified in all syndromic cases and in most cases the causal variant had been previously described. Exome sequencing is often successful at resolving Mendelian diseases (Bamshad et al., 2011) due to the strong inheritance patterns.

The genetic aetiology of isolated forms of CLP/CPO tends to involve multiple variants with varying penetrance, as well as environmental factors (Cobourne, 2004), acting additively or multiplicatively to produce to the phenotype. As with many complex traits and diseases, the importance of rare variation in phenotype manifestation is becoming apparent; 70% of rare variants that have been reported in NSCLP cases are not found in control samples (Leslie and Marazita, 2013, Leslie et al., 2013) suggesting they may be aetiological. Rare variant analysis in 12 Colombian NSCLP samples confirmed this result, with 59% of filtered variants not recorded in dbSNP135, the 1000G database, the NHLBI ESP database, or the Soton Exome database. The majority of these were also classified as potentially deleterious by at least one *in silico* prediction metric, supporting the hypothesis that most rare variants are expected to be pathogenic (Kryukov et al., 2007), although such variants will be less penetrant in complex traits than in Mendelian disorders (Leslie et al., 2013). Therefore, the spectrum of rare variation harboured by individuals affected with NSCLP is of great interest but is something that has not been explored.

The mutation profiles of the 12 sequenced NSCLP samples were fairly unique. Several genes that have been implicated in NSCLP risk were identified as harbouring rare variants in samples studied in this analysis: *CDH1* (Vogelaar et al., 2013), *COL11A2* (Nikopensius et al., 2010), *IRF6* (Zucchero et al., 2004), *MSX1* (Jezewski et al., 2003, Jagomägi et al., 2010), *MTHFR* (Jagomägi et al., 2010). These variants are therefore potential aetiological variants, contributing to the NSCLP phenotypes in the corresponding families. Further analysis of rare variation through use of the SKAT-O test identified a number of genes that showed significant association with the NSCLP phenotype (Table 5.4). Often the

significant result was driven by a single variant that had different genotype distributions in the NSCLP cases and non-NSCLP controls.

Some genes identified by the SKAT-O test have potential roles in craniofacial development, which may be relevant for NSCLP. The most significant gene identified by SKAT-O was *THBS4*, which encodes a member of the thrombospondin glycoprotein family. Thrombospondins mediate cell-cell and cell-matrix interactions (Adams and Lawler, 2011), which are crucial during embryogenesis to ensure correct foetal growth. Furthermore, thrombospondin-1 regulates TGF-β (Crawford et al., 1998) which is crucial during early embryogenesis and palate formation (Stanier and Moore, 2004), while thrombospodin-2 is involved in mouse embryonic palate development (Melnick et al., 2000), however, depletion of this protein in mice does not cause craniofacial abnormalities such as cleft palate (Kyriakides et al., 1998). Enrichment analysis of all significant genes identified a cluster associated with GTPase regulation (Table 5.6). Small GTPases may have roles to play in craniofacial development; one particular Rho GTPase activating protein, *ARHGAP29*, has been associated with NSCLP (Leslie et al., 2012, Letra et al., 2014). The *EFTUD2* gene encodes a spliceosomal GTPase and is responsible for mandibulofacial dysostosis (Lines et al., 2012), which often includes cleft palate in the phenotype. Mutations within proteins involved in regulating Rho GTPases have been linked to embryogenesis and resulting developmental disorders (Olson et al., 1996). GTPases are involved in cell junction formation and maintenance (Irie et al., 2004), cell migration through regulation of the actin cytoskeleton, and morphogenesis (Etienne-Manneville and Hall, 2002), disruption of these processes can lead to development of CLP/CPO (Mossey et al., 2009).

Although a number of significant genes were identified by the SKAT-O test it is possible that these represent population differences rather than actual disease-related differences. The families considered in this study were collected from Bogota, Colombia, which is likely to have a number of population specific variants, unrelated to CLP phenotypes. Twelve Colombian CLP patients were compared to 102 non-Colombian samples and one Colombian sample without a CLP phenotype. Therefore, in effect, the test is comparing Colombians to non-Colombians so it is likely that some of these significant results are likely to reflect Colombian-specific variants that are in no way related to disease

phenotype. To be able to identify any genes that have a different variant make-up in CLP samples requires a control population from the same ethnic background to neutralise any effects that are specific to the ethnic group. Furthermore, the case population in this study is small at only 12 samples, so our ability to identify any truly significant results is limited.

5.6 Conclusion

Whole-exome sequencing of individuals presenting with various cleft lip and/or palate phenotypes identified many rare and novel variants. Considering syndromic cases of CLP, which present with Mendelian patterns of inheritance and are often dominant, exome sequencing is successful at identifying the causative variant: in all 4 syndromic families strong candidates for the causal variant were identified. The results for nonsyndromic CLP are less clear-cut due to the likely polygenic nature of the trait, with multiple variants expected to contribute to the phenotype. Exploring rare variant association in NSCLP cases identified 60 genes with a significant association with the phenotype and when considered as a gene group demonstrated enrichment for GTPase regulatory functions, which have potential roles in craniofacial abnormalities.

Chapter 6: Identifying the Genetic Cause of Oculopharyngeal Muscular Dystrophy-like Disease in a Single Affected Family

6.1 Background

Oculopharyngeal muscular dystrophy (OPMD) is a late-onset autosomal disorder that is dominantly inherited in the majority of cases. All reported cases of OPMD, in many ethnic groups, are the result of mutations in the *PABPN1* gene, most commonly an expansion of the polyalanine tract at the N-terminus of the protein (Brais et al., 1998, Blumen et al., 2000, Mirabella et al., 2000, Nagashima et al., 2000, Hill et al., 2001, Müller et al., 2001, Rodríguez et al., 2005, Robinson et al., 2006, Robinson et al., 2011a, Scacheri et al., 1999, van der Sluijs et al., 2003).

There is evidence that the length of the polyalanine tract influences the age of onset of disease symptoms, with longer alanine tracts causing manifestation at an earlier age (Hill et al., 2001). There is little evidence for onset of OPMD before age 40; some recessive cases in a consanguineous family have been reported with onset in their mid-30s (Fried et al., 1975) and one case of a very early-onset recessive syndrome resembling OPMD with adolescent onset (Rose et al., 1997) although this syndrome was distinct from adult-onset OPMD.

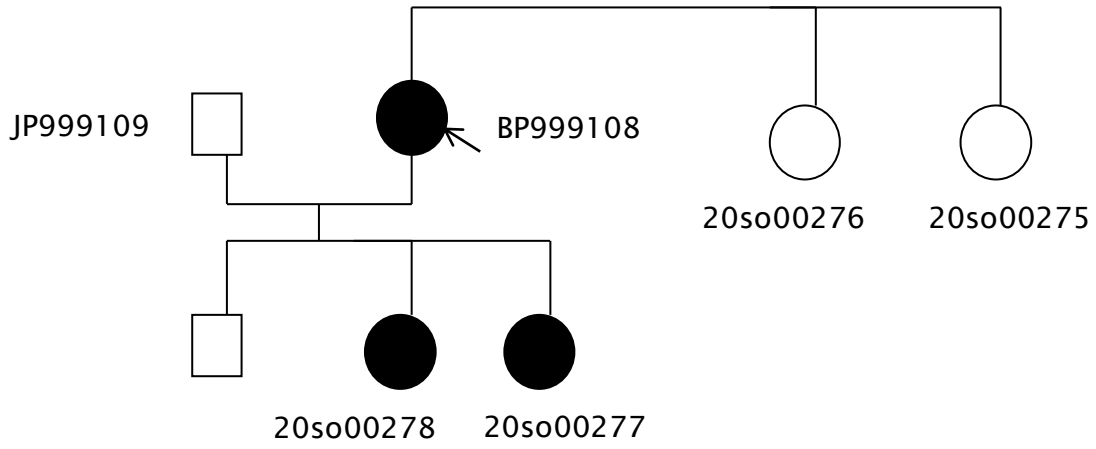


Figure 6.1. Pedigree of a family presenting with an OPMD-like phenotype

Genetic dissection of early-onset breast cancer and other diseases

A single family presenting with symptoms associated with autosomal dominant OPMD but no mutation in *PABPN1* were selected for exome sequencing due to a particularly unique clinical presentation of early-onset ophthalmoplegia and ptosis and a later-onset myopathy affecting the proximal and bulbar musculature. Three members of the family were affected (Figure 6.1) with symptom onset unusually early, in the third decade of life rather than the fifth or sixth. It is likely that the phenotype in this family is a clinically distinct entity and not OPMD although it shares many features with OPMD, suggestive of a similar disease mechanism.

6.2 Aim

To identify the mutation that is the cause of the OPMD-like phenotype observed in affected individuals of a single family. The causal mutation is suspected to be dominantly inherited and therefore present in only the affected family members.

6.3 Materials and Methods

6.3.1 Patients

A single family presenting with OPMD-like symptoms but no evidence of a *PABPN1* GCN expansion were selected for exome sequencing. Three members of the family were affected with suspected autosomal dominant disease (Figure 6.1). All three affected individuals presented with ophthalmoplegia (paralysis of the extraocular muscles) and ptosis at an unusually early age and the phenotype was considered clinically distinct from OPMD.

6.3.2 Data Processing and Exome Sequencing

Six individuals were selected for exome sequencing: the three affected family members (20so00277, 20so00278 and BP999108) and three unaffected family members (20so00275, 20so00276 and JP999109). Samples were sequenced in two batches at different times and locations: 20so00275, 20so00276, 20so00277 and 20so00278 were sequenced in the first batch at the Genetics and Genomics Lab, Royal Brompton and Harefield NHS trust using the 5500XL

SOLiD sequencer and Agilent SureSelect Human All Exon 50Mb capture kit; samples BP999108 and JP999109 were sequenced in the second batch at the Wellcome Trust Centre for Human Genetics using the Illumina HiSeq2000 and the Nextera Rapid Capture Expanded Exome capture kit. The rationale behind the selection of samples for the first sequencing batch was to include affected and unaffected samples to increase power to identify the causative mutation. Affected samples 20so00277 and 20so00278 are second degree relatives of both unaffected samples 20so00275 and 20so00276 and are thus expected to share approximately 50% of their genetic variants. Therefore, by analysing these four samples together it was hoped that the causative mutation would be identified. Over 35 variants unique to the affected samples were identified but none was a strong candidate for the causal mutation. Therefore, the second batch of samples were sequenced to further increase power to detect the causal variant by including another affected individual and another unaffected individual that was unrelated to affected BP999108.

Due to the use of two different sequencing platforms to sequence the two batches of exomes different versions of the Novoalign software were required for alignment of paired end sequence data to the human genome reference sequence 19 (hg19). Batch 1 was sequenced using the SOLiD 5500XL sequencer, requiring the SOLiD Colorspace specific version, NovoalignCSMPI. Trial version v1.02.02 was used. Batch 2 was sequenced using the Illumina HiSeq 2000 sequencer, thus the fully licensed version of NovoalignMPI v2.08.02 that is part of the standard Soton Exome Pipeline was used. All post-alignment steps of the Soton Exome Pipeline were applied to the data for both batches.

SAMtools v0.1.18 was used for variant calling with a base quality threshold of 20 applied. Variants with low read depth (< 4) were filtered out using the vcfutils module of SAMtools and variants with low PHRED quality were removed using custom scripts. Quality control metrics and coverage statistics were calculated using BEDTools and custom scripts.

Annotation of all remaining variants was carried out using ANNOVAR. All variants were annotated with alternative allele frequencies from the 1000 Genomes Project database and Exome Variant Server European American database. Nonsynonymous variants were annotated with scaled predictive

scores from dbNSFP for the algorithms PhyloP, 1-SIFT, Polyphen2, LRT, MutationTaster and GERP++ (Liu et al., 2011).

6.3.3 Variant Filtering

Variant filtering followed a tiered approach. (See Appendix XIX for breakdown of variants remaining after each step of the filtering procedure for each tier of the analysis).

6.3.3.1 Tier 1 Analysis

Samples 20so00275, 20so00276, 20so00277 and 20so00278 were analysed together in the first tier of analysis. The filtering pipeline was as follows:

1) Identify all variants common to the two affected siblings
2) Retain only variants with MAF ≤ 1% in Exome Variant Server
3) Retain variants only variants with MAF ≤ 1% in 1000 Genomes Project
4) Retain only heterozygous variants
5) Remove synonymous variants
6) Remove splicing, ncRNA splicing and 'unknown' variants
7) Remove variants in highly mutable genes (based on Fuentes Fajardo et al, (2012) and a custom program to identify genes with a higher number of mutations per base pair than expected)
8) Remove variants in homopolymer tracts or repeat regions
9) Retain only variants not recorded in dbSNP135
10) Filter out any variants also present in unaffected samples 20so00275 and 20so00276

6.3.3.2 Tier 2 Analysis

All six sequenced samples were considered in the second tier of analysis with focus on variants unique to the three affected samples. The filtering pipeline was as follows:

1) Identify all variants common to three affected samples
2) Retain only variants with MAF ≤ 1% in Exome Variant Server
3) Retain variants only variants with MAF ≤ 1% in 1000 Genomes Project
4) Retain only heterozygous variants

5) Remove variants in highly mutable genes
6) Retain only variants with MAF ≤ 5% in 46 Complete Genomics whole-genome samples
7) Remove variants in any Soton database samples without OPMD-like phenotype
8) Remove synonymous variants
9) Filter out any variants also present in unaffected samples 20so00275, 20so00276 and JP999109

6.3.3.3 Tier 3 Analysis

The final tier of analysis was applied to only BP999108. The same filtering pipeline was applied in tier 3 analysis as in tier 2 analysis except only sample BP999108 was considered. A final filtering step is applied in which all variants identified as 'disease-only' by the Soton pipeline were identified. In the context of this analysis a 'disease-only' variant is defined as unique to at least two of the affected members of this family.

6.3.4 Analysis of OPMD disease genes and genes related to Ptosis or Ophthalmoplegia

All genes identified as related to OPMD in HGMD Professional (release 2013.3, accessed Oct 2013) were interrogated to identify any variation in these genes in BP999108, which may explain the disorder in this family. Further genes associated with either of the eye muscle disorders ptosis or ophthalmoplegia, both of which are present in the phenotype of this family, were also interrogated to identify all variants present in these genes BP999108.

6.3.5 Analysis of *MYH2*

The early age of ptosis and ophthalmoplegia symptoms in the affected family members is unusual in OPMD. Early-onset ophthalmoplegia is a symptom of another similar myopathy; inclusion body myopathy 3 (IBM3). The myopathy is caused by mutations in the *MYH2* gene. The exome sequence data for all affected samples were screened for any variants in *MYH2*. BAM files for all six family members were manually inspected using IGV to check for evidence of each variant in any other sample.

6.4 Results

6.4.1 Sequencing Coverage and Quality Control Measures

Each exome had at least 65% of bases covered by at least 20 reads (Appendix XX). The mean read depth per sample was at least 43 reads (Appendix XX).

Quality control measures were applied to the data to identify potential sample contamination, confirm gender and check sample relatedness (Appendices XXI and XXII).

Contamination can be detected through elevated levels of heterozygosity. The four samples sequenced on the SOLiD platform exhibit higher levels of heterozygosity than would normally be expected (Appendix XXI); levels range from 65.03% to 79.00% compared to ~62% for the two samples sequenced on the Illumina platform. Samples 20so00275 and 20so00278 show extremely high levels of heterozygosity; levels as high as this have been observed in in-house exome samples previously and it was determined that these samples were contaminated. However, all previously exome sequenced samples were sequenced using the Illumina platform rather than the SOLiD platform that was used for these samples with high heterozygosity, therefore we cannot confirm whether these high levels are due to contamination or if this is due to the sequencing platform.

Levels of heterozygosity on the X chromosome can be used as a check for sample gender (Appendix XXI). Levels of X chromosome heterozygosity were calculated as the percentage of variants mapped to the X chromosome that were heterozygous. Males have a much reduced percentage of heterozygous variants on the X chromosome compared to females and to the genome overall. As with overall levels of heterozygosity, all samples sequenced on the SOLiD platform had elevated levels of heterozygosity and again this was particularly evident in samples 20so00275 and 20so00278. Nevertheless, gender was confirmed for all six samples.

Analysis of the percentage of variants shared between two individuals can indicate the relatedness between the individuals. Each sample was compared to all other samples individually and sample relatedness was calculated from the

number of variants shared between the individuals, as a proportion of the overall number of variants per individual. Results of the sample relatedness QC check on the six OPMD-like samples are lower than we would expect to see for related individuals (Appendix XXII); for first degree relatives it is expected that ~60% of genetic variants would be common to both samples. This reduces to ~50% for second degree relatives. The value of 58.20% for samples 20so00276 and BP999108 is the only value approaching what we would expect to see for first degree relatives. All other samples show very little relatedness to one another; the levels observed are similar to those one would expect for unrelated individuals.

6.4.2 Tier 1 Analysis

Tier 1 analysis focused on variants shared by affected siblings 20so00277 and 20so00278 and not shared with their unaffected maternal aunts 20so00275 and 20so00276. All four samples were sequenced using the SOLiD Colorspace platform. Inclusion of the unaffected aunts, who are second-degree relatives of the affected individuals, in the analysis allowed for exclusion of any variants shared with affected individuals. Tier 1 analysis identified 48 variants (Appendix XXIII). Thirty-six of these variants were unique to the two affected individuals; they were not recorded in the 1000 Genomes Project, Exome Variant Server, dbSNP135 or Soton Exome Database. The functions of all genes were investigated to check for any evidence of involvement in any eye disorders or muscular disorders. None of these genes is a likely candidate for the phenotype in this family.

The results of the first tier of analysis do not clearly implicate any variants in the phenotype in this family.

6.4.3 Tier 2 Analysis

All six sequenced family members were considered in the second tier of analysis. The inclusion of the affected mother and unaffected father made the search for the causal variant more powerful. Three variants unique to the three affected family members were identified (Table 6.1).

All variants had previously been identified in the Exome Variant Server at a frequency of 0.1% or more and had a dbSNP135 rsID. *In silico* functional scores identify the T>C nucleotide change in *SV2B* as likely to be damaging and present in a highly conserved region. 1-SIFT, PolyPhen2, LRT, MutationTaster and Grantham score all predict the variant to be a deleterious polymorphism while PhyloP and GERP++ indicate that the region is highly conserved. The GERP++ conservation score in particular is high at 5.2; the upper bound of the GERP++ score is 5.8, indicating that a region is perfectly conserved across all sequenced mammals (Davydov et al., 2010). The *SV2B* gene is likely to be involved in secretion in neural and endocrine cells but there is no clear evidence of a role for this gene in muscle cell function. Therefore, despite the likely damaging nature of this SNV, it is unlikely to be the causal variant in this disorder.

The G>C variant in *AKNA* is predicted to be damaging by three of the predictors but is not in a very highly conserved region. Grantham score is the only one of the functional annotation scores available for the variant in *PKHD1L1* and predicts that the variant is not likely to be deleterious.

The functions of the three genes do not have obvious roles in muscle or exhibit any muscular or eye-related phenotypes and were thus not considered to be likely candidates.

Based on this, all variants identified in step 8 of the filtering pipeline (30 variants; Appendix XXIV) were analysed for potential muscular or eye-related phenotypes. One gene, *TGFBI*, has been previously implicated in corneal dystrophies (Kannabiran and Klintworth, 2006). However, the G1513A (V505I) variant identified in this family is also present in one of BP999108's unaffected sisters; 20so00276. Therefore it is unlikely that this variant is related to the phenotype.

The results of tier 2 analyses do not clearly implicate any variants in the disease phenotype observed in this family.

6.4.4 Tier 3 Analysis

The quality of the exome sequence data for the patients sequenced on the SOLiD platform appears to be inferior to that from the individuals sequenced

on the Illumina platform (Appendices XX, XXI and XXII). It is therefore possible that the causal variant is not called in the affected samples 20so00277 and 20so00278. The third affected member of this family (BP999108) was sequenced on the Illumina platform using the Illumina Nextera Rapid Capture Expanded Exome capture kit and had high quality exome sequence data. Therefore, a third tier of analysis was used in which only variants present in BP999108 were examined.

A total of 223 rare or novel variants were identified in the BP999108 following filtering. Of these 12 were flagged as being unique to this family; they were found to be present in only the BP999108 plus one or both of her daughters (Table 6.2).

The functions of all 12 genes containing a variant unique to this family were investigated to identify any genes with potential roles in muscular disorders. Of particular interest were *MFN2* and *DYSF*, which are related to Charcot-Marie-Tooth disease (most commonly type 2A2) (Zuchner et al., 2004) and limb girdle type 2B muscular dystrophy (Liu et al., 1998) respectively. Both are muscular disorders that cause progressive loss of muscle tissue. In both cases the nucleotide change is unique to the affected members of this family; no evidence for either of these particular nucleotide changes in the 1000 Genomes Project, Exome Variant Server, dbSNP135, or Soton Exome Database. The *MFN2* variant is predicted to be both damaging and in a conserved region by all 7 *in silico* prediction algorithms. The *DYSF* variant is predicted to be damaging by 3 of the tools and both GERP++ and PhyloP indicate that the variant is in a highly conserved region of the genome.

The G1988T nucleotide change in *MFN2* is shared by samples BP999108 and 20so00278. To ensure that the variant is indeed not present in the third affected sample (20so00277), as opposed to simply not called by the pipeline due to poor quality reads, the location of the *MFN2* variant was inspected manually using IGV. There was no clear evidence for the variant in sample 20so00277. The data for unaffected samples 20so00275 and 20so00276 were also interrogated using IGV for any evidence of the variant; as with 20so00277, there was no clear evidence for presence of the variant in these individuals.

Table 6.1. Variants identified as unique to the three affected individuals in tier 2 analysis

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	BP999108	20so00277	20so00278
PKHD1L1	8	110478857	ns	C8464T	H2822Y	rs201478206	.	0.001	.	.	.	.	.	.	83	+	+	+
AKNA	9	117139642	ns	G445C	G149R	rs79864470	.	0.002	0.166	1.000	0.380	0.083	0.014	-1.94	125	+	+	+
SV2B	15	91795601	ns	T635C	I212T	rs145534909	0.001	0.001	0.998	1.000	0.795	1.000	1.000	5.2	89	+	+	+

ns – nonsynonymous variant

Underlined *In silico* scores are classified as damaging/deleterious by corresponding program.

+ indicates that the variant is present in a heterozygous state in that individual

Table 6.2. Variants identified as unique to this family and present in BP999108 in tier 3 analysis

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	BP999108	20so00277	20so00278
MFN2	1	12067225	ns	G1988T	R663L	.	.	.	0.999	0.96	0.993	1.000	1.000	4.22	102	+		+
DYSF	2	71887760	ns	C4961A	T1654K	.	.	.	0.999	0.88	0.997	1.000	0.883	5.43	78	+	+	
HNRNPA2B1	7	26232196	fsd	1002delA	G334fs	.	.	.	.	.	.	.	.	.	.	+	+	
AUH	9	93983148	ns	A782G	N261S	.	.	.	0.964	0.90	0.011	1.000	1.000	3.62	46	+		+
PALM2	9	112705395	ns	C926A	P309H	.	.	.	0.999	1.00	0.63	0.915	0.371	5.7	77	+	+	
C11orf21	11	2323019	ns	G73A	G25R	.	.	.	.	.	.	.	.	.	125	+		+
ABCC8	11	17428190	ns	G3308A	R1103Q	.	.	.	0.065	0.69	0.005	0.767	0.437	-4.79	43	+	+	
MLYCD	16	83945958	ns	G934A	V312I	.	.	.	0.084	0.61	0.088	0.999	0.934	-4.98	29	+	+	
VAT1	17	41168370	ns	A1052T	Q351L	.	.	.	0.997	0.99	0.206	1.000	0.916	4.77	113	+		+
PEX12	17	33902992	ns	T889G	L297V	.	.	.	0.200	0.77	0.688	0.858	0.988	0.13	32	+	+	

ns – nonsynonymous variant; fsd – frameshift deletion
Underlined *In silico* scores are classified as damaging/deleterious by corresponding program.
+ indicates that the variant is present in a heterozygous state in that individual

Affected individuals BP999108 and 20so00277 share the C4961A variant identified in the limb girdle dystrophy *DYSF* gene. As with the variant identified in *MFN2* all reads at the variant location were manually examined using IGV for the third affected individual, as well as the two unaffected sisters of BP999108. There was no evidence of the variant in affected sample 20so00278 or unaffected sample 20so00276. There was, however, some evidence of the variant in unaffected sample 20so00275; six of 57 good quality reads showed evidence of the alternate A allele. It is therefore unlikely that either of these variants is the causal mutation for the pedigree.

A third gene containing a variant unique to members of this family was identified as potentially interesting; *HNRNPA2B1*. The gene harbours a frameshift mutation at nucleotide 1002 in both BP999108 and 20so00277. *HNRNPA2B1* was identified as potentially important due to its role in mRNA trafficking; a function it shares with *PABPN1*, the gene responsible for OPMD. The in-house exome alignment and annotation pipeline did not call the variant in affected sample 20so00278, however, visualisation of the raw reads in IGV revealed that the variant was present in 16% of the reads spanning the nucleotide location. Presence of the variant was confirmed in affected samples BP999108 and 20so00277 in which the variant was called by the pipeline. Absence of the variant was confirmed in unaffected 20so00275 and 20so00276 using IGV.

Despite there being clear evidence for the presence of the frameshift deletion in *HNRNPA2B1* in sample 20so00278, the SAMTools software failed to call the variant. Further examination of the number of reads on both the forward and reverse strands in all three affected samples revealed that the majority of reference and alternate allele reads for 20so00278 were on the reverse strand, with only 2 reads of the reference allele and 2 reads of the alternate allele on the forward strand (Appendix XXV). In samples 20so00277 and BP999108 the minimum number of reads of the alternate allele on the forward strand is 4. Average PHRED base quality was calculated for each sample and was found to be good in all cases.

To test whether the variant was not called due to a low number of reads on the forward strands, a small SAM file and corresponding BAM file of all reads covering the variant location in sample 20so00278 were produced using the

SAMTools software. SAMTools variant calling was applied to the BAM file and confirmed that the variant could not be called. Reads in the SAM file were manipulated one at a time to increase evidence for alternate alleles on the forward strand; changing an alternate allele read on the reverse strand to an alternate allele on the forward strand. SamTools was unable to call the variant until the read distribution had been changed to 7 alternate allele reads on the forward strand and 24 on the reverse strand.

The total read depth at the variant location in 20so00278 was 190, but only 168 of these reads were good quality and included in the counts of reference and alternate alleles on the forward and reverse strands. To investigate whether the poor quality reads were preventing SAMTools from calling the variant, all reads in the SAM file that contained an unknown base in the neighbourhood of the variant location were deleted. The variant is the deletion of an A allele in a run of three A nucleotides. Therefore, if an unknown base was present within the AAA sequence (TTT with the SAM file) then the read was removed from the SAM file. This reduced the overall depth from 190 to 184. On its own this method did not allow SAMTools to call the variant. Combined with more evidence on the forward strand (6 reads in total) the variant could be called. Reads that contained an unknown base immediately before the AAA sequence were then also removed, reducing the overall read depth to 173. Combined with an increase of reads on the forward strand by two (4 reads in total) SAMTools was able to call the variant. Therefore, it appears that the pipeline was unable to call the variant in sample 20so00278 due to a reverse strand bias and a number of poor quality reads spanning the variant location.

To determine whether this *HNRNPA2B1* variant was present at low levels in any other exomed samples without OPMD or the OPMD-like symptoms, 120 samples (including the three affected individuals) were analysed together for evidence of the variant. A custom program was used to analyse the raw sequence VCF files of the 120 exome samples. A total of 128/37,722 reads spanning the variant location contained the variant, and all these reads were from the three affected individuals. Therefore, there is no evidence for this deletion in 117 exome samples without OPMD-like disease.

6.4.5 Analysis of Ptosis, Ophthalmoplegia and OPMD Disease-related Genes

All genes documented as related to either OPMD, ophthalmoplegia, or ptosis were identified using HGMD (Appendix XXVI). Twenty-eight genes were identified as associated with ptosis and 19 with ophthalmoplegia; a total of 41 unique genes. Sample BP999108 was filtered to identify any variation (except synonymous changes) present in any of these 41 genes. A total of 20 variants across 11 of the genes were identified (Appendix XXVII). Four of these variants are only present in BP999108 and one variant is common to the three affected samples. All other variants are present in at least one unaffected sample.

Each of the five variants present in only affected samples were recorded in the 1000 Genomes Project database and Exome Variant Server with a minimum frequency of 2%. Since four of these five variants were unique to BP999108 the BAM files containing the read data of the affected daughters were manually inspected using IGV, to determine whether there was any evidence for any of these variants. The BAM files for the three unaffected family members were also analysed to ensure that there was no evidence of these variants in these individuals (Appendix XXVIII).

Other than the variant in *OPA1*, which is shared by all three affected individuals and not present in the three unaffected individuals, none of the other variants shows clear evidence of presence in the affected samples and absence in the unaffected samples.

The variants in *DOK7* and *TH* do not appear in the affected daughters due to no coverage of these regions in the sequencing, however, both variants have high MAFs of 17% and 21% in 1000 Genomes project and NHLBI ESP respectively, making it unlikely that these variants are disease-related.

There was no evidence for the *NIPBL* non-frameshift deletion in any family members, thus the variant is unique to BP999108 and not responsible for the phenotype.

The *SPG7* variant is present at low levels in three family members other than BP999108; 20so00277, 20so00275, and JP999109. Therefore this cannot be

related to the OPMD-like phenotype since it is present in two unaffected family members.

Five genes implicated in the pathogenesis of OPMD were identified from HGMD. Variation was identified in only BP999108 in only one gene; *COMP*. There is no evidence of this variant in any other family member.

6.4.6 Variation in *MYH2*

All variants called in *MYH2* in the three affected samples were identified – a total of five variants (Appendix XXIX). None of these variants was called in BP999108 however. The BAM files for each sample were manually inspected using IGV for any evidence of each of the identified variants. No variants were shared by the three affected family members and all but one variant were shared with at least one unaffected family member. The only variant not present in any unaffected family members is a novel synonymous polymorphism in 20so00278. One of the variants (G2823T; E941D) is very rare in the 1000 Genomes Project database and NHLBI ESP but is shared with unaffected JP999109. The three other variants are all very common, with MAFs >40% in both the 1000 Genomes database and NHLBI ESP.

To ensure that no variants were filtered out during variant calling due to a small number of alternate allele reads or poor quality reads, the entire *MYH2* gene was manually inspected using IGV for all six samples. No variants were identified that were unique to the three affected individuals.

6.4.7 Segregation Analysis

Evidence from the exome sequence data of the members of this family strongly implicates the G334fs variant in *HNRNPA2B1* as the causal mutation for this phenotype. To further confirm that this variant is present in only the three affected family members, a segregation analysis was carried out in the Wessex Regional Genetics laboratory. Sanger sequencing was carried out on the three affected individuals, the three unaffected individuals and one further unaffected family member: the son of BP999108 and JP999109. This analysis confirmed that the mutation segregated with the phenotype, providing further supporting evidence that this variant is likely to be the causal mutation.

6.5 Discussion

Three individuals from a single family presenting with an OPMD-like phenotype were analysed to identify any potentially causal variation. The phenotype included symptoms of OPMD but was considered a clinically distinct entity. A major distinction from OPMD was the usually early onset of symptoms, in the third decade of life rather than the fifth.

As a result of three tiers of analysis, a number of potentially interesting variants were identified. The first finding was the presence of 'unique to family' variants in two genes previously identified as causal of two different muscular dystrophy disorders; *MFN2* and *DYSF*.

Mutations in the *MFN2* gene are responsible for type 2A Charcot-Marie-Tooth disease; a hereditary autosomal dominant myopathy. *MFN2* is highly conserved and codes for a mitochondrial GTPase (Feely et al., 2011). More than fifty pathogenic mutations have been observed in *MFN2* to date (Inherited Peripheral Neuropathies (IPN) Mutation Database, Accessed Jul 2014), located throughout the gene. The particular G>T mutation identified in this family at nucleotide 1988 is not recorded in the IPN database, nor is the arginine to leucine amino acid change at codon 663 recorded. It is noted that, in most cases, affected individuals are heterozygous for the pathogenic mutation (Polke et al., 2011), therefore if this polymorphism were pathogenic one would expect presentation of the phenotype. Further evidence that this variant was unlikely to be the causal mutation in this family came from the fact that the variant was only shared by two of the three affected family members.

Homozygous and heterozygous variants in *DYSF* have been linked to limb girdle type 2B muscular dystrophy (LGMD2B), distal myopathy, and Miyoshi muscular dystrophy (MM) in a number of pedigrees (Liu et al., 1998). The same variant can manifest as different forms of muscular dystrophy in different members of the same family (Liu et al., 1998). The C4961A (T1654K) variant identified here has not been previously reported as pathogenic in LGMD2B. As with the *MFN2* variant, this variant is only observed in two of the three affected samples from this family and therefore is likely to be unrelated to the phenotype.

Of all the variants detected in the affected samples one candidate gene was identified; *HNRNPA2B1*. A frameshift deletion was identified in the gene and was confirmed to segregate with the disease. As a frameshift deletion this variant has high potential of being damaging because it alters the entire amino acid sequence downstream of the deleted nucleotide. The variant occurs at the start of the final exon of the gene and causes the replacement of the final 19 amino acids of the protein sequence with an alternative 27 amino acids, resulting in an extension of the protein (Figure 6.2). The deletion occurs in a highly conserved region and many of the downstream amino acids, which are altered by the deletion, are also highly conserved.

The deletion is located within the M9 nuclear localisation sequence consisting of 40 amino acids, which is within a C-terminal glycine-rich region of the protein (UniProt, http://www.uniprot.org/uniprot/P22626; accessed Jul 2014; Figure 6.3). The M9 sequence acts as both a localisation and export sequence (Matthew Michael et al., 1995). The deletion mutates the M9 sequence from amino acid 28 onwards.

Characterisation of the M9 sequence in hnRNPA1 (which has high sequence similarity to M9 in hnRNPA2/B1) identified a 19 amino acid sequence that was both sufficient and necessary for nuclear import of the protein (Iijima et al., 2006), termed the M9 core. Furthermore, two particular domains of the sequence – residues 1-7 and 18-19 of the M9 core – are essential for nuclear import (Iijima et al., 2006). The deletion variant at G334 occurs just outside the M9 core region; 3 residues C-terminal to the end of the M9 core sequence (Figure 6.3). Therefore, the change to the protein does not affect the important region of the M9 sequence although it does alter the final 13 residues at the C-terminus of the M9 sequence. What effect this will ultimately have on the protein will require functional studies of the protein to fully understand.

There is no evidence of linkage studies from OPMD-like phenotype analysis that have identified any regions close to the location of the *HNRNPA2B1* gene that could link this gene to the phenotype. In fact, there are no reports of linkage studies carried out in any OPMD-like phenotypes, and therefore no genetic regions have been implicated in such phenotypes by linkage analysis. There is, however, a report of mutations identified in *HNRNPA2B1* in a single family presenting with IBMPFD2 (inclusion body myopathy with early-onset

Wild-type hnRNPA2/B1

GA	AAC	TAT	GGT	CCA	GGA	GGC	AGT	GGA	GGA	AGT	GGG	GGT	TAT	GGT	GGG	AGG	AGC	CGA	TAC	TGA
	N	Y	G	P	G	G	S	G	G	S	G	G	Y	G	G	R	S	R	Y	ST

Mutated hnRNPA2/B1

GA	ACT	ATG	GTC	CAG	GAG	GCA	GTG	GAG	GAA	GTG	GGG	GTT	ATG	GTG	GGA	GGA	GCC	GAT	ACT	GAG	CTT	CTT	CCT	ATT	TGC	CAT	GGG	TAA
	T	M	V	Q	E	A	V	E	E	V	G	V	M	V	G	G	A	D	T	E	L	L	P	I	C	H	G	ST

Figure 6.2. Amino acid sequence of exon 10 in wild-type hnRNPA2/B1 and G334fs mutated hnRNPA2/B1. The 1002delA variant mutates the amino acid sequence of wild-type hnRNPA2/B1 by replacing the 19 amino acids of the final exon with 27 different amino acids, causing an extension of the protein sequence. The deleted nucleotide is highlighted in yellow

1 MEKTLETVPLERKKREKEQFRKLFIGGLSFETTEESLRNYYEQWGKLTDCV
51 VMRDPASKRSRGFGFVTFSSMAEVDAAMAARPHSIDGRVVEPKRAVARE
101 ESGKPGAHVTVKKLFVGGIKEDTEEHHLRDYFEEYGKIDTIEIITDRQSG
151 KKRGFGFVTFDDHDPVDKIVLQKYHTINGHNAEVRKALSRQEMQEVQSSR
201 SGRGGNFGFGDSRGGGGNFGPGPGSNFRGGSDGYGSGRGFGDGYNGYGGG
251 PGGGNFGGSPGYGGGRGGYGGGGPGYGNQGGGYGGGYDNYGGGNYGSGNY
301 NDFGNYNQQPSNYGPMKSGNFGGSRNMGGPYGGGNYGPGGSGGSGGYGGR
351 SRY

Figure 6.3. Amino acid sequence of wild-type hnRNPB1 protein. Blue and black highlighting indicate the exons. Red residues overlap splice junctions between exons. Highlighted residue is G334 in which 1002Adel occurs. Black boxed region is the M9 nuclear localization sequence, red boxed region is the M9 core.

Paget disease with or without frontotemporal dementia 2) (Kim et al., 2013). The IBMPFD2 phenotype has a number of symptoms that are shared with OPMD, namely inclusion bodies, rimmed vacuoles and progressive disease.

The variant identified was a nonsynonymous A > T change at nucleotide 905, resulting in an aspartic acid to valine substitution (Kim et al., 2013). This same mutation was separately observed in another family presenting with an IBMPFD phenotype (Benatar et al., 2013).

IBMPFD2 is considered to be one of a number of multisystem proteinopathies (MSP); dominantly inherited degenerative disorders that primarily affect muscle, bone and the central nervous system. Analysis of 28 further patients presenting with various MSP phenotypes for mutations in *HNRNPA2B1* was unable to identify any variants, suggesting that *HNRNPA2B1* mutations are rare in MSP (Le Ber et al., 2014).

The hnRNPA2/B1 protein is ubiquitously expressed as two alternative isoforms: A2 and B1. These are heterogeneous nuclear ribonucleoproteins and are some of the most abundant of all hnRNPs (Kamma et al., 1999). The A2 isoform is most commonly occurring, accounting for approximately 90% of the protein in most tissue types (Kim et al., 2013).

The hnRNPA/B family of proteins appear to have important roles in the biogenesis and transport of mRNA molecules (Kamma et al., 1999). The

hnRNPA2/B1 protein has multiple roles related to mRNA processing and trafficking both within the nucleus and within the cytoplasm (Munro et al., 1999). When confined to the nucleus of neuronal cells, hnRNPA2/B1 (along with other hnRNPs) is involved in binding to and packaging pre-mRNA while in the cytoplasm, the protein is involved in mRNA transport (Munro et al., 1999). HnRNPA2/B1 also binds a very specific subset of miRNAs and is responsible for controlling their loading into exosomes for trafficking to other cells (Villarroya-Beltri et al., 2013). Modification of the hnRNPA2 response element to disrupt binding interferes with mRNA transport (Munro et al., 1999). Therefore, modification of the hnRNPA2/B1 protein is likely to impact on mRNA transport.

There is little work to date on the function of hnRNPA2/B1, therefore our ability to infer what effect the frameshift mutation has on cells, leading to the phenotype in this family, is limited. The run-on mutation identified in *HNRNPA2B1* is a good candidate for the phenotype presentation in this particularly unique family. The same gene has previously been implicated in other MSP disorders with muscular phenotypes. However, conclusively linking the deletion to the phenotype requires functional work and replication in other individuals presenting with a similar clinical phenotype. Forty-one samples with onset of OPMD-like features under the age of 50, and who did not have the *PABPN1* OPMD-causing (GCN) expansion, were sequenced for evidence of any mutations in *HNRNPA2B1.* No variants were identified although this is not surprising since the phenotype in this single family is so unique.

HnRNPs have been found to interact with the C-terminal domain of PABPN1, namely hnRNPA1 and hnRNPA/B (Fan et al., 2003). *In vitro* it has been shown that mutated PABPN1, hnRNPA1 and hnRNPA/B all co-localise to the protein aggregates while biopsies from OPMD patients have shown evidence of hnRNPA1 in the nuclear inclusions (Fan et al., 2003). The mutated protein forms nuclear inclusions in muscle tissue that act as mRNA traps and sequester other mRNA export proteins (Fan et al., 2003). The interaction between hnRNPA1 and PABPN1 is likely to be required for packaging of mRNA molecules ready for export (Fan et al., 2003).

6.6 Conclusion

Exome sequencing of individuals from a single family presenting with an OPMD-like phenotype allowed for the successful identification of a variant that is a very strong candidate for the causal mutation. Functional work is necessary to confirm that the *HNRNPA2B1* variant is disease-causal but genetic evidence shows that this variant is unique to the affected members of this family. Furthermore, the resultant protein has a role in mRNA trafficking, much like *PABPN1*, the causal gene for OPMD. The mechanism causing the muscular phenotype in this unique family is likely to be similar to the mechanism causing OPMD, however, current understanding of how mutations in *PABPN1* actually cause the specific craniofacial muscular phenotype of OPMD is lacking. It is likely that disruption of RNA processing in the nucleus, a mechanism that both PABPN1 and hnRNPA2/B1 are involved in, is in some way related to the phenotype.

Chapter 7: Thesis Summary and Future Research

Genetic variation is important in the aetiology of many human diseases. Analysis of genetic variation is crucial to increase understanding of disease susceptibility and progression. Since completion of the Human Genome Sequencing Project, analysis of genetic variation associated with disease phenotypes has become accessible to researchers. While there have been many successes in the field of genetics, identifying and characterising mutations for a range of rare and common diseases, both Mendelian and complex, there is still a huge amount of work to be done to improve understanding of disease. The genetic architectures of cancers and complex diseases are particularly complicated, involving many variants and therefore requiring extensive analysis. Without such research it will not be possible to understand the factors underlying disease presentation, which in turn will hamper our ability to successfully manage or cure affected individuals with appropriate treatments.

The aim of the research presented in this thesis was to explore genetic variation harboured by individuals presenting with disease phenotypes, with the aim of identifying genetic variation with potential roles in disease development and/or progression. Through analysis of both genomewide SNP data and whole-exome sequencing I was able to carry out various analyses on the genetic spectrum of individuals presenting with various complex and Mendelian disorders. There were some positive results that emerged from this analysis, however, on the whole it was difficult to resolve many of the phenotypes or draw any robust conclusions from the data, largely due to limitations arising from have too few samples available for analysis.

Sample size is incredibly important in the genetic analysis of disease, particularly when the disease is not Mendelian in nature and is likely to be influenced by many variants, with small associated risks, acting in unison to influence the phenotype. For genetic studies aimed at identifying potential genetic interactions in complex disease it has been suggested that it may be necessary to include approximately 500,000 samples (Zuk et al., 2012). Furthermore, a two-way SNP interaction study including breast cancer samples and control samples found that power to detect interactions between common SNPs (MAF > 10%) was over 90%, however, for SNPs with frequency less than

10% the power was much reduced (Milne et al., 2014b). This suggests that the sample size used in gene-gene interaction detection in chapter 3 is not sufficient to detect true interactions in SNP data. Currently, in GWAS studies it is recommended that thousands of samples are included in order to detect disease-associated variants, therefore, for epistasis detection analyses it is going to be necessary to include just as many samples.

Sample size in exome sequencing studies can limit one's ability to detect disease variants if too few samples are included in the analysis. If the disease under investigation is Mendelian and results from a single highly-penetrant variant then a small number of samples would be sufficient for variant detection. Sample size becomes more of a problem when the disease of interest is a complex trait that does not arise from a single variant, with some suggestion that thousands of samples will be necessary to identify rare variants associated with the phenotype (Kiezun et al., 2012, Cirulli and Goldstein, 2010). Therefore, the exome sequencing analyses that were applied to complex diseases in this thesis (early-onset breast cancer in chapter 4 and NSCLP in chapter 5) did not have sufficient sample sizes to identify disease-associated variants. In the case of the early-onset breast cancer sequencing, the samples that were selected were chosen because they presented with phenotypes that appeared to be Mendelian-like in inheritance or had a strong family history. Therefore, there was a greater likelihood of identifying disease variants than there would have been had eight samples without such phenotypes been selected. Ideally, exome sequencing studies should include many of affected samples presenting with the same phenotype in order for relevant variants to be detected.

Genetic analysis and the techniques available for such research has evolved and become more sophisticated over the years to reach the current situation where the entire genome of an individual can be sequenced. Before the advent of sophisticated sequencing methods, genetic studies took the form of linkage analysis. Such studies were incredibly important for identifying genomic regions associated with a number of diseases, however, these studies are most suited to Mendelian-type diseases resulting from mutations in highly-penetrant genes (Botstein and Risch, 2003). After sequencing of the majority of the human genome was completed in the early 2000s genome (Lander et al., 2001,

Sachidanandam et al., 2001), genomewide association studies were developed in order to detect variants associated with more complex traits and disorders.

Since the first GWAS was published in 2005 (Haines et al., 2005), association studies have been used widely in the field of genetic research to identify variants with disease association. The results of GWAS have been varied with some success but many results with unclear disease associations (McCarthy et al., 2008, Hirschhorn et al., 2002). Some researchers argue that GWAS have been successful and important for increasing understanding of the genetic architecture of many complex disorders and traits. Others argue that the success of GWAS has been limited, with many results that cannot be replicated and may therefore not be true disease associations. In my opinion, GWAS results have certainly be important for examining the overall spectrum of common variation within phenotypes and for detecting disease associations that would not have been possible with linkage analysis. Indeed, GWASs are still a popular analysis technique and associations are still being reported in diseases, including in breast cancer (Michailidou et al., 2013, Michailidou et al., 2015, Purrington et al., 2014). However, there is a major drawback associated with GWASs – the SNPs involved in any identified associations are very often not the variants that are responsible for the associations, they merely act as a marker for the associated variant. As a result further fine-mapping of the associated region is required to identify the variant that is responsible for the association.

As understanding of complex diseases has increased, it has become evident that interactions between genes may account for some of the missing heritability (Zuk et al., 2012). This is an area of research that is currently under explored and cannot be addressed through the use of a GWAS alone. It is possible, however, to use the SNP data harvested from GWASs in other analyses to search for potential interactions or relationships between variants that could be involved in disease development or progression. There is a wealth of GWAS data that has been produced over the past decade that needs to be made widely available and used by genetics researchers to really dig deeper into the architecture of many complex diseases and traits. This will require the use of new analysis techniques rather than simple association studies. Many of the strong disease associations have already been detected by GWAS meaning that studies need to include many thousands of samples to detect modest signals,

which leads to the question as to whether continuing with GWAS is an appropriate research strategy. Instead I believe it would be more appropriate to use the data that is already available with more sophisticated machine learning and analysis strategies to explore the data in a different way and identify underlying structure of the data that is not detectable with a typical GWAS. This reasoning underlies the decision to apply machine-learning models to early-onset breast cancer SNP data as described in chapters 2 and 3 of this thesis. There were just over 500 samples for which SNP data was available meaning that association testing was unlikely to identify true association signals. However, this presented a good opportunity to explore the structure of the data and search for potential interactions between SNPs.

As genetic studies have developed over the past few decades there has been a move away from GWA studies towards DNA sequencing studies due to the advantages associated with sequence data. Rather that detecting an association signal that then has to be further explored in order to attempt to characterise the variant responsible for the signal, all genetic variants are (in theory) captured in the DNA sequencing procedure, allowing for all variants to be analysed. Therefore, one of the major advantages of DNA sequencing technologies is the ability to report all variants in an individual, which can then be interrogated to identify potential causative variation.

Whole-exome sequencing is currently a popular approach to DNA sequencing because of the ability to capture the majority of the protein-coding region of the genome yet only a fraction of the entire genome actually needs to be sequenced. This massively reduces the cost of sequencing multiple individuals yet still ensures that the vast majority of disease-related variants will be detected. As such, exome sequencing is a good compromise, especially for studies of complex diseases and cancers where sequencing as many individuals as possible is desirable.

As DNA sequencing studies continue to become affordable the most preferable option will be to submit all samples of interest for whole-genome sequencing. The obvious advantage of genome sequencing over exome sequencing is the ability to capture all variation in an individual, be that protein-coding mutations in the exons of genes or regulatory mutations within intergenic regions. Currently there are limitations to whole-genome sequence analysis because the

role of non-protein-coding variants is hard to establish. Variants within exonic regions that are non-silent will alter the amino acid sequence of the coded protein, leading to a clear change that can be tested in functional studies to infer the potential consequence of such a variant in an individual. In the case of intergenic variants however, there is not such a clear role to be defined. It is suspected that many intergenic regions of the genome will have regulatory roles, so variants occurring within these regions could have effects on gene expression and splicing among others. Such effects could be much more important in disease-onset and progression but until there has been more thorough characterisation of these intergenic regions and potential roles determined, much of the data produced by genome sequencing will be unusable. Therefore, there is a real need in the field of genetics to characterise intergenic DNA regions. Research from the ENCODE project has successfully started to explore and characterise intergenic DNA regions (Birney et al., 2007, ENCODE Project Consortium, 2012, Schaub et al., 2012) but there needs to be a much wider effort in order to unlock the potential importance of such variants.

As discussed above, it may be informative to analyse genome-wide SNP data using novel methods in order to uncover hidden structures of the data that are not detectable using association studies. Therefore, a novel method was applied to SNP data from a cohort of ~ 500 early-onset breast cancer cases to investigate whether SNPs particularly associated with the estrogen receptor phenotype (this study included samples classified as either positive or negative for the estrogen receptor) could be used to distinguish between samples presenting with either of these phenotypes (Chapter 2). This method of analysis proved to be successful, with greater than 90% accuracy when classifying samples using a subset of 200 SNPs. The results of this analysis demonstrate the potential utility of machine-learning methods and novel approaches to genetic data in the understanding of disease phenotypes. This method allowed for the identification of ~140 genes implicated in the subtype distinction and functional analysis of these genes identified enrichment for an inflammatory response, which has a plausible biological role in oncogenesis and tumour maintenance. Further exploration of this result is necessary to assess the importance of genes that control the inflammatory response in the ER subtype classification.

As has been suggested for GWA studies, it is likely that to confirm the involvement of inflammatory genes in ER subtype determination will require the analysis of SNP variants in many thousands of samples. There is plenty of opportunity to extend and develop this study to further explore the genetic architecture underlying estrogen receptor subtype in early-onset breast cancer. To further investigate the result described in Chapter 2 it would be useful to obtain a test set of other early-onset breast cancer cases to which the SVM model could be applied to validate the strong classification accuracy observed. If the result were to be replicated in further datasets presenting with the same phenotypes it would provide strong evidence that these SNPs are indeed important in the receptor subtype distinction. This would suggest that the genes involved in this distinction should be studied much more closely to really understand how they might be responsible for the different phenotypes. It would also be interesting to compare the genetic landscape of these subtypes to those of control samples to identify other variants that may be involved in tumourigenesis. Indeed, analysis of this SNP data compared to control data has suggested that individuals with early-onset breast cancer have much more disordered genomes than non-breast cancer control samples (Smyth et al., 2015), indicating that there are many variants present in breast cancer cases that are not found in controls. Therefore, it would be interesting to explore whether specific sets of common genetic variants can allow for the distinction of breast cancer subtypes from controls.

Further analyses were carried out on the breast cancer SNP data to investigate whether there were any potential gene-gene interactions that could be underlying tumourigenesis or the receptor subtype distinction (Chapter 3). As with the analysis described in Chapter 2, this analysis was aimed at exploring the data in a different way so as to identify any potential structure to the data that would not be picked up by association analysis. The analyses did identify a number of gene-gene interactions for early-onset breast cancer overall as well as the receptor subtypes considered separately but the interpretation of these results was difficult due to the relatively small samples size and the suggestions in the field that many thousands of samples are probably necessary to detect true interactions (Milne et al., 2014b). To improve and extend this study it would be useful to include many more samples with early-onset breast cancer to further explore the spectrum of gene-gene interactions

in this disease. Comparing cancer samples to controls will potentially allow for the identification of interactions influencing early-onset breast cancer development and progression overall as well as interactions important for the specific subtypes.

Early-onset breast cancer is likely to be influenced by rare (MAF < 1%) and moderately rare (1% < MAF < 5%) variants as well as common variants. Such variants are not typed in a GWAS, so whole-exome sequencing was applied to eight early-onset breast cancer samples to explore the entire genetic spectrum in these individuals (analysis described in Chapter 4). There was a major limitation associated with this analysis however; the number of samples that underwent sequencing was not sufficient for disease variant detection in a complex phenotype such as this. Furthermore, four different breast cancer subtypes (based on receptor phenotypes; Table 4.1) were included in the analysis meaning that very little could be concluded about which variants may have been influencing disease in these individuals. The rationale behind sample selection was that samples with either a strong family history of early-onset disease (true for two of the samples) or what could be considered as Mendelian-like inheritance patterns (true for the remaining 6 samples), may be influenced by a small number of rare variants. If this were true, it may be possible to identify the causal variants despite the small number of samples. Thorough analysis of the exome data was unable to resolve the breast cancer cases, suggesting that these cases are in fact also polygenic in nature with multiple variants contributing to the phenotype.

Ideally, this study should have included samples presenting with the same phenotype, but even then, eight samples presenting with a complex phenotype is unlikely to be sufficient to identify relevant variants. This study could be improved if further samples were sequenced, however, it may take many hundreds to thousands of samples with the same phenotype to achieve sufficient power (Cirulli and Goldstein, 2010, Kiezun et al., 2012). In the immediate future progress could be made in this study by focussing analysis on the two samples presenting with a family history of early-onset breast cancer in at least one close female relative (Figures 4.1 and 4.2). To attempt to resolve the cancers presenting in these families it would have been more suitable if the affected members of each family and several unaffected members were sequenced and analysed together. This would have allowed for

the identification of variants shared by the affected individuals, which would have included the casual variant(s) (assuming the same mutations were responsible for the cases in the same family). In the case of the other six samples, it appears that there are not any obvious Mendelian-like variants that could be responsible for the phenotypes of these individuals. Therefore, it would be necessary to sequence many more samples with the same phenotypes to attempt to identify disease variants.

Despite the lack of success in applying exome sequencing to early-onset breast cancer, at least in terms of causal variant identification, whole-exome sequencing proved to be a successful approach for identifying causal variants in individuals presenting with syndromic forms of CLP (Chapter 5) and in a family presenting with an OPMD-like disease phenotype (Chapter 6).

Analysis of data from all syndromic cases of CLP led to the identification of the causal variant in each patient. All the syndromes present in these samples were associated with a candidate gene in which causal variants had been previously identified. Therefore, analysis of the candidate genes successfully resolved the phenotypes in all cases.

In the case of the OPMD-like affected family, whole-exome sequencing of affected and unaffected family members was successfully employed to identify a strong candidate variant in *HNRNPA2B1* that is likely to explain the disease in these individuals. There was no candidate gene associated with this phenotype but the inclusion of multiple affected and unaffected family members allowed for the identification of a small number of heterozygous candidate variants. This variant in particular was selected as the suspected causal variant because the *HNRNPA2B1* gene has a similar function to that of *PABPN1*, the gene responsible for OPMD phenotypes. Analysis of the effect of this identified mutation on protein function is necessary to provide supporting evidence that this mutation is the causative variant. Functional analysis of the protein containing the mutation *in vitro* will provide information about the potential effect of the mutation on the resultant protein. Further analysis of the mutation in mouse models would be informative and would increase the evidence in support of a pathogenic role for this mutation if the phenotype could be recapitulated in an animal model.

The results from the syndromic CLP and OPMD-like sequencing studies highlight the utility and importance of DNA sequencing methods for the identification of disease variants, particularly when the disorder is Mendelian in nature. Unfortunately, due to the complex nature of early-onset breast cancer and nonsyndromic forms of CLP, characterising disease variants underlying these phenotypes is far more complex. The main limitation of the exome sequencing studies carried out in early-onset breast cancer and NSCLP patient cohorts was the small sample size; the power to detect disease-associated variants in genetic studies very much depends on the number of samples included in the analysis. This is particularly true for rare variants that have small effects on disease since they will only be present in very few individuals presenting with the phenotype, meaning that a large number of samples are necessary to detect such mutations. The result of the small sample sizes of these studies was that very limited conclusions regarding disease-related variants could be drawn. Indeed, it was not appropriate to report any detected variants as potentially disease-related without further analysis including many more samples. This limitation does, however, indicate a clear extension to the exome analyses of early-onset breast cancer and NSCLP – exome sequencing of many more individuals would greatly increase the power of these studies and potentially identify disease-relevant variation.

A second potential limitation of the NSCLP study was the fact that all samples were Colombian, which could potentially complicate the search for disease variants. There are likely to be non-disease variants specific to the Colombian population that may be classified as rare by the analysis pipeline implemented here. As part of the filtering process all variants are compared to variant data from the 1000 Genomes Project and the Exome Variant Server Exome Sequencing Project. The 1000 Genome Project includes individuals from populations of European descent, African descent, Indian descent, Chinese descent and South American descent, while the Exome Sequencing Project has two arms, one containing data from North Americans of European descent and the other containing North Americans of African descent. By cross-referencing against these populations it was possible to obtain allele frequencies for each detected variant, however, these allele frequencies are not specific to the Colombian population and, furthermore, no Colombians are included in either of the databases meaning that the frequencies are not influenced in any way by

Colombian individuals. As a result there are likely to have been variants reported as very rare in these individuals that are in fact fairly common in the Colombian population and vice versa. Therefore, it is likely that disease-associated variants will be overlooked.

To resolve the nonsyndromic cleft lip and palate phenotypes described in Chapter 5 will probably require sequencing of further individuals. In particular, the inclusion of sequence data from control samples from the same population would be useful to exclude any population-specific variants. Furthermore, many of the NSCLP samples came from families in which there were multiple cases of NSCLP so it would be interesting and potentially very informative to include further affected and unaffected samples in order to identify variants shared by affected family members. Variants shared by all affected family members could represent disease-associated variants so inclusion of more affected individuals could be incredibly useful in resolving the phenotypes in these families. Furthermore, the inclusion of unaffected family members would make this analysis even stronger because it would allow for the exclusion of variants present in unaffected individuals, allowing for a small set of variants to be identified in each family. Such an approach does need to be applied with caution in a complex phenotype such as this however, because some disease-related variants may be carried by unaffected family members but may not present symptoms due to incomplete penetrance or they may require the presence of other variants for phenotype expression.

Overall, the analysis presented in this thesis has had varying results. The application of machine-learning methods to early-onset breast cancer genome-wide SNP data was useful for exploring the underlying structure of the data. These analyses allowed for the identification of genes potentially involved in the distinction between estrogen receptor-positive and –negative breast cancer samples as well as the identification of potential gene-gene interactions that may contribute to the estrogen receptor phenotypes. Whole-exome sequencing provided mixed results, with some Mendelian-like phenotypes successfully resolved (syndromic cleft lip and palate and an OPMD-like phenotype) while other more complex phenotypes could not be resolved due to sample size limitations. Despite this, whole-exome sequencing proved to be useful for exploring the underlying genetic spectrum of variants in samples with disease

phenotypes, demonstrating its potentially utility in genome studies, particularly in studies where larger sample sizes are available.

Appendices

Appendix I

200 SNPs which most strongly discriminate ER+ and ER- breast cancers used in the classification models

SNP rs ID	Chromosome	Base pair location in hg19	Gene or nearest gene	Full gene name	Gene function/description	Chi-square value	Weight from linear model
rs7713640	5	169099398	*DOCK2*	Dedicator of cytokinesis 2	Expressed in hematopoietic cells and is involved in remodelling of the actin cytoskeleton, which is necessary for lymphocyte migration.	12.54	-0.748
rs10930176	2	151465458	*RND3* (121278)	Rho family GTPase 3	Binds GTP but has no GTPase activity. Possible role as a negative regulator of cytoskeletal organisation leading to loss of adhesion.	12.44	-0.720
rs10030246	4	186541887	*SORBS2*	Sorbin and SH3 domain containing 2	Functions as an adapter protein that plays a role in the assembling of signaling complexes, and is a link between ABL kinases and actin cytoskeleton.	13.10	-0.630
rs1922987	1	101753101	*S1PR1* (46025)	Sphingosine-1-phosphate receptor 1	Receptor for sphingosine 1-phosphate ligand and suggested role in processes regulating endothelial cell differentiation. Activation of receptor induces cell-cell adhesion.	13.38	-0.583
rs1491477	8	126895276	*BX648371* (58102)	BX648371	Unknown	13.53	-0.537
rs575844	6	92395048	*BC037927*	BC037927	Unknown	12.91	-0.506
rs16946160	13	92203813	*GPC5*	Glypican 5	Cell surface heparan sulphate proteoglycan. Plays a role in cell division control and growth regulation.	13.18	-0.484
rs7313125	12	46464980	*SCAF11* (79077)	SR-related CTD-associated factor 11	Plays a role in pre-mRNA alternative splicing by regulating spliceosome assembly	13.16	-0.480
rs2034614	12	42929397	*PRICKLE1*	Prickle homolog 1 (Drosophila)	Nuclear receptor with potential role in negative regulation of Wnt signalling pathway. Implicated in nuclear trafficking of transcription repressors.	13.53	-0.461
rs4617179	8	4122597	*CSMD1*	CUB and Sushi multiple domains 1	Potential suppressor of squamous cell carcinomas	19.93	-0.460
rs11009375	10	33706590	*NRP1* (82584)	Neuropilin 1	Involved in several signalling pathways that control cell migration. Bind many ligands and co-receptors, influencing cell survival, migration, and attraction.	12.35	-0.460
rs7627289	3	167348838	*WDR49*	WD repeat domain 49	Member of the WD40-repeat family that are implicated in a wide range of processes, including signal transduction, transcription regulation, cell cycle control and apoptosis. Act as scaffolds for the assembly of protein complexes.	13.21	-0.439
rs3733236	4	76923998	*CXCL9*	Chemokine (C-X-C motif) ligand 9	Potential role in T cell trafficking.	12.45	-0.424
rs1052651	12	96052721	*NTN4*	Netrin 4	May play an important role in neural, kidney and vascular development. Promotes neurite elongation from olfactory bulb explants.	14.25	-0.416
rs1522993	2	45842064	*SRBD1* (3631)	S1 RNA binding domain 1	Unknown	12.57	-0.413
rs1320854	2	66363617	*AK131224* (51845)	AK131224	Unknown	17.24	-0.410
rs4923101	11	23520512	*ERV9-1* (438843)	Endogenous retrovirus group 9, member 1	Unknown	12.45	-0.379

rs10029313	4	40350147	*CHRNA9*	Cholinergic receptor, nicotinic, alpha 9 (neuronal)	Ligand-gated ionic channel family member. Forms divalent cation channels and is involved in cochlea hair cell development.	13.76	-0.367
rs1596623	12	46474068	*SCAF11* (88165)	SR-related CTD-associated factor 11	Plays a role in pre-mRNA alternative splicing by regulating spliceosome assembly	12.28	-0.365
rs6820738	4	102041236	*PPP3CA*	Protein phosphatase 3, catalytic subunit, alpha isozyme	Calcium-dependent, calmodulin-stimulated protein phosphatase. This subunit may have a role in the calmodulin activation of calcineurin.	13.67	-0.365
rs561545	11	88101247	*CTSC* (30306)	Cathepsin C	Lysosomal cysteine proteinase that activates serine proteases in immune/inflammatory cells.	14.33	-0.349
rs2129662	12	131084080	*RIMBP2*	RIMS binding protein 2	Plays a role in synaptic transmission.	12.67	-0.331
rs10498930	6	83366352	*UBE3D* (235834)	Ubiquitin protein ligase E3D	Accepts ubiquitin from specific E2 ubiquitin-conjugating enzymes, and transfers it to substrates, generally promoting their degradation by the proteasome	12.59	-0.329
rs1230064	17	43461460	*ARHGAP27* (9809)	Rho GTPase activating protein 27	GTPase-activating protein that inhibits Rho-like proteins.	12.57	-0.320
rs6082866	20	22803909	*CR618492* (93277)	CR618492	Unknown	19.19	-0.317
rs10505604	8	134027588	*TG*	Thyroglobulin	Glycoprotein that acts as a substrate for the synthesis of thyroxine and triiodothyronine as well as storage of the inactive forms of thyroid hormone and iodine.	12.39	-0.312
rs6766993	3	62433496	*CADPS*	Ca++-dependent secretion activator	Neural/endocrine-specific cytosolic and peripheral membrane protein required for the Ca2+-regulated exocytosis of secretory vesicles.	13.20	-0.311
rs7822472	8	6094406	*CR623475* (166673)	CR623475	Unknown	13.65	-0.307
rs6049391	20	24078128	*AK090900* (102275)	AK090900	Unknown	12.18	-0.306
rs4861065	4	40344395	*CHRNA9*	Cholinergic receptor, nicotinic, alpha 9 (neuronal)	Ligand-gated ionic channel family member. Forms divalent cation channels and is involved in cochlea hair cell development.	12.36	-0.285
rs2760415	9	5696021	*KIAA1432*	KIAA1432	Required for phosphorylation and localization of GJA1	12.35	-0.276
rs11980210	7	105448210	*ATXN7L1*	Ataxin 7-like 1	Unknown	13.34	-0.274
rs4288395	8	4122628	*CSMD1*	CUB and Sushi multiple domains 1	Potential suppressor of squamous cell carcinomas	18.79	-0.273
rs4782969	16	84492254	*ATP2C2*	ATPase, Ca++ transporting, type 2C, member 2	Magnesium-dependent enzyme that catalyzes the hydrolysis of ATP coupled with the transport of calcium	12.40	-0.271
rs12669163	7	50278585	*IKZF1* (65793)	IKAROS family zinc finger 1 (Ikaros)	Regulator of lymphocyte differentiation.	12.27	-0.271
rs2026604	1	153408831	*S100A7L2*	S100 calcium binding protein A7-like 2	Unknown	13.78	-0.253
rs638515	12	50303927	*LOC283332*	LOC283332	Unknown	12.32	-0.248
rs1515160	2	123929100	*AX747402* (840522)	AX747402	Unknown	15.59	-0.243
rs6706214	2	123929932	*AX747402* (839690)	AX747402	Unknown	15.59	-0.243

rs2052450	5	103925210	*CR610784* (327983)	CR610784	Unknown	14.26	-0.237
rs10906861	10	15252397	*FAM171A1* (1250)	Family with sequence similarity 171, member A1	Unknown	13.48	-0.233
rs1870583	17	12032944	*MAP2K4*	Mitogen-activated protein kinase kinase 4	Dual specificity protein kinase which acts as an essential component of the MAP kinase signal transduction pathway and the stress-activated protein kinase/c-Jun N-terminal kinase (SAP/JNK) signaling pathway.	14.77	-0.230
rs1230073	17	43467025	*ARHGAP27* (4244)	Rho GTPase activating protein 27	GTPase-activating protein that inhibits Rho-like proteins.	12.57	-0.230
rs4421124	5	165002334	*BC011998* (972913)	BC011998	Unknown	13.67	-0.220
rs11126074	2	66352599	*AK131224* (40827)	AK131224	Unknown	12.24	-0.218
rs4949788	1	77750699	*AK5*	Adenylate kinase 5	Regulates adenine nucleotide composition within cells by catalysing the transfer of phosphate groups.	14.17	-0.215
rs7717089	5	128643236	*ADAMTS19* (152867)	ADAM metallopeptidase with thrombospondin type 1 motif, 19	Member of the ADAMTS protein family. Suggested role in proteolysis	12.55	-0.210
rs11257101	10	11492868	*USP6NL* (9641)	USP6 N-terminal like	Acts as a GTPase-activating protein for RAB5A. Involved in receptor trafficking.	12.94	-0.200
rs10850783	12	110242933	*TRPV4*	Transient receptor potential cation channel, subfamily V, member 4	Non-selective calcium permeable cation channel probably involved in regulation of osmotic pressure.	15.59	-0.200
rs13261574	8	23168650	*LOXL2*	Lysyl oxidase-like 2	Mediates the post-translational oxidative deamination of lysine residues on target proteins and is essential to the biogenesis of connective tissue. Catalyses the first step in the formation of crosslinks in extracellular matrix proteins.	14.19	-0.199
rs4273857	8	23173053	*LOXL2*	Lysyl oxidase-like 2	Mediates the post-translational oxidative deamination of lysine residues on target proteins and is essential to the biogenesis of connective tissue. Catalyses the first step in the formation of crosslinks in extracellular matrix proteins.	13.66	-0.199
rs2454781	10	11493713	*USP6NL* (8796)	USP6 N-terminal like	Acts as a GTPase-activating protein for RAB5A. Involved in receptor trafficking.	14.45	-0.190
rs2707237	2	38095622	*AK057187* (39081)	AK057187	Unknown	12.26	-0.183
rs9945952	18	54702678	*WDR7* (5642)	WD repeat domain 7	Binds Rab3A GDP/GTP exchange and activating proteins which are regulators of the control of the calcium-dependent exocytosis of neurotransmitters.	13.63	-0.183
rs9520707	13	108666901	*FAM155A* (147441)	Family with sequence similarity 155, member A	Unknown	12.83	-0.182
rs211146	11	17981047	*SERGEF*	Secretion regulating guanine nucleotide exchange factor	Probable guanine nucleotide exchange factor (GEF), which may be involved in the secretion process.	13.58	-0.179
rs211141	11	17994314	*SERGEF*	Secretion regulating guanine nucleotide exchange factor	Probable guanine nucleotide exchange factor (GEF), which may be involved in the secretion process.	13.58	-0.179
rs2670765	11	17985273	*SERGEF*	Secretion regulating guanine nucleotide exchange factor	Probable guanine nucleotide exchange factor (GEF), which may be involved in the secretion process.	12.89	-0.179

rs4953908	2	134960601	*MGAT5* (51229)	Mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase	Important for regulation of the biosynthesis of glycoprotein oligosaccharides. Catalyzes the addition of beta-1,6- N-acetylglucosamine to the alpha-linked mannose of biantennary N-linked oligosaccharides.	12.17	-0.173
rs366592	21	17334568	*USP25* (82191)	Ubiquitin specific peptidase 25	Deubiquitinating enzyme that processes newly synthesized Ubiquitin, recycles ubiquitin molecules or edits polyubiquitin chains and prevents proteasomal degradation of substrates.	12.81	-0.173
rs2431111	5	103932079	*CR610784* (321114)	CR610784	Unknown	15.63	-0.169
rs2148713	1	153416036	*S100A7L2* (3533)	S100 calcium binding protein A7-like 2	Unknown	13.16	-0.163
rs4789799	17	80533079	*FOXK2*	Forkhead box K2	Binds purine-rich motifsin the HIV LTR and IL2 promoter. Possible role in regulating viral and cellular promoter elements.	12.71	-0.157
rs926745	20	22817120	*CR618492* (80066)	CR618492	Unknown	15.57	-0.155
rs4858909	3	162424485	*BC073807* (18042)	BC073807	Unknown	14.65	-0.153
rs6798611	3	162437102	*BC073807* (5425)	BC073807	Unknown	14.65	-0.153
rs502025	5	104022597	*CR610784* (230596)	CR610784	Unknown	13.34	-0.152
rs13006388	2	213825499	*IKZF2* (38914)	IKAROS family zinc finger 2 (Helios)	Associates with Ikaros and functions in early hematopoietic development.	17.97	-0.151
rs13006331	2	213825327	*IKZF2* (39086)	IKAROS family zinc finger 2 (Helios)	Associates with Ikaros and functions in early hematopoietic development.	17.72	-0.151
rs10515354	5	103947366	*CR610784* (305827)	CR610784	Unknown	14.71	-0.147
rs10515355	5	103947537	*CR610784* (305656)	CR610784	Unknown	14.71	-0.147
rs743562	5	131872383	*IL5* (4753)	Interleukin 5 (colony-stimulating factor, eosinophil)	Induces terminal differentiation of late-developing B-cells to immunoglobulin secreting cells	15.59	-0.140
rs440075	10	132190059	*GLRX3* (211419)	Glutaredoxin 3	Modulates the function of protein kinase C theta and may inhibit apoptosis and play a role in cell growth. Expression of this gene may be a marker for cancer.	12.58	-0.137
rs396186	10	132193177	*GLRX3* (214537)	Glutaredoxin 3	Modulates the function of protein kinase C theta and may inhibit apoptosis and play a role in cell growth. Expression of this gene may be a marker for cancer.	12.58	-0.137
rs2075713	11	124617939	*VSIG2*	V-set and immunoglobulin domain containing 2	Unknown	12.69	-0.137
rs12907348	15	49190740	*SHC4*	SHC (Src homology 2 domain containing) family, member 4	Activates both Ras-dependent and Ras-independent migratory pathways in melanomas.	14.78	-0.132

rs12039894	1	74464625	*LRRIQ3* (27079)	Leucine-rich repeats and IQ motif containing 3	Unknown	19.12	-0.128
rs297907	12	50316818	*BC034605*	BC034605	Unknown	19.93	-0.121
rs10501316	11	46090550	*PHF21A*	PHD finger protein 21A	Component of the BHC complex that represses transcription of neuron-specific genes.	14.11	-0.079
rs2655060	12	50298535	*FAIM2*	Fas apoptotic inhibitory molecule 2	Antiapoptotic protein which protects cells from Fas-induced apoptosis.	12.3	-0.077
rs10491618	9	33634591	*ANXA2* (9061)	Annexin A2	Calcium-dependent membrane-binding protein with potential role in signal transduction pathways and regulating cell growth.	12.22	-0.066
rs13016788	2	123930193	*AX747402* (839429)	AX747402	Unknown	13.82	-0.063
rs1519291	2	123954518	*AX747402* (815104)	AX747402	Unknown	13.38	-0.063
rs4845700	1	154981708	*ZBTB7B*	Zinc finger and BTB domain containing 7B	Transcription regulator and regulator of lineage commitment of immature T-cell precursors. Transcriptional repressor of the collagen COL1A1 and COL1A2 genes. May also function as a repressor of fibronectin and other extracellular matrix genes	14.88	-0.046
rs1555794	1	117602077	*TTF2*	Transcription termination factor, RNA polymerase II	Has dsDNA-dependent ATPase activity and RNA polymerase II termination activity. Associates with human splicing complexes and is involved in pre-mRNA splicing.	14.01	-0.042
rs10919584	1	198773357	*PTPRC* (46812)	Protein tyrosine phosphatase, receptor type, C	Protein tyrosine phosphatase that regulates T-cell and B-cell antigen receptor signalling. Suppresses JAK kinases, thus functioning as a cytokine receptor signalling regulator.	12.81	-0.041
rs1476689	7	22265585	*RAPGEF5*	Rap guanine nucleotide exchange factor (GEF) 5	Guanine nucleotide exchange factor (GEF) for RAP1A, RAP2A and MRAS/M-Ras-GTP.	15.03	-0.031
rs1577635	6	62040305	*KHDRBS2* (349560)	KH domain containing, RNA binding, signal transduction associated 2	RNA-binding protein involved in the regulation of alternative splicing, mRNA splice site selection and exon inclusion.	13.55	-0.025
rs2125111	12	103003530	*IGF1* (129152)	Insulin-like growth factor 1 (somatomedin C)	Involved in the mediation of growth and development.	13.04	-0.008
rs450798	12	50304949	*LOC283332*	LOC283332	Unknown	13.81	0.006
rs12897276	14	71111542	*TTC9*	Tetratricopeptide repeat domain 9	May play a role in cancer cell invasion and metastasis. Hormonally regulated in breast cancer cells	14.65	0.007
rs4884974	13	54895842	*BC044614* (76482)	BC044614	Unknown	12.31	0.021
rs16906788	8	138146684	*FAM135B* (995584)	Family with sequence similarity 135, member B	Unknown	12.45	0.024
rs9359489	6	82120647	*FAM46A* (334801)	Family with sequence similarity 46, member A	Unknown	15.25	0.030
rs10737381	1	34996413	*GJB5* (224308)	Gap junction protein, beta 5, 31.1kDa	Involved in intercellular communication related to epidermal differentiation and environmental sensing.	14.58	0.030
rs7328941	13	108564192	*FAM155A* (44732)	Family with sequence similarity 155, member A	Unknown	13.49	0.037
rs2711775	3	163180280	*LOC647107* (159191)	LOC647107	Unknown	12.13	0.040

rs10482869	21	16572437	*NRIP1* (135311)	Nuclear receptor interacting protein 1	Nuclear protein that interacts with the hormone-dependent domain of nuclear receptors. Modulates transcriptional activity of the estrogen receptor.	12.39	0.041
rs843055	3	162975100	*LOC647107*	LOC647107	Unknown	12.97	0.050
rs843044	3	162986427	*LOC647107*	LOC647107	Unknown	12.54	0.050
rs407921	21	17316672	*USP25* (64295)	Ubiquitin specific peptidase 25	Deubiquitinating enzyme that processes newly synthesized Ubiquitin, recycles ubiquitin molecules or edits polyubiquitin chains and prevents proteasomal degradation of substrates.	15.43	0.052
rs6677928	1	113771402	*AK123703* (22527)	AK123703	Unknown	12.16	0.053
rs394119	12	50305007	*LOC283332*	LOC283332	Unknown	17.81	0.056
rs2824703	21	19566451	*CHODL*	Chondrolectin	Encodes a type 1 membrane protein with a carbohydrate recognition domain characteristic.	13.85	0.057
rs2292354	12	110368201	*GIT2*	G protein-coupled receptor kinase interacting ArfGAP 2	GTPase-activating protein for the ADP ribosylation factor family.	15.22	0.070
rs7628408	3	151377883	*AADACL2* (73821)	Arylacetamide deacetylase-like 2	Proposed role in a metabolic process	21.47	0.076
rs1521590	3	151396537	*AADACL2* (55167)	Arylacetamide deacetylase-like 2	Proposed role in a metabolic process	21.47	0.076
rs6048372	20	22804426	*CR618492* (92760)	CR618492	Unknown	19.02	0.076
rs4679882	3	151355834	*AADACL2* (95870)	Arylacetamide deacetylase-like 2	Proposed role in a metabolic process	18.76	0.076
rs1331147	10	91224329	*SLC16A12*	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Catalyzes the rapid transport of monocarboxylates across the plasma membrane	14.53	0.079
rs17122097	10	91225140	*SLC16A12*	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Catalyzes the rapid transport of monocarboxylates across the plasma membrane	14.53	0.079
rs10828316	10	22838389	*PIP4K2A*	Phosphatidylinositol-5-phosphate 4-kinase, type II, alpha	Catalyzes the phosphorylation of phosphatidylinositol 5-phosphate to form phosphatidylinositol 4,5-bisphosphate.	12.79	0.083
rs10828317	10	22839628	*PIP4K2A*	Phosphatidylinositol-5-phosphate 4-kinase, type II, alpha	Catalyzes the phosphorylation of phosphatidylinositol 5-phosphate to form phosphatidylinositol 4,5-bisphosphate.	12.20	0.083
rs2675221	15	76125378	*UBE2Q2* (10244)	Ubiquitin-conjugating enzyme E2Q family member 2	Catalyses the covalent attachment of ubiquitin to other proteins.	19.61	0.095
rs584105	12	50306148	*BC034605*	BC034605	Unknown	17.51	0.096
rs2936610	11	126104559	*FAM118B*	Family with sequence similarity 118, member B	Unknown	12.98	0.099
rs11220405	11	126088824	*FAM118B*	Family with sequence similarity 118, member B	Unknown	12.86	0.099

rs2276312	11	126074192	*RPUSD4*	RNA pseudouridylate synthase domain containing 4	Proposed role in pseudouridine synthesis	12.47	0.099
rs254044	5	103961296	*CR610784* (291897)	CR610784	Unknown	15.56	0.101
rs8049005	16	48177062	*ABCC12*	ATP-binding cassette, sub-family C (CFTR/MRP), member 12	ABC transporter which transports molecules across extracellular and intracellular membranes. Increased expression of this gene is associated with breast cancer.	12.83	0.102
rs8062641	16	48183299	*ABCC12*	ATP-binding cassette, sub-family C (CFTR/MRP), member 12	ABC transporter which transports molecules across extracellular and intracellular membranes. Increased expression of this gene is associated with breast cancer.	12.83	0.102
rs10830841	11	88147573	*BC038205* (9976)	BC038205	Unknown	14.16	0.103
rs4146035	7	150152520	*GIMAP8*	GTPase, IMAP family member 8	Exerts an anti-apoptotic effect in the immune system and is involved in infection response	14.58	0.124
rs3760048	16	474180	*RAB11FIP3* (1488)	RAB11 family interacting protein 3 (class II)	Acts as a regulator of the formation, targeting and fusion of intracellular transport vesicles. Interacts with and regulates Rab GTPases.	13.20	0.125
rs161773	5	104086307	*CR610784* (166886)	CR610784	Unknown	13.90	0.126
rs807043	10	102847237	*TLX1NB* (1842)	TLX1 neighbor	Unknown	13.92	0.130
rs2823130	21	16566350	*NRIP1* (129224)	Nuclear receptor interacting protein 1	Nuclear protein that interacts with the hormone-dependent domain of nuclear receptors. Modulates transcriptional activity of the estrogen receptor.	13.03	0.131
rs4237446	10	15241923	*FAM171A1* (11724)	Family with sequence similarity 171, member A1	Unknown	12.41	0.141
rs12050778	15	76126371	*UBE2Q2* (9251)	Ubiquitin-conjugating enzyme E2Q family member 2	Catalyses the covalent attachment of ubiquitin to other proteins.	14.62	0.150
rs1487602	12	131103343	*RIMBP2*	RIMS binding protein 2	Plays a role in synaptic transmission.	12.46	0.151
rs4463750	10	14686790	*FAM107B*	Family with sequence similarity 107, member B	Unknown	12.31	0.156
rs970392	6	39229763	*KCNK5* (32512)	Potassium channel, subfamily K, member 5	pH-dependent, voltage insensitive, outwardly rectifying potassium channel.	13.52	0.156
rs7648055	3	151414394	*AADACL2* (37310)	Arylacetamide deacetylase-like 2	Proposed role in a metabolic process	20.09	0.156
rs7648113	3	151414310	*AADACL2* (37394)	Arylacetamide deacetylase-like 2	Proposed role in a metabolic process	19.22	0.156
rs4359642	2	222150749	*EPHA4* (132000)	EPH receptor A4	Receptor tyrosine kinase with a potential role in the mediation of developmental events, particularly in the nervous system.	12.52	0.157
rs2056246	11	18051446	*TPH1*	Tryptophan hydroxylase 1	Catalyses the biosynthesis of serotonin.	12.22	0.157
rs9889792	17	13626903	*AK123263* (53246)	AK123263	Unknown	13.48	0.159
rs8032239	15	58505122	*ALDH1A2*	Aldehyde dehydrogenase 1 family, member A2	Catalyses the synthesis of retinoic acid.	13.74	0.160
rs494734	20	56908563	*RAB22A*	RAB22A, member RAS oncogene family	May be involved in trafficking endosomal compartments.	13.16	0.162

rs568531	12	121579673	*P2RX7*	Purinergic receptor P2X, ligand-gated ion channel, 7	Receptor for ATP that acts as a ligand-gated ion channel. Responsible for ATP-dependent lysis of macrophages through the formation of membrane pores permeable to large molecules.	14.00	0.168
rs1653583	12	121598652	*P2RX7*	Purinergic receptor P2X, ligand-gated ion channel, 7	Receptor for ATP that acts as a ligand-gated ion channel. Responsible for ATP-dependent lysis of macrophages through the formation of membrane pores permeable to large molecules.	12.94	0.168
rs7038242	9	28805269	*LINGO2* (85966)	Leucine rich repeat and Ig domain containing 2	Unknown	12.78	0.174
rs7852834	9	28812865	*LINGO2* (93562)	Leucine rich repeat and Ig domain containing 2	Unknown	12.78	0.174
rs6799331	3	25703640	*TOP2B*	Topoisomerase (DNA) II beta 180kDa	Control of topological states of DNA during transcription. Catalyses the transient breakage and rejoining of DNA strands. Also involved in chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication.	12.41	0.178
rs2689158	1	238907046	*LOC339535* (257729)	LOC339535	Unknown	12.64	0.188
rs1861809	12	110245588	*TRPV4*	Transient receptor potential cation channel, subfamily V, member 4	Non-selective calcium permeable cation channel probably involved in regulation of osmotic pressure.	13.38	0.192
rs13340131	3	21794174	*ZNF385D* (1358)	Zinc finger protein 385D	Unknown	12.52	0.193
rs1523558	4	162002577	*CR595965* (297528)	CR595965	Unknown	13.14	0.196
rs3740030	10	91222287	*SLC16A12*	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Catalyzes the rapid transport of monocarboxylates across the plasma membrane	13.31	0.208
rs17122305	10	91237003	*SLC16A12*	Solute carrier family 16, member 12 (monocarboxylic acid transporter 12)	Catalyzes the rapid transport of monocarboxylates across the plasma membrane	12.65	0.208
rs10747353	1	77751193	*AK5*	Adenylate kinase 5	Regulates adenine nucleotide composition within cells by catalysing the transfer of phosphate groups.	12.72	0.209
rs9883543	3	162447818	*BC073807*	BC073807	Unknown	12.49	0.215
rs2293786	3	25666485	*TOP2B*	Topoisomerase (DNA) II beta 180kDa	Control of topological states of DNA during transcription. Catalyses the transient breakage and rejoining of DNA strands. Also involved in chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication.	12.27	0.229
rs11705878	3	25683930	*TOP2B*	Topoisomerase (DNA) II beta 180kDa	Control of topological states of DNA during transcription. Catalyses the transient breakage and rejoining of DNA strands. Also involved in chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication.	12.27	0.229

rs861157	20	42397767	*GTSF1L* (42125)	Gametocyte specific factor 1-like	Unknown	12.14	0.242
rs11962201	6	53489402	*AK126334* (7435)	AK126334	Unknown	12.25	0.245
rs11745512	5	164055678	*BC011998* (26257)	BC011998	Unknown	12.42	0.246
rs1577608	1	18708402	*IGSF21* (3426)	Immunoglobin superfamily, member 21	May act as a receptor in immune response pathways.	12.26	0.254
rs4983411	14	105916797	*MTA1*	Metastasis associated 1	May be involved in the regulation of transcription, which may result from chromatin remodelling.	12.83	0.259
rs995815	20	54530910	*CBLN4* (41587)	Cerebellin 4 precursor	Involved in regulation of neurexin signalling during synapse development.	13.18	0.279
rs4759493	12	131042460	*RIMBP2*	RIMS binding protein 2	Plays a role in synaptic transmission.	12.69	0.280
rs2836912	21	40506194	*PSMG1* (41196)	Proteasome (prosome, macropain) assembly chaperone 1	Chaperone protein which promotes assembly of the 20S proteasome as part of a heterodimer with PSMG2.	12.47	0.292
rs6850890	4	118387036	*TRAM1L1* (380300)	Translocation associated membrane protein 1-like 1	Required for the translocation of secretory proteins across the endoplasmic reticulum membrane	12.19	0.293
rs1034461	22	26331185	*MYO18B*	Myosin XVIIIB	May regulate muscle-specific genes and may influence intracellular trafficking, depending on localisation. May play a role in the control of tumour development and progression.	12.14	0.300
rs2442477	8	6354270	*MCPH1*	Microcephalin 1	Encodes a DNA damage response protein that may play a role in G2/M checkpoint arrest.	13.91	0.303
rs432519	7	151013263	*NUB1* (25595)	Negative regulator of ubiquitin-like proteins 1	Specific down-regulator of the NEDD8 conjugation system. Recruits NEDD8, UBD, and their conjugates to the proteasome for degradation.	14.47	0.305
rs2706399	5	131867702	*IL5* (9434)	Interleukin 5 (colony-stimulating factor, eosinophil)	Induces terminal differentiation of late-developing B-cells to immunoglobulin secreting cells	15.57	0.333
rs9555336	13	107904621	*FAM155A*	Family with sequence similarity 155, member A	Unknown	15.08	0.338
rs3136146	16	14028379	*ERCC4*	Excision repair cross-complementing rodent repair deficiency, complementation group 4	Complexes with ERCC1 to form a structure-specific DNA repair endonuclease responsible for the 5-prime incision during DNA repair.	13.25	0.351
rs9356859	6	23347228	*HDGFL1* (776479)	Hepatoma derived growth factor-like 1	Unknown	12.21	0.353
rs1174966	7	22046072	*CDCA7L* (60530)	Cell division cycle associated 7-like	Plays a role in transcriptional regulation and gene expression. Important for oncogenic role in mediating the full transforming effect of MYC in medulloblastoma cells. Involved in apoptotic signalling pathways	12.41	0.360
rs4074228	12	118987597	*SUDS3* (131758)	Suppressor of defective silencing 3 homolog (S. cerevisiae)	Subunit of the histone deactylase-dependent SIN3A corepressor complex. Potential role in tumour suppressor pathways.	12.19	0.366
rs4424536	1	59544435	*LOC729467* (53175)	LOC729467	Unknown	12.34	0.367
rs4726411	7	154074097	*DPP6*	Dipeptidyl-peptidase 6	Binds specific voltage-gated potassium channels and alters their expression and biophysical properties.	12.51	0.370

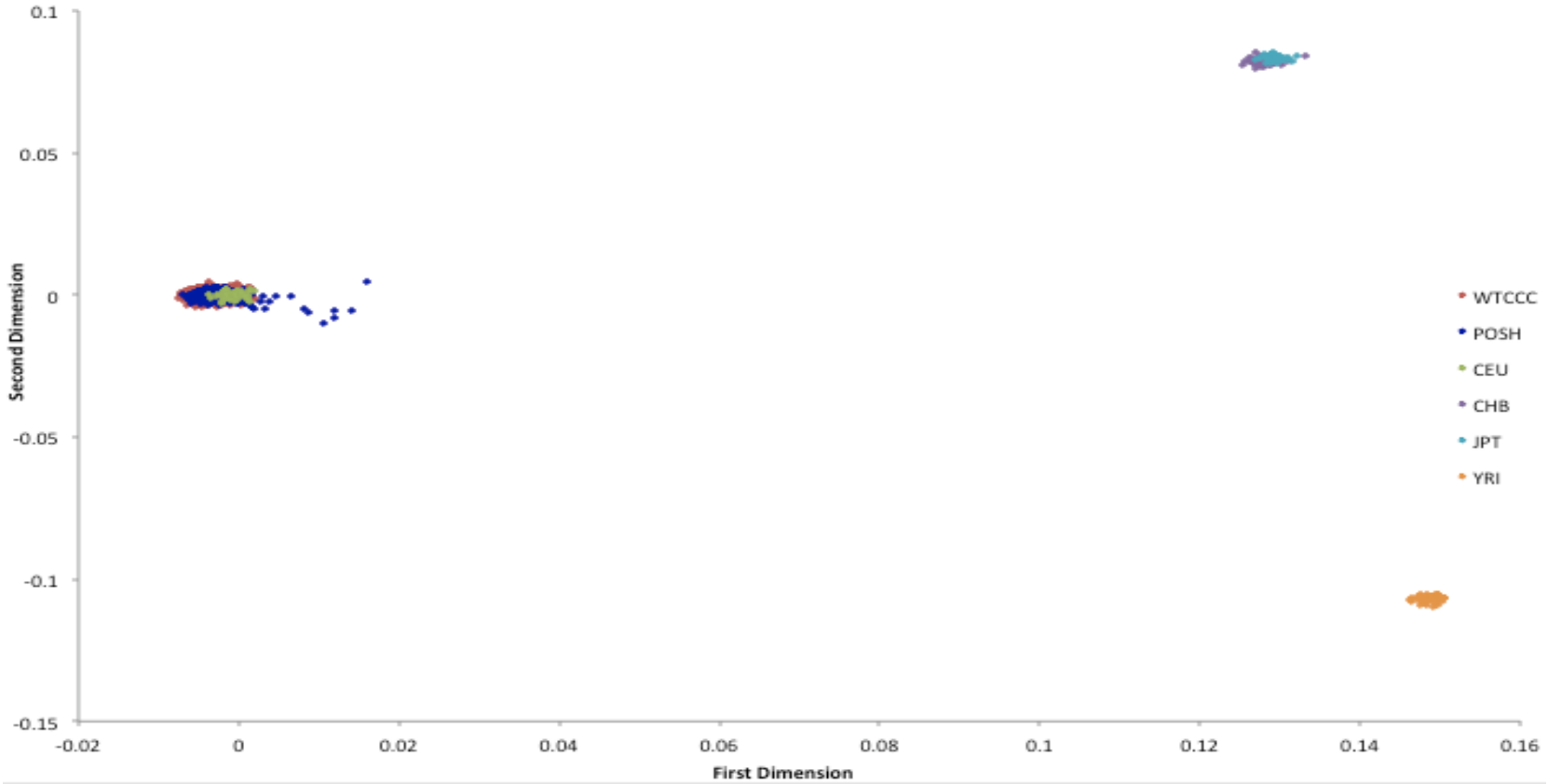
rs2513421	11	88133201	*BC038205* (24348)	BC038205	Unknown	17.40	0.370
rs2767326	1	117376448	*CD2* (64598)	CD2 molecule	T cell surface antigen.	12.51	0.382
rs16908031	10	57707708	*PCDH15* (320006)	Protocadherin-related 15	Calcium-dependent cell-adhesion protein. Essential for maintenance of normal retinal and cochlear function.	15.75	0.388
rs1367002	11	20871798	*NELL1*	NEL-like 1 (chicken)	Plays a role in the control of cell growth and differentiation.	12.38	0.392
rs1174965	7	22046015	*CDCA7L* (60473)	Cell division cycle associated 7-like	Plays a role in transcriptional regulation and gene expression. Has an oncogenic role in mediating the transforming effect of MYC in medulloblastoma cells. Involved in apoptotic signalling pathways	13.65	0.404
rs1033975	1	68567848	*AK096081/ AK124028*	AK096081/AK124028	Unknown	12.55	0.419
rs773620	1	113785301	*AK123703* (36426)	AK123703	Unknown	13.84	0.427
rs7096374	10	8484113	*BC031880* (173845)	BC031880	Unknown	12.51	0.430
rs10139234	14	71123560	*TTC9*	Tetratricopeptide repeat domain 9	May play a role in cancer cell invasion and metastasis. Hormonally regulated in breast cancer cells	13.49	0.432
rs6499323	16	70624483	*SF3B3* (12913)	Splicing factor 3b, subunit 3, 130kDa	Subunit of the splicing factor SF3B. Involved in binding pre-mRNA upstream of the intron's branch site. May function in chromatin modification, transcription, splicing, and DNA repair.	12.23	0.433
rs10021032	4	94671148	*GRID2*	Glutamate receptor, ionotropic, delta 2	Glutamate receptor that is one of the excitatory neurotransmitter receptors in the brain. Potential role in neuronal apoptosis.	12.96	0.440
rs1001776	16	4222512	*SRL* (16865)	Sarcalumenin	May be involved in the regulation of calcium transport.	13.95	0.444
rs4714562	6	42036850	*TAF8*	TAF8 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 43kDa	Subunit of the TFIID transcription factor complex. TFIID recognises the promoters of many genes and initiates assembly of a transcription preinitiation complex containing RNA polymerase II.	13.70	0.446
rs4665867	2	26739004	*OTOF*	Otoferlin	Potential role in vesicle membrane fusion. Mutations in OTOF cause neurosensory nonsyndromic recessive deafness.	13.31	0.462
rs4807753	19	5399090	*ZNRF4* (56336)	Zinc and ring finger 4	Unknown	12.53	0.496
rs10905371	10	8480044	*BC031880* (169776)	BC031880	Unknown	12.84	0.568
rs17096099	14	30332962	*PRKD1*	Protein kinase D1	Serine/threonine-protein kinase that regulates membrane receptor signalling, Golgi transport, protection from oxidative stress at the mitochondria, gene transcription, and regulation of cell shape, motility, and adhesion.	13.37	0.593
rs10854759	22	22889528	*abParts*	abParts	Unknown	13.17	0.599

rs7600426	2	2263952	*MYT1L*	Myelin transcription factor 1-like	May function as a panneural transcription factor associated with neuronal differentiation. May play a role in the development of neurons and oligodendrogalia in the CNS	15.37	0.613
rs9384805	6	112101725	*FYN*	FYN oncogene related to SRC, FGR, YES	Membrane-associated tyrosine kinase involved in cell growth control. Member of the protein-tyrosine kinase oncogene family.	12.97	0.642
rs290826	1	97473707	*DPYD* (69595)	Dihydropyrimidine dehydrogenase	Pyrimidine catabolic enzyme involved in catabolism of uracil and thymidine.	12.19	0.676
rs4146282	3	108497048	*RETNLB* (20918)	Resistin like beta	Probable hormone.	13.83	0.717
rs3796133	3	98517843	*DCBLD2*	Discoidin, CUB and LCCL domain containing 2	Proposed role in cell adhesion, wound healing, intracellular receptor mediated signalling pathway, and negative regulation of cell growth	12.52	0.758
rs3773162	3	14526031	*SLC6A6*	Solute carrier family 6 (neurotransmitter transporter, taurine), member 6	Required for the uptake of taurine.	12.14	0.783
rs4936947	11	124498785	*FLJ00213*	FLJ00213	Unknown	17.24	0.869
rs4732990	8	29536211	*LINC00589* (42567)	Long intergenic non-protein coding RNA 589	Unknown	17.16	1.176

SNP weights are taken from a linear model built using one iteration of 10-fold cross-validation in the WEKA Explorer. Classification accuracy for this model was 92.4%. Magnitude of SNP weights indicates importance of the SNP for classifying cases; absolute values of larger weights are more important. Positive SNP weights relate to classifying ER+ cases while negative SNP weights relate to classifying ER- cases. For those SNPs that are not located within a gene the nearest gene is given and the distance of the SNP from this gene is indicated by d=.

Appendix II

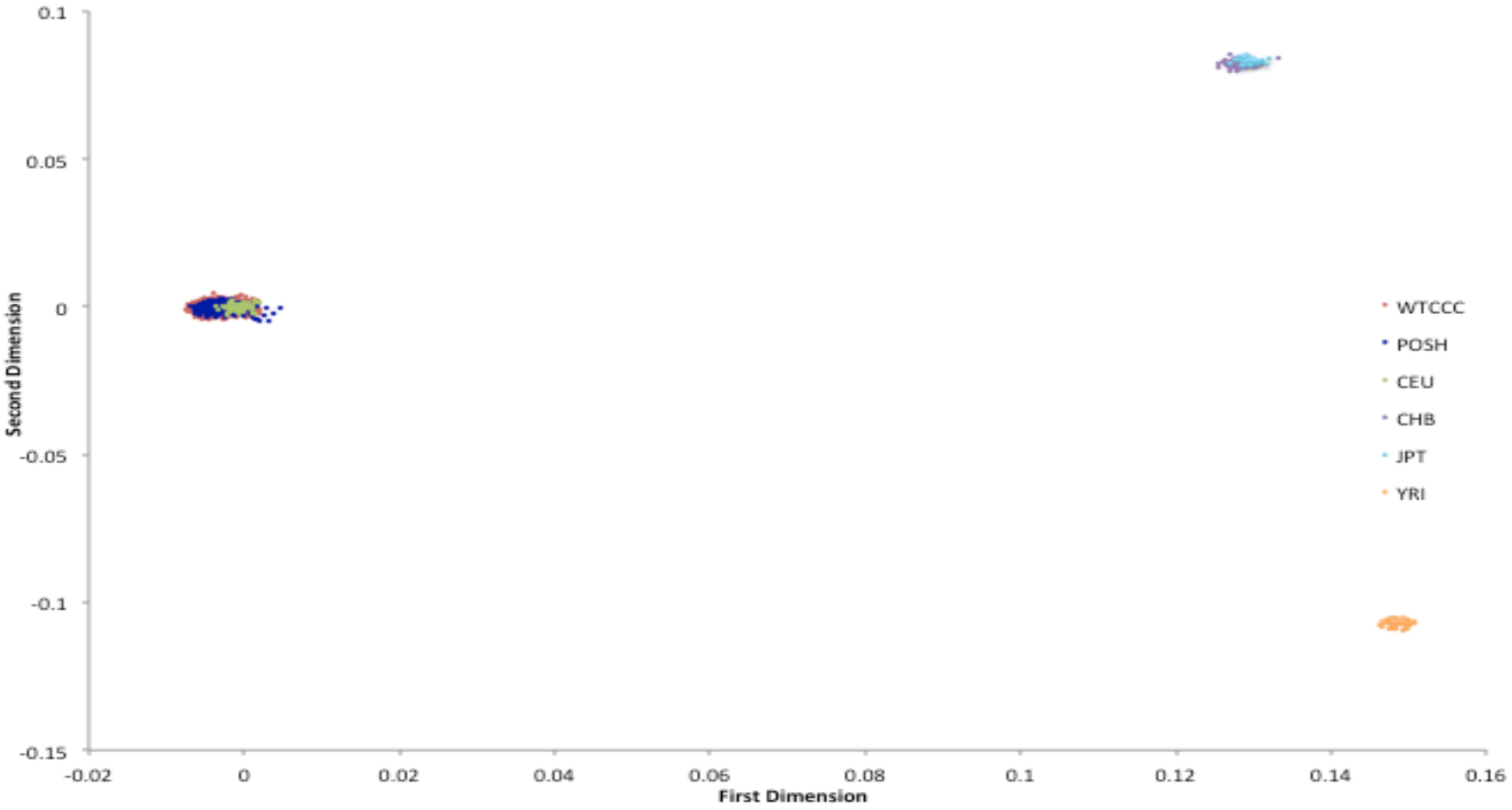
MDS plot of POSH samples, WTCCC controls and HapMap2 reference populations



POSH, WTCCC and CEU (Utah residents with European ancestry) populations cluster tightly, indicating their genetic similarity. Han Chinese (CHB) and Japanese (JPT) populations cluster and the Yoruba (YRI) population cluster separately, demonstrating the genetic differences between Caucasian, Asian and African populations. Eight POSH samples demonstrate ethnic admixture and were therefore removed from all downstream analysis.

Appendix III

MDS plot of POSH samples, WTCCC controls and HapMap2 reference populations after removal of ethnic outliers.



MDS plot of POSH samples, WTCCC controls and HapMap2 reference populations after removal of eight POSH samples that were ethnic outliers. No other POSH or WTCCC samples are obvious outliers, thus all remaining samples were retained for the analysis.

Appendix IV

87 breast cancer susceptibility genes

ADAM29	*FGFR2*	*NTN4*
ANKRD16	*FOXQ1*	*OVOL1*
ARHGEF5	*FTO*	*PALB2*
ATM	*H19*	*PAX9*
BABAM1	*HCN1*	*PDE4D*
BCL2L15	*HNF4G*	*PEX14*
BRCA1	*IGFBP2*	*PTHLH*
BRCA2	*IGFBP5*	*RAB3C*
BRIP1	*ISYNA1*	*RAD23B*
CASC16	*ITPR1*	*RAD51B*
CASP8	*KCNN4*	*RALY*
CCDC88C	*KLF4*	*RANBP9*
CCND1	*LGR6*	*RHBDD3*
CDCA7	*LSP1*	*SLC25A21*
CDKN2A	*MAP3K1*	*SLC4A7*
CDKN2B	*MAPKAP5*	*SSBP4*
CDYL2	*MDM4*	*STXBP4*
CHEK2	*MEIR3*	*TBX3*
CHST9	*METAP1D*	*TCF7L2*
CLPTM1L	*MIR1208*	*TERT*
COX11	*MIR1972-2*	*TET2*
DIRC3	*MKL1*	*TGFBR2*
DNAJC1	*MLLT10*	*TOX3*
EBF1	*MRPS30*	*TP53*
EGOT	*MYC*	*TPN2*
ELL	*NEK10*	*TPN22*
EMID1	*NOBOX*	*ZMIZ1*
ESR1	*NOTCH2*	*ZNF283*
FCGR1B	*NRIP1*	*ZNF365*

Appendix V

49 breast cancer-associated SNPs identified from GWAS

SNP	Chromosome	Base pair location (hg19)	Type	Gene
rs616488	1	10566215	Intronic	*PEX14*
rs11249433	1	121280613	ncRNA intronic	*EMBP1*
rs6678914	1	202187176	Intronic	*LGR6*
rs4849887	2	121245122	Intergenic	*LOC84931* (21197)
rs2016394	2	172972971	Intergenic	*DLX2* (5493)
rs1550623	2	174212894	Intergenic	*CDCA7* (6667)
rs10931936	2	202143928	Intronic	*CASP8*
rs13387042	2	217905832	Intergenic	*TNP1* (181050)
rs6762644	3	4742276	Intronic	*ITPR1*
rs4973768	3	27416013	UTR3	*SLC4A7*
rs9790879	5	44899885	Intergenic	*MRPS30* (84267)
rs889312	5	56031884	Intergenic	*MAP3K1* (79016)
rs1353747	5	58337481	Intronic	*PDE4D*
rs1432679	5	158244083	Intronic	*EBF1*
rs11242675	6	1318878	Intergenic	*FOXQ1* (3885)
rs204247	6	13722523	Intergenic	*RANBP9* (10727)
rs17530068	6	82193109	Intergenic	*FAM46A* (262338)
rs3757318	6	151914113	Intronic	*CCDC170*
rs720475	7	144074929	Intronic	*ARHGEF5*
rs9693444	8	29509616	Intergenic	*LINC00589* (69160)
rs13281615	8	128355618	Intergenic	*POU5F1B* (72239)
rs1562430	8	128387852	Intergenic	*POU5F1B* (40005)
rs11780156	8	129194641	Intergenic	*MIR1208* (32207)
rs1011970	9	22062134	ncRNA intronic	*CDKN2B-AS1*
rs865686	9	110888478	Intergenic	*KLF4* (636431)
rs2380205	10	5886734	Intergenic	*ANKRD16* (16955)
rs7072776	10	22032942	Downstream	*MLLT10*
rs10995190	10	64278682	Intronic	*ZNF365*
rs704010	10	80841148	Intronic	*ZMIZ1*
rs7904519	10	114773927	Intronic	*TCF7L2*
rs11199914	10	123093901	Intergenic	*FGFR2* (143943)
rs2981579	10	123337335	Intronic	*FGFR2*
rs3817198	11	1909006	Intronic	*LSP1*
rs909116	11	1941946	Intronic	*TNNT3*
rs614367	11	69328764	Intergenic	*CCND1* (127109)
rs17356907	12	96027759	Intergenic	*NTN4* (23824)
rs1292011	12	115836522	Intergenic	*MED13L* (559859)
rs999737	14	69034682	Intronic	*RAD51B*
rs8009944	14	69039588	Intronic	*RAD51B*
rs12443621	16	52548037	Intronic	*TOX3*
rs3803662	16	52586341	ncRNA exonic	*CASC16*
rs13329835	16	80650805	Intronic	*CDYL2*
rs1436904	18	24570667	Intronic	*CHST9*
rs8170	19	17389704	Exonic	*BABAM1*
rs2363956	19	17394124	Exonic	*ANKLE1*
rs4808801	19	18571141	Intronic	*ELL*
rs3760982	19	44286513	Intergenic	*KCNN4* (1104)
rs2284378	20	32588095	Intronic	*RALY*
rs2823093	21	16520832	Intergenic	*NRIP1* (83706)

Intergenic SNPs are annotated with the nearest gene; distance from the gene (bp) is given in parentheses.

Appendix VI

Websites used to construct a comprehensive list of 372 genes that are candidate members of the *TP53* pathway.

All websites were accessed in September 2011.

Wikipedia webpage for p53 protein, 'Interactions' subsection:
http://en.wikipedia.org/wiki/P53

KEGG pathway for p53 signaling:
(http://www.genome.jp/kegg-bin/show_pathway?hsa04115).

p53 pathways subsection of 'The TP53 Web Site':
http://p53.free.fr/p53_info/p53_Pathways.html

TP53 tumor protein p53 page of NCBI website:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=7157#geneInteractions

Appendix VII

327 genes of the *TP53* pathway

ABL1	*CCL18*	*CSNK1D*	*GTSE1*	*MAPK8*	*PBK*	*RCHY1*	*SP3*	*TP53BP2*
ACTA1	*CCNA2*	*CSNK2A1*	*GUSBP1*	*MAPK9*	*PCDHA4*	*RECQL4*	*SSTR3*	*TP53I3*
AIMP2	*CCNG1*	*CSNK2B*	*HABP4*	*MAPKAPK5*	*PERP*	*RFWD2*	*ST13*	*TP53INP1*
ANKRD2	*CCNH*	*CUL9*	*HDAC1*	*MDC1*	*PHB*	*RPA1*	*STEAP3*	*TP53RK*
ANXA3	*CCT5*	*DAXX*	*HDAC2*	*MDM2*	*PIAS1*	*RPL11*	*STK11*	*TP63*
APEX1	*CD40LG*	*DDX5*	*HDAC3*	*MDM4*	*PIAS4*	*RPRM*	*STK4*	*TP73*
APTX	*CD82*	*DHCR24*	*HIF1A*	*MED1*	*PIDD*	*RRM2*	*STRA13*	*TRAF6*
ARID3A	*CDC14A*	*DVL2*	*HIPK1*	*MED17*	*PIN1*	*RRM2B*	*STX5*	*TRIAP1*
ARIH2	*CDC14B*	*E2F1*	*HIPK2*	*MIF*	*PLAGL1*	*S100A2*	*SULT1E1*	*TSC2*
ARL3	*CDC25C*	*E4F1*	*HMGB1*	*MNAT1*	*PLK1*	*S100A4*	*SUMO1*	*TSG101*
ATF3	*CDC42*	*EEF2*	*HMGB2*	*MNDA*	*PLK3*	*S100A8*	*SYVN1*	*TXN*
ATM	*CDK1*	*EFEMP2*	*HSBP1*	*MPHOSPH6*	*PMAIP1*	*S100B*	*TADA3*	*UBB*
ATR	*CDK2*	*EGR1*	*HSP90AA1*	*MRE11A*	*PML*	*SAT1*	*TAF1A*	*UBC*
AURKA	*CDK5*	*EI24*	*HSP90AB1*	*MSH2*	*PNP*	*SCAMP1*	*TAF1B*	*UBE2A*
BAI1	*CDK7*	*EIF2AK2*	*HSPA1A*	*MSH6*	*POLA1*	*SERPINB5*	*TAF1C*	*UBE2I*
BAK1	*CDK9*	*EIF2S2*	*HSPA8*	*MSX1*	*PPA1*	*SERPINB9*	*TAF9*	*UBE2K*
BANP	*CDKN1A*	*ELL*	*HSPA9*	*MTA2*	*PPM1D*	*SERPINE1*	*TAF9B*	*UBE3A*
BARD1	*CDKN2A*	*EP300*	*HTT*	*MUC1*	*PPP1CA*	*SESN1*	*TBP*	*USP7*
BAX	*CDKN2C*	*EPHA3*	*HUWE1*	*NAP1L1*	*PPP1CC*	*SESN2*	*TEC*	*VRK1*
BBC3	*CEBPZ*	*ERCC3*	*IFI16*	*NCL*	*PPP1R13B*	*SESN3*	*TEP1*	*WDR33*
BCL2	*CFLAR*	*ERCC6*	*IGFBP3*	*NDN*	*PPP1R13L*	*SETD7*	*TFAP2A*	*WRAP53*
BCL2L1	*CHD3*	*ERH*	*ING1*	*NEDD8*	*PPP2CA*	*SFN*	*TFAP2C*	*WRN*
BCR	*CHEK1*	*ESR1*	*ING4*	*NFKBIA*	*PPP2R2B*	*SHISA5*	*TFDP1*	*WT1*
BLM	*CHEK2*	*FAM173A*	*ING5*	*NFYA*	*PRIM1*	*SIAH1*	*THAP8*	*WWOX*
BNIP3L	*CHUK*	*FAS*	*KAT2B*	*NFYB*	*PRKCA*	*SIN3A*	*THBS1*	*XRCC6*
BRCA1	*COPS2*	*FBXO11*	*KAT5*	*NKAP*	*PRKDC*	*SIRT1*	*THRB*	*YBX1*
BRCA2	*COPS3*	*FOXO3*	*KLF4*	*NMT1*	*PRKRA*	*SMAD2*	*TK1*	*YPEL3*
BRCC3	*COPS4*	*FXYD6*	*KLF6*	*NMT2*	*PRMT1*	*SMAD3*	*TNFRSF10B*	*YWHAG*
BRE	*COPS5*	*GADD45A*	*KPNA2*	*NPM1*	*PSMD11*	*SMARCA4*	*TNFRSF10C*	*YWHAZ*
BRF1	*COPS6*	*GADD45B*	*KPNB1*	*NQO1*	*PSME3*	*SMARCB1*	*TOP1*	*YY1*
BTBD2	*COPS7A*	*GADD45G*	*LAMA4*	*NR3C1*	*PTEN*	*SMG1*	*TOP2A*	*ZCCHC10*
BTK	*COPS8*	*GNL3*	*MAD2L1BP*	*NTHL1*	*PTGS2*	*SMN1*	*TOP2B*	*ZHX1*
CABLES1	*COX17*	*GPS1*	*MAGEB18*	*NUMB*	*PTK2*	*SMN2*	*TOPORS*	*ZMAT3*
CABLES2	*CR2*	*GPS2*	*MAPK1*	*OPN1LW*	*PTTG1*	*SMYD2*	*TP53*	*ZNF148*
CAPN1	*CREB1*	*GSK3B*	*MAPK10*	*PAFAH1B3*	*RAB4A*	*SNRPN*	*TP53AIP1*	*ZNF24*
CARM1	*CREBBP*	*GSTM4*	*MAPK3*	*PARP1*	*RAD51*	*SP1*	*TP53BP1*	*ZNHIT1*
CCDC106	*CSNK1A1*	*GTF2H1*						

Appendix VIII

GATK quality criteria annotations

QD – Qual By Depth. Calculated from the variant confidence divided by the unfiltered depth of non-reference samples. Low scores suggest a false positive call or artefact of the data.

MQ – Root Mean Square of Mapping Quality. Estimation of the overall mapping quality of all reads overlapping a variant site, averaged over all samples

HaplotypeScore – defined as the consistency of the site with two (and only two) segregating haplotypes. Higher scores are suggestive of regions with poor sequence alignment, which leads to SNP or indel calls that are artifacts of this poor alignment.

MQRankSum – Mapping Quality Rank Sum Test. Defined as a U-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities.

ReadPosRankSum – Read Pos Rank Sum Test. Defined as a U-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of a read of an alternate allele call. If the alternate allele call is only seen close to the end of sequence reads this is suggestive of an error.

Appendix IX

Mapping and coverage summary statistics for exome sequencing

Sample	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
Total no. sequence reads	98,207,532	103,485,536	95,229,626	94,298,228	99,731,496	85,944,740	78,256,592	95,424,926
Total no. aligned reads	94,137,095	99,522,188	91,331,286	87,072,415	95,792,743	82,191,545	75,080,439	91,540,596
Total no. unique alignments	88,392,929	93,445,028	85,681,256	81,776,349	90,064,624	77,301,387	70,484,755	86,061,321
Mapped to target reads +/- 150bp (%)	84.57	84.84	84.84	83.49	84.48	84.4	84.61	84.06
Mapped to target reads (%)	73.87	74.07	74.12	73.39	73.27	73.15	73.31	72.66
Target bases with coverage >1 (%)	94.11	93.97	93.93	94.85	94.16	95.12	93.52	94.04
Target bases with coverage >5 (%)	86.19	85.93	85.85	87.71	86.2	87.95	84.96	85.82
Target bases with coverage >10 (%)	81.12	80.84	80.67	83.24	81.09	83.64	79.17	80.47
Target bases with coverage >20 (%)	73.02	72.81	72.41	76.06	72.93	77.02	69.66	71.96
Mean read depth across exome	90.16	98.05	92.67	86.11	93.1	117.14	74.6	88.04

Appendix X

Potential compound heterozygous splicing or non-damaging variants identified from all genes of the exome

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
HMCN1	30	1	ns	184242922	C4515G	D1505E	.	.	0.002	0.56	B	45	C	-1.284				+				
HMCN1	100	1	sp	184402554	15440-9C>T	.	.	.	0.000	.	.	.	.	.				+				
CYP39A1	9	6	sp	46671677	1065+6C>G	.	.	.	0.003	.	.	.	.	.	+							
CYP39A1	5	6	ns	46712104	C713G	S238C	.	.	0.006	0.14	D	112	MR	3.323	+							
C6orf221	3	6	ns	74130138	A488G	Q163R	N	.	.	1.00	B	43	C	.								+
C6orf221	3	6	ns	74130144	T494C	V165A	N	.	.	0.74	B	64	MC	.								+
COL5A1	2	9	ns	136722662	C193T	R65W	.	.	0.003	0.00	D	101	MR	-2.962	+							
COL5A1	54	9	sp	136849503	4230+5C>T	.	.	.	0.003	.	.	.	.	.	+							
DMBT1	16	10	sp	124348611	1783+8T>C	.	.	.	0.000	.	.	.	.	.		+						
DMBT1	51	10	ns	124386687	C6424A	Q2142K	.	.	0.000	0.03	P	53	MC	-1.284		+						
ADAMTS20	36	12	sp	42056120	5312+7G>C	.	N	.	.	.	.	.	.	.		+						
ADAMTS20	24	12	ns	42109760	A3416T	K1139I	N	.	.	0.12	B	102	MR	2.750		+						
LRP1	36	12	sp	55863816	5795-9C>T	.	N	.	.	.	.	.	.	.							+	
LRP1	41	12	ns	55865848	G6731A	R2244Q	.	.	0.000	0.24	P	43	C	-2.962							+	
FREM2	1	13	ns	38163545	G4064A	R1355Q	N	.	.	0.63	B	43	C	.			+					
FREM2	18	13	sp	38346732	8281+9C>T	.	N	.	.	.	.	.	.	.			+					
THSD4	15	15	sp	69844595	2769+3A>G	.	N	.	.	.	.	.	.	.		+						
THSD4	17	15	ns	69856727	G3017A	R1006H	.	.	0.000	0.18	D	29	C	.		+						
NECAB2	10	16	sp	82589302	850-9G>A	.	N	.	.	.	.	.	.	.	+							
NECAB2	10	16	sp	82589305	850-6C>T	.	.	.	0.003	.	.	.	.	.	+							
ABCC11	.	16	sp	46806451	.	.	.	.	.	.	.	.	.	.							+	
ABCC11	2	16	ns	46823327	A7G	R3G	.	.	0.000	N/A	B	125	MR	-2.216							+	

Where a specific variant is present in a sample this is indicated by +

B, benign; C, conservative; D, probably damaging; MC, moderately conservative; MR, moderately radical; N, novel; ns, nonsynonymous; P, possibly damaging; sp, intronic splice variant

Appendix XI

Potential recessive splicing or non-damaging variants identified from all genes of the exome

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
PTPRC	.	1	sp	196983842	.	.	N	.	.	.	.	.	.	.								+
RAB3GAP2	.	1	sp	218436369	.	.	N	.	.	.	.	.	.	.							+	+
SLC1A4	1	2	ns	65070321	G40A	A14T	.	.	0.007	1	B	58	MC	.				+				
PDE1A	.	2	sp	182741266	.	.	N	.	.	.	.	.	.	.	+							
COL4A3	.	2	sp	227867905	.	.	.	.	.	.	.	.	.	.			+					
ITIH4	13	3	sp	52830040	1679+7G>T	.	.	.	0.001	.	.	.	.	.		+						
IFT80	.	3	sp	161459127	.	.	N	.	.	.	.	.	.	.			+				+	
MYL5	7	4	ns	665710	T449C	I150T	.	.	0.009	0.31	B	89	MC	.				+				
IPO11	.	5	sp	61781495	.	.	N	.	.	.	.	.	.	.								+
BRD8	.	5	sp	137535004	.	.	N	.	.	.	.	.	.	.					+			
CPNES	.	6	sp	36897919	.	.	.	.	.	.	.	.	.	.							+	
TMEM209	.	7	sp	129630926	.	.	.	.	.	.	.	.	.	.		+		+				
FAM90A10	4	8	ns	7666231	G706T	G236C	N	.	.	N/A	.	159	R	.						+		
RRM2B	9	8	sp	103294301	790-8C>A	.	N	.	.	.	.	.	.	.		+						
CELF2	.	10	sp	11407787	.	.	.	.	.	.	.	.	.	.					+			
CSGALNACT2	.	10	sp	42991385	.	.	N	.	.	.	.	.	.	.			+					
ZNF215	.	11	sp	6904204	.	.	.	.	.	.	.	.	.	.			+					
TRIM49L2	3	11	ns	89408169	T142A	F48I	.	.	.	N/A	B	21	C	.				+				
UBTFL1	1	11	ns	89459632	G867C	K289N	N	.	.	N/A	.	94	MC	.			+					
KRT83	.	12	sp	50994358	.	.	N	.	.	.	.	.	.	.						+		
ATG2B	34	14	sp	95839354	4843-9C>T	.	.	.	0.001	.	.	.	.	.		+						
ATG2B	30	14	ns	95842917	T4393C	S1465P	.	.	0.001	0.34	B	74	MC	-2.962		+						
HSD3B7	6	16	sp	30905653	532-9T>C	.	.	.	0.000	.	.	.	.	.		+						
CES4A	8	16	sp	65594910	901-9C>G	.	N	.	.	.	.	.	.	.		+						
KIAA0355	.	19	sp	39516423	.	.	N	.	.	.	.	.	.	.	+							
COL6A2	.	21	sp	46369798	.	.	.	.	.	.	.	.	.	.							+	
SLC9A6	.	X	sp	134908290	.	.	N	.	.	.	.	.	.	.					+			

Where a specific variant is present in a sample this is indicated by +
B, benign; C, conservative; MC, moderately conservative; N, novel; ns, nonsynonymous; R, radical; sp intronic splice variant

Appendix XII

Variants identified in known cancer genes catalogued in the COSMIC database

Gene	Exon	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8
EPS15	8	1	ns	51637315	G1087A	A363T	N	.	.	0.88	B	58	MC	-2.216				+				
BCL9	10	1	ns	145562428	A3325G	M1109V	N	.	.	0.33	B	21	C	.		+						
ABL2	10	1	ns	177346184	T1573G	S525A	N	.	.	0.2	B	99	MC	0.749			+					
TPR	33	1	ns	184572374	C4582G	L1528V	N	.	.	0.27	B	32	C	3.934		+						
TPR	20	1	sp	184587873	2335-8A>-	.	.	.	.	.	.	.	.	.						+		
MLF1	8	3	ns	159805590	C637G	Q213E	N	.	.	0.21	B	29	C	.							+	
NSD1	.	5	sp	176494703	.	.	.	.	0.000	.	.	.	.	.		+						
NSD1	7	5	sp	176595561	3114+9C>T	.	N	.	.	.	.	.	.	.				+				
AKAP9	13	7	ns	91489514	G3864A	M1288I	N	.	.	0.42	B	10	C	3.507		+						
WHSC1L1	14	8	sp	38282126	2243-6T>C	.	.	.	0.001	.	.	.	.	.							+	
MEN1	6	11	sp	64331737	670-9C>T	.	N	.	.	.	.	.	.	.	+							
MLL2	10	12	ns	47731149	C2584T	P862S	N	.	.	N/A	B	74	MC	-2.962								+
NKX2-1	2	14	ns*	36058321	G83A	R28H	N	.	.	0.21	B	29	C	-2.962		+						
TCL1A	.	14	sp	95247902	.	.	.	.	.	.	.	.	.	.	+							
TCL1A	.	14	sp	95247902	.	.	.	.	.	.	.	.	.	.		+						
TSC2	23	16	sp	2065904	2639+10C>T	.	N	.	.	.	.	.	.	.							+	
MYH11	15	16	sp	15757696	1749+3A>C	.	.	.	0.004	.	.	.	.	.					+			
MLLT6	10	17	ns	34126448	G1339A	A447T	N	.	.	0.45	B	58	MC	-0.568						+		
ASPSCR1	5	17	sp	77545979	375-7C>G	.	N	.	.	.	.	.	.	.					+			
RARA	4	17	sp	35759712	469+9C>T	.	.	.	0.006	.	.	.	.	.			+					
GATA1	.	X	sp	48534437	.	.	N	.	.	.	.	.	.	.	+							
KDN5C	.	X	sp	53243790	.	.	N	.	.	.	.	.	.	.							+	
MSN	12	X	sp	64875554	1345-3C>T	.	N	.	.	.	.	.	.	.		+						
MED12	30	X	ns	70268766	C4238A	T1413N	N	.	.	0.58	B	65	MC	-2.962		+						

Where a specific variant is present in a sample this is indicated by +
B, benign; C, conservative; MC, moderately conservative; N, novel; ns, nonsynonymous; ns*, potential exonic splice site; sp intronic splice variant

Appendix XIII

Variants identified as disease-causing and catalogued in HGMD

Gene	Exon	Chromosome	Variant	Base pair location in hg18	Nucleotide change	Protein change	Novel or Clinical	Frequency in 1000G	Frequency in NHLBI ESP	SIFT score	PolyPhen2 prediction	Grantham score	Grantham prediction	Exonic splice site score	DE1	DE2	DE3	DE4	DE5	DE6	DE7	DE8	Associated disease from HGMD
RHD	9	1	ns	25521008	G1195A	A399T	.	.	0.009	0.57	B	58	MC	-1.284	+								Reduced expression (weak D)
C7	17	5	sp	41015770	2350+2T>C	.	.	.	0.000	.	.	.	.	.				+						Complement C7 deficiency
PKHD1	36	6	ns	51932767	A5768T	Q1923L	.	.	0.002	0.58	B	113	MR	-1.284	+								Polycystic kidney disease
BBS10	2	12	ns	75265131	G765A	M255I	.	.	0.001	1.00	B	10	C	.				+					Bardet-Biedl syndrome
MEFV	10	16	ns	3233258	G2230T	A744S	Cl	.	0.002	1.00	B	99	MC	.							+		Mediterranean fever, familial
SGSH	8	17	ns	75799216	A1139G	Q380R	.	.	0.000	1.00	B	43	C	.								+	Sanfilippo syndrome A
USP26	1	X	ns	131988178	G1737A	M579I	.	.	0.006	0.34	B	10	C	.						+			Azoospermia / oligozoospermia

Where a specific variant is present in a sample this is indicated by +
B, benign; C, conservative; Cl, clinical; MC, moderately conservative; MR, moderately radical; ns, nonsynonymous; sp intronic splice variant

Appendix XIV

Potential post-zygotic variants in *BRCA1*, *BRCA2* and *TP53*

Gene	Chromosome	Variant type	Base pair location in hg18	Nucleotide change	Novel	Homopolymer region	DE1		DE2		DE3		DE4		DE5		DE6		DE7		DE8	
							Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads	Ref reads	Alt reads
BRCA1	17	intr_ins	41197939	insT	N	ATTTTTTTTTTTTTTTTT	0	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
BRCA1	17	intr_del	41231805	delA	N	TAAAAAAAAAAAAAAAA	.	.	.	.	.	.	.	.	.	.	12	17	.	.	.	.
BRCA2	13	intr_ins	32950658	insA	N	CAAAAAAAAAAAAAA	.	.	.	.	0	6	.	.	.	.	.	.	4	4	.	.
TP53	17	intr_ns	7579653	T > C	N	.	.	.	.	.	.	.	.	.	.	.	47	5	.	.	.	.
TP53	17	intr_ns	7579658	G > T	.	.	.	.	.	.	.	.	.	.	.	.	46	7	.	.	.	.

Alt reads, number of alternate reads; intr_del, intronic deletion; intr_ins, intronic insertion; intr_ns, intronic nonsynonymous; Ref reads, number of reference reads;

Appendix XV

Criteria used to predict effect of amino acid substitution for each prediction algorithm.

SIFT: scores < 0.05 are predicted to affect protein function

PolyPhen2 HumVar: scores ≤ 0.446 considered 'benign'; scores between 0.447 and 0.908 considered 'possibly damaging'; scores ≥ 0.909 considered 'probably damaging'. The HumVar dataset is based on all known human disease-causing mutations from UniProtKB, along with common human nsSNPs (MAF > 1%) without known disease involvement, considered as non-damaging

PolyPhen2 HumDiv: scores ≤ 0.452 considered 'benign'; scores between 0.453 and 0.956 considered 'possibly damaging'; scores ≥ 0.957 considered 'probably damaging'. The HumDiv dataset is based on all damaging alleles in the UniProtKB database known to affect molecular function and cause Mendelian diseases. Differences between human proteins and closely related mammalian homologs were also included and considered to be non-damaging.

LRT: variants are predicted deleterious if they are: (i) from a codon considered to be significantly constrained; (ii) from a site with alignments in at least 10 eutherian mammal species; and (iii) the alternative amino acid is not observed in any other eutherian mammal species. All other variants are classified as neutral or unknown

MutationTaster: variants with scores >0.95 considered damaging

GERP++: scores range from < 0 to 6.17, with higher scores indicating stronger constraint, a score of 6.17 indicates perfect conservation across all sequenced mammals

PhyloP: larger positive scores represent conserved sites while negative scores indicate non-conserved sites

Grantham score: ≤ 50 for conservative amino acid changes; scores between 51 and 100 classified as moderately conservative; scores between 101 and 150 classified as moderately radical; > 150 for radical amino acid changes

PHRED-scaled CADD: higher scores indicate that a variant is more likely to be deleterious

Logit: the conditional probability that a variant is Mendelian disease-causing given prediction scores from 13 programs, including SIFT, PolyPhen2, LRT, MutationTaster, PhyloP, GERP++ and CADD, under a logistic regression model

Appendix XVI

Candidate genes for cleft lip and/or palate phenotypes

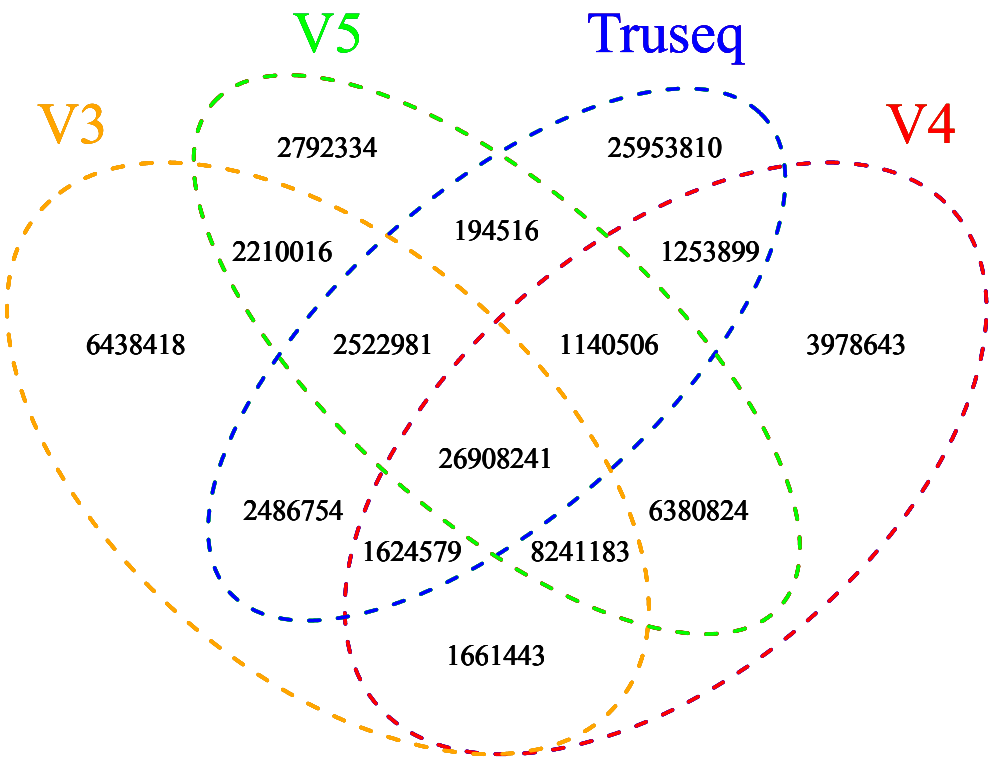
ABCA12	*APC*	*BCS1L*	*CEP89*	*CPT2*	*DLX5*	*EPHA7*	*FIG4*	*GK*
ABCA3	*ARHGAP29*	*BEST1*	*CERS3*	*CREBBP*	*DLX6*	*EPHX1*	*FLNA*	*GLE1*
ABCA4	*ARHGAP31*	*BLM*	*CFDP1*	*CRIPT*	*DNM2*	*ERBB3*	*FLNB*	*GLI2*
ABCB1	*ARHGEF9*	*BMP1*	*CFTR*	*CRISPLD2*	*DNMT3B*	*ERCC1*	*FLRT3*	*GLI3*
ABCC6	*ARID1B*	*BMP2*	*CHD2*	*CRLF1*	*DOCK6*	*ERCC2*	*FLVCR2*	*GMPPB*
ABCD4	*ARL6*	*BMP4*	*CHD6*	*CTAG1A*	*DOK7*	*ERCC5*	*FMN1*	*GNAI3*
ABO	*ARNT*	*BMPER*	*CHD7*	*CTAG1B*	*DPAGT1*	*ERCC6*	*FMR1*	*GNAS*
ACAN	*ARX*	*BMPR1A*	*CHM*	*CTAG2*	*DPM1*	*ERF*	*FOXC1*	*GNAS-AS1*
ACHE	*ASNS*	*BMPR1B*	*CHRNA1*	*CTCF*	*DPM2*	*ESCO2*	*FOXC2*	*GNAT2*
ACOX1	*ASPH*	*BRAF*	*CHRNA7*	*CTDP1*	*DPYD*	*ESR1*	*FOXE1*	*GNPAT*
ACTA1	*ASXL1*	*BRIP1*	*CHRND*	*CUL4B*	*DPYS*	*EVC*	*FOXF1*	*GNPTAB*
ACTB	*ASXL3*	*BUB1B*	*CHRNE*	*CUL7*	*DURS1*	*EVC2*	*FOXF2*	*GNRH1*
ADAMTS10	*ATD*	*C12orf57*	*CHRNG*	*CYP11A1*	*DUSP22*	*EXPH5*	*FOXL2*	*GNRHR*
ADAMTS17	*ATIC*	*C15orf41*	*CHST14*	*CYP17A1*	*DUSP6*	*EYA1*	*FOXP2*	*GOSR2*
ADAMTS2	*ATL1*	*C5orf42*	*CHST3*	*CYP19A1*	*DYNC1H1*	*EZH2*	*FRAS1*	*GPC3*
ADAMTSL2	*ATP6V0A2*	*C6*	*CHSY1*	*CYP1B1*	*DYNC2H1*	*F13A1*	*FREM1*	*GPC6*
ADH1A	*ATP7A*	*CACNA1C*	*CHUK*	*CYP26B1*	*DYRK1A*	*F8*	*FREM2*	*GPR143*
ADNP	*ATPAF2*	*CAMK2G*	*CLAM*	*CYP26C1*	*EARS2*	*FAF1*	*FTO*	*GPSM2*
AFF2	*ATR*	*CANT1*	*CLCF1*	*D2HGDH*	*EBM*	*FAM111A*	*G6PC3*	*GRB10*
AGPS	*ATRNL1*	*CAPN5*	*CLPTM1*	*DAB1*	*EBP*	*FAM111B*	*GAA*	*GREM1*
AHDC1	*ATRX*	*CASK*	*CLPTM1L*	*DCHS1*	*ECE1*	*FAM20C*	*GABRB3*	*GRHL3*
AIC	*AUTS2*	*CASP7*	*CNTNAP2*	*DDR1*	*ECEL1*	*FAM58A*	*GAD1*	*GRIA3*
AKT1	*B3GALT6*	*CAV3*	*COG1*	*DDR2*	*EDA2R*	*FANCA*	*GAS1*	*GRIN1*
AKT3	*B3GALTL*	*CBFB*	*COG7*	*DDX11*	*EDAR*	*FANCE*	*GATA2*	*GRIN2A*
ALDH18A1	*B3GAT3*	*CC2D2A*	*COL11A1*	*DDX59*	*EDN1*	*FAT4*	*GATA3*	*GRIP1*
ALG1	*B4GALT7*	*CCBE1*	*COL11A2*	*DEAF1*	*EEC1*	*FBLN1*	*GBA*	*GSC*
ALG12	*B9D1*	*CCM2*	*COL17A1*	*DHCR24*	*EFEMP2*	*FBN1*	*GCK*	*GSK3B*
ALG6	*BANF1*	*CD96*	*COL1A1*	*DHCR7*	*EFNB1*	*FBXL4*	*GCLC*	*GTF2IRD1*
ALG9	*BBIP1*	*CDAN1*	*COL1A2*	*DHODH*	*EFTUD2*	*FERMT1*	*GDF5*	*GUSB*
ALX1	*BBS1*	*CDC6*	*COL26A1*	*DIH1*	*EIF4A3*	*FGD1*	*GDF6*	*H19*
ALX3	*BBS10*	*CDH1*	*COL2A1*	*DIS3L2*	*ELP4*	*FGF10*	*GGH*	*HAMP*
ALX4	*BBS12*	*CDH15*	*COL6A2*	*DISC1*	*EMG1*	*FGF16*	*GH1*	*HCN4*
AMER1	*BBS2*	*CDH3*	*COL7A1*	*DISP1*	*EMX2*	*FGF17*	*GHRHR*	*HDAC4*
ANK1	*BBS4*	*CDH8*	*COLEC11*	*DKC1*	*ENG*	*FGF8*	*GJA1*	*HDAC6*
ANKH	*BBS5*	*CDKN1C*	*COLQ*	*DLG1*	*EP300*	*FGFR1*	*GJB2*	*HDAC8*
ANKRD11	*BBS7*	*CDON*	*COMP*	*DLG2*	*EPCAM*	*FGFR2*	*GJB3*	*HESX1*
ANTXR1	*BBS9*	*CEP290*	*COMT*	*DLX2*	*EPG5*	*FGFR3*	*GJB4*	*HFE*
ANTXR2	*BCOR*	*CEP57*	*COQ2*	*DLX3*	*EPHA3*	*FGFRL1*	*GJB6*	*HIC1*

HMBS	*IRX5*	*LAMB2*	*MIPOL1*	*NIPBL*	*PDE4D*	*PKLR*	*RAB28*	*RPS27A*
HMX1	*ITCH*	*LHX8*	*MIR140*	*NKX3-2*	*PDE6D*	*PLCB4*	*RAB3GAP1*	*RPS6KA3*
HOXA13	*ITGB3*	*LIPH*	*MIR17HG*	*NNT*	*PDGFC*	*PLEC*	*RAD21*	*RPS7*
HOXA2	*JAG1*	*LMBR1*	*MKKS*	*NOG*	*PDGFRA*	*PLOD1*	*RAF1*	*RREB1*
HOXB1	*JAG2*	*LMNA*	*MKS1*	*NOTCH2*	*PDHA1*	*PLOD3*	*RAI1*	*RUNX2*
HOXB6	*KAL1*	*LMX1B*	*MKX*	*NPAS4*	*PDR*	*PMM2*	*RAPSN*	*RYK*
HOXD1	*KANSL1*	*LRBA*	*MLH1*	*NPHS2*	*PEPD*	*POC1A*	*RARB*	*RYR1*
HOXD13	*KAT6B*	*LRP4*	*MMP2*	*NPR2*	*PEX1*	*POFUT1*	*RB1*	*SALL1*
HPE1	*KCNJ11*	*LRP8*	*MOGS*	*NR0B1*	*PEX10*	*POGLUT1*	*RBBP8*	*SALL4*
HPGD	*KCNJ13*	*LRPPRC*	*MPP7*	*NR2F1*	*PEX12*	*POLE*	*RBFOX1*	*SATB2*
HRAS	*KCNJ2*	*LTBP2*	*MRPS16*	*NRAS*	*PEX13*	*POLG*	*RBM10*	*SC5D*
HS6ST1	*KCNK9*	*LTBP4*	*MRXS11*	*NRN1*	*PEX14*	*POLR1C*	*RBM28*	*SCARF2*
HSD17B4	*KCNQ1OT1*	*LYST*	*MSX1*	*NRXN1*	*PEX16*	*POLR1D*	*RBPJ*	*SCD5*
HSPG2	*KCNV2*	*LZTFL1*	*MSX2*	*NSD1*	*PEX19*	*POMGNT1*	*RDH12*	*SCLT1*
HYAL1	*KCTD1*	*MAB21L2*	*MTHFR*	*NSDHL*	*PEX2*	*POMT1*	*RECQL4*	*SCN1A*
HYLS1	*KDM6A*	*MAFB*	*MTND3P1*	*NSMF*	*PEX26*	*POMT2*	*RELA*	*SCN2A*
ICK	*KIAA0196*	*MALT1*	*MTR*	*OASD*	*PEX3*	*PORCN*	*RELN*	*SCRIB*
IDS	*KIAA1279*	*MAP2*	*MUC5B*	*OCA2*	*PEX5*	*PPOX*	*RFX3*	*SCZD1*
IDUA	*KIF11*	*MAP2K1*	*MYCN*	*OCLN*	*PEX6*	*PQBP1*	*RIEG2*	*SEC23A*
IFT122	*KIF22*	*MAP2K2*	*MYH3*	*OFC1*	*PEX7*	*PRDM16*	*RIN2*	*SEMA3A*
IFT140	*KIF7*	*MAPT*	*MYH8*	*OFCC1*	*PGAP2*	*PREPL*	*RIPK4*	*SEMA3E*
IFT172	*KIRREL3*	*MASP1*	*MYMY1*	*OFD1*	*PGAP3*	*PRICKLE1*	*RIT1*	*SEPN1*
IFT27	*KISS1*	*MBD5*	*NAA10*	*OPHN1*	*PGK1*	*PRKAR1A*	*RMRP*	*SEPT9*
IFT43	*KISS1R*	*MBOAT1*	*NALCN*	*ORC1*	*PGM1*	*PROKR2*	*RNF135*	*SERPINC1*
IFT80	*KIT*	*MBS1*	*NBAS*	*ORC4*	*PHF21A*	*PRPH2*	*RNU4ATAC*	*SERPINH1*
IGBP1	*KITLG*	*MBTPS2*	*NBN*	*OSR2*	*PHF8*	*PRRX1*	*ROGDI*	*SETBP1*
IGF1R	*KL*	*MC2R*	*NDN*	*OTX2*	*PHGDH*	*PRSS12*	*ROR2*	*SETD5*
IHH	*KLHL41*	*MCM4*	*NDUFAF2*	*PACS1*	*PIEZO2*	*PTCH1*	*RPGRIP1L*	*SF3B4*
IKBKG	*KLHL7*	*MCPH1*	*NEB*	*PAFAH1B1*	*PIGA*	*PTCH2*	*RPL11*	*SFTPA1*
IL11RA	*KLK1*	*MECP2*	*NEBL*	*PAK3*	*PIGL*	*PTDSS1*	*RPL15*	*SFTPA2*
IL17RD	*KMT2A*	*MED12*	*NEK1*	*PALB2*	*PIGN*	*PTEN*	*RPL26*	*SFTPB*
IL1RAPL1	*KMT2D*	*MEF2C*	*NF1*	*PAPSS2*	*PIGV*	*PTH1R*	*RPL35A*	*SFTPC*
IL21	*KRAS*	*MEGF10*	*NFATC2*	*PAX1*	*PIK3CA*	*PTHLH*	*RPL36*	*SH3BP2*
IMPAD1	*KRIT1*	*MEGF8*	*NFIA*	*PAX2*	*PIK3R1*	*PTPN11*	*RPL5*	*SH3PXD2B*
INHBA	*KRT1*	*MEIS2*	*NFIX*	*PAX3*	*PIK3R2*	*PUF60*	*RPS10*	*SHANK3*
INPP5E	*KRT10*	*MELK*	*NFKB1*	*PAX6*	*PITX1*	*PUS1*	*RPS15*	*SHFM1*
INPPL1	*KRT14*	*MEN1*	*NFKB2*	*PAX7*	*PITX2*	*PVRL1*	*RPS17*	*SHFM2*
INSR	*KRT5*	*MEOX1*	*NHEJ1*	*PAX9*	*PKD1*	*PVRL4*	*RPS19*	*SHFM5*
IQSEC2	*KRT9*	*MEOX2*	*NHS*	*PCNT*	*PKD2*	*PYCR1*	*RPS24*	*SHH*
IRF6	*L1CAM*	*MID1*	*NIPAL4*	*PCYT1A*	*PKHD1*	*RAB23*	*RPS26*	*SHOX*

SIL1	*SMAD3*	*SOX3*	*STK11*	*TBX5*	*THADA*	*TRIM37*	*VAX1*	*XIST*
SIX3	*SMAD4*	*SOX9*	*STRA6*	*TCF12*	*THAP1*	*TRPS1*	*VPS33B*	*XPC*
SIX6	*SMARCA2*	*SP8*	*STX16*	*TCF4*	*THAS*	*TRPV4*	*VSX1*	*YAP1*
SKI	*SMARCAD1*	*SPECC1L*	*STXBP1*	*TCOF1*	*THRB*	*TSC2*	*WDPCP*	*YPEL1*
SLC12A6	*SMARCB1*	*SPG20*	*SUCLA2*	*TCTN2*	*TINF2*	*TTC21B*	*WDR11*	*YWHAE*
SLC19A1	*SMC1A*	*SPG23*	*SUFU*	*TCTN3*	*TMCO1*	*TTC8*	*WDR19*	*ZBTB16*
SLC1A3	*SMC3*	*SPINT2*	*SUMO1*	*TDGF1*	*TMEM216*	*TTI2*	*WDR34*	*ZBTB24*
SLC25A1	*SMOC1*	*SPRED1*	*SUOX*	*TERC*	*TMEM67*	*TUBB2B*	*WDR35*	*ZC4H2*
SLC26A2	*SMS*	*SPRY2*	*SZT2*	*TFAP2A*	*TNNI2*	*TWIST1*	*WDR60*	*ZEB2*
SLC29A3	*SNAI2*	*SPRY4*	*TACR3*	*TFAP2B*	*TNNT3*	*TWIST2*	*WDR62*	*ZFHX4*
SLC2A10	*SNAP29*	*SPTAN1*	*TALDO1*	*TFDP1*	*TOR1A*	*TYMS*	*WNK1*	*ZFP37*
SLC38A8	*SNIP1*	*SRCAP*	*TBC1D24*	*TFR2*	*TP63*	*UBA1*	*WNT10B*	*ZFP57*
SLC6A8	*SNRPN*	*SRY*	*TBC1D32*	*TGFA*	*TPM1*	*UBB*	*WNT3*	*ZFP90*
SLC7A9	*SNTG1*	*ST3GAL5*	*TBCE*	*TGFB1*	*TPM2*	*UBE3B*	*WNT4*	*ZIC2*
SLC9A6	*SOS1*	*ST5*	*TBX1*	*TGFB3*	*TPM3*	*UFD1L*	*WNT5A*	*ZIC3*
SLCO2A1	*SOST*	*STAC3*	*TBX10*	*TGFBR1*	*TRAF6*	*UGT1A9*	*WNT7A*	*ZMPSTE24*
SLURP1	*SOX10*	*STAMBP*	*TBX15*	*TGFBR2*	*TRAPPC9*	*UROD*	*WNT9B*	*ZNF335*
SLX4	*SOX2*	*STAR*	*TBX22*	*TGIF1*	*TRIM32*	*USB1*	*WT1*	*ZNF81*
SMAD2								

Appendix XVII

Venn diagram of the intersection of the number of bases covered by four exome capture kits.



All exome sequenced samples included in the SKAT-O analysis were sequenced using Truseq, Agilent V3, Agilent V4, or Agilent V5.

Appendix XVIII

Mapping and coverage summary statistics for exome sequencing

Sample	CL001_1	CL001_2	CL002_1	CL002_2	CL003_1	CL004_1	CL004_2	CL005_1	CL006_1
Total no. sequence reads	54,956,672	74,355,880	53,630,678	46,509,220	43,363,414	41,622,776	46,087,918	45,257,626	46,341,632
Total no. aligned reads	54,558,769	73,326,222	53,245,840	46,179,091	43,032,505	41,312,749	45,762,033	44,906,405	45,995,191
Total no. unique alignments	53,763,584	72,878,819	52,373,414	45,401,788	42,300,126	40,618,226	44,987,196	44,151,854	45,230,995
Mapped to target reads +/- 150bp (%)	91.13	95.19	88.21	87.93	87.89	88.20	87.56	85.90	86.47
Mapped to target reads (%)	76.81	84.36	74.84	75.31	75.34	75.50	74.41	72.96	82.91
Target bases with coverage >1 (%)	99.34	99.87	99.32	99.2	99.18	99.16	99.18	99.25	99.27
Target bases with coverage >5 (%)	98.84	99.33	98.69	98.41	98.27	98.2	98.39	98.26	98.40
Target bases with coverage >10 (%)	97.74	98.13	97.14	96.42	95.72	95.48	96.28	95.83	96.28
Target bases with coverage >20 (%)	92.92	93.67	90.68	88.26	85.87	85.04	87.83	86.65	87.80
Mean read depth across exome	68.84	94.55	65.22	57.33	52.62	50.94	56.61	54.92	56.88

CL007_1	CL010_1	CL010_2	CL012_1	CL012_2	CL014_1	CL014_2	CL018_1	CL018_2	CL018_3
49,254,976	100,068,158	99,974,938	88,367,648	66,687,586	95,428,304	97,299,918	102,525,932	92,357,604	102,409,746
48,875,805	99,055,310	98,831,948	87,481,021	66,633,760	94,564,501	96,408,611	101,596,662	91,426,256	101,335,363
48,031,749	97,692,151	97,384,564	86,218,963	62,443,807	93,168,229	95,022,993	100,154,970	90,117,254	99,904,666
84.46	91.54	87.17	88.57	79.90	88.30	89.82	89.49	89.12	89.77
71.99	86.61	81.75	83.31	72.97	84.74	85.70	85.25	84.60	85.52
99.31	99.35	99.38	99.27	98.49	99.25	99.27	99.28	99.36	99.38
98.48	99.06	99.13	98.97	96.76	98.96	98.99	99.02	99.07	99.12
96.36	98.62	98.75	98.52	93.55	98.54	98.60	98.64	98.59	98.73
88.18	96.93	97.38	96.94	83.34	97.14	97.34	97.38	96.88	97.40
59.21	109.24	119.05	108.03	56.79	118.47	123.11	127.87	113.70	127.25

Appendix XIX

Breakdown of variants remaining after each filtering step in three tiers of data analysis

Filtering step	Tier 1 Analysis		Tier 2 Analysis		Tier 3 Analysis	
1	Variants common to affected siblings	11,217	Variants common to all affected samples	6522	Variants in proband	24,351
2	MAF <= 1% in NHLBI ESP	1717	MAF <= 1% in NHLBI ESP	831	MAF <= 1% in NHLBI ESP	4426
3	MAF <= 1% in 1000G	1076	MAF <= 1% in 1000G	465	MAF <= 1% in 1000G	2815
4	Heterozygous	974	Heterozygous	378	Heterozygous	2307
5	Remove synonymous	729	Remove variants in highly mutable genes	180	Remove variants in highly mutable genes	1487
6	Remove splicing, ncRNA splicing and 'unknown'	512	MAF <= 5% CG46	131	MAF <= 5% CG46	1273
7	Remove variants in highly mutable genes	288	Remove variants in Soton database samples	43	Remove variants in Soton database samples	490
8	Remove variants in homopolymer tracts and repeat regions	269	Remove synonymous	30	Remove synonymous	363
9	Not in dbSNP135	113	Remove variants also in 20so00275 and 20so00276	3	Remove variants also in 20so00275 and 20so00276	223
10	Remove variants also in 20so00275 and 20so00276	47	Disease-only	3	Disease-only	12

NHLBI ESP – National Heart, Lung and Blood Institute Exome Sequencing Project Variant Server; 1000G – 1000 Genomes Project; CG46 – Complete Genomics Whole Genomes

Appendix XX

Mapping and coverage statistics for exome sequence data

Sample	20so00275	20so00276	20so00277	20so00278	BP999108	JP999109
Total no. sequence reads	534,863,042	322,518,818	287,329,190	423,028,830	87,454,478	133,140,382
Total no. aligned reads	447,728,771	251,855,188	224,124,216	328,844,385	65,026,970	107,824,035
Total no. unique alignments	397,041,844	220,805,234	197,480,053	292,734,010	59,586,225	98,541,776
Mapped to target reads +/- 150bp (%)	90.75	91.05	90.91	90.49	71.01	68.09
Mapped to target reads (%)	66.04	70.27	65.42	63.69	58.97	54.79
Target bases with coverage >1 (%)	90.36	92.28	91.41	88.50	97.10	98.49
Target bases with coverage >5 (%)	82.13	87.23	85.63	77.95	92.11	95.20
Target bases with coverage >10 (%)	75.42	83.47	81.28	69.69	87.01	92.21
Target bases with coverage >20 (%)	65.54	76.90	73.76	58.63	72.99	85.45
Mean read depth across exome	61.70	78.97	66.43	47.08	43.12	65.94

Appendix XXI

Quality control statistics

Sample	Variants	% Het	% X Het	Gender	SNP Fingerprinting
20so00275	25,306	78.41*	76.81*	Female	N/A
20so00276	22,541	65.19†	65.80†	Female	N/A
20so00277	21,787	65.03†	66.67†	Female	N/A
20so00278	23,599	79.00*	78.33*	Female	N/A
BP999108	24,351	62.28	61.24	Female	Identity confirmed
JP999109	24,799	61.77	19.07	Male	Identity confirmed

Variants – total called variants; % Het – genome wide % of heterozygous calls; % X Het - % heterozygous calls mapped to X chromosome; Gender – apparent gender based upon % X Het; SNP Fingerprinting – individuals that were fingerprinted using the SNP panel to confirm identity.
Samples marked * exhibit levels of heterozygosity that are much higher than we would expect to see; normally this follows a very tight distribution. Values as high as those observed have previously indicated sample contamination.
Samples marked † also exhibit higher than expected levels of heterozygosity, although it is not as extreme. All four samples with high levels of heterozygosity were sequenced using SOLiD. Results from the two samples sequenced with Illumina are as we would expect.

Appendix XXII

Percentage of variants shared between individuals

	20so00275	20so00276	20so00277	20so00278	BP999108	JP999109
20so00275	100.00	48.29	41.35	47.93	45.51	35.03
20so00276	54.21	100.00	47.91	44.81	58.20	39.36
20so00277	48.03	49.57	100.00	51.42	50.84	52.14
20so00278	51.40	42.81	47.47	100.00	41.95	42.76
BP999108	47.29	53.87	45.49	40.66	100.00	42.73
JP999109	35.75	35.77	45.81	40.69	41.96	100.00

Appendix XXIII

Variants identified as unique to the two affected siblings in tier 1 analysis

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	Novel	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	20so00277	20so00278
COPA	1	160302336	ns	G398C	G133A	.	.	N	**0.999**	**1**	**0.998**	**1**	**1.000**	4.61	60	+	+
PLA2G2D	1	20442950	ns	G61A	G21R	.	0.000	.	0.816	0.9	**0.871**	0.846	0.355	0.03	**125**	+	+
PRR21	2	240982285	ns	T115C	W39R	.	.	N	0.040	0.81	**0.676**	0.609	0.000	-2.81	**101**	+	+
DNAH1	3	52431772	ns	T11837C	I3946T	.	0.000	.	.	.	.	.	.	5.28	89	+	+
CLDN16	3	190106074	fsd	166delG	A56fs	.	.	N	.	.	.	.	.		.	+	+
HTT	4	3174102	ns	T3920C	M1307T	.	.	N	**0.997**	0.29	**0.999**	**1.000**	**0.988**	4.29	81	+	+
ZNF732	4	265813	ns	T830A	F277Y	.	.	N	.	.	.	.	.	-1.87	22	+	+
FGA	4	155507560	ns	A1021C	T341P	.	.	N	0.004	0.14	0.002	0.007	0.000	-9.12	38	+	+
ZNF718	4	155719	ns	A1244G	K415R	.	.	N	.	.	.	.	.	.	26	+	+
SPATA4	4	177106011	fsd	835_838del	279_280del	.	.	.	.	.	.	.	.	.	.	+	+
SPDYE1	7	44046934	ns	G700A	G234R	.	.	N	0.799	0.79	0.011	0.985	**0.903**	0.96	**125**	+	+
PKHD1L1	8	110478857	ns	C8464T	H2822Y	.	0.001	.	.	.	.	.	.	6.16	83	+	+
MPDZ	9	13168455	ns	T3164G	L1055W	.	.	N	.	.	.	.	.	5.32	61	+	+
FAM75A3	9	40705431	ns	G3088A	A1030T	.	.	N	.	.	.	.	.	-3.48	58	+	+
SLC46A2	9	115652029	ns	G933C	E311D	.	.	N	0.095	0.43	0.067	0.961	**0.936**	-4.7	45	+	+
FAM75A3	9	40705785	ns	G3442A	V1148I	.	.	N	.	.	.	.	.	-5.08	29	+	+

DUX4L2, DUX4L3, DUX4L6	10	135491068	ns	G679A	A227T	.	.	N	.	.	.	.	.	.	58	+	+
WNK1	12	1005633	ns	C6736A	P2246T	.	.	N	**0.999**	**0.99**	0.568	**1.000**	0.051	5.04	38	+	+
RIMBP2	12	130926735	ns	G1111A	V371I	.	0.000	.	0.889	0.82	0	0.924	0.124	0.09	29	+	+
SELPLG	12	109017374	ns	A710C	E237A	.	.	N	0.023	0.00	0	0.053	0.000	-6.4	**107**	+	+
POSTN	13	38156522	ns	G1373A	R458K	.	.	N	**0.999**	**1.00**	**0.976**	**1.000**	**0.974**	5.64	26	+	+
CPSF2	14	92625491	ns	G1986A	M662I	.	.	N	**0.999**	0.60	0.023	**1.000**	**0.935**	4.65	10	+	+
SAMD4A	14	55169155	ns	A572G	Q191R	.	.	N	**0.999**	0.05	0.023	**1.000**	**0.994**	4.59	43	+	+
CDAN1	15	43022361	ns	C2227G	L743V	.	0.000	.	**0.965**	**0.96**	0.107	**1.000**	**0.526**	2.61	32	+	+
SEC14L5	16	5064931	ns	C2051T	S684F	.	0.000	.	.	.	.	.	.	1.31	**155**	+	+
CLTC	17	57762427	ns	C4445T	T1482I	.	.	N	**0.999**	0.95	**0.963**	**1.000**	**0.973**	5.14	89	+	+
KCNH6	17	61613357	ns	G1429A	V477M	.	0.000	.	**0.999**	**1.00**	**0.991**	**1.000**	**1.000**	3.89	21	+	+
ZNF549	19	58050044	sg	G1672T	E558X	.	0.000	.	0.903	0.90	0.717	0.987	**1.000**	0.87	.	+	+
ZNF30	19	35434672	ns	G805T	A269S	.	.	N	.	.	.	.	.	0.60	99	+	+
ZNF773	19	58018741	ns	C1278A	H426Q	.	.	N	0.944	0.00	0.015	0.855	0.008	0.03	24	+	+
ZNF430	19	21240085	ns	G971A	R324K	.	.	N	0.082	0.00	0.001	0.618	0.001	-1.54	26	+	+
ZNF429	19	21720133	ns	T1278A	N426K	.	.	N	0.016	0.00	0	0.600	0.003	-1.6	94	+	+
ZNF676	19	22363008	ns	G1511C	R504P	.	.	N	0.014	0.67	0	0.574	0.007	-1.6	**103**	+	+

CCDC8	19	46915031	ns	G1037A	G346E	.	.	N	0.030	0.69	0	0.800	0.001	-1.72	98	+	+
ZNF493	19	21607526	ns	A1681G	K561E	.	.	N	0.026	0.00	0	0.602	0.001	-1.87	56	+	+
ZNF773	19	58018717	ns	G1254T	R418S	.	.	N	0.192	0.59	0.146	0.634	0.017	-2.62	**110**	+	+
ZNF439	19	11979292	ns	A1408C	K470Q	.	.	N	0.003	0.00	0	0.596	0.000	-2.89	53	+	+
ZNF586	19	58290663	ns	C581T	A194V	.	.	N	.	.	.	.	.	-3.31	64	+	+
ZNF665	19	53668864	ns	G879T	K293N	.	.	N	.	.	.	.	.	-4.67	94	+	+
KRTAP10-12	21	46117403	ns	C287T	T96I	.	.	N	0.832	0.70	0.340	0.813	0.000	-2.36	89	+	+
KRTAP10-11	21	46067030	ns	C655T	P219S	.	.	N	0.013	0.67	0.001	0.589	0.116	-5.12	74	+	+
MORC4	X	106186131	ns	G1990A	D664N	.	.	N	**0.973**	**1.00**	**0.899**	0.313	0.069	2.91	23	+	+
TFDP3	X	132351861	ns	G427A	A143T	.	0.007	.	.	.	.	.	.	0.23	58	+	+
ACRC	X	70823916	ns	A789T	E263D	.	.	N	0.002	0.58	0.130	0.579	0.000	-1.01	45	+	+
GPR50	X	150349560	nfd	1505_1516del	502_506del	.	.	.	.	.	.	.	.	.	.	+	+
ZFP92	X	152686347	ns	G512A	R171H	.	.	N	.	.	.	.	.	.	29	+	+
ZFP92	X	152686349	ns	A514G	I172V	.	.	N	.	.	.	.	.	.	29	+	+

In silico scores in bold are classified as damaging/deleterious by corresponding program.
+ indicates that the variant is present in a heterozygous state in that individual
fsd, frameshift deletion; N, novel - variant is unique to these two individuals; ns, nonsynonymous; nfd, non-frameshift deletion; sg, stopgain

Appendix XXIV

All variants identified in step 8 of tier 2 filtering

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	BP999108	20so00277	20so00278	20so00275	20so00276
EPHX1	1	226027735	ns	G928A	V310I	rs143164964	0.00	0.000	1.000	0.69	0.137	1.000	0.439	5.11	29	+	+	+		+
GALNT2	1	230401082	ns	T1409G	V470G	.	.	.	0.997	1.00	0.753	1.000	1.000	4.91	109	+	+	+		+
MAGI1	3	65376834	ns	G2399A	S800N	.	.	.	0.999	0.98	0.956	1.000	0.998	5.67	46	+	+	+	+	
FILIP1L	3	99568059	ns	G1741A	A581T	.	.	.	0.985	0.7	0.000	0.991	0.094	4.55	58	+	+	+	+	
DBR1	3	137880785	ns	G1581T	R527S	rs201982495	.	0.000	0.982	1.00	0.992	1.000	0.999	3.21	110	+	+	+	+	
CLSTN2	3	140178464	ns	G1075T	D359Y	.	.	0.000	1.000	0.99	0.966	1.000	1.000	5.06	160	+	+	+	+	
TGFBI	5	135391471	ns	G1513A	V505I	rs201775031	.	0.000	.	.	.	.	.	.	29	+	+	+		+
CDC25C	5	137621421	ns	G1382A	R461Q	rs139145068	0.00	0.005	0.100	0.87	0.022	0.830	0.002	-2.12	43	+	+	+		+
HMGXB3	5	149431421	ns	G3545A	G1182D	rs201643282	.	0.001	.	.	.	.	.	.	94	+	+	+		+
CYFIP2	5	156741400	ns	C1159T	R387C	.	.	.	.	.	.	.	.	.	180	+	+	+		+
DOCK2	5	169127014	sp	1133-4G>T	.	rs201185913	.	0.000	.	.	.	.	.	.	.	+	+	+		+
MAP7	6	136709581	ns	G542A	R181H	rs148680029	0.00	0.000	0.999	1.00	0.781	1.000	0.966	5.71	29	+	+	+	+	
CTTNBP2	7	117365316	sp	4056-5G>T		rs200533770	.	0.000	.	.	.	.	.	.	.	+	+	+	+	

RGS22	8	101059715	ns	C1799T	S600F	.	.	.	0.992	1.00	0.940	0.932	**0.655**	3.47	155	+	+	+	+	
PKHD1L1	8	110478857	ns	C8464T	H2822Y	rs201478206	.	0.001	.	.	.	.	.	.	83	+	+	+		
AKNA	9	117139642	ns	G445C	G149R	rs79864470	.	0.002	0.166	1.00	0.380	0.083	0.014	-1.94	125	+	+	+		
PLAU	10	75673800	ns	C743T	T248M	.	.	0.000	0.981	0.90	0.490	0.989	0.017	3.66	81	+	+	+		+
WASF3	13	27250750	ns	G605A	R202H	.	.	.	1.000	0.99	0.922	1.000	1.000	5.3	29	+	+	+	+	+
KHNYN	14	24902050	ns	A1472G	K491R	rs201372060	.	0.000	0.199	0.91	0.030	0.999	0.473	0.58	26	+	+	+	+	
SLC12A1	15	48537081	ns	G1432A	G478R	.	.	.	.	.	.	.	.	.	125	+	+	+	+	+
PARP16	15	65563344	ns	C241T	R81W	rs146581565	0.00	0.000	0.144	0.93	0.000	0.992	0.005	-4.97	101	+	+	+	+	+
SV2B	15	91795601	ns	T635C	I212T	rs145534909	0.00	0.001	0.998	1.00	0.795	1.000	1.000	5.2	89	+	+	+		
AMFR	16	56438922	ns	G739A	E247K	rs149927445	.	0.000	0.999	0.34	0.486	1.000	0.999	5.71	56	+	+	+	+	
ANKRD11	16	89348713	ns	G4237A	E1413K	rs140373729	.	0.001	0.977	0.62	0.960	0.994	0.664	4.13	56	+	+	+	+	
GEMIN4	17	650826	ns	T457A	S153T	rs201063045	0.00	0.002	.	.	.	.	.	.	58	+	+	+	+	+
MPO	17	56348110	nd	2143_2145del	715_715del	.	.	.	.	.	.	.	.	.	.	+	+	+	+	
CCDC57	17	80121126	ns	G1990A	G664R	.	.	0.000	.	.	.	.	.	.	125	+	+	+		+
CCDC151	19	11545746	ns	A92G	E31G	.	.	.	0.249	1.00	0.053	0.641	0.001	0.60	98	+	+	+		+
ZNF737	19	20728510	ns	C499T	H167Y	rs191587895	0.00	.	.	.	.	.	.	.	83	+	+	+	+	+
CCNE1	19	30312916	ns	G719A	R240H	rs146040933	.	0.000	0.148	0.89	0.001	0.946	0.154	-2.65	29	+	+	+	+	+
EP300	22	41572542	sp	5061+10G>A	.	rs78432056	0.00	0.002	.	.	.	.	.	.	.	+	+	+	+	

In silico scores in bold are classified as damaging/deleterious by corresponding program.
+ indicates that the variant is present in a heterozygous state in that individual
ns, nonsynonymous; nd, non-frameshift deletion; sp, intronic splice variant

Appendix XXV

Reference and alternate allele reads in three affected individuals for G334fs variant in *HNRNPA2B1*

Sample	Total read depth	Total ref. reads	Ref. reads on forward strand	Ref. reads on reverse strand	Average PHRED score	Total alt. reads	Alt. reads on forward strand	Alt. reads on reverse strand	Average PHRED score
BP999108	37	16	8	8	29	16	7	9	31
20so00277	222	116	5	111	48	81	4	77	47
20so00278	190	137	2	135	42	31	2	29	59

Ref. – reference allele; Alt. – alternate allele; +ve strand – reads on the positive/forward strand; -ve strand – reads on the negative/reverse strand; Average ref./alt. read PHRED – average PHRED score for reference/alternate allele reads

Appendix XXVI

Genes related to OPMD, Ptosis, or Ophthalmoplegia

Gene names	Related disorder
ACTA1	Ophthalmoplegia
ALX3	Ptosis
C10orf2	Ophthalmoplegia; Ptosis
C12orf65	Ophthalmoplegia
CHAT	Ptosis
CHRND	Ptosis
COLQ	Ptosis
COMP	OPMD
DNM2	Ptosis
DOK7	Ptosis
EBP	Ptosis
FGFR2	Ptosis
FLNB	Ophthalmoplegia; Ptosis
FOXC2	Ptosis
FOXL2	Ptosis
GATA3	Ptosis
KARS	Ophthalmoplegia
KIF21A	Ptosis
MNGIE	OPMD
MTM1	Ptosis
MYH2	Ophthalmoplegia
MYOD1	OPMD
NIPBL	Ptosis
NPC1	Ophthalmoplegia
NPC2	Ophthalmoplegia
OPA1	Ophthalmoplegia
PABPN1	Ptosis; OPMD
PAX6	Ptosis
PHOX2A	Ophthalmoplegia; Ptosis
POLG	Ophthalmoplegia; Ptosis
POLG2	Ophthalmoplegia
RAPSN	Ptosis
RRM2B	Ophthalmoplegia
RYR1	Ophthalmoplegia
SACS	Ophthalmoplegia
SKIP	OPMD
SLC25A4	Ophthalmoplegia
SPG7	Ptosis
TH	Ptosis
TK2	Ophthalmoplegia
TPM2	Ptosis
TPM3	Ptosis
TWIST1	Ptosis
TYMP	Ophthalmoplegia; Ptosis
UBE3B	Ptosis

Appendix XXVII

Variants identified in the proband (BP999108) in 45 OPMD, ptosis and ophthalmoplegia disease-related genes

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID/Novel	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	BP999108	20so00277	20so00278
Variants not present in any of the unaffected family members																		
OPMD genes:																		
COMP	19	18897440	ns	A1156G	N386D	rs61739916	0.02	0.047	0.989	0.98	0.998	1.000	0.926	3.00	**23**	+		
Ptosis genes:																		
DOK7	4	3495095	ns	G1382A	G461D	rs9684786	0.17	0.184	0.953	1.00	0.979	0.998	0.065	1.73	94	+		
NIPBL	5	37063938	nd	7907_7909del	2636_2637del	N	.	.	.	.	.	.	.	.	.	+		
TH	11	2187855	sp	977+8C>G	.	rs12419447	0.21	0.331	.	.	.	.	.	.	.	+		
SPG7	16	89620328	ns	G2063A	R688Q	rs12960	0.12	0.194	0.999	0.66	0.203	1.000	0.998	4.86	43	+		
Ophthalmoplegia genes:																		
OPA1	3	193336676	ns	C629T†	A210V†	rs34307082	0.02	0.029	0.999	0.84	0.578	0.364	0.981	5.35	64	+	+	+
Variants present in at least one unaffected family member																		
Ptosis Genes:																		
CHRND	2	233396375	sp	1002+9T>C	.	rs3762528	0.08	0.067	.	.	.	.	.	.	.	+		+

NIPBL	5	36985303	ns	A2021G	N674S	rs3822471	0.12	0.125	0.852	0.91	0.028	1.000	0.618	1.18	46	+		+
CHAT	10	50856652	ns	G1027A	V343M	rs4838544	0.97	0.999	0.865	0.66	0.000	1.000	0.000	3.59	21	◇	◇	
UBE3B	12	109937534	ns	G1037A	R346Q	rs7298565	0.52	0.532	0.980	0.53	0.000	1.000	0.000	5.29	43	◇	◇	+
SPG7	16	89597221	sp	987+5A>G	.	rs4785691	0.52	0.445	.	.	.	.	.	.	.	+	+	+
SPG7	16	89613123	ns	A1507G	T503A	rs2292954	0.12	0.194	0.972	0.50	0.001	1.000	0.000	3.67	58	+	+	+
MTM1	X	149826503	sp	1260+3G>A	.	rs222410	0.58	0.499	.	.	.	.	.	.	.	+	+	+
Ophthalmoplegia genes:																		
OPA1	3	193334991	ns	G473A	S158N	rs7624750	0.46	0.464	0.839	0.66	0.352	0.993	0.000	2.77	46	+	+	◇
OPA1	3	193355074	sp	924+4T>C	.	rs166850	0.91	0.826	.	.	.	.	.	.	.	◇	+	◇
NPC1	18	21120444	ns	A2572G	I858V	rs1805082	0.50	0.471	0.999	0.76	0.031	1.000	0.999	5.72	29	+		+
NPC1	18	21124945	ns	G1926C	M642I	rs1788799	0.82	0.645	0.995	0.00	0.000	1.000	0.000	5.56	10	◇	+	◇
NPC1	18	21140432	ns	A644G	H215R	rs1805081	0.24	0.403	0.765	0.41	0.000	0.758	0.075	-0.07	29	+		
Ptosis and ophthalmoplegia genes:																		
FLNB	3	58109162	ns	G3469A	D1157N	rs1131356	0.51	0.244	0.995	0.97	0.996	1.000	1.000	4.32	23	+	+	
FLNB	3	58118555	ns	G4411A	V1471M	rs12632456	0.56	0.245	0.885	1.00	0.003	1.000	0.998	1.64	21	+	+	
FLNB	3	58145348	ns	T6956C	I2319T	rs116826041	0.00	0.013	0.998	0.99	0.561	1.000	0.996	5.39	89	+		

In silico scores in bold are classified as damaging/deleterious by corresponding program.
N, novel – variant is not recorded in 1000 Genomes Project, Exome variant server, dbSNP135, or Soton database
nd, non-frameshift deletion; ns, nonsynonymous; sp, intronic splice variant
+ indicates that the variant is heterozygous
◇ indicates that the variant is homozygous

Appendix XXVIII

Evidence for variants in OPMD, Ptosis and Ophthalmoplegia genes in all family members from manual inspection of reads using IGV

Gene	BP999108	20so00277	20so00278	20so00275	20so00276	JP999109
OPA1	31/59	31/89	12/55	.	.	.
DOK7	14/28	Not covered	Not covered	Not covered	.	.
NIPBL	2/15	.	.	.	.	.
TH	24/57	Not covered	Not covered	Not covered	Not covered	Not covered
SPG7	41/100	1/2	.	1/3	.	2/146
COMP	12/26	.	.	.	.	.

x/y – x is the count of alternate allele reads out of the total count of reads y

Samples with no evidence of the alternate allele denoted by '.'

Not covered – no reads span the variant location

Appendix XXIX

Variants identified in *MYH2* across all six samples

Gene	Chromosome	Base pair location in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID/Novel	MAF in 1000 Genomes Project	MAF in Exome Sequencing Project	PhyloP	1-SIFT	PolyPhen2	LRT	MutationTaster	GERP++	Grantham score	BP999108	20so00277	20so00278	20so00275	20so00276	JP999109
MYH2	17	10427897	sy	G5061T	L1687L	N	.	.	.	.	.	.	.	.	.			+			
MYH2	17	10433266	ns	G2823T	E941D	rs138206136	.	0.001	0.920	0.64	0.0	0.999	0.013	1.72	45			+			+
MYH2	17	10446182	sp	904+10G>A		rs719277	0.51	0.407	.	.	.	.	.	.	.		+			+	+
MYH2	17	10448769	sy	T399A	P133P	rs11078850	0.51	0.408	.	.	.	.	.	.	.		+		+	+	+
MYH2	17	10450816	sy	A324G	E108E	rs12600539	0.53	0.413	.	.	.	.	.	.	.		+	+	+	+	+

In silico scores in bold are classified as damaging/deleterious by corresponding program.

N, novel – variant is not recorded in 1000 Genomes Project, Exome variant server, dbSNP135, or Soton database

ns, nonsynonymous; sp, intronic splice variant; sy, synonymous

+ indicates that the variant is heterozygous

References

2012. Comprehensive molecular portraits of human breast tumours. *Nature,* 490**,** 61-70.

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

Abkevich, V., Zharkikh, A., Deffenbaugh, A. M., Frank, D., Chen, Y., Shattuck, D., Skolnick, M. H., Gutin, A. & Tavtigian, S. V. 2004. Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. *Journal of Medical Genetics,* 41**,** 492-507.

Abu-Baker, A., Messaed, C., Laganiere, J., Gaspar, C., Brais, B. & Rouleau, G. A. 2003. Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. *Human Molecular Genetics,* 12**,** 2609-2623.

Abu-Baker, A. & Rouleau, G. A. 2007. Oculopharyngeal muscular dystrophy: Recent advances in the understanding of the molecular pathogenic mechanisms and treatment strategies. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease,* 1772**,** 173-185.

Abuzzahab, M. J., Schneider, A., Goddard, A., Grigorescu, F., Lautier, C., Keller, E., Kiess, W., Klammt, J., Kratzsch, J., Osgood, D., Pfäffle, R., Raile, K., Seidel, B., Smith, R. J. & Chernausek, S. D. 2003. IGF-I Receptor Mutations Resulting in Intrauterine and Postnatal Growth Retardation. *New England Journal of Medicine,* 349**,** 2211-2222.

Adams, J. C. & Lawler, J. 2011. The Thrombospondins. *Cold Spring Harbor Perspectives in Biology,* 3.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* United States.

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., Morrison, J., Maranian, M., Pooley, K. A., Luben, R., Eccles, D., Evans, D. G., Fletcher, O., Johnson, N., dos Santos Silva, I., Peto, J., Stratton, M. R., Rahman, N., Jacobs, K., Prentice, R., Anderson, G. L., Rajkovic, A., Curb, J. D., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Diver, W. R., Bojesen, S., Nordestgaard, B. G., Flyger, H., Dork, T., Schurmann, P., Hillemanns, P., Karstens, J. H., Bogdanova, N. V., Antonenkova, N. N., Zalutsky, I. V., Bermisheva, M., Fedorova, S., Khusnutdinova, E., Kang, D., Yoo, K.-Y., Noh, D. Y., Ahn, S.-H., Devilee, P., van Asperen, C. J., Tollenaar, R. A. E. M., Seynaeve, C., Garcia-Closas, M., Lissowska, J., Brinton, L., Peplonska, B., Nevanlinna, H., Heikkinen, T., Aittomaki, K., Blomqvist, C., Hopper, J. L., Southey, M. C., Smith, L., Spurdle, A. B., Schmidt, M. K., Broeks, A., van Hien, R. R., Cornelissen, S., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Schmutzler, R. K., Burwinkel, B., Bartram, C. R., Meindl, A., Brauch, H., Justenhoven, C., Hamann, U., Chang-Claude, J., Hein, R., Wang-Gohrke, S., Lindblom, A., Margolin, S., Mannermaa, A., Kosma, V.-M., Kataja, V., Olson, J. E., Wang, X., Fredericksen, Z., Giles, G. G., Severi, G., Baglietto, L., English, D. R., Hankinson, S. E., Cox, D. G., Kraft, P., Vatten, L. J., Hveem, K., Kumle, M., et al. 2009. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet,* 41**,** 585-590.

Ahsan, H., Halpern, J., Kibriya, M. G., Pierce, B. L., Tong, L., Gamazon, E., McGuire, V., Felberg, A., Shi, J., Jasmine, F., Roy, S., Brutus, R., Argos, M., Melkonian, S., Chang-Claude, J., Andrulis, I., Hopper, J. L., John, E. M., Malone, K., Ursin, G., Gammon, M. D., Thomas, D. C., Seminara, D., Casey, G., Knight, J. A., Southey, M. C., Giles, G. G., Santella, R. M., Lee, E., Conti, D., Duggan, D., Gallinger, S., Haile, R., Jenkins, M., Lindor, N. M., Newcomb, P., Michailidou, K., Apicella, C., Park, D. J., Peto, J., Fletcher, O., dos Santos Silva, I., Lathrop, M., Hunter, D. J., Chanock, S. J., Meindl, A., Schmutzler, R. K., Müller-Myhsok, B., Lochmann, M., Beckmann, L., Hein, R., Makalic, E., Schmidt, D. F., Bui, Q. M., Stone, J., Flesch-Janys, D., Dahmen, N., Nevanlinna, H., Aittomäki, K., Blomqvist, C., Hall, P., Czene, K., Irwanto, A., Liu, J., Rahman, N., Turnbull, C., for the Familial Breast Cancer, S., Dunning, A. M., Pharoah, P., Waisfisz, Q., Meijers-Heijboer, H., Uitterlinden, A. G., Rivadeneira, F., Nicolae, D., Easton, D. F., Cox, N. J. & Whittemore, A. S. 2014. A Genome-wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age. *Cancer Epidemiology Biomarkers & Prevention*.

Allred, D. C., Brown, P. & Medina, D. 2004. The origins of estrogen receptor alpha-positive and estrogen receptor alpha-negative human breast cancer. *Breast Cancer Research,* 6**,** 240-257.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. & Müller-Myhsok, B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics,* 131**,** 1541-1554.

Altshuler, D., Daly, M. J. & Lander, E. S. 2008. Genetic mapping in human disease. *Science,* 322**,** 881-8.

Anders, C. K., Hsu, D. S., Broadwater, G., Acharya, C. R., Foekens, J. A., Zhang, Y., Wang, Y., Marcom, P. K., Marks, J. R., Febbo, P. G., Nevins, J. R., Potti, A. & Blackwell, K. L. 2008. Young Age at Diagnosis Correlates With Worse Prognosis and Defines a Subset of Breast Cancers With Shared Patterns of Gene Expression. *Journal of Clinical Oncology,* 26**,** 3324-3330.

Anderson, W., Chatterjee, N., Ershler, W. & Brawley, O. 2002. Estrogen Receptor Breast Cancer Phenotypes in the Surveillance, Epidemiology, and End Results Database. *Breast Cancer Research and Treatment,* 76**,** 27-36.

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tulinius, H., Thorlacius, S., Eerola, H., Nevanlinna, H., Syrjäkoski, K., Kallioniemi, O. P., Thompson, D., Evans, C., Peto, J., Lalloo, F., Evans, D. G. & Easton, D. F. 2003. Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *The American Journal of Human Genetics,* 72**,** 1117-1130.

Antoniou, A. C. & Easton, D. F. 2006. Models of genetic susceptibility to breast cancer. *Oncogene,* 25**,** 5898-5905.

Antoniou, A. C., Wang, X., Fredericksen, Z. S., McGuffog, L., Tarrell, R., Sinilnikova, O. M., Healey, S., Morrison, J., Kartsonaki, C., Lesnick, T., Ghoussaini, M., Barrowdale, D., Peock, S., Cook, M., Oliver, C., Frost, D., Eccles, D., Evans, D. G., Eeles, R., Izatt, L., Chu, C., Douglas, F., Paterson, J., Stoppa-Lyonnet, D., Houdayer, C., Mazoyer, S., Giraud, S.,

Lasset, C., Remenieras, A., Caron, O., Hardouin, A., Berthet, P., Hogervorst, F. B., Rookus, M. A., Jager, A., van den Ouweland, A., Hoogerbrugge, N., van der Luijt, R. B., Meijers-Heijboer, H., Gomez Garcia, E. B., Devilee, P., Vreeswijk, M. P., Lubinski, J., Jakubowska, A., Gronwald, J., Huzarski, T., Byrski, T., Gorski, B., Cybulski, C., Spurdle, A. B., Holland, H., Goldgar, D. E., John, E. M., Hopper, J. L., Southey, M., Buys, S. S., Daly, M. B., Terry, M. B., Schmutzler, R. K., Wappenschmidt, B., Engel, C., Meindl, A., Preisler-Adams, S., Arnold, N., Niederacher, D., Sutter, C., Domchek, S. M., Nathanson, K. L., Rebbeck, T., Blum, J. L., Piedmonte, M., Rodriguez, G. C., Wakeley, K., Boggess, J. F., Basil, J., Blank, S. V., Friedman, E., Kaufman, B., Laitman, Y., Milgrom, R., Andrulis, I. L., Glendon, G., Ozcelik, H., Kirchhoff, T., Vijai, J., Gaudet, M. M., Altshuler, D., Guiducci, C., Loman, N., Harbst, K., Rantala, J., Ehrencrona, H., Gerdes, A. M., Thomassen, M., Sunde, L., Peterlongo, P., Manoukian, S., Bonanni, B., Viel, A., Radice, P., et al. 2010. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet,* 42**,** 885-92.

Anvar, S. Y., Raz, Y., Verway, N., van der Sluijs, B., Venema, A., Goeman, J. J., Vissing, J., van der Maarel, S. M., t Hoen, P. A., van Engelen, B. G. & Raz, V. 2013. A decline in PABPN1 induces progressive muscle weakness in oculopharyngeal muscle dystrophy and in muscle aging. *Aging (Albany NY),* 5**,** 412-26.

Apponi, L. H., Leung, S. W., Williams, K. R., Valentini, S. R., Corbett, A. H. & Pavlath, G. K. 2010. Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. *Human Molecular Genetics,* 19**,** 1058-1065.

Aradhya, S., Bardaro, T., Galgoczy, P., Yamagata, T., Esposito, T., Patlan, H., Ciccodicola, A., Munnich, A., Kenwrick, S., Platzer, M., D'Urso, M. & Nelson, D. L. 2001. Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum Mol Genet,* 10**,** 2557-67.

Aschard, H., Chen, J., Cornelis, M. C., Chibnik, L. B., Karlson, E. W. & Kraft, P. 2012. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am J Hum Genet,* 90**,** 962-72.

Balmain, A., Gray, J. & Ponder, B. 2003. The genetics and genomics of cancer. *Nat Genet,* 33**,** 238-244.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. & Shendure, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* 12**,** 745-755.

Ban, H.-J., Heo, J. Y., Oh, K.-S. & Park, K.-J. 2010. Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC Genetics,* 11**,** 26.

Banerjee, A., Apponi, L. H., Pavlath, G. K. & Corbett, A. H. 2013. PABPN1: molecular function and muscle disease. *FEBS Journal,* 280**,** 4230-4250.

Baron, M. 2001. The search for complex disease genes: fault by linkage or fault by association? *Molecular psychiatry,* 6**,** 143-149.

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., Liang, K. Y., Wu, T., Murray, T. & Fallin, M. D. 2010. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature genetics,* 42**,** 525-529.

Ben-Hur, A. & Weston, J. 2010. A User's Guide to Support Vector Machines. *In:* Carugo, O. & Eisenhaber, F. (eds.) *Data Mining Techniques for the Life Sciences.* Humana Press.

Benatar, M., Wuu, J., Fernandez, C., Weihl, C. C., Katzen, H., Steele, J., Oskarsson, B. & Taylor, J. P. 2013. Motor neuron involvement in multisystem proteinopathy: Implications for ALS. *Neurology,* 80**,** 1874-1880.

Berliner, J. L. & Fay, A. M. 2007. Risk assessment and genetic counseling for hereditary breast and ovarian cancer: recommendations of the National Society of Genetic Counselors. *J Genet Couns,* 16**,** 241-60.

Bernier, F. P., Caluseriu, O., Ng, S., Schwartzentruber, J., Buckingham, K. J., Innes, A. M., Jabs, E. W., Innis, J. W., Schuette, J. L., Gorski, J. L., Byers, P. H., Andelfinger, G., Siu, V., Lauzon, J., Fernandez, B. A., McMillin, M., Scott, R. H., Racher, H., Majewski, J., Nickerson, D. A., Shendure, J., Bamshad, M. J. & Parboosingh, J. S. 2012. Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome. *Am J Hum Genet,* 90**,** 925-33.

Bharat, A., Aft, R. L., Gao, F. & Margenthaler, J. A. 2009. Patient and tumor characteristics associated with increased mortality in young women (≤40 years) with breast cancer. *Journal of Surgical Oncology,* 100**,** 248-251.

Bi, J., Bennett, K., Embrechts, M., Breneman, C. & Song, M. 2003. Dimensionality reduction via sparse support vector machines. *The Journal of Machine Learning Research,* 3**,** 1229-1243.

Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., Baluardo, C., Ferrian, M., Almeida de Assis, N., Alblas, M. A., Barth, S., Freudenberg, J., Lauster, C., Schmidt, G., Scheer, M., Braumann, B., Berge, S. J., Reich, R. H., Schiefke, F., Hemprich, A., Potzsch, S., Steegers-Theunissen, R. P., Potzsch, B., Moebus, S., Horsthemke, B., Kramer, F.-J., Wienker, T. F., Mossey, P. A., Propping, P., Cichon, S., Hoffmann, P., Knapp, M., Nothen, M. M. & Mangold, E. 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet,* 41**,** 473-477.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447**,** 799-816.

Blanton, S. H., Burt, A., Garcia, E., Mulliken, J. B., Stal, S. & Hecht, J. T. 2010. Ethnic Heterogeneity of IRF6 AP-2a Binding Site Promoter SNP Association With Nonsyndromic Cleft Lip and Palate. *The Cleft Palate-Craniofacial Journal,* 47**,** 574-577.

Blanton, S. H., Cortez, A., Stal, S., Mulliken, J. B., Finnell, R. H. & Hecht, J. T. 2005. Variation in IRF6 contributes to nonsyndromic cleft lip and palate. *American Journal of Medical Genetics Part A,* 137A**,** 259-262.

Blumen, S. C., Korczyn, A. D., Lavoie, H., Medynski, S., Chapman, J., Asherov, A., Nisipeanu, P., Inzelberg, R., Carasso, R. L., Bouchard, J. P., Tomé, F. M. S., Rouleau, G. A. & Brais, B. 2000. Oculopharyngeal MD among Bukhara Jews is due to a founder (GCG)9 mutation in the PABP2 gene. *Neurology,* 55**,** 1267-1270.

Boardman, L. A., Thibodeau, S. N., Schaid, D. J., Lindor, N. M., McDonnell, S. K., Burgart, L. J., Ahlquist, D. A., Podratz, K. C., Pittelkow, M. & Hartmann, L. C. 1998. Increased Risk for Cancer in Patients with the Peutz-Jeghers Syndrome. *Annals of Internal Medicine,* 128**,** 896-899.

Bodmer, W. & Bonilla, C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet,* 40**,** 695-701.

Bodmer, W. & Tomlinson, I. 2010. Rare genetic variants and the risk of cancer. *Current Opinion in Genetics & Development,* 20**,** 262-267.

Botstein, D. & Risch, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G. & Bustamante, C. D. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet,* 4**,** e1000083.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition,* 30**,** 1145-1159.

Brais, B., Bouchard, J.-P., Xie, Y.-G., Rochefort, D. L., Chretien, N., Tome, F. M. S., Lafrentere, R. G., Rommens, J. M., Uyama, E., Nohira, O., Blumen, S., Korcyn, A. D., Heutink, P., Mathieu, J., Duranceau, A., Codere, F., Fardeau, M. & Rouleau, G. A. 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet,* 18**,** 164-167.

Brais, B., Xie, Y.-G., Sanson, M., Morgan, K., Weissenbach, J., Korczyn, A. D., Blumen, S. C., Fardeau, M., Tomé, F. M. S., Bouchard, J.-P. & Rouleau, G. A. 1995. The oculopharyngeal muscular dystrophy locus maps to the region of the cardiac α and β myosin heavy chain genes on chromosome 14q11.2–q13. *Human Molecular Genetics,* 4**,** 429-434.

Bray, F., Ren, J.-S., Masuyer, E. & Ferlay, J. 2013. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer,* 132**,** 1133-1145.

Breiman, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

Bull, L. 2000. Genetics, Mutations, and Polymorphisms. In: Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. & Van Eerdewegh, P. 2005. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology,* 28**,** 171-182.

Burges, C. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2**,** 121-167.

Burgess, D. J. 2013. Tumour evolution: Weighed down by passengers? *Nat Rev Cancer,* 13**,** 219-219.
Bush, W. S. & Moore, J. H. 2012. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol,* 8**,** e1002822.
Calabrò, A., Beissbarth, T., Kuner, R., Stojanov, M., Benner, A., Asslaber, M., Ploner, F., Zatloukal, K., Samonigg, H., Poustka, A. & Sültmann, H. 2009. Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Research and Treatment,* 116**,** 69-77.
Calado, A., Tomé, F. M. S., Brais, B., Rouleau, G. A., Kühn, U., Wahle, E. & Carmo-Fonseca, M. 2000. Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. *Human Molecular Genetics,* 9**,** 2321-2328.
Camargo, M., Rivera, D., Moreno, L., Lidral, A. C., Harper, U., Jones, M., Solomon, B. D., Roessler, E., Vélez, J. I., Martinez, A. F., Chandrasekharappa, S. C. & Arcos-Burgos, M. 2012. GWAS reveals new recessive loci associated with non-syndromic facial clefting. *European Journal of Medical Genetics,* 55**,** 510-514.
Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. & Lander, E. S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet,* 22**,** 231-8.
Cattaert, T., Calle, M. L., Dudek, S. M., Mahachie John, J. M., Van Lishout, F., Urrea, V., Ritchie, M. D. & Van Steen, K. 2011. Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case–control data in the presence of noise. *Annals of Human Genetics,* 75**,** 78-89.
Ch'ng, E. S., Tuan Sharif, S. E. & Jaafar, H. 2013. In human invasive breast ductal carcinoma, tumor stromal macrophages and tumor nest macrophages have distinct relationships with clinicopathological parameters and tumor angiogenesis. *Virchows Arch,* 462**,** 257-67.
Chamary, J. V. & Hurst, L. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology,* 6**,** R75.
Chartier, A., Benoit, B. & Simonelig, M. 2006. A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. *The EMBO Journal,* 25**,** 2253-2262.
Chen, R., Alvero, A. B., Silasi, D.-A. & Mor, G. 2007. Inflammation, Cancer and Chemoresistance: Taking Advantage of the Toll-Like Receptor Signaling Pathway. *American Journal of Reproductive Immunology,* 57**,** 93-107.
Chen, S. H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B. L., Zheng, S. L., Gronberg, H., Xu, J. & Hsu, F. C. 2008. A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol,* 32**,** 152-67.
Christensen, K. & Murray, J. C. 2007. What genome-wide association studies can do for medicine. *N Engl J Med,* 356**,** 1094-1097.
Chun, S. & Fay, J. C. 2009. Identification of deleterious mutations within three human genomes. *Genome Res,* 19**,** 1553-61.
Cirulli, E. T. & Goldstein, D. B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet,* 11**,** 415-425.
Clarke, A. J. & Cooper, D. N. 2010. GWAS: heritability missing in action[quest]. *Eur J Hum Genet,* 18**,** 859-861.

Cobourne, M. T. 2004. The complex genetics of cleft lip and palate. *The European Journal of Orthodontics,* 26**,** 7-16.

Collins, A., Arias, L., Pengelly, R., Martínez, J., Briceño, I. & Ennis, S. 2014. The potential for next-generation sequencing to characterise the genetic variation underlying non-syndromic cleft lip and palate phenotypes.

Cooper, D., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics,* 132**,** 1077-1130.

Corbeil-Girard, L.-P., Klein, A. F., Sasseville, A. M.-J., Lavoie, H., Dicaire, M.-J., Saint-Denis, A., Pagé, M., Duranceau, A., Codère, F., Bouchard, J.-P., Karpati, G., Rouleau, G. A., Massie, B., Langelier, Y. & Brais, B. 2005. PABPN1 overexpression leads to upregulation of genes encoding nuclear proteins that are sequestered in oculopharyngeal muscular dystrophy nuclear inclusions. *Neurobiology of Disease,* 18**,** 551-567.

Cordell, H. J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics,* 11**,** 2463-2468.

Cordell, H. J. 2009. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet,* 10**,** 392-404.

Cortes, C. & Vapnik, V. 1995. Support-Vector Networks. *Machine Learning,* 20**,** 273-297.

Crawford, S. E., Stellmach, V., Murphy-Ullrich, J. E., Ribeiro, S. M. F., Lawler, J., Hynes, R. O., Boivin, G. P. & Bouck, N. 1998. Thrombospondin-1 Is a Major Activator of TGF-β1 In Vivo. *Cell,* 93**,** 1159-1170.

Cruz, J. A. & Wishart, D. S. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform,* 2**,** 59-77.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A.-L., Brenton, J. D., Tavare, S., Caldas, C. & Aparicio, S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* 486**,** 346-352.

Davies, J. E., Sarkar, S. & Rubinsztein, D. C. 2008. Wild-type PABPN1 is anti-apoptotic and reduces toxicity of the oculopharyngeal muscular dystrophy mutation. *Human Molecular Genetics,* 17**,** 1097-1108.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. & Batzoglou, S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol,* 6**,** e1001025.

de Jong, M. M., Nolte, I. M., te Meerman, G. J., van der Graaf, W. T. A., Oosterwijk, J. C., Kleibeuker, J. H., Schaapveld, M. & de Vries, E. G. E. 2002. Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. *Journal of Medical Genetics,* 39**,** 225-242.

Dear, P. H. 2001. Genome Mapping.

Delous, M., Baala, L., Salomon, R., Laclef, C., Vierkotten, J., Tory, K., Golzio, C., Lacoste, T., Besse, L., Ozilou, C., Moutkine, I., Hellman, N. E., Anselme, I., Silbermann, F., Vesque, C., Gerhardt, C., Rattenberry, E., Wolf, M. T., Gubler, M. C., Martinovic, J., Encha-Razavi, F., Boddaert, N., Gonzales, M., Macher, M. A., Nivet, H., Champion, G., Bertheleme, J. P., Niaudet, P., McDonald, F., Hildebrandt, F., Johnson, C. A., Vekemans, M., Antignac, C., Ruther, U., Schneider-Maunoury, S., Attie-Bitach, T. & Saunier, S.

2007. The ciliary gene RPGRIP1L is mutated in cerebello-oculo-renal syndrome (Joubert syndrome type B) and Meckel syndrome. *Nat Genet,* 39**,** 875-81.

DeNardo, D. G. & Coussens, L. M. 2007. Inflammation and breast cancer. Balancing immune response: crosstalk between adaptive and innate immune cells during breast cancer progression. *Breast Cancer Res,* 9**,** 212.

Deroo, B. J. & Korach, K. S. 2006. Estrogen receptors and human disease. *The Journal of Clinical Investigation,* 116**,** 561-570.

Dixon, M. J., Marazita, M. L., Beaty, T. H. & Murray, J. C. 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet,* 12**,** 167-178.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R. & Gingeras, T. R. 2012. Landscape of transcription in human cells. *Nature,* 489**,** 101-108.

Easton, D. F., Deffenbaugh, A. M., Pruss, D., Frye, C., Wenstrup, R. J., Allen-Brady, K., Tavtigian, S. V., Monteiro, A. N. A., Iversen, E. S., Couch, F. J. & Goldgar, D. E. 2007a. A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer–Predisposition Genes. *The American Journal of Human Genetics,* 81**,** 873-883.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C.-Y., Wu, P.-E., Wang, H.-C., Eccles, D., Evans, D. G., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K.-Y., Noh, D.-Y., Ahn, S.-H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y.-L., Bogdanova, N., Schurmann, P., Dork, T., Tollenaar, R. A. E. M., Jacobi, C. E., Devilee, P., Klijn, J. G. M., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., MacPherson, G., Reed, M. W. R., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCredie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko,

Y.-D., Spurdle, A. B., Beesley, J., Chen, X., Mannermaa, A., Kosma, V.-M., Kataja, V., Hartikainen, J., Day, N. E., et al. 2007b. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature,* 447**,** 1087-1093.

Eccles, D., Gerty, S., Simmonds, P., Hammond, V., Ennis, S., Altman, D. & the, P. s. g. 2007. Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH): study protocol. *BMC Cancer,* 7**,** 160.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet,* 11**,** 446-50.

ENCODE Project Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Esseghir, S., Kennedy, A., Seedhar, P., Nerurkar, A., Poulsom, R., Reis-Filho, J. S. & Isacke, C. M. 2007. Identification of NTN4, TRA1, and STC2 as prognostic markers in breast cancer in a screen for signal sequence encoding proteins. *Clin Cancer Res,* 13**,** 3164-73.

Etienne-Manneville, S. & Hall, A. 2002. Rho GTPases in cell biology. *Nature,* 420**,** 629-635.

Fabian, C. J. & Kimler, B. F. 2005. Selective Estrogen-Receptor Modulators for Primary Prevention of Breast Cancer. *Journal of Clinical Oncology,* 23**,** 1644-1655.

Fan, X., Dion, P., Laganiere, J., Brais, B. & Rouleau, G. A. 2001. Oligomerization of polyalanine expanded PABPN1 facilitates nuclear protein aggregation that is associated with cell death. *Human Molecular Genetics,* 10**,** 2341-2351.

Fan, X., Messaed, C., Dion, P., Laganiere, J., Brais, B., Karpati, G. & Rouleau, G. A. 2003. HnRNP A1 and A/B interaction with PABPN1 in oculopharyngeal muscular dystrophy. *The Canadian Journal of Neurological Sciences,* 30**,** 244-251.

Fan, X. & Rouleau, G. A. 2003. Progress in understanding the pathogenesis of oculopharyngeal muscular dystrophy. *The Canadian Journal of Neurological Sciences,* 30**,** 8-14.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

Feely, S. M. E., Laura, M., Siskind, C. E., Sottile, S., Davis, M., Gibbons, V. S., Reilly, M. M. & Shy, M. E. 2011. MFN2 mutations cause severe phenotypes in most patients with CMT2A. *Neurology,* 76**,** 1690-1696.

FitzGerald, M. G., Marsh Dj Fau - Wahrer, D., Wahrer D Fau - Bell, D., Bell D Fau - Caron, S., Caron S Fau - Shannon, K. E., Shannon Ke Fau - Ishioka, C., Ishioka C Fau - Isselbacher, K. J., Isselbacher Kj Fau - Garber, J. E., Garber Je Fau - Eng, C., Eng C Fau - Haber, D. A. & Haber, D. A. 1998. Germline mutations in PTEN are an infrequent cause of genetic predisposition to breast cancer.

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C., Coupland, B., Broderick, P., Schoemaker, M., Jones, M., Williamson, J., Chilcott-Burns, S., Tomczyk, K., Simpson, G., Jacobs, K. B., Chanock, S. J., Hunter, D. J., Tomlinson, I. P., Swerdlow, A., Ashworth, A., Ross, G., dos Santos Silva, I., Lathrop, M., Houlston, R. S. & Peto, J. 2011. Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. *Journal of the National Cancer Institute,* 103**,** 425-435.

Fontoura, C., Silva, R. M., Granjeiro, J. M. & Letra, A. 2012. Further evidence of association of the ABCA4 gene with cleft lip/palate. *European Journal of Oral Sciences,* 120**,** 553-557.

Forsberg, L. A., Absher, D. & Dumanski, J. P. 2012. Non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. *Journal of Medical Genetics*.

Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. 2009. Human genetic variation and its contribution to complex traits. *Nat Rev Genet,* 10**,** 241-51.

Fried, K., Arlozorov, A. & Spira, R. 1975. Autosomal recessive oculopharyngeal muscular dystrophy. *Journal of Medical Genetics,* 12**,** 416-418.

Fuentes Fajardo, K. V., Adams, D., Program, N. C. S., Mason, C. E., Sincan, M., Tifft, C., Toro, C., Boerkoel, C. F., Gahl, W. & Markello, T. 2012. Detecting false-positive signals in exome sequencing. *Human Mutation,* 33**,** 609-613.

Fusco, F., Paciolla, M., Napolitano, F., Pescatore, A., D'Addario, I., Bal, E., Lioi, M. B., Smahi, A., Miano, M. G. & Ursini, M. V. 2012. Genomic architecture at the Incontinentia Pigmenti locus favours de novo pathological alleles through different mechanisms. *Hum Mol Genet,* 21**,** 1260-71.

Garcia-Closas, M., Couch, F. J., Lindstrom, S., Michailidou, K., Schmidt, M. K., Brook, M. N., Orr, N., Rhie, S. K., Riboli, E., Feigelson, H. S., Le Marchand, L., Buring, J. E., Eccles, D., Miron, P., Fasching, P. A., Brauch, H., Chang-Claude, J., Carpenter, J., Godwin, A. K., Nevanlinna, H., Giles, G. G., Cox, A., Hopper, J. L., Bolla, M. K., Wang, Q., Dennis, J., Dicks, E., Howat, W. J., Schoof, N., Bojesen, S. E., Lambrechts, D., Broeks, A., Andrulis, I. L., Guenel, P., Burwinkel, B., Sawyer, E. J., Hollestelle, A., Fletcher, O., Winqvist, R., Brenner, H., Mannermaa, A., Hamann, U., Meindl, A., Lindblom, A., Zheng, W., Devillee, P., Goldberg, M. S., Lubinski, J., Kristensen, V., Swerdlow, A., Anton-Culver, H., Dork, T., Muir, K., Matsuo, K., Wu, A. H., Radice, P., Teo, S. H., Shu, X.-O., Blot, W., Kang, D., Hartman, M., Sangrajrang, S., Shen, C.-Y., Southey, M. C., Park, D. J., Hammet, F., Stone, J., Veer, L. J. V. t., Rutgers, E. J., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Peto, J., Schrauder, M. G., Ekici, A. B., Beckmann, M. W., dos Santos Silva, I., Johnson, N., Warren, H., Tomlinson, I., Kerin, M. J., Miller, N., Marme, F., Schneeweiss, A., Sohn, C., Truong, T., Laurent-Puig, P., Kerbrat, P., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Milne, R. L., Perez, J. I. A., Menendez, P., Muller, H., Arndt, V., Stegmaier, C., Lichtner, P., Lochmann, M., Justenhoven, C., et al. 2013. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet,* 45**,** 392-398.

Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K. & Arias, J. I. 2008. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS genetics,* 4**,** e1000054.

Gasco, M., Shami, S. & Crook, T. 2002. The p53 pathway in breast cancer. *Breast Cancer Res,* 4**,** 70-76.

Gautheron, J., Pescatore, A., Fusco, F., Esposito, E., Yamaoka, S., Agou, F., Ursini, M. V. & Courtois, G. 2010. Identification of a new NEMO/TRAF6 interface affected in incontinentia pigmenti pathology. *Hum Mol Genet,* 19**,** 3138-49.

Ghassibe, M., Bayet, B., Revencu, N., Verellen-Dumoulin, C., Gillerot, Y., Vanwijck, R. & Vikkula, M. 2005. Interferon regulatory factor-6: a gene

predisposing to isolated cleft lip with or without cleft palate in the Belgian population. *Eur J Hum Genet,* 13**,** 1239-1242.

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., Dennis, J., Wang, Q., Humphreys, M. K., Luccarini, C., Baynes, C., Conroy, D., Maranian, M., Ahmed, S., Driver, K., Johnson, N., Orr, N., dos Santos Silva, I., Waisfisz, Q., Meijers-Heijboer, H., Uitterlinden, A. G., Rivadeneira, F., Hall, P., Czene, K., Irwanto, A., Liu, J., Nevanlinna, H., Aittomaki, K., Blomqvist, C., Meindl, A., Schmutzler, R. K., Muller-Myhsok, B., Lichtner, P., Chang-Claude, J., Hein, R., Nickels, S., Flesch-Janys, D., Tsimiklis, H., Makalic, E., Schmidt, D., Bui, M., Hopper, J. L., Apicella, C., Park, D. J., Southey, M., Hunter, D. J., Chanock, S. J., Broeks, A., Verhoef, S., Hogervorst, F. B. L., Fasching, P. A., Lux, M. P., Beckmann, M. W., Ekici, A. B., Sawyer, E., Tomlinson, I., Kerin, M., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Guenel, P., Truong, T., Cordina-Duverger, E., Menegaux, F., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Milne, R. L., Alonso, M. R., Gonzalez-Neira, A., Benitez, J., Anton-Culver, H., Ziogas, A., Bernstein, L., Dur, C. C., Brenner, H., Muller, H., Arndt, V., Stegmaier, C., Justenhoven, C., Brauch, H., Bruning, T., Wang-Gohrke, S., Eilber, U., Dork, T., Schurmann, P., Bremer, M., Hillemanns, P., Bogdanova, N. V., Antonenkova, N. N., Rogov, Y. I., Karstens, J. H., Bermisheva, M., Prokofieva, D., Khusnutdinova, E., Lindblom, A., Margolin, S., Mannermaa, A., et al. 2012. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet,* 44**,** 312-318.

Ghoussaini, M., Pharoah, P. D. P. & Easton, D. F. 2013. Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning? *The American Journal of Pathology,* 183**,** 1038-1051.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B. & Shen, Y. 2003. The international HapMap project. *Nature,* 426**,** 789-796.

Gilissen, C., Arts, H. H., Hoischen, A., Spruijt, L., Mans, D. A., Arts, P., van Lier, B., Steehouwer, M., van Reeuwijk, J., Kant, S. G., Roepman, R., Knoers, N. V. A. M., Veltman, J. A. & Brunner, H. G. 2010. Exome Sequencing Identifies WDR35 Variants Involved in Sensenbrenner Syndrome. *The American Journal of Human Genetics,* 87**,** 418-423.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. 2011. Unlocking Mendelian disease using exome sequencing. *genome,* 11**,** 64.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. 2012. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet,* 20**,** 490-497.

Ginsburg, O., Akbari, M., Aziz, Z., Young, R., Lynch, H., Ghadirian, P., Robidoux, A., Londono, J., Vasquez, G., Gomes, M., Costa, M., Dimitrakakis, C., Gutierrez, G., Pilarski, R., Royer, R. & Narod, S. 2009. The prevalence of germ-line TP53 mutations in women diagnosed with breast cancer before age 30. *Familial Cancer,* 8**,** 563-567.

Gonzalez-Angulo, A. M., Broglio, K., Kau, S.-W., Eralp, Y., Erlichman, J., Valero, V., Theriault, R., Booser, D., Buzdar, A. U., Hortobagyi, G. N. & Arun, B. 2005. Women age ≤ 35 years with primary breast carcinoma. *Cancer,* 103**,** 2466-2472.

Gopalakrishna, A., Jinka, R., Kumar, T. S., Khan, B. A. & Mevada, K. 2014. Joubert syndrome with cleft palate. *Journal of Cleft Lip Palate and Craniofacial Anomalies,* 1**,** 59.

Grant, S. F. A., Wang, K., Zhang, H., Glaberson, W., Annaiah, K., Kim, C. E., Bradfield, J. P., Glessner, J. T., Thomas, K. A., Garris, M., Frackelton, E. C., Otieno, F. G., Chiavacci, R. M., Nah, H.-D., Kirschner, R. E. & Hakonarson, H. 2009. A Genome-Wide Association Study Identifies a Locus for Nonsyndromic Cleft Lip with or without Cleft Palate on 8q24. *The Journal of Pediatrics,* 155**,** 909-913.

Grantham, R. 1974. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science,* 185**,** 862-864.

Grivennikov, S. I., Greten, F. R. & Karin, M. 2010. Immunity, inflammation, and cancer. *Cell,* 140**,** 883-99.

Grosen, D., Chevrier, C., Skytthe, A., Bille, C., Mølsted, K., Sivertsen, Å., Murray, J. C. & Christensen, K. 2010. A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. *Journal of medical genetics,* 47**,** 162-168.

Gudmundsdottir, K. & Ashworth, A. 2006. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene,* 25**,** 5864-5874.

Guenard, F., Pedneault Cs Fau - Ouellette, G., Ouellette G Fau - Labrie, Y., Labrie Y Fau - Simard, J., Simard, J. & Durocher, F. 2010. Evaluation of the contribution of the three breast cancer susceptibility genes CHEK2, STK11, and PALB2 in non-BRCA1/2 French Canadian families with high risk of breast cancer.

Gupta, P. B., Proia, D., Cingoz, O., Weremowicz, J., Naber, S. P., Weinberg, R. A. & Kuperwasser, C. 2007. Systemic Stromal Effects of Estrogen Promote the Growth of Estrogen Receptor–Negative Cancers. *Cancer Research,* 67**,** 2062-2071.

Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., Wang, X., Ademuyiwa, F., Ahmed, S., Ambrosone, C. B., Baglietto, L., Balleine, R., Bandera, E. V., Beckmann, M. W., Berg, C. D., Bernstein, L., Blomqvist, C., Blot, W. J., Brauch, H., Buring, J. E., Carey, L. A., Carpenter, J. E., Chang-Claude, J., Chanock, S. J., Chasman, D. I., Clarke, C. L., Cox, A., Cross, S. S., Deming, S. L., Diasio, R. B., Dimopoulos, A. M., Driver, W. R., Dunnebier, T., Durcan, L., Eccles, D., Edlund, C. K., Ekici, A. B., Fasching, P. A., Feigelson, H. S., Flesch-Janys, D., Fostira, F., Forsti, A., Fountzilas, G., Gerty, S. M., Giles, G. G., Godwin, A. K., Goodfellow, P., Graham, N., Greco, D., Hamann, U., Hankinson, S. E., Hartmann, A., Hein, R., Heinz, J., Holbrook, A., Hoover, R. N., Hu, J. J., Hunter, D. J., Ingles, S. A., Irwanto, A., Ivanovich, J., John, E. M., Johnson, N., Jukkola-Vuorinen, A., Kaaks, R., Ko, Y.-D., Kolonel, L. N., Konstantopoulou, I., Kosma, V.-M., Kulkarni, S., Lambrechts, D., Lee, A. M., Le Marchand, L., Lesnick, T., Liu, J., Lindstrom, S., Mannermaa, A., Margolin, S., Martin, N. G., Miron, P., Montgomery, G. W., Nevanlinna, H., Nickels, S., Nyante, S., Olswold, C., Palmer, J., Pathak, H., Pectasides, D., Perou, C. M., Peto, J., Pharoah, P. D. P., Pooler, L. C., Press, M. F., Pylkas, K., Rebbeck, T. R., Rodriguez-Gil, J. L., Rosenberg, L., Ross, E., Rudiger, T., Silva, I. d. S., et al. 2011. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet,* 43**,** 1210-1214.

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., Schnetz-Boutaud, N., Agarwal, A., Postel, E. A. & Pericak-Vance, M. A. 2005.

Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science,* 308**,** 419-421.
Hall, J. M., Friedman, L., Guenther, C., Lee, M. K., Weber, J. L., Black, D. M. & King, M. C. 1992. Closing in on a breast cancer gene on chromosome 17q. *Am J Hum Genet,* 50**,** 1235-42.
Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. & King, M. C. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science,* 250**,** 1684-1689.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter,* 11**,** 10-18.
Hanahan, D. & Weinberg, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57-70.
Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.
Haraksingh, R. R. & Snyder, M. P. 2013. Impacts of Variation in the Human Genome on Gene Regulation. *Journal of Molecular Biology,* 425**,** 3970-3977.
Harburg, G. C. & Hinck, L. 2011. Navigating breast cancer: axon guidance molecules as breast cancer tumor suppressors and oncogenes. *J Mammary Gland Biol Neoplasia,* 16**,** 257-70.
Healey, C. S., Dunning, A. M., Dawn Teare, M., Chase, D., Parker, L., Burn, J., Chang-Claude, J., Mannermaa, A., Kataja, V., Huntsman, D. G., Pharoah, P. D. P., Luben, R. N., Easton, D. F. & Ponder, B. A. J. 2000. A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. *Nat Genet,* 26**,** 362-364.
Hilbers, F. S. M., Vreeswijk, M. P. G., van Asperen, C. J. & Devilee, P. 2013. The impact of next generation sequencing on the analysis of breast cancer susceptibility: a role for extremely rare genetic variation? *Clinical genetics,* 84**,** 407-414.
Hill, M. E., Creed, G. A., McMullan, T. F. W., Tyers, A. G., Hilton-Jones, D., Robinson, D. O. & Hammans, S. R. 2001. Oculopharyngeal muscular dystrophy. *Brain,* 124**,** 522-526.
Hindorff, L. A Catalog of Published Genome-Wide Association Studies.
Hinnebusch, A. G. 2011. Molecular Mechanism of Scanning and Start Codon Selection in Eukaryotes. *Microbiology and Molecular Biology Reviews,* 75**,** 434-467.
Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. 2002. A comprehensive review of genetic association studies. *Genet Med,* 4**,** 45-61.
Huang, D. W., Sherman, B. T. & Lempicki, R. A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols,* 4**,** 44-57.
Huang, D. W., Sherman, B. T. & Lempicki, R. A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research,* 37**,** 1-13.
Hunter, D. J. 2005. Gene-environment interactions in human diseases. *Nat Rev Genet,* 6**,** 287-298.
Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R. N., Thomas, G. & Chanock, S. J. 2007. A genome-wide association study

identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet,* 39**,** 870-874.
Ibrahim, Z. M., Newhouse, S. & Dobson, R. Detecting epistasis in the presence of linkage disequilibrium: A focused comparison. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on, 16-19 April 2013 2013. 96-103.
Iijima, M., Suzuki, M., Tanabe, A., Nishimura, A. & Yamada, M. 2006. Two motifs essential for nuclear import of the hnRNP A1 nucleocytoplasmic shuttling sequence M9 core. *FEBS Letters,* 580**,** 1365-1370.
Irie, K., Shimizu, K., Sakisaka, T., Ikeda, W. & Takai, Y. 2004. Roles and modes of action of nectins in cell–cell adhesion. *Seminars in Cell & Developmental Biology,* 15**,** 643-656.
Jagomägi, T., Nikopensius, T., Krjutškov, K., Tammekivi, V., Viltrop, T., Saag, M. & Metspalu, A. 2010. MTHFR and MSX1 contribute to the risk of nonsyndromic cleft lip/palate. *European Journal of Oral Sciences,* 118**,** 213-220.
Jakobsen, L. P., Ullmann, R., Christensen, S. B., Jensen, K. E., Molsted, K., Henriksen, K. F., Hansen, C., Knudsen, M. A., Larsen, L. A., Tommerup, N. & Tumer, Z. 2007. Pierre Robin sequence may be caused by dysregulation of SOX9 and KCNJ2. *J Med Genet.* England.
Jezewski, P. A., Vieira, A. R., Nishimura, C., Ludwig, B., Johnson, M., O'Brien, S. E., Daack-Hirsch, S., Schultz, R. E., Weber, A., Nepomucena, B., Romitti, P. A., Christensen, K., Orioli, I. M., Castilla, E. E., Machida, J., Natsume, N. & Murray, J. C. 2003. Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. *Journal of Medical Genetics,* 40**,** 399-407.
Jugessur, A., Rahimov, F., Lie, R. T., Wilcox, A. J., Gjessing, H. K., Nilsen, R. M., Nguyen, T. T. & Murray, J. C. 2008. Genetic variants in IRF6 and the risk of facial clefts: single-marker and haplotype-based analyses in a population-based case-control study of facial clefts in Norway. *Genetic Epidemiology,* 32**,** 413-424.
Kamma, H., Horiguchi, H., Wan, L., Matsui, M., Fujiwara, M., Fujimoto, M., Yazawa, T. & Dreyfuss, G. 1999. Molecular Characterization of the hnRNP A2/B1 Proteins: Tissue-Specific Expression and Novel Isoforms. *Experimental Cell Research,* 246**,** 399-411.
Kannabiran, C. & Klintworth, G. K. 2006. TGFBI gene mutations in corneal dystrophies. *Human Mutation,* 27**,** 615-625.
Karchin, R. 2009. Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics,* 10**,** 35-52.
Kheirelseid, E., Boggs, J., Curran, C., Glynn, R., Dooley, C., Sweeney, K. & Kerin, M. 2011. Younger age as a prognostic indicator in breast cancer: A cohort study. *BMC Cancer,* 11**,** 383.
Kiezun, A., Garimella, K., Do, R., Stitziel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. F., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y. Y., Price, A. L., de Bakker, P. I. W., Purcell, S. M. & Sunyaev, S. R. 2012. Exome sequencing and the genetic basis of complex traits. *Nat Genet,* 44**,** 623-630.
Kim, H. J., Kim, N. C., Wang, Y.-D., Scarborough, E. A., Moore, J., Diaz, Z., MacLea, K. S., Freibaum, B., Li, S., Molliex, A., Kanagaraj, A. P., Carter, R., Boylan, K. B., Wojtas, A. M., Rademakers, R., Pinkus, J. L., Greenberg, S. A., Trojanowski, J. Q., Traynor, B. J., Smith, B. N., Topp, S., Gkazi, A.-S., Miller, J., Shaw, C. E., Kottlors, M., Kirschner, J., Pestronk, A., Li, Y. R., Ford, A. F., Gitler, A. D., Benatar, M., King, O. D., Kimonis, V. E., Ross, E.

D., Weihl, C. C., Shorter, J. & Taylor, J. P. 2013. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature,* 495**,** 467-473.

Kim, Y.-J., Noguchi, S., Hayashi, Y. K., Tsukahara, T., Shimizu, T. & Arahata, K. 2001. The product of an oculopharyngeal muscular dystrophy gene, poly(A)-binding protein 2, interacts with SKIP and stimulates muscle-specific gene expression. *Human Molecular Genetics,* 10**,** 1129-1139.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet,* 46**,** 310-5.

Klagsbrun, M. & Eichmann, A. 2005. A role for axon guidance receptors and ligands in blood vessel development and tumor angiogenesis. *Cytokine Growth Factor Rev,* 16**,** 535-48.

Knight, W. A., Livingston, R. B., Gregory, E. J. & McGuire, W. L. 1977. Estrogen Receptor as an Independent Prognostic Factor for Early Recurrence in Breast Cancer. *Cancer Research,* 37**,** 4669-4671.

Kollias, J., Elston, C. W., Ellis, I. O., Robertson, J. F. & Blamey, R. W. 1997. Early-onset breast cancer--histopathological and prognostic considerations. *Br J Cancer,* 75**,** 1318-23.

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., Howard, E., Ferreira de Lima, R. L. L., Daack-Hirsch, S., Sander, A., McDonald-McGinn, D. M., Zackai, E. H., Lammer, E. J., Aylsworth, A. S., Ardinger, H. H., Lidral, A. C., Pober, B. R., Moreno, L., Arcos-Burgos, M., Valencia, C., Houdayer, C., Bahuau, M., Moretti-Ferreira, D., Richieri-Costa, A., Dixon, M. J. & Murray, J. C. 2002. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet,* 32**,** 285-289.

Kraus, M. H., Popescu, N. C., Amsbaugh, S. C. & King, C. R. 1987. Overexpression of the EGF receptor-related proto-oncogene erbB-2 in human mammary tumor cell lines by different molecular mechanisms. *Embo j,* 6**,** 605-10.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics,* 22**,** 139-144.

Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. 2007. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics,* 80**,** 727-739.

Kumar, S., Dudley, J. T., Filipski, A. & Liu, L. 2011. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet,* 27**,** 377-86.

Kyriakides, T. R., Zhu, Y.-H., Smith, L. T., Bain, S. D., Yang, Z., Lin, M. T., Danielson, K. G., Iozzo, R. V., LaMarca, M., McKinney, C. E., Ginns, E. I. & Bornstein, P. 1998. Mice That Lack Thrombospondin 2 Display Connective Tissue Abnormalities That Are Associated with Disordered Collagen Fibrillogenesis, an Increased Vascular Density, and a Bleeding Diathesis. *The Journal of Cell Biology,* 140**,** 419-430.

Kühn, U. & Wahle, E. 2004. Structure and function of poly(A) binding proteins. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression,* 1678**,** 67-84.

Lacroix, M., Toillon, R.-A. & Leclercq, G. 2006. p53 and breast cancer, an update. *Endocrine-Related Cancer,* 13**,** 293-325.

Lalloo, F. & Evans, D. G. 2012. Familial Breast Cancer. *Clinical Genetics,* 82**,** 105-114.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Lander, E. S. & Schork, N. J. 1994. Genetic dissection of complex traits. *Science,* 265**,** 2037-2048.

Lange, C. A. & Yee, D. 2008. Progesterone and breast cancer. *Women's Health,* 4**,** 151-162.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segre, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J.-H., Yang, J., Gudbjartsson, D., Heard-Costa, N. L., Randall, J. C., Qi, L., Vernon Smith, A., Magi, R., Pastinen, T., Liang, L., Heid, I. M., Luan, J. a., Thorleifsson, G., Winkler, T. W., Goddard, M. E., Sin Lo, K., Palmer, C., Workalemahu, T., Aulchenko, Y. S., Johansson, A., Carola Zillikens, M., Feitosa, M. F., Esko, T., Johnson, T., Ketkar, S., Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N. L., Hayward, C., Hottenga, J.-J., Jacobs, K. B., Knowles, J. W., Kutalik, Z., Monda, K. L., Polasek, O., Preuss, M., Rayner, N. W., Robertson, N. R., Steinthorsdottir, V., Tyrer, J. P., Voight, B. F., Wiklund, F., Xu, J., Hua Zhao, J., Nyholt, D. R., Pellikka, N., Perola, M., Perry, J. R. B., Surakka, I., Tammesoo, M.-L., Altmaier, E. L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D. I., Chen, C., Coin, L., Cooper, M. N., Dixon, A. L., Gibson, Q., Grundberg, E., Hao, K., Juhani Junttila, M., Kaplan, L. M., Kettunen, J., Konig, I. R., Kwan, T., Lawrence, R. W., Levinson, D. F., Lorentzon, M., McKnight, B., Morris, A. P., Muller, M., Suh Ngwa, J., Purcell, S., Rafelt, S., Salem, R. M., Salvi, E., et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-838.

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., The Geuvadis,

C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Hasler, R., Syvanen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I. G., Estivill, X. & Dermitzakis, E. T. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature,* 501**,** 506-511.

Le Ber, I., Van Bortel, I., Nicolas, G., Bouya-Ahmed, K., Camuzat, A., Wallon, D., De Septenville, A., Latouche, M., Lattante, S., Kabashi, E., Jornea, L., Hannequin, D. & Brice, A. 2014. hnRNPA2B1 and hnRNPA1 mutations are rare in patients with "multisystem proteinopathy" and frontotemporal lobar degeneration phenotypes. *Neurobiology of Aging,* 35**,** 934.e5-934.e6.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M. & Lin, X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet,* 91**,** 224-37.

Lee, Sang H., Wray, Naomi R., Goddard, Michael E. & Visscher, Peter M. 2011. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics,* 88**,** 294-305.

Leek, R. D. & Harris, A. L. 2002. Tumor-associated macrophages in breast cancer. *J Mammary Gland Biol Neoplasia,* 7**,** 177-89.

Leek, R. D., Lewis, C. E., Whitehouse, R., Greenall, M., Clarke, J. & Harris, A. L. 1996. Association of macrophage infiltration with angiogenesis and prognosis in invasive breast carcinoma. *Cancer Res,* 56**,** 4625-9.

Lennon, C. J., Birkeland, A. C., Nuñez, J. A. P., Su, G. H., Lanzano, P., Guzman, E., Celis, K., Eisig, S. B., Hoffman, D., Rendon, M. T. G., Ostos, H., Chung, W. K. & Haddad, J. 2012. Association of candidate genes with nonsyndromic clefts in Honduran and Colombian populations. *The Laryngoscope,* 122**,** 2082-2087.

Leslie, E. J., Mansilla, M. A., Biggs, L. C., Schuette, K., Bullard, S., Cooper, M., Dunnwald, M., Lidral, A. C., Marazita, M. L., Beaty, T. H. & Murray, J. C. 2012. Expression and mutation analyses implicate ARHGAP29 as the etiologic gene for the cleft lip with or without cleft palate locus identified by genome-wide association on chromosome 1p22. *Birth Defects Research Part A: Clinical and Molecular Teratology,* 94**,** 934-942.

Leslie, E. J. & Marazita, M. L. 2013. Genetics of cleft lip and cleft palate. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics,* 163**,** 246-258.

Leslie, E. J. & Murray, J. C. 2013. Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clinical Genetics,* 84**,** 496-500.

Leslie, E. J., Standley, J., Compton, J., Bale, S., Schutte, B. C. & Murray, J. C. 2013. Comparative analysis of IRF6 variants in families with Van der Woude syndrome and popliteal pterygium syndrome using public whole-exome databases. *Genet Med,* 15**,** 338-344.

Letra, A., Bjork, B., Cooper, M. E., Szabo-Rogers, H., Deleyiannis, F. W. B., Field, L. L., Czeizel, A. E., Ma, L., Garlet, G. P., Poletta, F. A., Mereb, J. C., Lopez-Camelo, J. S., Castilla, E. E., Orioli, I. M., Wendell, S., Blanton, S. H., Liu, K., Hecht, J. T., Marazita, M. L., Vieira, A. R. & Silva, R. M. 2012. Association of AXIN2 with Non-syndromic Oral Clefts in Multiple Populations. *Journal of Dental Research,* 91**,** 473-478.

Letra, A., Maili, L., Mulliken, J. B., Buchanan, E., Blanton, S. H. & Hecht, J. T. 2014. Further evidence suggesting a role for variation in ARHGAP29 variants in nonsyndromic cleft lip/palate. *Birth Defects Research Part A: Clinical and Molecular Teratology,* 100**,** 679-685.

Li, M. X., Gui, H. S., Kwan, J. S., Bao, S. Y. & Sham, P. C. 2012. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res,* 40**,** e53.

Li, M. X., Kwan, J. S., Bao, S. Y., Yang, W., Ho, S. L., Song, Y. Q. & Sham, P. C. 2013. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet,* 9**,** e1003143.

Li, W.-H., Wu, C.-I. & Luo, C.-C. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution,* 21**,** 58-71.

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., Grarup, N., Guo, Y., Hellman, I., Jin, X., Li, Q., Liu, J., Liu, X., Sparso, T., Tang, M., Wu, H., Wu, R., Yu, C., Zheng, H., Astrup, A., Bolund, L., Holmkvist, J., Jorgensen, T., Kristiansen, K., Schmitz, O., Schwartz, T. W., Zhang, X., Li, R., Yang, H., Wang, J., Hansen, T., Pedersen, O., Nielsen, R. & Wang, J. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet,* 42**,** 969-972.

Lin, E. Y., Li, J.-F., Gnatovskiy, L., Deng, Y., Zhu, L., Grzesik, D. A., Qian, H., Xue, X.-n. & Pollard, J. W. 2006. Macrophages Regulate the Angiogenic Switch in a Mouse Model of Breast Cancer. *Cancer Research,* 66**,** 11238-11246.

Lines, Matthew A., Huang, L., Schwartzentruber, J., Douglas, S. L., Lynch, Danielle C., Beaulieu, C., Guion-Almeida, Maria L., Zechi-Ceide, Roseli M., Gener, B., Gillessen-Kaesbach, G., Nava, C., Baujat, G., Horn, D., Kini, U., Caliebe, A., Alanay, Y., Utine, Gulen E., Lev, D., Kohlhase, J., Grix, Arthur W., Lohmann, Dietmar R., Hehr, U., Böhm, D., Majewski, J., Bulman, Dennis E., Wieczorek, D. & Boycott, Kym M. 2012. Haploinsufficiency of a Spliceosomal GTPase Encoded by EFTUD2 Causes Mandibulofacial Dysostosis with Microcephaly. *The American Journal of Human Genetics,* 90**,** 369-377.

Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R. & Zanke, B. 2004. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. *Clinical Cancer Research,* 10**,** 2725-2737.

Liu, J., Aoki, M., Illa, I., Wu, C., Fardeau, M., Angelini, C., Serrano, C., Urtizberea, J. A., Hentati, F., Hamida, M. B., Bohlega, S., Culper, E. J., Amato, A. A., Bossie, K., Oeltjen, J., Bejaoui, K., McKenna-Yasek, D., Hosler, B. A., Schurr, E., Arahata, K., de Jong, P. J. & Brown, R. H. 1998. Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. *Nat Genet,* 20**,** 31-36.

Liu, W., Sun, X., Braut, A., Mishina, Y., Behringer, R. R., Mina, M. & Martin, J. F. 2005. Distinct functions for Bmp signaling in lip and palate fusion in mice. *Development,* 132**,** 1453-1461.

Liu, X., Jian, X. & Boerwinkle, E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat,* 32**,** 894-9.

Liu, X., Jian, X. & Boerwinkle, E. 2013. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat,* 34**,** E2393-402.

Lucek, P. R. & Ott, J. 1997. Neural network analysis of complex traits. *Genetic Epidemiology,* 14**,** 1101-1106.

Ludwig, K. U., Mangold, E., Herms, S., Nowak, S., Reutter, H., Paul, A., Becker, J., Herberz, R., AlChawa, T., Nasser, E., Bohmer, A. C., Mattheisen, M., Alblas, M. A., Barth, S., Kluck, N., Lauster, C., Braumann, B., Reich, R. H., Hemprich, A., Potzsch, S., Blaumeiser, B., Daratsianos, N., Kreusch, T., Murray, J. C., Marazita, M. L., Ruczinski, I., Scott, A. F., Beaty, T. H., Kramer, F.-J., Wienker, T. F., Steegers-Theunissen, R. P., Rubini, M., Mossey, P. A., Hoffmann, P., Lange, C., Cichon, S., Propping, P., Knapp, M. & Nothen, M. M. 2012. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet,* 44**,** 968-971.

Mahmoud, S. M., Paish, E. C., Powe, D. G., Macmillan, R. D., Grainge, M. J., Lee, A. H., Ellis, I. O. & Green, A. R. 2011. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J Clin Oncol,* 29**,** 1949-55.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. 2011. What can exome sequencing do for you? *J Med Genet,* 48**,** 580-9.

Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A. & et, a. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science,* 250**,** 1233-1238.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N. A., Chawa, T. A., Mattheisen, M., Steffens, M., Barth, S., Kluck, N., Paul, A., Becker, J., Lauster, C., Schmidt, G., Braumann, B., Scheer, M., Reich, R. H., Hemprich, A., Potzsch, S., Blaumeiser, B., Moebus, S., Krawczak, M., Schreiber, S., Meitinger, T., Wichmann, H.-E., Steegers-Theunissen, R. P., Kramer, F.-J., Cichon, S., Propping, P., Wienker, T. F., Knapp, M., Rubini, M., Mossey, P. A., Hoffmann, P. & Nothen, M. M. 2010. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet,* 42**,** 24-26.

Mangold, E., Ludwig, K. U. & Nöthen, M. M. 2011. Breakthroughs in the genetics of orofacial clefting. *Trends in Molecular Medicine,* 17**,** 725-733.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

Mantovani, A., Allavena, P., Sica, A. & Balkwill, F. 2008. Cancer-related inflammation. *Nature,* 454**,** 436-444.

Marazita, M. L., Murray, J. C., Lidral, A. C., Arcos-Burgos, M., Cooper, M. E., Goldstein, T., Maher, B. S., Daack-Hirsch, S., Schultz, R., Mansilla, M. A., Field, L. L., Liu, Y.-e., Prescott, N., Malcolm, S., Winter, R., Ray, A., Moreno, L., Valencia, C., Neiswanger, K., Wyszynski, D. F., Bailey-Wilson, J. E., Albacha-Hejazi, H., Beaty, T. H., McIntosh, I., Hetmanski, J. B., Tunçbilek, G., Edwards, M., Harkin, L., Scott, R. & Roddick, L. G. 2004.

Meta-Analysis of 13 Genome Scans Reveals Multiple Cleft Lip/Palate Genes with Novel Loci on 9q21 and 2q32-35. *The American Journal of Human Genetics,* 75**,** 161-173.

Marchini, J., Donnelly, P. & Cardon, L. R. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet,* 37**,** 413-417.

Matthew Michael, W., Choi, M. & Dreyfuss, G. 1995. A nuclear export signal in hnRNP A1: A signal-mediated, temperature-dependent nuclear protein export pathway. *Cell,* 83**,** 415-422.

Mazoyer, S., Dunning, A. M., Serova, O., Dearden, J., Puget, N., Healey, C. S., Gayther, S. A., Mangion, J., Stratton, M. R., Lynch, H. T., Goldgar, D. E., Ponder, B. A. J. & Lenoir, G. M. 1996. A polymorphic stop codon in BRCA2. *Nat Genet,* 14**,** 253-254.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. & Hirschhorn, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-369.

McClellan, J. & King, M.-C. 2010. Genetic Heterogeneity in Human Disease. *Cell,* 141**,** 210-217.

McDonald, M. T. & Gorski, J. L. 1993. Nager acrofacial dysostosis. *J Med Genet,* 30**,** 779-82.

Melnick, M., Chen, H., Zhou, Y. & Jaskoll, T. 2000. Thrombospondin-2 gene expression and protein localization during embryonic mouse palate development. *Archives of Oral Biology,* 45**,** 19-25.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., Rahman, N., Turnbull, C., Bocs, Fletcher, O., Peto, J., Gibson, L., dos-Santos-Silva, I., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Eilber, U., Behrens, S., Nevanlinna, H., Muranen, T. A., Aittomaki, K., Blomqvist, C., Khan, S., Aaltonen, K., Ahsan, H., Kibriya, M. G., Whittemore, A. S., John, E. M., Malone, K. E., Gammon, M. D., Santella, R. M., Ursin, G., Makalic, E., Schmidt, D. F., Casey, G., Hunter, D. J., Gapstur, S. M., Gaudet, M. M., Diver, W. R., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Berg, C. D., Chanock, S. J., Figueroa, J., Hoover, R. N., Lambrechts, D., Neven, P., Wildiers, H., van Limbergen, E., Schmidt, M. K., Broeks, A., Verhoef, S., Cornelissen, S., Couch, F. J., Olson, J. E., Hallberg, E., Vachon, C., Waisfisz, Q., Meijers-Heijboer, H., Adank, M. A., van der Luijt, R. B., Li, J., Liu, J., Humphreys, K., Kang, D., Choi, J.-Y., Park, S. K., Yoo, K.-Y., Matsuo, K., Ito, H., Iwata, H., Tajima, K., Guenel, P., Truong, T., Mulot, C., Sanchez, M., Burwinkel, B., Marme, F., Surowy, H., Sohn, C., Wu, A. H., Tseng, C.-c., Van Den Berg, D., Stram, D. O., Gonzalez-Neira, A., et al. 2015. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet,* 47**,** 373-380.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Fletcher, O., Peto, J., Gibson, L., dos Santos Silva, I., Nevanlinna, H., Muranen, T. A., Aittomaki, K., Blomqvist, C., Czene, K., Irwanto, A., Liu, J., Waisfisz, Q.,

Meijers-Heijboer, H., Adank, M., van der Luijt, R. B., Hein, R., Dahmen, N., Beckman, L., Meindl, A., Schmutzler, R. K., Muller-Myhsok, B., Lichtner, P., Hopper, J. L., Southey, M. C., Makalic, E., Schmidt, D. F., Uitterlinden, A. G., Hofman, A., Hunter, D. J., Chanock, S. J., Vincent, D., Bacot, F., Tessier, D. C., Canisius, S., Wessels, L. F. A., Haiman, C. A., Shah, M., Luben, R., Brown, J., Luccarini, C., Schoof, N., Humphreys, K., Li, J., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Couch, F. J., Wang, X., Vachon, C., Stevens, K. N., Lambrechts, D., Moisse, M., Paridaens, R., Christiaens, M.-R., Rudolph, A., Nickels, S., Flesch-Janys, D., Johnson, N., Aitken, Z., Aaltonen, K., Heikkinen, T., Broeks, A., Veer, L. J. V. t., van der Schoot, C. E., Guenel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Zamora, M. P., Perez, J. I. A., Pita, G., Alonso, M. R., Cox, A., Brock, I. W., Cross, S. S., Reed, M. W. R., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Henderson, B. E., et al. 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet,* 45**,** 353-361.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W. & et, a. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science,* 266**,** 66-71.

Milne, R. L., Burwinkel, B., Michailidou, K., Arias-Perez, J.-I., Zamora, M. P., Menéndez-Rodríguez, P., Hardisson, D., Mendiola, M., González-Neira, A., Pita, G., Alonso, M. R., Dennis, J., Wang, Q., Bolla, M. K., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M., Ko, Y.-D., Brauch, H., Hamann, U., The, G. N., Andrulis, I. L., Knight, J. A., Glendon, G., Tchatchou, S., kConFab, I., Australian Ovarian Cancer Study, G., Matsuo, K., Ito, H., Iwata, H., Tajima, K., Li, J., Brand, J. S., Brenner, H., Dieffenbach, A. K., Arndt, V., Stegmaier, C., Lambrechts, D., Peuteman, G., Christiaens, M.-R., Smeets, A., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Hartman, M., Hui, M., Yen Lim, W., Wan Chan, C., Marme, F., Yang, R., Bugert, P., Lindblom, A., Margolin, S., García-Closas, M., Chanock, S. J., Lissowska, J., Figueroa, J. D., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Hooning, M. J., Kriege, M., van den Ouweland, A. M. W., Koppert, L. B., Fletcher, O., Johnson, N., dos-Santos-Silva, I., Peto, J., Zheng, W., Deming-Halverson, S., Shrubsole, M. J., Long, J., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Cox, A., Cross, S. S., Reed, M. W. R., Schmidt, M. K., Broeks, A., Cornelissen, S., Braaf, L., Kang, D., Choi, J.-Y., Park, S. K., Noh, D.-Y., Simard, J., Dumont, M., Goldberg, M. S., Labrèche, F., Fasching, P. A., Hein, A., Ekici, A. B., et al. 2014a. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. *Human Molecular Genetics,* 23**,** 6096-6111.

Milne, R. L., Herranz, J., Michailidou, K., Dennis, J., Tyrer, J. P., Zamora, M. P., Arias-Perez, J. I., González-Neira, A., Pita, G., Alonso, M. R., Wang, Q., Bolla, M. K., Czene, K., Eriksson, M., Humphreys, K., Darabi, H., Li, J., Anton-Culver, H., Neuhausen, S. L., Ziogas, A., Clarke, C. A., Hopper, J. L., Dite, G. S., Apicella, C., Southey, M. C., Chenevix-Trench, G., kConFab, I., Australian Ovarian Cancer Study, G., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Andrulis, I. L., Knight, J. A., Glendon, G., Mulligan, A. M., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Chang-

Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Wang, X., Olson, J. E., Vachon, C., Purrington, K., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Dunning, A. M., Shah, M., Guénel, P., Truong, T., Sanchez, M., Mulot, C., Brenner, H., Dieffenbach, A. K., Arndt, V., Stegmaier, C., Lindblom, A., Margolin, S., Hooning, M. J., Hollestelle, A., Collée, J. M., Jager, A., Cox, A., Brock, I. W., Reed, M. W. R., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., Haiman, C. A., Henderson, B. E., Schumacher, F., Le Marchand, L., Simard, J., Dumont, M., Soucy, P., Dörk, T., Bogdanova, N. V., Hamann, U., Försti, A., Rüdiger, T., Ulmer, H.-U., Fasching, P. A., Häberle, L., Ekici, A. B., Beckmann, M. W., Fletcher, O., Johnson, N., et al. 2014b. A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46 450 cases and 42 461 controls from the breast cancer association consortium. *Human Molecular Genetics,* 23**,** 1934-1946.

Minić, S., Trpinac, D., Gabriel, H., Gencik, M. & Obradović, M. 2013. Dental and oral anomalies in incontinentia pigmenti: a systematic review. *Clinical Oral Investigations,* 17**,** 1-8.

Mirabella, M., Silvestri, G., de Rosa, G., Di Giovanni, S., Di Muzio, A., Uncini, A., Tonali, P. & Servidei, S. 2000. GCG genetic expansions in Italian patients with oculopharyngeal muscular dystrophy. *Neurology,* 54**,** 608-608.

Mitri, Z., Constantine, T. & O'Regan, R. 2012. The HER2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy. *Chemotherapy research and practice,* 2012.

Mooney, S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics,* 6**,** 44-56.

Moore, J. H. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity,* 56**,** 73-82.

Moore, J. H., Asselbergs, F. W. & Williams, S. M. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics,* 26**,** 445-455.

Moore, J. H. & Ritchie, M. D. 2004. THe challenges of whole-genome approaches to common diseases. *JAMA,* 291**,** 1642-1643.

Moore, J. H. & Williams, S. M. 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays,* 27**,** 637-646.

Moore, J. H. & Williams, S. M. 2009. Epistasis and Its Implications for Personal Genetics. *The American Journal of Human Genetics,* 85**,** 309-320.

Moore, M. R., Spence, J. B., Kiningham, K. K. & Dillon, J. L. 2006. Progestin inhibition of cell death in human breast cancer cell lines. *The Journal of Steroid Biochemistry and Molecular Biology,* 98**,** 218-227.

Mor, G., Yue, W., Santen, R. J., Gutierrez, L., Eliza, M., Berstein, L. M., Harada, N., Wang, J., Lysiak, J., Diano, S. & Naftolin, F. 1998. Macrophages, estrogen and the microenvironment of breast cancer. *J Steroid Biochem Mol Biol,* 67**,** 403-11.

Mort, M., Ivanov, D., Cooper, D. N. & Chuzhanova, N. A. 2008. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat,* 29**,** 1037-47.

Morton, N. E. 1955. Sequential tests for the detection of linkage. *American journal of human genetics,* 7**,** 277.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J. & Shaw, W. C. 2009. Cleft lip and palate. *The Lancet,* 374**,** 1773-1785.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. 2008. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology,* 32**,** 325-340.

Munro, T. P., Magee, R. J., Kidd, G. J., Carson, J. H., Barbarese, E., Smith, L. M. & Smith, R. 1999. Mutational Analysis of a Heterogeneous Nuclear Ribonucleoprotein A2 Response Element for RNA Trafficking. *Journal of Biological Chemistry,* 274**,** 34389-34395.

Ménard, S., Tomasic, G., Casalini, P., Balsari, A., Pilotti, S., Cascinelli, N., Salvadori, B., Colnaghi, M. I. & Rilke, F. 1997. Lymphoid infiltration as a prognostic variable for early-onset breast carcinomas. *Clinical Cancer Research,* 3**,** 817-819.

Müller, T., Schröder, R. & Zierz, S. 2001. GCG Repeats and phenotype in oculopharyngeal muscular dystrophy. *Muscle & Nerve,* 24**,** 120-122.

Nagashima, T., Kato, H., Kase, M., Maguchi, S., Mizutani, Y., Matsuda, K., Chuma, T., Mano, Y., Goto, Y.-i., Minami, N., Nonaka, I. & Nagashima, K. 2000. Oculopharyngeal muscular dystrophy in a Japanese family with a short GCG expansion (GCG)11 in PABP2 gene. *Neuromuscular Disorders,* 10**,** 173-177.

Nagy, R., Sweet, K. & Eng, C. 2004. Highly penetrant hereditary cancer syndromes. *Oncogene,* 23**,** 6445-70.

Nelen, M. R., Padberg, G. W., Peeters, E. A. J., Lin, A. Y., Helm, B. v. d., Frants, R. R., Goulon, V., Goldstein, A. M., Reen, M. M. M. v., Eastern, D. F., Eeles, R. A., Hodgson, S., Mulvihill, J. J., Murday, V. A., Tucker, M. A., Mariman, E. C. M., Starink, T. M., Ponder, B. A. J., Ropers, H. H., Kremer, H., Longy, M. & Eng, C. 1996. Localization of the gene for Cowden disease to chromosome 10q22-23. *Nat Genet,* 13**,** 114-116.

Ng, P. C. & Henikoff, S. 2001. Predicting Deleterious Amino Acid Substitutions. *Genome Research,* 11**,** 863-874.

Ng, P. C. & Henikoff, S. 2002. Accounting for Human Polymorphisms Predicted to Affect Protein Function. *Genome Research,* 12**,** 436-446.

Ng, P. C. & Henikoff, S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research,* 31**,** 3812-3814.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J. & Bamshad, M. J. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet,* 42**,** 30-35.

Nikopensius, T., Ambrozaityte, L., Ludwig, K. U., Birnbaum, S., Jagomägi, T., Saag, M., Matuleviciene, A., Linkeviciene, L., Herms, S. & Knapp, M. 2009. Replication of novel susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24 in Estonian and Lithuanian patients. *Am J Med Genet A,* 149**,** 2551-2553.

Nikopensius, T., Jagomägi, T., Krjutškov, K., Tammekivi, V., Saag, M., Prane, I., Piekuse, L., Akota, I., Barkane, B., Krumina, A., Ambrozaitytė, L., Matulevičienė, A., Kučinskienė, Z. A., Lace, B., Kučinskas, V. & Metspalu, A. 2010. Genetic variants in COL2A1, COL11A2, and IRF6 contribute risk to nonsyndromic cleft palate. *Birth Defects Research Part A: Clinical and Molecular Teratology,* 88**,** 748-756.

Nowak, D. M., Pitarque, J. A., Molinari, A., Bejjani, B. A. & Gajecka, M. 2012. Linkage Analysis as an Approach for Disease-related Loci Identification. *Computational Methods in Science and Technology,* 18**,** 95-101.

Olson, M. F., Pasteris, N. G., Gorski, J. L. & Hall, A. 1996. Faciogenital dysplasia protein (FGD1) and Vav, two related proteins required for normal

embryonic development, are upstream regulators of Rho GTPases. *Current Biology,* 6**,** 1628-1633.
Otero, L., Gutiérrez, S., Cháves, M., Vargas, C. & Bérmudez, L. 2007. Association of MSX1 With Nonsyndromic Cleft Lip and Palate in a Colombian Population. *The Cleft Palate-Craniofacial Journal,* 44**,** 653-656.
Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J. & Chatterjee, N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet,* 42**,** 570-575.
Park, J. W., McIntosh, I., Hetmanski, J. B., Jabs, E. W., Kolk, C. A. V., Wu-Chou, Y.-H., Chen, P. K., Chong, S. S., Yeow, V., Jee, S. H., Park, B. Y., Fallin, M. D., Ingersoll, R., Scott, A. F. & Beaty, T. H. 2007. Association between IRF6 and nonsyndromic cleft lip with or without cleft palate in four populations. *Genet Med,* 9**,** 219-227.
Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A.-L., Brown, P. O. & Botstein, D. 2000. Molecular portraits of human breast tumours. *Nature,* 406**,** 747-752.
Peto, J., Collins, N., Barfoot, R., Seal, S., Warren, W., Rahman, N., Easton, D. F., Evans, C., Deacon, J. & Stratton, M. R. 1999. Prevalence of BRCA1 and BRCA2 Gene Mutations in Patients With Early-Onset Breast Cancer. *Journal of the National Cancer Institute,* 91**,** 943-949.
Pharoah, P. D. P., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F. & Ponder, B. A. J. 2002. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet,* 31**,** 33-36.
Phillips, P. C. 2008. Epistasis [mdash] the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet,* 9**,** 855-867.
Polke, J. M., Laurá, M., Pareyson, D., Taroni, F., Milani, M., Bergamin, G., Gibbons, V. S., Houlden, H., Chamley, S. C., Blake, J., DeVile, C., Sandford, R., Sweeney, M. G., Davis, M. B. & Reilly, M. M. 2011. Recessive axonal Charcot-Marie-Tooth disease due to compound heterozygous mitofusin 2 mutations. *Neurology,* 77**,** 168-173.
Prescott, N. J., Lees, M. M., Winter, R. M. & Malcolm, S. 2000. Identification of susceptibility loci for nonsyndromic cleft lip with or without cleft palate in a two stage genome scan of affected sib-pairs. *Human Genetics,* 106**,** 345-350.
Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-909.
Pritchard, J. K. 2001. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics,* 69**,** 124-137.
Pritchard, J. K. & Cox, N. J. 2002. The allelic architecture of human disease genes: common disease–common variant…or not? *Human Molecular Genetics,* 11**,** 2417-2423.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. & Sham, P. C. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics,* 81**,** 559-575.

Purrington, K. S., Slager, S., Eccles, D., Yannoukakos, D., Fasching, P. A., Miron, P., Carpenter, J., Chang-Claude, J., Martin, N. G., Montgomery, G. W., Kristensen, V., Anton-Culver, H., Goodfellow, P., Tapper, W. J., Rafiq, S., Gerty, S. M., Durcan, L., Konstantopoulou, I., Fostira, F., Vratimos, A., Apostolou, P., Konstanta, I., Kotoula, V., Lakis, S., Dimopoulos, M. A., Skarlos, D., Pectasides, D., Fountzilas, G., Beckmann, M. W., Hein, A., Ruebner, M., Ekici, A. B., Hartmann, A., Schulz-Wendtland, R., Renner, S. P., Janni, W., Rack, B., Scholz, C., Neugebauer, J., Andergassen, U., Lux, M. P., Haeberle, L., Clarke, C., Pathmanathan, N., Rudolph, A., Flesch-Janys, D., Nickels, S., Olson, J. E., Ingle, J. N., Olswold, C., Slettedahl, S., Eckel-Passow, J. E., Anderson, S. K., Visscher, D. W., Cafourek, V. L., Sicotte, H., Prodduturi, N., Weiderpass, E., Bernstein, L., Ziogas, A., Ivanovich, J., Giles, G. G., Baglietto, L., Southey, M., Kosma, V.-M., Fischer, H.-P., The, G. N., Reed, M. W. R., Cross, S. S., Deming-Halverson, S., Shrubsole, M., Cai, Q., Shu, X.-O., Daly, M., Weaver, J., Ross, E., Klemp, J., Sharma, P., Torres, D., Rüdiger, T., Wölfing, H., Ulmer, H.-U., Försti, A., Khoury, T., Kumar, S., Pilarski, R., Shapiro, C. L., Greco, D., Heikkilä, P., Aittomäki, K., Blomqvist, C., Irwanto, A., Liu, J., Pankratz, V. S., Wang, X., Severi, G., Mannermaa, A., Easton, D., Hall, P., Brauch, H., et al. 2014. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis,* 35**,** 1012-1019.

Putti, T. C., El-Rehim, D. M. A., Rakha, E. A., Paish, C. E., Lee, A. H. S., Pinder, S. E. & Ellis, I. O. 2004. Estrogen receptor-negative breast carcinomas: a review of morphology and immunophenotypical analysis. *Mod Pathol,* 18**,** 26-35.

Péqueux, C., Raymond-Letron, I., Blacher, S., Boudou, F., Adlanmerini, M., Fouque, M.-J., Rochaix, P., Noël, A., Foidart, J.-M., Krust, A., Chambon, P., Brouchet, L., Arnal, J.-F. & Lenfant, F. 2012. Stromal Estrogen Receptor-α Promotes Tumor Growth by Normalizing an Increased Angiogenesis. *Cancer Research,* 72**,** 3010-3019.

Rafiq, S., Tapper, W., Collins, A., Khan, S., Politopoulos, I., Gerty, S., Blomqvist, C., Couch, F. J., Nevanlinna, H., Liu, J. & Eccles, D. 2013. Identification of Inherited Genetic Variations Influencing Prognosis in Early-Onset Breast Cancer. *Cancer Research,* 73**,** 1883-1891.

Rahimov, F., Jugessur, A. & Murray, J. C. 2011. Genetics of Nonsyndromic Orofacial Clefts. *The Cleft Palate-Craniofacial Journal,* 49**,** 73-91.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C., Melbye, M., Jugessur, A., Lie, R. T., Wilcox, A. J., Fitzpatrick, D. R., Green, E. D., Mossey, P. A., Little, J., Steegers-Theunissen, R. P., Pennacchio, L. A., Schutte, B. C. & Murray, J. C. 2008. Disruption of an AP-2[alpha] binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet,* 40**,** 1341-1347.

Rapakko, K., Allinen, M., Syrjakoski, K., Vahteristo, P., Huusko, P., Vahakangas, K., Eerola, H., Kainu, T., Kallioniemi, O. P., Nevanlinna, H. & Winqvist, R. 2001. Germline TP53 alterations in Finnish breast cancer families are rare and occur at conserved mutation-prone sites. *Br J Cancer,* 84**,** 116-9.

Raz, V., Routledge, S., Venema, A., Buijze, H., van der Wal, E., Anvar, S., Straasheijm, K. R., Klooster, R., Antoniou, M. & van der Maarel, S. M. 2011. Modeling Oculopharyngeal Muscular Dystrophy in Myotube Cultures Reveals Reduced Accumulation of Soluble Mutant PABPN1 Protein. *The American Journal of Pathology,* 179**,** 1988-2000.

Reed, M. J. & Purohit, A. 1997. Breast cancer and the role of cytokines in regulating estrogen synthesis: an emerging hypothesis. *Endocr Rev,* 18**,** 701-15.

Reich, D. E. & Lander, E. S. 2001. On the allelic spectrum of human disease. *Trends in Genetics,* 17**,** 502-510.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. & Moore, J. H. 2001. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics,* 69**,** 138-147.

Robertson, L., Hanson, H., Seal, S., Warren-Perry, M., Hughes, D., Howell, I., Turnbull, C., Houlston, R., Shanley, S., Butler, S., Evans, D. G., Ross, G., Eccles, D., Tutt, A. & Rahman, N. 2012. BRCA1 testing should be offered to individuals with triple-negative breast cancer diagnosed below 50 years. *Br J Cancer,* 106**,** 1234-1238.

Robinson, D., Hammans, S., Read, S. & Sillibourne, J. 2005. Oculopharyngeal muscular dystrophy (OPMD): analysis of the PABPN1 gene expansion sequence in 86 patients reveals 13 different expansion types and further evidence for unequal recombination as the mutational mechanism. *Human Genetics,* 116**,** 267-271.

Robinson, D. O., Hilton-Jones, D., Mansfield, D., Hildebrand, G. D., Marks, S., Mechan, D. & Ramsay, J. 2011a. Two cases of oculopharyngeal muscular dystrophy (OPMD) with the rare PABPN1 c.35G > C; p.Gly12Ala point mutation. *Neuromuscular Disorders,* 21**,** 809-811.

Robinson, D. O., Wills, A. J., Hammans, S. R., Read, S. P. & Sillibourne, J. 2006. Oculopharyngeal muscular dystrophy: a point mutation which mimics the effect of the PABPN1 gene triplet repeat expansion mutation. *Journal of Medical Genetics,* 43**,** e23-e23.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011b. Integrative genomics viewer. *Nat Biotech,* 29**,** 24-26.

Rodríguez, M., Camejo, C., Bertoni, B., Braida, C., Rodríguez, M. M., Brais, B., Medici, M. & Roche, L. 2005. (GCG)11 founder mutation in the PABPN1 gene of OPMD Uruguayan families. *Neuromuscular Disorders,* 15**,** 185-190.

Rojas-Martinez, A., Reutter, H., Chacon-Camacho, O., Leon-Cachon, R. B. R., Munoz-Jimenez, S. G., Nowak, S., Becker, J., Herberz, R., Ludwig, K. U., Paredes-Zenteno, M., Arizpe-Cantú, A., Raeder, S., Herms, S., Ortiz-Lopez, R., Knapp, M., Hoffmann, P., Nöthen, M. M. & Mangold, E. 2010. Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. *Birth Defects Research Part A: Clinical and Molecular Teratology,* 88**,** 535-537.

Rose, M. R., Landon, D. N., Papadimitriou, A. & Morgan-Hughes, J. A. 1997. A rapidly progressive adolescent-onset oculopharyngeal somatic syndrome with rimmed vacuoles in two siblings. *Annals of Neurology,* 41**,** 25-31.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D. N., Renwick, A., Seal, S. & Ramsay, E. 2013. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature,* 493**,** 406-410.

Ryan, J. G., Masters, S. L., Booty, M. G., Habal, N., Alexander, J. D., Barham, B. K., Remmers, E. F., Barron, K. S., Kastner, D. L. & Aksentijevich, I. 2010. Clinical features and functional significance of the P369S/R408Q variant

in pyrin, the familial Mediterranean fever protein. *Ann Rheum Dis,* 69**,** 1383-8.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409**,** 928-33.

Sanger, F., Nicklen, S. & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences,* 74**,** 5463-5467.

Sapkota, Y., Mackey, J. R., Lai, R., Franco-Villalobos, C., Lupichuk, S., Robson, P. J., Kopciuk, K., Cass, C. E., Yasui, Y. & Damaraju, S. 2013. Assessing SNP-SNP Interactions among DNA Repair, Modification and Metabolism Related Pathway Genes in Breast Cancer Susceptibility. *PLoS ONE,* 8**,** e64896.

Sartorius, C. A., Harvell, D. M. E., Shen, T. & Horwitz, K. B. 2005. Progestins Initiate a Luminal to Myoepithelial Switch in Estrogen-Dependent Human Breast Tumors without Altering Growth. *Cancer Research,* 65**,** 9779-9788.

Sasseville, M.-J. A., Caron, A. W., Bourget, L., Klein, A. F., Dicaire, M.-J., Rouleau, G. A., Massie, B., Langelier, Y. & Brais, B. 2006. The dynamism of PABPN1 nuclear inclusions during the cell cycle. *Neurobiology of Disease,* 23**,** 621-629.

Sauna, Z. E. & Kimchi-Sarfaty, C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet,* 12**,** 683-691.

Scacheri, P. C., Garcia, C., Hébert, R. & Hoffman, E. P. 1999. Unique PABP2 mutations in “Cajuns” suggest multiple founders of oculopharyngeal muscular dystrophy in populations with French ancestry. *American Journal of Medical Genetics,* 86**,** 477-481.

Scapoli, L., Palmieri, A., Martinelli, M., Pezzetti, F., Carinci, P., Tognon, M. & Carinci, F. 2005. Strong Evidence of Linkage Disequilibrium between Polymorphisms at the IRF6 Locus and Nonsyndromic Cleft Lip With or Without Cleft Palate, in an Italian Population. *The American Journal of Human Genetics,* 76**,** 180-183.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Research,* 22**,** 1748-1759.

Scheubert, L., Lustrek, M., Schmidt, R., Repsilber, D. & Fuellen, G. 2012. Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. *BMC Bioinformatics,* 13**,** 266.

Schliekelman, P. & Slatkin, M. 2002. Multiplex Relative Risk and Estimation of the Number of Loci Underlying an Inherited Disease. *The American Journal of Human Genetics,* 71**,** 1369-1385.

Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev,* 19**,** 212-9.

Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth,* 7**,** 575-576.

Serrano-Fernández, P., Dębniak, T., Górski, B., Bogdanova, N., Dörk, T., Cybulski, C., Huzarski, T., Byrski, T., Gronwald, J., Wokołorczyk, D., Narod, S. & Lubiński, J. 2009. Synergistic interaction of variants in CHEK2 and BRCA2 on breast cancer risk. *Breast Cancer Research and Treatment,* 117**,** 161-165.

Shendure, J. & Ji, H. 2008. Next-generation DNA sequencing. *Nat Biotech,* 26**,** 1135-1145.

Siepel, A., Pollard, K. & Haussler, D. 2006. New Methods for Detecting Lineage-Specific Selection. *In:* Apostolico, A., Guerra, C., Istrail, S., Pevzner, P. & Waterman, M. (eds.) *Research in Computational Molecular Biology.* Springer Berlin Heidelberg.

Sivertsen, Å., Wilcox, A. J., Skjærven, R., Vindenes, H. A., Åbyholm, F., Harville, E. & Lie, R. T. 2008. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *BMJ : British Medical Journal,* 336**,** 432-434.

Smyth, C., Špakulová, I., Cotton-Barratt, O., Rafiq, S., Tapper, W., Upstill-Goddard, R., Hopper, J. L., Makalic, E., Schmidt, D. F., Kapuscinski, M., Fliege, J., Collins, A., Brodzki, J., Eccles, D. M. & MacArthur, B. D. 2015. Quantifying the cumulative effect of low-penetrance genetic variants on breast cancer risk. *Molecular Genetics & Genomic Medicine***,** n/a-n/a.

Solinas, G., Germano, G., Mantovani, A. & Allavena, P. 2009. Tumor-associated macrophages (TAM) as major players of the cancer-related inflammation. *J Leukoc Biol,* 86**,** 1065-73.

Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R. & Borresen-Dale, A.-L. 2006. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics,* 7**,** 127.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L. & Liu, E. T. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences,* 100**,** 10393-10398.

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., Aben, K. K., Strobbe, L. J., Albers-Akkers, M. T., Swinkels, D. W., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Millastre, E., Andres, R., Godino, J., Garcia-Prats, M. D., Polo, E., Tres, A., Mouy, M., Saemundsdottir, J., Backman, V. M., Gudmundsson, L., Kristjansson, K., Bergthorsson, J. T., Kostic, J., Frigge, M. L., Geller, F., Gudbjartsson, D., Sigurdsson, H., Jonsdottir, T., Hrafnkelsson, J., Johannsson, J., Sveinsson, T., Myrdal, G., Grimsson, H. N., Jonsson, T., von Holst, S., Werelius, B., Margolin, S., Lindblom, A., Mayordomo, J. I., Haiman, C. A., Kiemeney, L. A., Johannsson, O. T., Gulcher, J. R., Thorsteinsdottir, U., Kong, A. & Stefansson, K. 2007. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet,* 39**,** 865-869.

Stanier, P. & Moore, G. E. 2004. Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts. *Human Molecular Genetics,* 13**,** R73-R81.

Stewart, R. E., Funderburk, S. & Setoguchi, Y. 1979. A malformation complex of ectrodactyly, clefting and hypomelanosis of Ito (incontinentia pigmenti achromians). *Cleft Palate J,* 16**,** 358-362.

Stratton, M. R. 1997. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. *The Lancet,* 349**,** 1505-1510.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-724.

Stratton, M. R. & Rahman, N. 2008. The emerging landscape of breast cancer susceptibility. *Nat Genet,* 40**,** 17-22.

Sunyaev, S., Ramensky, V. & Bork, P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet,* 16**,** 198-200.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E. & Børresen-Dale, A.-L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences,* 98**,** 10869-10874.

Tabor, H. K., Risch, N. J. & Myers, R. M. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet,* 3**,** 391-397.

Tan, T. Y., Kilpatrick, N. & Farlie, P. G. 2013. Developmental and genetic perspectives on Pierre Robin sequence. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics,* 163**,** 295-305.

Tavanez, J. P., Calado, P., Braga, J., Lafarga, M. & Carmo-Fonseca, M. 2005. In vivo aggregation properties of the nuclear poly(A)-binding protein PABPN1. *RNA,* 11**,** 752-762.

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O. & Caldas, C. 2007. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol,* 8**,** R157.

The International HapMap, C. 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-1320.

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K., Chatterjee, N., Garcia-Closas, M., Gonzalez-Bosquet, J., Prokunina-Olsson, L., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Diver, R., Prentice, R., Jackson, R., Kooperberg, C., Chlebowski, R., Lissowska, J., Peplonska, B., Brinton, L. A., Sigurdson, A., Doody, M., Bhatti, P., Alexander, B. H., Buring, J., Lee, I. M., Vatten, L. J., Hveem, K., Kumle, M., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R. N., Chanock, S. J. & Hunter, D. J. 2009. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet,* 41**,** 579-584.

Thomas, P. D. & Kejariwal, A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 15398-15403.

Thornton-Wells, T. A., Moore, J. H. & Haines, J. L. 2004. Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics,* 20**,** 640-647.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics,* 14**,** 178-192.

Tomé, F. M. S., Chateau, D., Helbling-Leclerc, A. & Fardeau, M. 1997. Morphological changes in muscle fibers in oculopharyngeal muscular dystrophy. *Neuromuscular Disorders,* 7, Supplement 1**,** S63-S69.

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C. S., Hughes, D., Warren-Perry, M., Tapper, W., Eccles, D., Evans, D. G., Hooning, M., Schutte, M., van den Ouweland, A., Houlston, R., Ross, G., Langford, C., Pharoah, P. D. P., Stratton, M. R., Dunning, A. M., Rahman, N. & Easton, D. F. 2010. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet,* 42**,** 504-507.

Turnbull, C. & Rahman, N. 2008. Genetic Predisposition to Breast Cancer: Past, Present, and Future. *Annual Review of Genomics and Human Genetics,* 9**,** 321-345.

Turnbull, C., Seal, S., Renwick, A., Warren-Perry, M., Hughes, D., Elliott, A., Pernet, D., Peock, S., Adlard, J. W., Barwell, J., Berg, J., Brady, A. F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P. J., Paterson, J., Porteous, M., Rogers, M. T., Shanley, S., Walker, L., Breast Cancer Susceptibility Collaboration, E., Ahmed, M., Eccles, D., Evans, D. G., Donnelly, P., Easton, D. F., Stratton, M. R. & Rahman, N. 2012. Gene–gene interactions in breast cancer susceptibility. *Human Molecular Genetics,* 21**,** 958-962.

Twyman, R. 2003. *Mutation or polymorphism?* [Online]. http://genome.wellcome.ac.uk/doc_WTD020780.html. [Accessed October 2014].

UK, C. R. *Breast cancer incidence statistics* [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive - heading-One [Accessed July 2015].

van der Sluijs, B. M., van Engelen, B. G. M. & Hoefsloot, L. H. 2003. Oculopharyngeal muscular dystrophy (OPMD) due to a small duplication in the PABPN1 gene. *Human Mutation,* 21**,** 553-553.

Vanderas, A. P. 1987. Incidence of cleft lip, cleft palate, and cleft lip and palate among races: a review. *Cleft palate J,* 24**,** 216-225.

Venkitaraman, A. R. 2002. Cancer Susceptibility and the Functions of BRCA1 and BRCA2. *Cell,* 108**,** 171-182.

Vieira, A. R., Cooper, M. E., Marazita, M. L., Orioli, I. M. & Castilla, E. E. 2007. Interferon regulatory factor 6 (IRF6) is associated with oral-facial cleft in individuals that originate in South America. *American Journal of Medical Genetics Part A,* 143A**,** 2075-2078.

Villarroya-Beltri, C., Gutiérrez-Vázquez, C., Sánchez-Cabo, F., Pérez-Hernández, D., Vázquez, J., Martin-Cofreces, N., Martinez-Herrera, D. J., Pascual-Montano, A., Mittelbrunn, M. & Sánchez-Madrid, F. 2013. Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs. *Nat Commun,* 4.

Visscher, P. M., Hill, W. G. & Wray, N. R. 2008. Heritability in the genomics era [mdash] concepts and misconceptions. *Nat Rev Genet,* 9**,** 255-266.

Vogelaar, I. P., Figueiredo, J., van Rooij, I. A. L. M., Simões-Correia, J., van der Post, R. S., Melo, S., Seruca, R., Carels, C. E. L., Ligtenberg, M. J. L. & Hoogerbrugge, N. 2013. Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. *Human Molecular Genetics,* 22**,** 919-926.

Waddell, M., Page, D. & Shaughnessy Jr, J. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. 2005 2005. ACM, 21-28.

Walker, R. A., Lees, E., Webb, M. B. & Dearing, S. J. 1996. Breast carcinomas occurring in young women (< 35 years) are different. *Br J Cancer,* 74**,** 1796-800.

Wall, B. L. & Elser, J. K. Imputation of Missing Data for input to Support Vector Machines.

Walsh, T., Lee, M. K., Casadei, S., Thornton, A. M., Stray, S. M., Pennil, C., Nord, A. S., Mandell, J. B., Swisher, E. M. & King, M.-C. 2010. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences,* 107**,** 12629-12633.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S. & Yu, W. 2010a. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics,* 87**,** 325-340.

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L. S. & Yu, W. 2010b. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics,* 26**,** 30-37.

Wang, K., Li, M. & Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research,* 38**,** e164-e164.

Wang, Y., Liu, G., Feng, M. & Wong, L. 2011. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics,* 27**,** 2936-2943.

Ward, A. J. & Cooper, T. A. 2010. The pathobiology of splicing. *The Journal of Pathology,* 220**,** 152-163.

Wenham, R. M., Schildkraut, J. M., McLean, K., Calingaert, B., Bentley, R. C., Marks, J. & Berchuck, A. 2003. Polymorphisms in BRCA1 and BRCA2 and Risk of Epithelial Ovarian Cancer. *Clinical Cancer Research,* 9**,** 4396-4403.

Whitford, P., Mallon, E. A., George, W. D. & Campbell, A. M. 1990. Flow cytometric analysis of tumour infiltrating lymphocytes in breast cancer. *Br J Cancer,* 62**,** 971-5.

Winter, R., Kühn, U., Hause, G. & Schwarz, E. 2012. Polyalanine-independent Conformational Conversion of Nuclear Poly(A)-binding Protein 1 (PABPN1). *Journal of Biological Chemistry,* 287**,** 22662-22671.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J. a., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., Karjalainen, J., Lo, K. S., Locke, A. E., Magi, R., Mihailov, E., Porcu, E., Randall, J. C., Scherag, A., Vinkhuyzen, A. A. E., Westra, H.-J., Winkler, T. W., Workalemahu, T., Zhao, J. H., Absher, D., Albrecht, E., Anderson, D., Baron, J., Beekman, M., Demirkan, A., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Fraser, R. M., Goel, A., Gong, J., Justice, A. E., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Lui, J. C., Mangino, M., Leach, I. M., Medina-Gomez, C., Nalls,

M. A., Nyholt, D. R., Palmer, C. D., Pasko, D., Pechlivanis, S., Prokopenko, I., Ried, J. S., Ripke, S., Shungin, D., Stancakova, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Afzal, U., Arnlov, J., Arscott, G. M., Bandinelli, S., Barrett, A., Bellis, C., Bennett, A. J., Berne, C., Bluher, M., Bolton, J. L., Bottcher, Y., Boyd, H. A., Bruinenberg, M., Buckley, B. M., Buyske, S., Caspersen, I. H., Chines, P. S., Clarke, R., Claudi-Boehm, S., Cooper, M., Daw, E. W., De Jong, P. A., Deelen, J., Delgado, G., et al. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet,* advance online publication.

Woolfe, A., Mullikin, J. & Elnitski, L. 2010. Genomic features defining exonic variants that modulate splicing. *Genome Biology,* 11**,** R20.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. & Micklem, G. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature,* 378**,** 789-92.

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D. & et, a. 1994. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science,* 265**,** 2088-2090.

Wu, K., Hinson, S. R., Ohashi, A., Farrugia, D., Wendt, P., Tavtigian, S. V., Deffenbaugh, A., Goldgar, D. & Couch, F. J. 2005. Functional Evaluation and Cancer Risk Assessment of BRCA2 Unclassified Variants. *Cancer Research,* 65**,** 417-426.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet,* 89**,** 82-93.

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H. & Yu, W. 2009. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics,* 25**,** 504-511.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. & Visscher, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-569.

Yell, J. A., Walshe, M. & Desai, S. N. 1991. Incontinentia pigmenti associated with bilateral cleft lip and palate. *Clinical and experimental dermatology,* 16**,** 49-50.

Yoshida, K. & Miki, Y. 2004. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Science,* 95**,** 866-871.

Younkin, S., Scharpf, R., Schwender, H., Parker, M., Scott, A., Marazita, M., Beaty, T. & Ruczinski, I. 2014. A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk. *BMC Genetics,* 15**,** 24.

Zeggini, E. & Ioannidis, J. P. 2009. Meta-analysis in genome-wide association studies. *Pharmacogenomics,* 10**,** 191-201.

Zhang, Z. & Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research,* 31**,** 5338-5348.

Zheng, W., Cai, Q., Signorello, L. B., Long, J., Hargreaves, M. K., Deming, S. L., Li, G., Li, C., Cui, Y. & Blot, W. J. 2009. Evaluation of 11 Breast Cancer Susceptibility Loci in African-American Women. *Cancer Epidemiology Biomarkers & Prevention*.

Zucchero, T. M., Cooper, M. E., Maher, B. S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., Caprau, D., Christensen, K., Suzuki, Y. & Machida, J. 2004. Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *New England Journal of Medicine,* 351**,** 769-780.

Zuchner, S., Mersiyanova, I. V., Muglia, M., Bissar-Tadmouri, N., Rochelle, J., Dadali, E. L., Zappia, M., Nelis, E., Patitucci, A., Senderek, J., Parman, Y., Evgrafov, O., Jonghe, P. D., Takahashi, Y., Tsuji, S., Pericak-Vance, M. A., Quattrone, A., Battologlu, E., Polyakov, A. V., Timmerman, V., Schroder, J. M. & Vance, J. M. 2004. Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A. *Nat Genet,* 36**,** 449-451.

Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences,* 109**,** 1193-1198.

Genetic dissection of early-onset breast cancer and other genetic diseases

Published Papers

Genetic dissection of early-onset breast cancer and other genetic diseases

CLINICAL GENETICS An International Journal of Genetics, Molecular and Personalized Medicine

Clin Genet 2014
Printed in Singapore. All rights reserved

CLINICAL GENETICS
doi: 10.1111/cge.12547

Original Article

Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing

Pengelly R.J., Upstill-Goddard R., Arias L., Martinez J., Gibson J., Knut M., Collins A.L., Ennis S., Collins A., Briceno I. Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing. Clin Genet 2014. © John Wiley & Sons A/S. Published by John Wiley & Sons Ltd, 2014

Individuals from three families ascertained in Bogota, Colombia, showing syndromic phenotypes, including cleft lip and/or palate, were exome-sequenced. In each case, sequencing revealed the underlying causal variation confirming or establishing diagnoses. The findings include very rare and novel variants providing insights into genotype and phenotype relationships. These include the molecular diagnosis of an individual with Nager syndrome and a family exhibiting an atypical incontinentia pigmenti phenotype with a missense mutation in *IKBKG*. *IKBKG* mutations are typically associated with preterm male death, but this variant is associated with survival for 8–15 days. The third family exhibits unusual phenotypic features and the proband received a provisional diagnosis of Pierre Robin sequence (PRS). Affected individuals share a novel deleterious mutation in *IRF6*. Mutations in *IRF6* cause Van der Woude and popliteal pterygium syndrome and contribute to nonsyndromic cleft lip phenotypes but have not previously been associated with a PRS phenotype. Exome sequencing followed by *in silico* screening to identify candidate causal variant(s), and functional assay in some cases offers a powerful route to establishing molecular diagnoses. This approach is invaluable for conditions showing phenotypic and/or genetic heterogeneity including cleft lip and/or palate phenotypes where many underlying causal genes have not been identified.

Conflict of interest

The authors declare no conflicts of interest.

R.J. Pengelly[a,†], R. Upstill-Goddard[a,†], L. Arias[b,†], J. Martinez[b,†], J. Gibson[c], M. Knut[a], A.L. Collins[d], S. Ennis[a], A. Collins[a] and I. Briceno[b]

[a]Genetic Epidemiology and Genomic Informatics, Faculty of Medicine, University of Southampton, Southampton, UK, [b]Department of Biomedical Sciences, Medical School, Universidad de La Sabana, Bogota, Colombia, [c]Centre for Biological Sciences, Faculty of Natural & Environmental Sciences, University of Southampton, Southampton, UK, and [d]Department of Clinical Genetics, Southampton General Hospital, Southampton, UK

[†]These authors equally contributed.

Key words: cleft lip and palate – exome sequencing – incontinentia pigmenti – Nager syndrome – Pierre Robin sequence – syndromic disease

Corresponding author: Prof Andrew Collins, Genetic Epidemiology and Genomic Informatics, Faculty of Medicine, University of Southampton, Duthie Building (808), Tremona Road, Southampton, SO166YD, UK.
Tel.: +44(0)2381206939;
fax: +44(0)2380794264;
e-mail: arc@soton.ac.uk

Received 15 October 2014, revised and accepted for publication 26 November 2014

Cleft lip and/or palate (CLP) is a phenotypic feature of at least 275 genetic syndromes that arise through single-gene mutations, chromosomal abnormalities or teratogens (1). The syndromic designation refers to the presence of additional physical or cognitive abnormalities, along with CLP. About 75% of CLP syndromes have a known genetic cause and include many arising through Mendelian inheritance at single genetic loci.

An important example is Van der Woude syndrome (VWS), and its allelic disorder popliteal pterygium syndrome (PPS), which is caused by mutations in the *IRF6* gene (2). VWS is the most common cause of syndromic clefting accounting for 2% of CLP cases. This gene functions as a transcriptional activator and shows high allelic heterogeneity as 100s of mutations in *IRF6* have been reported to cause these disorders. The DNA-binding domains of *IRF6* are particularly enriched for causal mutations, but mutations are also found extensively throughout the protein-binding domain (3). Furthermore, *IRF6* mutations have also been linked to nonsyndromic forms of CLP (4).

Because high allelic heterogeneity underlies many syndromes, targeted next-generation sequencing (NGS) of individual genes or panels of genes provides a route to establish molecular diagnoses that inform clinical management. A number of standard gene sequencing panels have been developed and provide particularly cost-effective routes to exploiting these technologies in a clinical setting. In contrast to standard panels, 'custom' gene panels are more expensive when used for a few samples but less so for larger sample sizes (5). However, developing optimal gene panels to screen samples representing genetically and phenotypically heterogeneous diseases or syndromes can be difficult particularly where there are ambiguous genotype–phenotype correlations. A cost-effective alternative strategy is to employ whole-exome sequencing (WES) which identifies the majority of coding variants in a DNA sample. For syndromic conditions, where underlying mutations are most likely to show Mendelian patterns of inheritance, exome sequencing, which screens up to 3% of the genome, is a powerful strategy to identify causal variation. As part of a study into the genetic basis of CLP phenotypes in Colombia, the utility of the WES strategy for CLP syndromes in families with particularly rare and/or atypical clinical phenotypes is considered here. We describe the results from the exome sequencing of six individuals from three families. Results confirm that genetic variants underlying CLP phenotypes in these Colombian families comprise both known and novel variants and establish new variant: phenotypic relationships.

Materials and methods

Patients

Families were ascertained at Operation Smile, Bogota, Colombia. All affected individuals or their parents gave written informed consent for the study. Ethical permission was obtained from the Research Ethics Committee at the Universidad de La Sabana, Bogota. The main clinical findings are summarized in Table 1, and the pedigrees in Figs. 1 and 2.

Exome sequencing

Samples from six individuals with syndromic phenotypes from three families were exome-sequenced. DNA derived from peripheral blood was sequenced for five individuals during 2012–2013 on an Illumina HiSeq 2000 sequencer at the Wellcome Trust Centre for Human Genomics using the Agilent SureSelect v5 capture kit encompassing 51 Mb of genome sequence. A sample from an additional individual (CLP2, II10, the half-uncle of the proband) was exome-sequenced at the Beijing Genomics Institute (BGI) during 2014. Paired-end exome sequence reads were aligned to the hg19 human reference genome using novoalignMPI (V2.08.02, Novocraft Technologies, Selangor, Malaysia). Picard (v1.34) and SAMtools (v0.01.18) were used to merge, sort and manipulate format aligned sequence files (Sam and Bam files) and create a 'pileup' of reads for each sample. Coverage statistics were calculated using BEDTools (v2.13.2) and are described in Table S1, Supporting Information. The mean read depths across the exomes are in the range 57–128. Sample provenance was ensured using an optimized panel of 24 SNPs (6). For filtering of single nucleotide polymorphism (SNP) variants and indel calls, we established a comprehensive list of genes previously implicated in any form of CLP phenotype including search terms related to the clinical diagnoses made for the patients. First, we queried the Human Gene Mutation Database (HGMD professional http://www.hgmd.org/), in July 2014, using the following search terms: cleft lip, cleft palate, cleft, syndactyly, brachydactyly, Pierre Robin, incontinentia pigmenti, Nager syndrome, hyperpigmentation, craniofacial, clubbing, dysmorphic, dysmorphia and micrognathia. This list comprised 363 genes. Additional genes were included after a corresponding interrogation of OMIM (http://omim.org/, accessed July 2014), and a small number of additional CLP-related genes from the review

Table 1. Description of the probands

Family number	Proband diagnosis/ *N* affected in family/*N* exomes sequenced	Phenotype
CLP1	Nager syndrome/1/1	Age at examination: 9 years; history of swallowing disorder because of retrognathia; bilateral dacryostenosis; micrognathia; atresia of the right external auditory canal; agenesis of first finger (bilateral); and normal external genitalia.
CLP2	Incontinentia pigmenti/11/2	Age at examination: 34 years; cleft lip palate on left side; nail hyperpigmentation; nail clubbing; and cutaneous syndactyly.
CLP3	Pierre Robin syndrome /5/3	Age at examination: 18 months; history of respiratory failure because of micrognathia; and cleft palate.

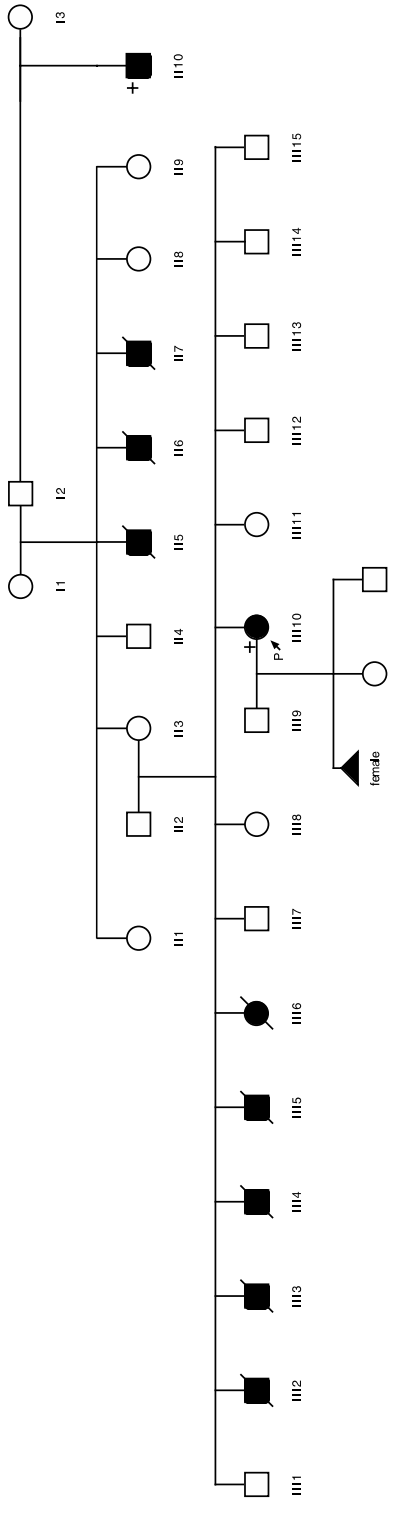


Fig. 1. Pedigree of family CLP2. II5, II6, II7 facial clefting, cause of death uncertain; II10 (half-uncle of proband) facial clefting, syndactyly, proximal thumbs, brachydactyly (exome-sequenced) ; III2, III3, III4, II5 (males) post-natal death at 8–15 days and facial clefting; III6 (female) post-natal death at 8 days and cleft lip and palate; IV1 pre-natal death and facial clefting; III10 (proband, exome-sequenced), unilateral (left side) cleft lip and palate, clubbing, nail hyperpigmentation, cutaneous syndactyly.

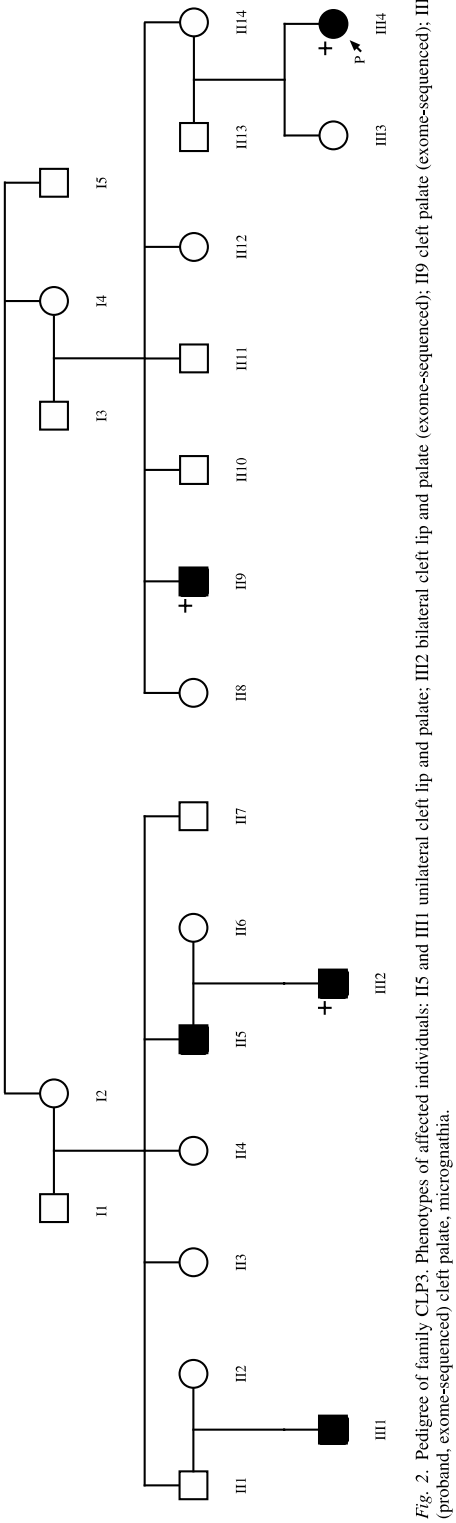


Fig. 2. Pedigree of family CLP3. Phenotypes of affected individuals: II5 and III1 unilateral cleft lip and palate; III2 bilateral cleft lip and palate (exome-sequenced); II9 cleft palate (exome-sequenced); III4 (proband, exome-sequenced) cleft palate, micrognathia.

were also included by Collins et al. (7). The complete list of 865 genes considered in variant filtering is given in Table S2. We filtered the lists of called variants to identify all novel non-synonymous (NS), stopgain, stoploss, splicing and indel variants in these genes as well as known rare variants with an allele frequency of less than 1% in the 1000 Genomes Project database (http://www.1000genomes.org/) (Table 2 and Table S3). More frequent variants were excluded from further consideration as unlikely causes of rare syndromic disease. For NS variants, we used the scaled predictive scores (8, 9) from dbNSFP v2 (10, 11) and only considered NS variants classed as deleterious or damaging by any of: PhyloP (larger positive scores represent conserved sites while negative scores indicate non-conserved sites) (12); SIFT (scores <0.05 are predicted to affect protein function) (13); Polyphen2 HumVar (scores ≤0.446 considered 'benign'; scores between 0.447 and 0.908 considered 'possibly damaging'; scores ≥0.909 considered 'probably damaging') (14); LRT for which variants are predicted deleterious if they are: (i) from a codon considered to be significantly constrained, (ii) from a site with alignments in at least 10 eutherian mammal species, and (iii) the alternative amino acid is not observed in any other eutherian mammal species with other variants classified as neutral or unknown (15); MutationTaster (variants with scores >0.95 considered damaging; 16) and GERP++ (scores range from <0 to 6.17, with higher scores indicating stronger constraint, a score of 6.17 indicates perfect conservation across all sequenced mammals) (17, 18). Grantham scores were also assigned to all NS substitutions (50 or below for conservative amino acid changes, scores for moderate changes 51–100, and radical changes >100) (19). All variants were also annotated with combined scores for deleteriousness: PHRED-scaled CADD (higher scores indicate that a variant is more likely to be deleterious) (20); Logit (the conditional probability that a variant is Mendelian disease-causing given prediction scores from 13 programs, including sift, PolyPhen2, LRT, MutationTaster, PhyloP, GERP++ and CADD, under a logistic regression model) (10, 11). We also produced a combined rank for variants with PhylopP, GERP++, CADD and Logit scores based on the summed ranks across all four scores (Table 2 and Table S3).

We excluded variants found in homopolymer/repeat regions that can arise through miss-alignment between the sequenced reads and reference sequence. Any variants with read depth of <10 or in genes considered to be 'highly mutable' (21) were removed from further consideration. All identified variants were cross-referenced with an in-house database of exome-sequenced samples and variants present in any of these exomes, which included 12 Colombian samples with nonsyndromic CLP, and were excluded from further consideration as unlikely to contribute to syndromic phenotypes. For family CLP3, only variants shared by all three individuals were tested, which were included in Table 2. The three families display distinct phenotypes, and we considered it unlikely that causal variants would be common to more than one family. We therefore excluded variants present in more than one of the three families as likely to reflect local population variation or artefacts from the sequencing batch.

Sanger sequencing

A number of candidate causal variants identified in the exome data were validated by Sanger sequencing (Table 3) using primers detailed in Table S4.

Results

Table 2 and Table S3 lists a total of 35 variants that met the filtering criteria across the 6 exome-sequenced individuals of which 30 are NS SNPs, 3 are splicing variants, and there are single stop gain and frameshift insertions. Table 2 lists the 13 variants from this set predicted as most likely to be deleterious (frameshift and highest ranked from combined scores). Table S3 lists the 22 variants predicted as least damaging. Analysis on each family suggests causal variation in each case, as described below.

Family CLP1

The proband was diagnosed as a potential Nager syndrome patient. Nager syndrome is extremely rare, and fewer than 100 cases have been reported (22, 23). Nager syndrome belongs to a group of conditions displaying acrofacial dysostosis, characterized by association between craniofacial and limb malformations (24). The patient phenotype (Table 1) shows features associated with this condition including micrognathia, auditory canal defects and malformed fingers. The patient represents a sporadic isolated case with no known cases among relatives.

Exome sequencing of the proband identified novel heterozygous NS variants in the *IFT172* (rank 4, Table 2), *ERCC2* (rank 7) and *PROKR2* genes (rank 10). More significantly, sequencing also identified the known R354fs (1060_1061insC) frameshift mutation in exon 5 of the *SF3B4* gene (Table 2). This variant was confirmed as present by Sanger sequencing (Table 3). Exome sequencing has previously established mutations in the *SF3B4* gene (splicing factor 3B, subunit 4) as responsible for autosomal dominant Nager syndrome (23). *SF3B4* encodes a highly conserved protein involved in mRNA splicing and bone morphogenic protein (BMP) signalling. The latter presumably contributes largely to the skeletal phenotype in this syndrome. However, *SF3B4* testing is negative in approximately one-third of Nager cases, for example, in 16 of 41 individuals (23); 5 of 14 families tested (24); 5 of 12 families, (25). Most patients who are negative for *SF3B4* mutations are phenotypically identical, indicating genetic heterogeneity. The variant identified in this patient corresponds to the same frameshift mutation identified as *de novo* in family 'I' by Bernier et al. (23) and Petit et al. (24) in their 'case 13'. The identification of the same mutation in three independent studies suggests that this may be one of the

Table 2. Rare and novel variants from exome sequencing predicted as the most damaging[a,b]

Gene	Chromosome	Exon	Base pair in hg19	Variant type	Nucleotide change	Amino acid change	dbSNP135 rsID	MAF (1000 genomes)	MAF (exome variant server)	SIFT score	PolyPhen2	LRT	Mutation Taster	Grantham score	PhyloP	GERP++	CADD score	Logit score	Rank	CLP1_proband	CLP2_III10	CLP2_II10	CLP3_III2	CLP3_II9	CLP3_III4
SF3B4	1	5	149895760	Fi	1060_1061insC	R354fs	–	–	–					–			–	–	–	◇					
IFT140	16	20	1576628	Ns	G2569A	G857S	rs200876696	–	–	0.07	0.995	0.000	1.00	56	7.661	5.22	34	0.275	1		◇				
RPGRIP1L	16	6	53720397	Sg	G724T	E242X	–	–	–	0.14	–	0.000	1.00	–	4.463	5.87	36	0.164	2		◇				
IRF6	1	5	209964011	Ns	G604A	V202I	–	–	–	–	0.916	0.000	1.00	29	7.311	6.17	27.1	0.119	3				◇	◇	◇
IFT172	2	33	27676956	Ns	G3604T	V1202L	–	–	–	–	–	–	1.00	32	4.362	5.7	28.3	0.148	4	◇					
IDUA	4	7	995942	Ns	T965A	V322E	rs76722191	0.002	0.0001	0.00	0.999	0.000	1.00	121	5.962	5.15	23.3	0.151	5		◇				
SH3PXD2B	5	13	171765821	Ns	C2288T	P763L	–	–	–	–	0.997	0.000	1.00	98	7.565	5.29	19.54	0.116	6			◇			
ERCC2	19	20	45856006	Ns	A1900G	K634E	–	–	–	0.00	0.925	0.000	1.00	56	5.182	5.13	28.3	0.123	7	◇					
IKBKG	X	2	153780386	Ns	G169A	E57K	rs148695964	–	0.0018	0.16	0.997	0.000	0.99	56	5.105	5.6	21.9	0.094	8		◇				
NKX3-2	4	2	13544126	Ns	G493C	D165H	rs61795263	–	–	0.17	0.419	0.022	1.00	81	3.045	5.31	21.4	0.212	9		◇				
PROKR2	20	2	5283122	Ns	C719T	T240I	–	–	–	0.23	0.841	0.000	1.00	89	5.246	5.16	23.1	0.096	10	◇					
COL1A2	7	48	94056566	Ns	C3226T	P1076S	–	–	–	0.02	0.063	0.000	1.00	74	3.858	5.32	16.87	0.147	11		◇				
PGM1	1	5	64100551	Ns	C143T	A48V	–	–	–	0.13	0.025	0.000	1.00	64	7.651	5.13	19.05	0.072	12		◇				

Ns, non-synonymous; sp, splicing; fi, frameshift insertion; sg, stopgain; rank is based on sum of ranks for variants with PhyloP, GERP++, CADD and Logit scores and range from (predicted) most to least deleterious.

[a]Scores for variants predicted as potentially deleterious underlined.

[b]◇ indicates heterozygous variant.

Table 3. Sequencing results by individual

Pedigree individual	Phenotype	Causal variant(s) from exome sequencing/confirmed by Sanger sequencing
CLP1 (proband)	Nager syndrome	*SF3B4 R354fs*/Confirmed
CLP2 III10 (proband)	Incontinentia pigmenti	*IKBKG E57K*/NT
CLP2 II10 (half-uncle of proband)	Facial clefting, syndactyly, brachydactyly, proximal thumbs	Negative for *IKBKG E57K*/NT
CLP3 III4 (proband)	Pierre Robin syndrome	*IRF6 G604A*/NT
CLP3 II14	Unaffected (carrier)	NT/Confirmed carrier of *IRF6 G604A*
CLP3 II9	Cleft palate	*IRF6 G604A*/NT
CLP3 III2	Bilateral cleft lip and palate	*IRF6 G604A*/NT
CLP3 III1	Unilateral cleft lip and palate	NT/Confirmed carrier of *IRF6 G604A*

CLP, cleft lip and/or palate; NT, not tested.

more frequent mutations in Nager syndrome; however, causal mutations have been identified in all six exons of the gene. Phenotypic differences between patients with and without *SF3B4* mutations are poorly defined. Czeschik et al. (25) noted that a cleft palate occurs more frequently in *SF3B4* mutation-positive patients (86% *vs* 20%). Larger patient cohorts will be required to better establish the phenotype–genotype relationships.

Family CLP2

The female proband presented with bilateral CLP together with a catalogue of other syndromic features (Table 1), including abnormal nail pigmentation and cutaneous syndactyly. The family pedigree (Fig. 1) suggests an X-linked disorder associated with lethality at a post-natal stage in males, but also in one female. Interestingly, the half-uncle of the proband (II10) shows some shared phenotypic features, including syndactyly. Incontinentia pigmenti (IP) was the clinical diagnosis for the proband, but this is usually lethal prenatally in males (OMIM), whereas in this family affected males are known to have survived for 8–15 days. Facial clefting is a feature of the family phenotype (Fig. 1), although a case of IP associated with bilateral CLP was described as 'unique' (26). Familial IP is a rare condition arising approximately in 1 of 50,000 newborns (27), and the most conspicuous phenotypic feature is a progressive skin pigmentation abnormality resulting in linear or hypopigmented patches. However, the phenotypic expression is highly variable. Hadj-Rabia et al. (28) studied the phenotypes of 40 IP cases of which 7 had been misdiagnosed because of similarity to other pigmentation disorders. IP is an X-linked dominant disorder that causes skewed X-inactivation in female patients but affected male IP conceptuses typically fail to survive the second trimester.

Exome sequencing of the proband (III10, Table 2 and Table S3) reveals 15 rare and novel variants in different genes that include *IFT140* (combined score rank 1), *RPGRIP1L* (rank 2), *IDUA* (rank 5) and *IKBKG* (rank 8). Both variants in *IFT140* and *IDUA* are known in dbSNP135 and have not previously been linked to clinical phenotypes. The second ranked variant in Table 2 is a heterozygous stop gain in the *RPGRIP1L* gene on chromosome 16. This variant is classed as damaging by most predictive metrics, including a very high GERP++ score of 5.87 suggesting a highly deleterious variant. Homozygous and compound heterozygous mutations in *RPGRIP1L* are associated with Joubert syndrome and Meckel syndrome (29). However, there is no evidence thus far that heterozygous variants in this gene are pathogenic and the patient's phenotype does not overlap characteristic features of these syndromes. However, the patient also carries the E57K (Glu57Lys) missense mutation in exon 2 of the *IKBKG* gene on chromosome X. Smahi et al. (30) showed that cells of IP patients lack NF-kB function due to mutations in the *IKBKG* (NEMO) gene (NF-kB essential modulator). Swaroop et al. (27) identified 277 patients with *IKBKG* mutations from a sample of 357 unrelated patients. A total of 248 of the 277 patients (90%) exhibited an identical deletion that eliminates exons 4–10. Their study also revealed that 29 of 357 patients had smaller mutations including microdeletions, substitutions and duplications. The Glu57Lys mutation found here is a substitution also reported by Swaroop et al. as only one of the two (of 29) small mutations that changed the amino acid identity. Swaroop et al. also identified *IKBKG* polymorphisms in unaffected members of IP pedigrees but all were in untranslated or intronic regions suggesting that an undisrupted *IKBKG* sequence is usually essential for normal function. Conte et al. (31) point out that IP is most frequently a sporadic condition with 65% of *IKBKG* mutations occurring *de novo*. However, the missense mutation identified here was also reported in a familial case by Swaroop et al., (27). Conte et al. (31) consider genotype and phenotype correlations in IP and recognize that the clinical phenotype is highly variable, and there is an expectation that some missense mutations might only slightly affect *IKBKG* function. The missense E57K (Glu57Lys) mutation we have identified here is described as presenting a 'milder' IP phenotype (27, 31), although Swaroop et al. indicate there is no evidence that it is compatible with male survival. The family presented here establishes that this missense mutation is compatible with male survival but only just beyond full term whereas the majority of *IKBKG* mutations

do not permit survival beyond the second trimester. The pedigree also features a phenotypically normal transmitting mother (II3) and a female post-natal death at 8 days (III6). Differences in X-inactivation are known to produce variation in the degree of clinical expression and this variability may explain the diversity of female phenotypes in this pedigree. We exome-sequenced the half-uncle of the proband (II10), who also shows a facial clefting and a syndactyly phenotype. As expected, he does not carry the *IKBKG* mutation that is associated with male death. Assuming the shared syndactyly features have a common genetic basis, the heterozygous stop gain in the *RPGRIP1L* gene (shared by both individuals) is a possible cause. However, this is speculative in the absence of evidence for clinical phenotypes arising from heterozygous mutations in this gene and functional assays may be required to establish causality.

Family CLP3

The family (Fig. 2) shows a complex pattern with very variable penetrance (including unaffected presumed transmitting relatives) with unilateral and bilateral CLP and cleft palate. Unlike other members of the pedigree the proband, (Table 1), shows micrognathia and, as a result, needed ventilator support in the ICU at birth and was treated with oral surgery (mandibuloplasty). Pierre Robin syndrome was diagnosed based on paediatric clinical history of respiratory failure as a consequence of micrognathia. Physical examination did not reveal congenital heart abnormalities or developmental delay to suggest 22q11 deletion. The Pierre Robin syndrome is characterized by cleft palate and micrognathia resulting in glossoptosis arising when the tongue obstructs the airway causing feeding and respiratory problems in the early post-natal period (32). It represents a causally heterogeneous series of events (micrognathia causing glossoptosis preventing palatal shelves to fuse) and is often referred to as the Pierre Robin sequence (PRS). Tan et al. (33) describe the highly heterogeneous nature of genetic factors that underlie the PRS phenotype. Mutations in the *SOX9* gene are known to explain a proportion of PRS cases but a number of other genes have been implicated (33).

Exome analysis (Table 2) identifies a novel p.V202I missense mutation in the *IRF6* gene (c.G604A) in exon 5 shared by all three affected relatives tested. This variant is damaging by most predictive metrics (including the highest GERP++ score of 6.17) and has the third highest rank in the table for the combined scores. *IRF6* mutations underlie VWS and 80% of the causal mutations are found in exons 3, 4, 7 and 9, whereas mutations underlying PPS are more frequent in exon 4 (3). Wu-Chou et al. (34) found exon 5 mutations in 2 of 13 VWS cases. However, the CLP3 family exhibits variable PRS features and lacks lip pits that are characteristic of VWS. Nikopensius et al. (35) were able to show that mutations in *IRF6* also underlie susceptibility to some nonsyndromic CLP cases, so mutations in this gene are associated with considerable phenotypic heterogeneity. Interestingly Tan et al. highlight the evidence for shared genetic aetiology between dental anomalies and clefts. Vieira (36) describes positive associations of clefting with hypodontia with *IRF6*, although the role of this gene in PRS has not been previously described. Sanger sequencing (Table 3) confirmed carrier status for unaffected (transmitting) relatives II14 and III1.

Discussion

The three families considered present a diversity of phenotypes for which molecular diagnosis by exome sequencing has achieved greater understanding of the respective conditions. Family CLP2 presents an atypical IP phenotype that we have established arises through a rare missense mutation in the *IKBKG* gene. Missense mutations in this gene might be linked to less severe phenotypes (37), but there is limited information. As the vast majority of hemizygous male mutation carriers die *in utero* survival to full term, as seen in this family, is atypical. However, Swaroop et al. (38) report two duplication mutations in a cytosine tract in exon 10 of *IKBKG* also associated with male survival. Another feature of the IP phenotypic spectrum is associated CLP that has been has been considered rare with few reports until recently but including unilateral CLP phenotypes (39) and bilateral (26). However, Minic et al. (40) reviewed 1286 published IP cases and found that from cases classified between 1993–2010, 1.55% of IP patients had cleft palate. Clefts of the lip and palate were approximately ten times more frequent in IP patients than in the general population. The authors suggest that cleft palate could be classified as a minor criterion for IP diagnosis.

The CLP3 family provides further insights into the spectrum of phenotypes associated with mutations in the *IRF6* gene. Here, the proband was clinically defined as displaying a PRS phenotype. Exome analysis and Sanger sequencing confirm the presence of a novel damaging *IRF6* mutation in affected and carrier individuals in the pedigree linking mutations in this gene to PRS for the first time.

Yang et al. (41) evaluated WES as a strategy for diagnosis of Mendelian disorders. They presented data from the first 250 unselected consecutive probands from referring physicians who ordered WES. They identified 86 mutant alleles that were highly likely to be causal in 62 of the patients thereby achieving a 25% molecular diagnostic rate amongst cases not molecularly diagnosed by conventional tests such as karyotyping. This rate is higher than the positive rates of other genetic tests including karyotype analysis (5–15%), microarray analysis (15–20%) and Sanger sequencing for single genes (on average lower but gene/phenotype dependent). The 25% diagnostic rate is likely to increase in future through the rapid growth in recognition of clinically relevant genes and, for example, improved resolution of copy-number variation. It is also notable that molecular diagnoses for 25% of the 62 patients were based on disease gene discoveries made in the previous 2 years, suggesting that many Mendelian disease genes remain undiscovered. Interestingly, Yang et al. highlight four patients who were shown by exome sequencing to have

Noonan spectrum disorder. For one patient, a sequencing panel of Noonan-related genes failed to identify the causal variant. WES showed that a mutation in the *CBL* gene was causal, but this had not been used in the previous gene panel as it has only been relatively recently linked to the disorder. The three other patients had atypical clinical phenotypes, and Noonan spectrum disorders were not in the clinical diagnosis. As sequencing resolves more patients with atypical presentation of known genetic diseases, the spectrum of associated phenotypes will expand.

Difficulties in interpretation plague the analysis of exome data. Here, we adopted a strategy of interrogating all genes previously associated with any form of CLP phenotype. This produced a comprehensive list of more than 800 genes. Although this is a large number of genes to consider, this strategy still carries the risk of missing causal variants in novel genes not previously linked to these disorders. The very large number of variants revealed by sequencing presents difficulties for adequately excluding neutral variants, and querying the most likely subset of genes is one route to reduce this complexity. Filtering of variants and ranking those most likely to underlie a phenotype are facilitated by a range of predictive scores of which a number are presented in Table 2 and Table S3. The four scores used to produce a single rank are closely correlated, although the composite measures are not independent in every case. Improvements in predictive tools, facilitated by recognition of more disease variants and definition of disease pathways, along with development of functional assays in some cases, will enable further advances in interpretation of these complex data sets.

The limitations of exome sequencing include lack of coverage of non-coding regions that may contain regulatory variants influencing disease. Technical limitations can include poor coverage of some coding regions of the genome excluding some causal coding variants. Approximately 5% of the coding regions may be in this category (41). Whole-genome sequencing offers resolution to these coverage issues but at high cost and with considerable challenges for data management and analysis, making this less practical at this time.

The Colombian families include cases from remote areas with a high degree of consanguinity and extended pedigrees presenting interesting possibilities for sequencing studies. Analysis of pedigree CLP2 indicates that two genes are involved in defining the partly overlapping phenotypic features of the proband and half-uncle. The evidence suggests a possible role for heterozygous stop gain in the *RPGRIP1L* gene, alongside the X-linked *IKBKG* gene. Future analyses in this and other populations may provide insights into possible involvement of this gene in related phenotypic contexts. Exome analyses in these three Colombian families have identified novel variants involved in the clinical syndromes alongside new genotype–phenotype relationships. The future potential for genetic dissection of syndromic disease using NGS technology is clearly demonstrated.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Acknowledgements

The authors acknowledge the support for this research from the Newlife Foundation for Disabled Children. They would also like to gratefully acknowledge Operation Smile, Colombia, and the patients who participated in the study.

References

1. Leslie EJ, Marazita ML. Genetics of cleft lip and cleft palate. Am J Med Genet C Semin Med Genet 2013: 163 (4): 246–258.
2. Kondo S, Schutte C, Richardson RJ et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat Genet 2002: 32 (2): 285–289.
3. de Ferreira Lima RLL, Hoper SA, Ghassibe M et al. Prevalence and nonrandom distribution of exonic mutations in interferon regulatory factor 6 in 307 families with Van der Woude syndrome and 37 families with popliteal pterygium syndrome. Genet Med 2009: 11 (4): 241–247.
4. Blanton SH, Cortez A, Stal S, Mulliken JB, Finnell RH, Hecht JT. Variation in IRF6 contributes to nonsyndromic cleft lip and palate. Am J Med Genet A 2005: 137 (3): 259–262.
5. Altmüller J, Budde BS, Nürnberg P. Enrichment of target sequences for next-generation sequencing applications in research and diagnostics. Biol Chem 2014: 395 (2): 231–237.
6. Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med 2013: 5 (9): 89.
7. Collins A, Arias L, Pengelly R, Martínez J, Briceño I, Ennis S. The potential for next generation sequencing to characterise the genetic variation underlying nonsyndromic cleft lip and palate phenotypes. OA Genetics 2013: 1 (1): 10.
8. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011: 32 (8): 894–899.
9. Liu X, Jian X, Boerwinkle E. dbNSFP v2. 0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 2013: 34 (9): E2393–E2402.
10. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 2012: 40 (7): e53.
11. Li MX, Kwan JS, Bao SY et al. Predicting Mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet 2013: 9 (1): e1003143.
12. Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. Berlin, Heidelberg: Springer-Verlag, 2006: 190–205.
13. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003: 31 (13): 3812–3814.
14. Adzhubei IA, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations. Nat Methods 2010: 7: 248–249.
15. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009: 19 (9): 1553–1561.
16. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010: 7 (8): 575–576.
17. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 2010: 6 (12): e1001025.
18. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 2011: 12 (9): 628–640.
19. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974: 185: 862–864.
20. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014: 46 (3): 310–315.
21. Fuentes Fajardo KV, Adams D, Mason CE et al. Detecting false-positive signals in exome sequencing. Hum Mutat 2012: 33 (4): 609–613.

22. Schlieve T, Almusa M, Miloro M, Kolokythas A. Temporomandibular joint replacement for ankylosis correction in Nager syndrome: case report and review of the literature. J Oral Maxillofac Surg 2012: 70 (3): 616–625.
23. Bernier FP, Caluseriu O, Ng S et al. Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes nager syndrome. Am J Hum Genet 2012: 90 (5): 925–933.
24. Petit F, Escande F, Jourdain A-S et al. Nager syndrome: confirmation of SF3B4 haploinsufficiency as the major cause. Clin Genet 2014: 86 (3): 246–251.
25. Czeschik JC, Voigt C, Alanay Y et al. Clinical and mutation data in 12 patients with the clinical diagnosis of Nager syndrome. Hum Genet 2013: 132 (8): 885–898.
26. Yell JA, Walshe M, Desai SN. Incontinentia pigmenti associated with bilateral cleft lip and palate. Clin Exp Dermatol 1991: 16 (1): 49–50.
27. Swaroop A, Woffendin H, Jakins T et al. A recurrent deletion in the ubiquitously expressed NEMO (IKK-γ) gene accounts for the vast majority of incontinentia pigmenti mutations. Hum Mol Genet 2001: 10 (19): 2171–2179.
28. Hadj-Rabia S, Froidevaux D, Bodak N et al. Clinical study of 40 cases of incontinentia pigmenti. Arch Dermatol 2003: 139 (9): 1163–1170.
29. Delous M, Baala L, Salomon R et al. The ciliary gene RPGRIP1L is mutated in cerebello-oculo-renal syndrome (Joubert syndrome type B) and Meckel syndrome. Nat Genet 2007: 39: 875–881.
30. Smahi A, Courtois G, Vabres P et al. The International Incontinentia Pigmenti (IP) Consortium. Genomic rearrangement in NEMO impairs NF-kappaB activation and is a cause of incontinentia pigmenti. Nature 2000: 405 (6785): 466–472.
31. Conte MI, Pescatore A, Paciolla M et al. Insight into IKBKG/NEMO locus: report of new mutations and complex genomic rearrangements leading to incontinentia pigmenti disease. Hum Mutat 2014: 35 (2): 165–177.
32. Jakobsen LP, Knudsen MA, Lespinasse J et al. The genetic basis of the Pierre Robin Sequence. Cleft Palate Craniofac J 2006: 43 (2): 155–159.
33. Tan TY, Kilpatrick N, Farlie PG. Developmental and genetic perspectives on Pierre Robin sequence. Am J Med Genet C Semin Med Genet 2013: 163 (4): 295–305.
34. Wu-Chou YH, Lo LJ, Chen KTP, Chang CSF, Chen YR. A combined targeted mutation analysis of IRF6 gene would be useful in the first screening of oral facial clefts. BMC Med Genet 2013: 14 (1): 37.
35. Nikopensius T, Jagomägi T, Krjutškov K et al. Genetic variants in COL2A1, COL11A2, and IRF6 contribute risk to nonsyndromic cleft palate. Birth Defects Res A Clin Mol Teratol 2010: 88 (9): 748–756.
36. Vieira AR. Unraveling human cleft lip and palate research. J Dent Res 2008: 87 (2): 119–125.
37. Fusco F, Bardaro T, Fimiani G et al. Molecular analysis of the genetic defect in a large cohort of IP patients and identification of novel NEMO mutations interfering with NF-kappaB activation. Hum Mol Genet 2004: 13: 1763–1773.
38. Swaroop A, Courtois G, Rajkovic A et al. Atypical forms of incontinentia pigmenti in male individuals result from mutations of a cytosine tract in exon 10 of NEMO(IKK-γ). Am J Hum Genet 2001: 68 (3): 765–771.
39. Stewart RE, Funderburk S, Setoguchi Y. A malformation complex of ectrodactyly, clefting and hypomelanosis of Ito (incontinentia pigmenti achromians). Cleft Palate J 1979: 16: 358–362.
40. Minić S, Trpinac D, Gabriel H, Gencik M, Obradović M. Dental and oral anomalies in incontinentia pigmenti: a systematic review. Clin Oral Investig 2013: 17 (1): 1–8.
41. Yang Y, Muzny DM, Reid JG et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. N Engl J Med 2013: 369 (16): 1502–1511.

Genetic dissection of early-onset breast cancer and other genetic diseases

Molecular Genetics & Genomic Medicine

Open Access

ORIGINAL ARTICLE

Quantifying the cumulative effect of low-penetrance genetic variants on breast cancer risk

Conor Smyth[1], Iva Špakulová[1], Owen Cotton-Barratt[1], Sajjad Rafiq[2], William Tapper[3], Rosanna Upstill-Goddard[3], John L. Hopper[4], Enes Makalic[4], Daniel F. Schmidt[4], Miroslav Kapuscinski[4], Jörg Fliege[1], Andrew Collins[3], Jacek Brodzki[1], Diana M. Eccles[2] & Ben D. MacArthur[1,5,6]

[1]Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom
[2]Cancer Sciences Academic Unit and University of Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton and University Hospital Southampton Foundation Trust, Tremona Road, Southampton, SO16 6YA, United Kingdom
[3]Human Genetics, Faculty of Medicine, University of Southampton, Tremona Road, Southampton, SO16 6YA, United Kingdom
[4]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, School of Population and Global Health, The University of Melbourne, Carlton, Victoria, Australia
[5]Human Development and Health, Faculty of Medicine, University of Southampton, Tremona Road, Southampton, SO16 6YA, United Kingdom
[6]Institute for Life Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom

Keywords
breast cancer, polygenic disorder, information theory

Correspondence
Ben D. MacArthur, Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ United Kingdom. Tel: +44 (0)23 8059 4255; Fax: +44 (0)23 8059 3858; E-mail: B.D.MacArthur@soton.ac.uk

Funding Information
Engineering and Physical Sciences Research Council grant EP/I016945/1.

Received: 27 August 2014; Revised: 28 November 2014; Accepted: 4 December 2014

Molecular Genetics & Genomic Medicine **2015; 3(3): 182–188**

doi: 10.1002/mgg3.129

Abstract

Many common diseases have a complex genetic basis in which large numbers of genetic variations combine with environmental factors to determine risk. However, quantifying such polygenic effects has been challenging. In order to address these difficulties we developed a global measure of the information content of an individual's genome relative to a reference population, which may be used to assess differences in global genome structure between cases and appropriate controls. Informally this measure, which we call relative genome information (RGI), quantifies the relative "disorder" of an individual's genome. In order to test its ability to predict disease risk we used RGI to compare single-nucleotide polymorphism genotypes from two independent samples of women with early-onset breast cancer with three independent sets of controls. We found that RGI was significantly elevated in both sets of breast cancer cases in comparison with all three sets of controls, with disease risk rising sharply with RGI. Furthermore, these differences are not due to associations with common variants at a small number of disease-associated loci, but rather are due to the combined associations of thousands of markers distributed throughout the genome. Our results indicate that the information content of an individual's genome may be used to measure the risk of a complex disease, and suggest that early-onset breast cancer has a strongly polygenic component.

Introduction

Accumulating evidence suggests that many common diseases have a polygenic basis, in which large numbers of genetic variations combine with environmental and lifestyle factors to determine risk (Khoury et al. 2013). While genome-wide association studies (GWAS), and more recently exome and whole-genome sequencing projects, have found hundreds of genetic variants associated with disease, the ability to predict susceptibility from these associations is generally low because the contribution of individual variants to risk is often very modest. In the case of breast cancer, published GWAS have identified markers (single-nucleotide polymorphisms, or SNPs) in more than 70 independent regions (loci), the majority with odd ratios less than 1.1 (Bogdanova et al. 2013). Collectively these loci explain, in the statistical but not causative sense, approximately 15% of the familial relative risk which, when combined with the approximately 21% attributed to moderate- to high-penetrance variants (typically very rare mutations) in a dozen or so susceptibility genes, leaves almost two-thirds of the familial basis of the disease unaccounted for (Antoniou and Easton 2006; Bogdanova et al. 2013). It is likely that additional genes

that explain a proportion of this missing heritability will be found using both whole-exome/genome and candidate gene sequencing of familial and young-onset cases, where the genetic component of risk is likely to be greatest (Hopper and Carlin 1992; Manolio et al. 2009; Park et al. 2012; Ruark et al. 2013; Akbari et al. 2014). Nevertheless, our current understanding of the genetic basis of breast cancer is still far from complete.

While most studies to date have focussed on individual genes or gene mutations and their contribution to disease, there has been limited effort to quantify the cumulative effect of variation across the whole genome on disease risk. This is partly due to the historical lack of sufficient data to appropriately quantify normal genomic variation within control populations, and the absence of the statistical techniques needed to analyze such large-scale variation. However, recent years have seen concerted effort to collect and collate the large numbers of genomes (for example the UK Department of Health's 100K initiative http://www.genomicsengland.co.uk) and there is now a need to develop the accompanying methodological tools to assess genomic variation (Yang et al. 2011; Zhou et al. 2013).

In order to begin to address this issue we describe here a measure of the extent to which a set of case genomes differ from a set of control genomes in their global structure. Our method uses ideas from information theory to provide a measure of the information content of an individual's genome with reference to a control population. The procedure first uses the reference population to estimate a probability measure on the space of all genomes, and then uses the estimated probability measure to assess how unusual an individual's genome is with respect to the reference population, as quantified by its self-information (also known in information theory as "surprisal") (Cover and Thomas 1991). Formally, the resulting measure, which we refer to as the relative genome information (RGI), is the amount of information, measured in bits, required to specify the observed genome with respect to the unique encoding that minimizes the expected number of bits required to specify the genome of an individual drawn at random from the reference population. Informally, the RGI measures how unusual a genome is with respect to the reference population or, since we construct an information-theoretic measure closely related to the Shannon entropy, how "disordered" it is. Thus, someone with a higher RGI has a more unusual genome, either having less common alleles more often than expected, or having some particularly rare alleles. By contrast a lower RGI corresponds to having more common alleles more often, and therefore a less surprising genome.

We hypothesized that global measures of genome variation, such as RGI, might quantify the polygenic basis of complex diseases more completely than GWAS analyses that seek to find statistically significant associations of particular markers with disease. In order to test this hypothesis we compared the RGI of two independent samples of women with early-onset breast cancer genotyped for SNPs relative to three independent samples of unaffected controls.

Methods

Data sets and quality control

SNP genotypes obtained from blood samples from the following three independent studies were considered: (*i*) The Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) cohort (Eccles et al. 2007). The POSH cohort consists of approximately 3000 women aged 40 years or younger at breast cancer diagnosis from which 574 cases were genotyped on the Illumina (San Diego, CA, USA) 660-Quad SNP array. Genotyping was conducted in two batches at the Mayo Clinic, Rochester, MN (274 samples) and the Genome Institute of Singapore, National University of Singapore (300 samples). A total of 536 samples that passed quality control filters were considered in this study (Rafiq et al. 2013). (*ii*) The Wellcome Trust Case Control Consortium (WTCCC, http://www.wtccc.org.uk/). The WTCCC consists of two independent sets of disease-free controls: 2699 individuals from the 1958 British Birth Cohort and 2501 individuals from the UK National Blood Service (NBS) Collection. Genotyping of both sets was conducted using the Illumina 1.2M chip. (*iii*) The Australian Breast Cancer Family Study (ABCFS) (McCredie et al. 1998; Dite et al. 2003). Cases were a subset of 204 of women aged 40 years or younger at breast cancer diagnosis from the ABCFS; controls were 287 unaffected women aged 40 years and older from the Australian Mammographic Density Twins and Sisters Study (Odefrey et al. 2010). Genotyping was conducted at the Australian Genome Research Facility using the Illumina 610-Quad SNP array. A summary of all data sets is given in Table 1.

Only autosomes were considered and SNPs were excluded from each data set if they failed any of the following quality control filters: minor allele frequencies <1%; genotyping call rate <99%; significant deviation from Hardy–Weinberg equilibrium ($P < 0.0001$). All quality control filters were implemented using the software package PLINK (Purcell et al. 2007). In total, approximately 475,000 SNPs were genotyped in all five data sets. When comparing data sets and computing RGI only these shared SNPs were considered.

Individuals with evidence of ethnic admixture were excluded by performing multi-dimensional scaling (MDS) analysis. Firstly, linkage disequilibrium (LD)-based pruning

($r^2 > 0.5$) of genotypes was undertaken using PLINK to generate a reduced set of approximately independent SNPs. In total there were approximately 133,000 LD-pruned SNPs common to all samples. The HapMap data for the African, Asian, and Caucasian populations (Gibbs et al. 2003) were then used to provide reference population genotypes against which the genotype data of the cases and controls were compared (Fig. 1A). We identified eight POSH and ten ABCFS samples that showed evidence of mixed ethnicity that did not cluster well with the HapMap Caucasian population reference sample, and these were excluded from further analysis. Since they only form a small subset of the total samples considered, the conclusions of our analysis do not differ without removal of these samples. However, we expect that, in general, significant ethnic variation within either the case or control populations would confound the results of our method.

Quantifying relative genome information

Let L denote a set of locations in the genome (loci), and let $\Lambda = \{A, C, G, T\}$ be the alphabet of possible alleles at each locus $l \in L$. Let $\Pi_l(\lambda, \mu)$ denote the likelihood of finding the unordered allele pair $(\lambda, \mu) \in \Lambda \times \Lambda$ at locus $l \in L$ in

Table 1. Overview of case and control data sets.

Data set	Size	Size after QC	Gender	Ethnicity	Genotyping platform
ABCFS cases	204	201	Female	Caucasian[1]	Illumina 610-Quad SNP array
POSH cases	574	536	Female	Caucasian[1]	Illumina 660-Quad SNP array
ABCFS control	287	280	Female	Caucasian[1]	Illumina 610-Quad SNP array
NBS control	2501	2501	Both	Caucasian	Illumina 1.2M chip
1958 control	2699	2699	Both	Caucasian	Illumina 1.2M chip

[1]post-QC.

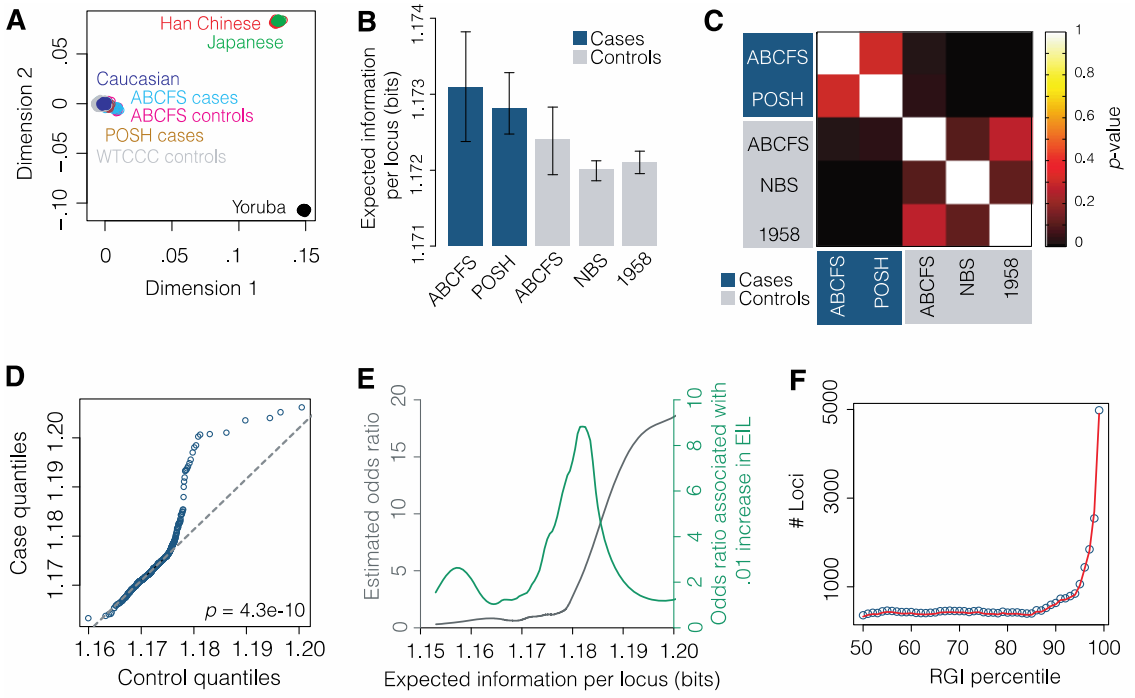


Figure 1. Breast cancer risk is associated with increased genome-wide disorder. (A) Multidimensional scaling plot of all samples and HapMap2 populations genotyped for ~133,000 SNPs. (B) Expected information per locus (EIL) for each of the different data sets. Median ± 95% confidence intervals are shown. (C) Matrix of FDR adjusted *P*-values for comparisons of medians (two-sided Wilcoxon rank-sum test). (D) Q-Q plot of EIL in cases versus controls. *P*-value from a two-sample Kolmogorov–Smirnov test is shown. (E) Estimated odds ratio as a function of EIL. (F) Median number of loci required to account for the differences in EIL observed between cases and controls by percentile. 95% confidence intervals are within the markers, so are not shown.

the reference population and let Π be the product measure of Π_l over all $l \in L$. Thus, Λ^{2L} denotes the space of all possible genomes, and Π represents the probability measure on Λ^{2L}. Now let $X \in \Lambda^{2L}$ be a genome with allele pair $X_t \in \Lambda \times \Lambda$ at locus $l \in L$. We define the relative local information (RLI) $\mathrm{I}_l(X_l) = -\log_2 \Pi_l(X_l)$ at each locus $l \in L$ in the genome X and the RGI $\mathrm{I}(X) = \sum_{l \epsilon L} \mathrm{I}_l(X_l)$ for each genome X of interest. For the purposes of comparison it is also convenient to normalize the RGI by n, the number of loci genotyped, to give the expected information per locus (EIL), $\mathbb{E}_n(\mathrm{I}_l) = \frac{1}{n}\sum_{l \epsilon L} \mathrm{I}_l(X_l)$. When comparing sequences of the same length the EIL and RGI are equivalent up to a normalizing factor. However, by normalizing by the number of loci sampled, the EIL allows comparison of relative information content of sequences of different lengths (for instance, comparison of relative information content of different chromosomes). The RLI is the natural information-theoretic measure of the "surprisal" of observing allele pair $X_l \in \Lambda \times \Lambda$ at locus $l \in L$ given the probability measure Π_l (Cover and Thomas 1991). Similarly, the RGI is the natural information-theoretic measure of the "surprisal" of observing the genome X, given the probability measure Π.

In practice Π is not known a priori and must be estimated from an appropriate reference sample of similar ethnic background to that of the cases. Here, we estimated Π using the WTCCC 1958 birth cohort since it was the largest reference sample available. In all calculations, Π_l was estimated for each locus $l \in L$ using all available genotypes in the reference population at that locus. Once Π had been estimated, the RGI was calculated for each genome in each of the remaining four (test) samples (POSH cases, ABCFS cases, ABCFS controls, NBS controls). The two additional independent sets of controls (ABCFS and NBS) were included in order to assess the robustness of the approximation of the background probability measure Π from the 1958 control cohort alone. For each of the four test samples, missing genotype data at each locus $l \in L$ were assigned the expected value of Π_l (i.e., the Shannon entropy $-\sum_{X_l} \Pi_l(X_l) \log_2 \Pi_l(X_l)$ of Π_l). This method of imputation minimizes the influence of missing data on the calculation of RGI. We also conducted all calculations using only those loci for which there were no missing readings in any of the data sets, and results obtained with and without imputation did not differ qualitatively. A brief worked example illustrating how Π was estimated, and the RLI and RGI were calculated, is given in the Data S1. Estimation of RGI for N case genomes takes $O(n(m + N))$ computational time, where n is the number of loci and m is the number of genomes in the control population, and can be conducted on a desktop PC for moderate sample sizes (thousands of samples and hundreds of thousands of genotyped loci).

Statistical analysis

All analysis was conducted in *R* and Matlab (Natick, MA, USA) using custom written scripts. The association between EIL and disease odds was estimated using a logistic generalized additive model (Hastie et al. 2009). Tests for significant differences between groups were assessed using Wilcoxon rank-sum tests (two-sided tests were used when testing the null hypothesis of no difference in EIL between cases and controls against the alternative hypothesis that EIL differs in cases and controls; one-sided tests were used when testing the null hypothesis of no difference in EIL between cases and controls against the alternative hypothesis that EIL is raised in cases). All *P*-values were false-discovery rate (FDR) adjusted using the Benjamini and Hochberg (1995) procedure.

Results

We did not observe any difference in EIL (RGI normalized by the number of loci genotyped, EIL) between the three different control sets (1958, NBS and ABCFS controls) indicating that the background measure Π was reliably estimated; similarly, no difference in EIL between the POSH and ABCFS cases was observed (Fig. 1B and C). However, EIL was significantly higher in both the POSH and ABCFS cases than the three sets of reference controls (FDR adjusted $P < 0.01$, two-sided Wilcoxon rank-sum test) (Fig. 1B and C). Since significant differences within case and control sets were not observed, we amalgamated samples to form one case set (consisting of the ABCFS and POSH cases) and one control set (consisting of the ABCFS, NBS and 1958 controls) for further analysis. Comparison of the distribution of RGI in amalgamated case set and amalgamated control set revealed significant differences in distribution structure ($P = 4.3 \times 10^{-10}$, two-sample Kolmogorov–Smirnov test) with the case distribution having a substantially heavier tail than the control distribution, indicating a greater proportion of samples with higher EIL (Fig. 1D). To investigate further we conducted regression using a logistic generalized additive model (Hastie et al. 2009) in order to estimate the relationship between disease odds and EIL (Fig. 1E). Consistent with the heavy-tailed nature of the case distribution we observed a strong positive association between odds ratio and EIL. In particular, the odds ratio increased sharply for EIL above 1.75, with the highest percentile EIL (above 1.183) having an odds ratio greater than 12 by comparison with the lower 99% ($P < 1 \times 10^{-16}$, Fisher's exact test). These results indicate that EIL is significantly elevated in breast cancer cases, with the highest percentiles EIL conferring a substantially increased risk.

In order to investigate the genetic basis for these observations we sought to assess whether the differences observed were associated with particular genomic loci or SNP annotations. We began by estimating the number of loci required to account for observed differences at each percentile using random resampling with replacement (1×10^4 times) from the case genomes until the required difference was achieved. Differences in median EIL between cases and controls were found to be due to contributions from an estimated 327 distinct loci (median, 95% confidence intervals [306, 349]) (Fig. 1F). The expected number of loci required to account for differences between cases and controls sharply increased with percentile, with differences in the 99th percentile (which conferred the greatest disease risk) requiring an estimated 4954 loci (median, 95% confidence intervals [4921, 5000]) (Fig. 1F). These results indicate that observed differences in EIL are not due to high-penetrance variations at a small number of disease-associated loci, but rather are due to widespread variation at thousands of genomic loci.

In order to investigate this further we assessed the EIL on individual chromosomes. We found that EIL was consistently elevated in the cases by comparison with the controls on 19 of 22 chromosomes (Fig. 2A), and significantly so on 12 of 22 chromosomes (FDR adjusted $P < 0.05$, one-sided Wilcoxon rank-sum test), indicating that differences in EIL are distributed throughout the genome. We also observed notable variations in EIL by SNP annotation, with the lowest EIL (and therefore the least variation within the samples) occurring in the 5′/3′ untranslated and exonic regions, and the highest EIL (and therefore the greatest variation within the samples) occurring in the intergenic regions (Fig. 2B). This is consistent with previous assessment of relative mutation rates and suggests that 5′/3′ UTRs and exonic regions are subject to stronger negative selection than intergenic regions, in accordance with their phenotypic importance (Ward and Kellis 2012a,b; Khurana et al. 2013). In all annotation categories, we again observed a significant increase in EIL in the cases (FDR adjusted $P < 0.05$, one-sided Wilcoxon rank-sum test) (Fig. 2B). These results indicate observed differences in EIL are not localized to distinct regions of the genome (either chromosomes or SNP annotations) but rather are due to widespread variation distributed throughout the genome.

Discussion

Genetic factors that contribute to breast cancer risk range from rare highly penetrant functionally deleterious

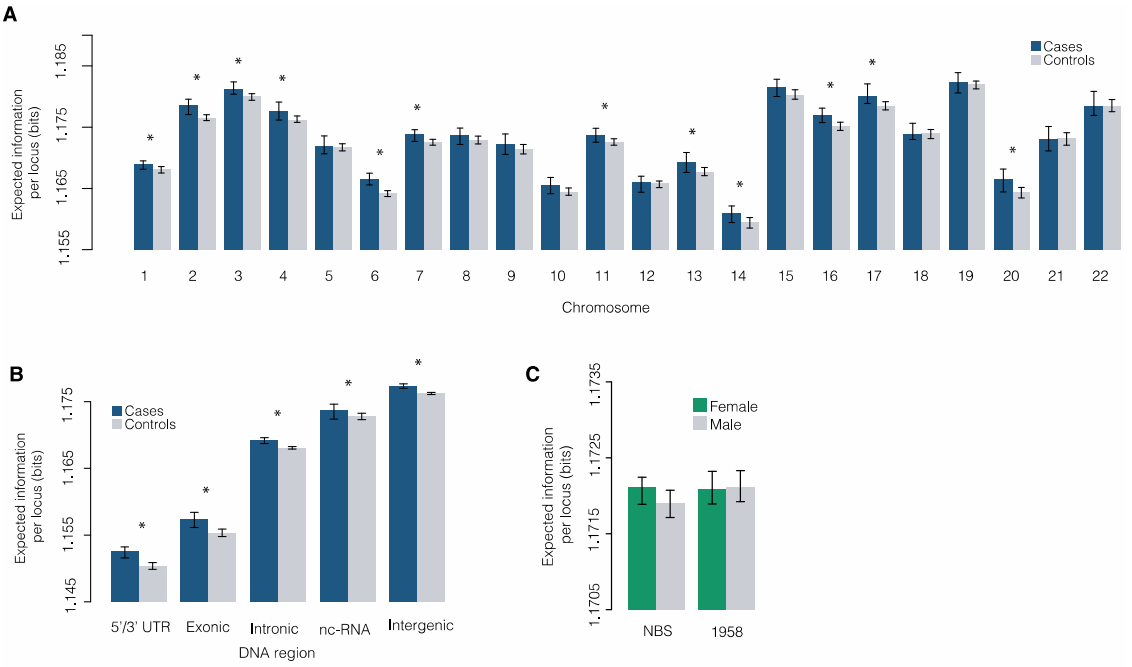


Figure 2. Disorder is not localized to specific regions of the genome. (A) Expected information per locus (EIL) by chromosome. (B) EIL by SNP annotation. (C) EIL in males and females in the controls. In all panels, median ± 95% confidence intervals are shown. Stars indicate significant changes at FDR adjusted $P < 0.05$ by one-sided Wilcoxon rank-sum test.

mutations in genes like BRCA1 and BRCA2 to genetic variants that are relatively frequently observed and are associated with small increases in risk (Mavaddat et al. 2010). However, we do not yet have a complete understanding of the genetic basis of breast cancer. Much of the missing heritability may be either very rare highly penetrant genes not currently known or, more likely, hundreds to thousands of rare genetic variants with small effect sizes. Current approaches to discovering low-penetrance genetic susceptibility alleles using GWAS rely on risk alleles being relatively common in the population. Even with case–control studies involving hundreds of thousands of individuals, identifying all the genes responsible for susceptibility is likely to prove difficult if important effects relate to the accumulation of rare low-penetrance alleles. By comparing individual genetic sequences with that expected from a control population our approach assesses the cumulative effect of low-penetrance alleles on disease risk. Our results suggest that such cumulative effects are a significant component of the missing heritability in breast cancer. Prior to analysis all genotyping data were subjected to stringent quality assurance and we observed no association between sex, sequencing platform, time/place of sequencing and EIL, indicating that poor data quality or variation in genotype due to ethnicity or sex are unlikely to explain our results (Figs. 1B, C, and 2C). Rather, changes in EIL appear to quantify statistically significant differences in allele frequencies between breast cancer cases and controls.

Taken together our analysis indicates that early-onset breast cancer has a strongly polygenic component, involving variation at thousands of markers distributed throughout the genome. Thus, along with assessment of known risk-associated variants, the information content of an individual's genome is likely to be a useful predictor of breast cancer susceptibility. Further analysis of the relationship between global genome structure and disease risk may reveal a similarly polygenic basis for a variety of other complex diseases.

Conflict of Interest

None declared.

References

Akbari, M. R., P. Lepage, B. Rosen, J. McLaughlin, H. Risch, M. Minden, et al. 2014. PPM1D mutations in circulating white blood cells and the risk for ovarian cancer. J. Natl. Cancer I 106: djt323.

Antoniou, A. C., and D. F. Easton. 2006. Models of genetic susceptibility to breast cancer. Oncogene 25:5898–5905.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B Met. 57:289–300.

Bogdanova, N., S. Helbig, and T. Dork. 2013. Hereditary breast cancer: ever more pieces to the polygenic puzzle. Hered. Cancer Clin. Pr. 11:12.

Cover, T. M., and J. A. Thomas. 1991. Elements of information theory. Wiley and Sons, New York, NY.

Dite, G. S., M. A. Jenkins, M. C. Southey, J. S. Hocking, G. G. Giles, M. R. E. McCredie, et al. 2003. Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. J. Natl. Cancer I 95:448–457.

Eccles, D., S. Gerty, P. Simmonds, V. Hammond, S. Ennis, D. G. Altman, et al. 2007. Prospective study of outcomes in sporadic versus hereditary breast cancer (POSH): study protocol. BMC Cancer 7:160.

Gibbs, R. A., J. W. Belmont, P. Hardenbol, T. D. Willis, F. L. Yu, H. M. Yang, et al. 2003. The international hapmap project. Nature 426:789–796.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Springer, New York, NY.

Hopper, J. L., and J. B. Carlin. 1992. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. Am. J. Epidemiol. 136:1138–1147.

Khoury, M. J., A. C. J. W. Janssens, and D. F. Ransohoff. 2013. How can polygenic inheritance be used in population screening for common diseases? Geneti. Med. 15:437–443.

Khurana, E., Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342:1235587.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Mavaddat, N., A. C. Antoniou, D. F. Easton, and M. Garcia-Closas. 2010. Genetic susceptibility to breast cancer. Mol. Oncol. 4:174–191.

McCredie, M. R. E., G. S. Dite, G. G. Giles, and J. L. Hopper. 1998. Breast cancer in Australian women under the age of 40. Cancer Cause Control 9:189–198.

Odefrey, F., J. Stone, L. C. Gurrin, G. B. Byrnes, C. Apicella, G. S. Dite, et al. 2010. Common genetic variants associated with breast cancer and mammographic density measures that predict disease. Cancer Res. 70:1449–1458.

Park, D. J., F. Lesueur, T. Nguyen-Dumont, M. Pertesi, F. Odefrey, F. Hammet, et al. 2012. Rare mutations in XRCC2 increase the risk of breast cancer. Am. J. Hum. Genet. 90:734–739.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Rafiq, S., W. Tapper, A. Collins, S. Khan, I. Politopoulos, S. Gerty, et al. 2013. Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. Cancer Res. 73:1883–1891.

Ruark, E., K. Snape, P. Humburg, C. Loveday, I. Bajrami, R. Brough, et al. 2013. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature 493:406–410.

Ward, L. D., and M. Kellis. 2012a. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337:1675–1678.

Ward, L. D., and M. Kellis. 2012b. Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. 30:1095–1106.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82.

Zhou, X., P. Carbonetto, and M. Stephens. 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 9:e1003264.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Supplementary Materials.

OPEN ACCESS Freely available online

PLOS ONE

A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis

Sajjad Rafiq[1], Sofia Khan[4], William Tapper[1], Andrew Collins[1], Rosanna Upstill-Goddard[1], Susan Gerty[2], Carl Blomqvist[5], Kristiina Aittomäki[6], Fergus J. Couch[3], Jianjun Liu[7], Heli Nevanlinna[4☯], Diana Eccles[8*☯]

1 Genetic Epidemiology and Bioinformatics Research Group, Human Genetics, Faculty of Medicine, University of Southampton, Southampton General Hospital, Hants, United Kingdom, **2** Clinical Trials Unit, Faculty of Medicine, University of Southampton, Hants, United Kingdom, **3** Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, United States of America, **4** Department of Obstetrics and Gynaecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland, **5** Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland, **6** Department of Clinical Genetics, Helsinki University Central Hospital, Helsinki, Finland, **7** Human Genetics, Genome Institute of Singapore, Singapore, **8** Cancer Sciences Division, University of Southampton, School of Medicine, Southampton General Hospital, Hants, United Kingdom

Abstract

Objective: Genome wide association studies (GWAs) of breast cancer mortality have identified few potential associations. The concordance between these studies is unclear. In this study, we used a meta-analysis of two prognostic GWAs and a replication cohort to identify the strongest associations and to evaluate the loci suggested in previous studies. We attempt to identify those SNPs which could impact overall survival irrespective of the age of onset.

Methods: To facilitate the meta-analysis and to refine the association signals, SNPs were imputed using data from the 1000 genomes project. Cox-proportional hazard models were used to estimate hazard ratios (HR) in 536 patients from the POSH cohort (Prospective study of Outcomes in Sporadic versus Hereditary breast cancer) and 805 patients from the HEBCS cohort (Helsinki Breast Cancer Study). These hazard ratios were combined using a Mantel-Haenszel fixed effects meta-analysis and a p-value threshold of 5×10^{-8} was used to determine significance. Replication was performed in 1523 additional patients from the POSH study.

Results: Although no SNPs achieved genome wide significance, three SNPs have significant association in the replication cohort and combined p-values less than 5.6×10^{-6}. These SNPs are; rs421379 which is 556 kb upstream of *ARRDC3* (HR = 1.49, 95% confidence interval (CI) = 1.27–1.75, $P = 1.1\times10^{-6}$), rs12358475 which is between *ECHDC3* and *PROSER2* (HR = 0.75, CI = 0.67–0.85, $P = 1.8\times10^{-6}$), and rs1728400 which is between *LINC00917* and *FOXF1*.

Conclusions: In a genome wide meta-analysis of two independent cohorts from UK and Finland, we identified potential associations at three distinct loci. Phenotypic heterogeneity and relatively small sample sizes may explain the lack of genome wide significant findings. However, the replication at three SNPs in the validation cohort shows promise for future studies in larger cohorts. We did not find strong evidence for concordance between the few associations highlighted by previous GWAs of breast cancer survival and this study.

Citation: Rafiq S, Khan S, Tapper W, Collins A, Upstill-Goddard R, et al. (2014) A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis. PLoS ONE 9(12): e101488. doi:10.1371/journal.pone.0101488

Editor: Xiaoping Miao, MOE Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

Received December 13, 2013; **Accepted** June 9, 2014; **Published** December 19, 2014

Funding: This research was supported in part by National Institutes of Health grant CA128978, and grants from the Komen Foundation for the Cure and the Breast Cancer Research Foundation (BCRF) to F.J.C. The POSH study is supported by Breast Cancer Campaign grant number 2010NovPR62. Funding for the POSH study was also provided by The Wessex Cancer Trust and Cancer Research UK (grant refs A7572, A11699, C22524). The Helsinki study was financially supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (132473), the Finnish Cancer Society, The Nordic Cancer Union, and the Sigrid Juselius Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: d.m.eccles@soton.ac.uk

☯ These authors contributed equally to this work.

Introduction

Although the incidence of breast cancer has been relatively stable since 2003, at 157 new cases per 100,000, it remains the most common cancer in the UK and accounts for 31% of new cancer cases in women. The latest age-standardised survival rate for breast cancer in England is predicted to be 85% at 5 years, falling to 65% at 20 years [1]. Traditionally prognostic information is derived from tumour phenotypic characteristics including tumour size, stage, and grade. These tumour phenotypes and cancer cell surface receptors such as oestrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) are also used to guide treatment. Although the breast cancer survival rate has

Table 1. Clinical characteristics of Study participants from the discovery and replication sets.

Study	Number of breast cancer deaths	Total number of Breast cancer patients	Estrogen Receptor (ER) status-Negative (%)	Average age at Diagnosis (±SD)	Follow-up time in years (±SD)	N-stage	M-stage	T-stage
POSH stage-1 (Discovery)	236	536	370 (69.2%)	35.7 (3.8)	4.1 (2.0)	N0–248 N1–262 NA-26	M0–481 M1–50 NA-5	T1–227 T2–207 T3–20 T4–31 NA-51
HEBCS (Discovery)	301	805	230 (30.0%) NA-39	56.8 (12.4)	7.2 (2.9)	N0–338 N1–446 NA-21	M0–740 M1–57 NA-8	T1–390 T2–304 T3–50 T4–47 NA-14
POSH stage-2 (Replication)	221	1415	362 (23.7%)	35.8 (3.5)	5.2 (1.7)	N0–705 N1–810 NA-8	M0–1506 M1–18 NA-1	T1–692 T2–494 T3–49 T4–34 NA-254

HEBCS: Helsinki Breast Cancer Study; NA = not available, HER2 = Human Epidermal Growth Factor Receptor 2, N-stage = metastasis to lymph node, M-stage = metastasis.
doi:10.1371/journal.pone.0101488.t001

improved, the response to treatment and longevity of patients is often unpredictable even between those with similar tumours and general health. More recently tumour genomic profiling experiments have suggested cancer molecular signatures may give more accurate prognostic information [2–4]. These signatures may predict outcome better than conventional histopathology based risk algorithms but are not in routine clinical use [5].

Familial studies suggest a genetic component for breast cancer prognosis [6,7]. The familial contribution to prognosis may arise as a result of the background genotype affecting acquired tumour characteristics which influence prognosis. Indeed high penetrance predisposition genes which lead to the consistent development of specific breast tumour sub-types have been identified [8,9]. Low penetrance risk SNPs tend to be associated with either ER positive or ER negative breast cancer but often not both [10–14]. In addition there may be pharmacogenomic effects of background genotype on response to cancer treatment. It is anticipated that genome wide association studies (GWAs) with sufficient sample size and genetic coverage may lead to novel insights into common inherited genetic variants which influence prognosis.

In the past few years several GWAs of breast cancer survival have been reported. These studies have had limited success and none of them have identified variants that are associated at genome wide levels of significance [15–19]. While small sample sizes are likely to be of one of the main factors responsible for the modest levels of significance and lack of concordance between the GWAs; small effect sizes, incomplete genetic coverage, and phenotypic heterogeneity could also contribute and need to be addressed.

In this study, we used a meta-analysis to combine evidence from two GWAs consisting of 536 patients from the POSH cohort (Prospective study of Outcomes in Sporadic versus Hereditary breast cancer) and 805 patients from the HEBCS cohort (Helsinki Breast Cancer Study). A further 1523 patients from the POSH cohort were used to validate the most significant SNPs. With a combined sample size of 2864 participants, this analysis has 81% power to detect effects of modest sizes (HR$\geq$1.25, p = 0.05) and with relatively rare SNPs (MAF = 10%). The cohorts used in this analysis have a high incidence of breast cancer related mortality and well documented tumour and treatment data which make them ideal for the purpose of exploring genetic factors influencing prognosis. In addition, these cohorts are similar in terms of their patient recruitment from regional medical centres, duration of prospective follow-up, and documentation of breast cancer related mortality.

Materials and Methods

All participants from POSH and Helsinki gave written informed consent, all were female. The POSH study received approval from the South and West Multi-centre Research Ethics Committee (MREC 00/6/69). The Helsinki breast cancer study received approval from the Ethical Committee of the Departments of Oncology and Obstetrics and Gynaecology, Helsinki University Central Hospital.

Breast cancer patients and genotyping

Breast cancer cases were selected from the POSH study and the Helsinki breast cancer family Study (HEBCS). POSH study participants were diagnosed with invasive breast cancer and were aged forty or younger at diagnosis, the mean age at diagnosis in this cohort is 36 years. Recruitments to the POSH cohort were made between January 2000 and January 2008 from oncology clinics across the UK and the majority (98%) of patients presented symptomatically. The recruitment, data collection and follow up procedures for the POSH study participants are described in detail elsewhere [20].

The HEBCS samples were collected in Helsinki, Finland and are representative of breast cancer case series at the recruitment centre during the collection periods (unselected sporadic and familial cases collected between 1997 and 2004). All of the cases used in the meta-analysis had histopathological and survival data. Detailed information on the patient series and data collection has previously been published [21]. The mean age at diagnosis was 56.8 years.

Stage 1 discovery dataset

In stage-1, 574 participants from the POSH study were selected for the discovery phase of the analysis aimed at hypothesis generation [20]. In keeping with a recent GWAS which identified five new breast cancer susceptibility loci by enriching cases by recruiting individuals with family history of breast cancer [22],

Table 2. SNPs representing the 25 most significant associations in the discovery sets (after excluding SNPs in relative LD≥r^2 of 0.60 and associated with less significant p-value with lead SNP at a locus) and their association estimates (adjusted for ER-status).

Lead SNP	Chr	Position	Alleles	MAF	POSH stage-1 pre meta-analysis HR (95% CI)	HEBCS pre meta-analysis HR (95% CI)	POSH stage-1 and HEBCS meta-analysis HR (95% Confidence Intervals)	POSH stage-1 and HEBCS meta-analysis p-value	Genes
rs12026014	1	39060495	G/A	0.39	0.70 (0.57–0.86)	0.80 (0.68–0.95)	0.76 (0.69–0.84)	2.84×10^{-5}	POU3F1: LOC400750
rs12735344	1	111559848	G/T	0.22	0.71 (0.58–0.89)	0.75 (0.61–0.88)	0.74 (0.67–0.82)	3.46×10^{-5}	CCNT2P1
rs1149185	1	111546531	C/T	0.50	1.51 (1.26–1.80)	1.21 (1.03–1.42)	1.34 (1.22–1.46)	2.14×10^{-6}	C1orf103: TMEM77
rs1578790	1	111575921	G/T	0.50	1.47 (1.23–1.77)	1.16 (0.99–1.36)	1.28 (1.16–1.52)	3.31×10^{-5}	C1orf103: TMEM77
rs11723068	4	7797435	G/A	0.12	1.99 (1.51–2.64)	1.23 (0.99–1.54)	1.48 (1.18–1.99)	9.83×10^{-6}	AFAP1
rs7441398	4	63653135	G/T	0.13	1.43 (1.14–1.79)	1.43 (1.14–1.79)	1.43 (1.16–1.85)	1.22×10^{-5}	LPHN3: LOC644548
rs10457678	6	139122240	A/G	0.24	1.39 (1.15–1.72)	1.30 (1.10–1.55)	1.34 (1.21–1.47)	9.38×10^{-6}	ECT2L
rs1525677	7	110302695	T/C	0.31	1.31 (1.09–1.58)	1.31 (1.10–1.56)	1.31 (1.18–1.44)	2.74×10^{-5}	IMMP2L
rs13274039	8	8111659	A/G	0.28	1.15 (0.96–1.39)	1.41 (1.05–1.47)	1.29 (1.17–1.41)	2.92×10^{-5}	FLJ10661: PRAGMIN
rs12358475	10	11848792	G/A	0.23	0.73 (0.59–0.90)	0.71 (0.58–0.86)	0.72 (0.65–0.80)	6.77×10^{-6}	ECHDC3: C10orf47
rs2921923	10	55662089	A/G	0.49	1.40 (1.16–1.68)	1.26 (1.09–1.49)	1.32 (1.20–1.44)	4.73×10^{-6}	PCDH15
rs10777864	12	97838685	A/C	0.40	0.79 (0.65–0.95)	0.74 (0.62–0.88)	0.76 (0.63–0.89)	3.28×10^{-5}	RMST
rs1499384	14	43049048	A/G	0.04	1.10 (0.76–1.59)	1.87 (1.44–2.42)	1.56 (1.35–1.78)	3.56×10^{-5}	LRFN5: FSCB
rs8060556	16	6868511	C/T	0.23	1.29 (1.04–1.59)	1.43 (1.17–1.74)	1.36 (1.14–1.69)	2.95×10^{-5}	RBFOX1
rs1728400	16	86434446	C/A	0.38	1.37 (1.15–1.64)	1.25 (1.06–1.47)	1.30 (1.13–1.55)	1.40×10^{-5}	LOC732275: FOXF1
rs8045253	16	86437767	T/C	0.34	1.28 (1.07–1.52)	1.32 (1.11–1.57)	1.30 (1.18–1.43)	2.82×10^{-5}	LOC732275: FOXF1
rs9978224	21	41309823	G/A	0.29	1.26 (1.03–1.54)	1.38 (1.16–1.65)	1.33 (1.13–1.61)	2.61×10^{-5}	TMPRSS3
rs421379	5	91275313	G/A	0.08	1.98 (1.46–2.70)	1.24 (0.91–1.68)	1.55 (1.25–1.93)	7.3×10^{-5}	ARRDC3

doi:10.1371/journal.pone.0101488.t002

Table 3. Replication of most significant associations from the discovery set meta-analysis in the replication samples.

Lead SNP	Chr	Position	Alleles	MAF	Stage-2 replication HR (95% Confidence Interval)	Stage-2 replication p-values	All stages meta-analysis HR (95% Confidence Interval)	All stages meta-analysis p-value	p-value for Q-statistic	Genes
rs7441398	4	63653135	G/T	0.13	1.12 (0.90–1.39)	0.28	1.31 (1.15–1.49)	3.3×10^{-5}	0.21	LPHN3: LOC644548
rs1525677	7	110302695	T/C	0.31	1.08 (0.92–1.27)	0.35	1.22 (1.10–1.34)	0.0001	0.18	IMMP2L
rs12358475	10	11848792	G/A	0.23	0.82 (0.67–1.00)	0.05	0.75 (0.67–0.84)	1.8×10^{-6}	0.57	ECHDC3: C10orf47
rs2921923	10	55662089	A/G	0.49	1.03 (0.88–1.21)	0.69	1.20 (1.10–1.33)	0.0001	0.04	PCDH15
rs10777864	12	97838685	A/C	0.41	0.91 (0.77–1.07)	0.25	0.82 (0.74–0.90)	8.0×10^{-5}	0.22	RMST
rs8060556	16	6868511	C/T	0.23	1.01 (0.84–1.22)	0.75	1.22 (1.09–1.36)	0.001	0.04	RBFOX1
rs1728400	16	86434446	C/A	0.38	1.16 (0.99–1.37)	0.07	1.25 (1.13–1.39)	5.6×10^{-6}	0.39	LOC732275: FOXF1
rs8045253	16	86437767	T/C	0.34	1.04 (0.88–1.23)	0.65	1.17 (1.05–1.31)	0.003	0.04	LOC732275: FOXF1
rs421379	5	91275313	G/A	0.08	1.41 (1.11–1.8)	0.005	1.49 (1.27–1.75)	1.1×10^{-6}	0.09	ARRDC3

Results are presented for those SNPs which remained associated in the same direction in the validation set as in the discovery set (adjusted for ER-status).

doi:10.1371/journal.pone.0101488.t003

sample selection for stage-1 utilised an "extreme phenotype" approach, this included selection of triple negative cases genotyped in a collaboration aimed at risk associated SNPs in triple negative breast cancer [11] and a second group enriched for exceptionally short survival genotyped as described previously [23]. We observed 236 breast cancer specific deaths in the POSH discovery set patients.

In HEBCS, 805 cases were selected from the patient series described earlier [22], including 423 unselected cases collected between years 1997 and 2000 as well as 140 cases collected between years 2001 and 2004, with 242 additional familial cases. The GWAS series was specifically enriched for cases with reduced survival, in the form of distant metastasis or death at the time of the initiation of the study in 2008, resulting in 301 breast cancer specific deaths at the time of analysis.

Stage-2 replication Samples

A further 1523 breast cancer patients from the POSH study [20] unselected for any survival differential were used for replication in stage-2. At stage 2, there were 293 breast cancer specific deaths.

Genome wide genotyping

Genotyping of 574 POSH phase-1 breast cancer cases was conducted using the Illumina 660-Quad SNP array. Genotyping was conducted in two separate batches at two locations. The Mayo Clinic (Rochester, Minnesota, USA) genotyped 274 triple negative breast cancers (negative for ER, PR and HER2) [11]. The remaining 300 POSH patients were genotyped at the Genome Institute of Singapore (GIS), National University of Singapore; these were selected based on either short duration of breast cancer specific survival (<2 years) or long duration of breast cancer specific survival (>4 years). In order to ensure complete harmonisation of genotype calling, the intensity data from GIS and MAYO were combined and the genotyping module of Illumina's Genome Studio software was used to generate genotypes. A GenCall threshold of 0.15 was selected and the HumanHap660 annotation file was used. Of the 300 samples genotyped in Singapore, 3 were excluded from analysis because they had sample call rates lower than 95%. No individuals among the two hundred and seventy four triple negative cohort genotyped at the Mayo clinic were excluded from analysis based on poor call rate. The genotyping accuracy for SNPs genotyped by GIS and Mayo were over 99%.

Genotyping of the HEBCS samples was conducted using the Illumina 550 platform as previously described [24]. SNP quality control (QC) measures were implemented using Plink. The initial sample size of 832 was reduced to 805, following quality control measures to remove patients with; unidentified affectation status and gender discordance (n = 6), familial relationships and poor SNP call rate (<95% n = 18), and missing phenotype information (n = 1). Genotypes were determined using the Genome Studio, a GenCall threshold of 0.15, and the HumanHap550-duo v3 annotation file.

Further quality control of the genotypic data from POSH and HEBCS was used to exclude rare SNPs with a MAF ≤0.01, and SNPs with significant deviation from Hardy-Weinberg equilibrium (HWE) p-value≤0.0001. To select SNPs for generation of pairwise identity by state (IBS) estimates, we used plink to perform genome wide linkage disequilibrium (LD)-based pruning with an r^2 cut-off of 0.5 and a window of 50 SNPs. Multi-dimensional scaling (MDS) plots were generated on the basis of a square matrix of IBS values between all pairs of individuals. To act as a reference, individuals with known African, Asian, and Caucasian ancestry from

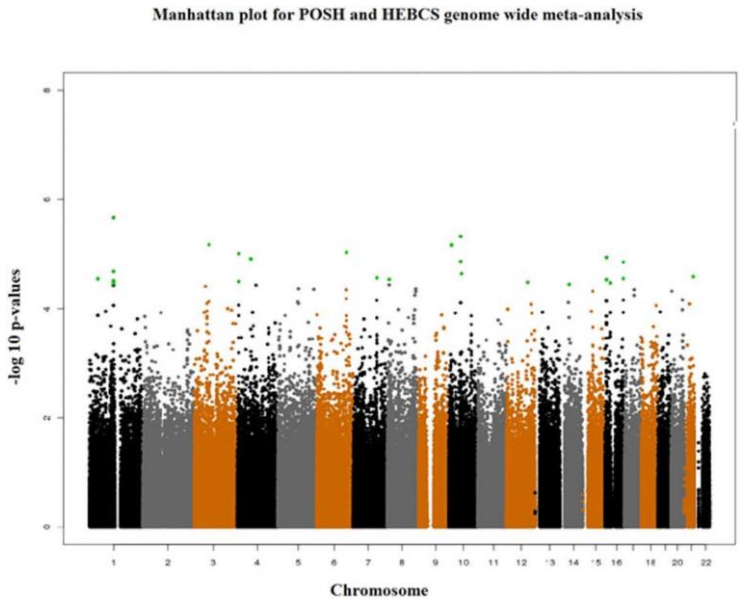


Figure 1. Manhattan plot of results from genome wide meta-analysis of POSH stage-1 and HEBCS hazard ratios and 95% confidence intervals. The 25 most associated SNPs are highlighted in green.
doi:10.1371/journal.pone.0101488.g001

HapMap were also used for the MDS analysis [25]. The MDS analysis excluded 35 cases from POSH and no cases from Helsinki whose genotypes did not concur with a European ancestry.

Statistical Analysis

We used GenABEL [26] in R.2.14.0 environment to perform survival analysis using post-QC genome wide SNP data. Follow-up time was calculated as the difference between the date of diagnosis of breast cancer and the date of death due to breast cancer or the date of last follow-up if still alive and right-censored at 10 years. Distant disease free interval was calculated as the time from diagnosis to occurrence of metastasis. We excluded patients with contralateral or ipsilateral cancers for testing association with distant disease free interval. All the Cox-proportional hazard models were adjusted for ER-status. Kaplan-Meier plots were generated using STATA v11.0 and IBM SPSS statistics 19. Mantel-Haenszel Fixed effects meta-analysis was performed using the metan module in STATA v11.0 [27]. For multivariate models we used ER-status, metastasis stage (0 or 1), nodal stage (1 = no nodes positive, 2 = 1–3 nodes positive, 3 = more than 3 nodes positive) and tumour size (centimetres) as covariates.

Cochran's Q-statistic and the resultant p-value was used to detect heterogeneity in association estimates between POSH and HEBCS. Genome wide meta-analysis was performed using MetABEL [28].

Genome wide imputation and meta-analysis

We imputed genome wide SNP information in POSH and HEBCS based on European phase 1 and release version 3 haplotypes. The reference haplotypes are derived from the 1000 genomes project which is the most comprehensive catalogue of human genetic variation including SNP, Indels and CNVs. Quality control measures applied to imputed data included excluding SNPs with HWE p-value$<1\times10^{-6}$, MAF <5%; and genotyping call rate <90% and individuals call rate <90%. Genome wide survival analysis of imputed information was performed in R-2.14.0 using GenABEL. Meta-analysis of results from GenABEL was performed using MetABEL. For imputing data we used MACH (http://www.sph.umich.edu/csg/abecasis/MACH/index.html). We used VCFtools - v0.1.9.0 to generate plink format files from output files generated by MACH. The reference haplotypes for 1000 genomes project were downloaded from MACH software's download page. We used Phase I version 3 European reference haplotypes for imputation analysis.

Manhattan and Regional plots

Manhattan and QQ-plots were generated in R using the plot command. Regional plots were generated using LocusZoom [29].

Sample size calculations

Sample size calculations were performed in R.2.14.2 using survSNP package. The event rate used for power calculations was 0.29 and a two-sided alpha of 0.05 was applied.

Gene Expression variation by SNP

We used Genevar 3.2.0 to study variation in expression levels by SNP genotypes available from the MuTHER pilot project while using NCBI Build 36 Ensembl 54 as reference [30]. Twin pairs were divided into two groups of unrelated individuals. Expression data from Lymphoblastoid cell lines are reported here. In addition we used SNP and CNV annotation database (scandb) [31] that uses the lymphoblastoid cell line expression data derived from 90 HapMap CEU samples in trios [32].

Prediction of transcription factor binding site changes

The putative changes on transcription factor binding sites caused by the variants were predicted *in silico* with MatInspector

Table 4. Associations of the most significantly associated SNPs from the discovery set s with disease free survival (adjusted for ER-status).

Lead SNP	Chr	Position	Alleles	MAF	Stage −1 association (95% Confidence Interval)	HEBC associations	Stage-2 association (95% Confidence Interval)	All stages meta-analysis p-value	p-value for Q-statistic
rs7441398	4	63653135	G/T	0.13	1.41 (1.12–1.77)	1.11 (0.88–1.42)	1.07 (0.91–1.26)	0.01	0.15
rs1525677	7	110302695	T/C	0.31	1.21 (1.0–1.45)	1.14 (0.96–1.35)	1.05 (0.88–1.24)	0.02	0.56
rs12358475	10	11848792	G/A	0.23	0.83 (0.67–1.02)	0.78 (0.65–0.94)	0.86 (0.70–1.05)	0.001	0.77
rs2921923	10	55662089	A/G	0.49	1.42 (1.19–1.71)	1.31 (1.12–1.52)	1.08 (0.91–1.27)	3.9×10^{-6}	0.08
rs10777864	12	97838685	A/C	0.41	0.80 (0.66–0.96)	0.72 (0.60–0.85)	1.07 (0.91–1.26)	0.005	0.003
rs8060556	16	6868511	C/T	0.23	1.20 (0.97–1.48)	1.30 (1.07–1.57)	0.94 (0.77–1.15)	0.03	0.06
rs1728400	16	86434446	C/A	0.38	1.27 (1.07–1.52)	1.21 (1.03–1.41)	1.17 (0.99–1.38)	5.7×10^{-5}	0.81
rs8045253	16	86437767	T/C	0.34	1.27 (1.07–1.51)	1.26 (1.07–1.48)	0.90 (0.75–1.06)	0.009	0.006
rs421379	5	91275313	G/A	0.08	1.69 (1.24–2.30)	1.04 (0.76–1.43)	1.32 (0.90–1.93)	0.003	0.10

Results are presented for those SNPs which remained associated in the same direction in the replication set as in the discovery set.

doi:10.1371/journal.pone.0101488.t004

within Genomatix software suite v2.5 (Genomatix Software GmbH) [33].

Results

POSH stage-1 and HEBCS meta-analysis

Genome wide genotype data were available from 536Caucasian participants of the POSH study and 805 Caucasian participants of the HEBCS study. A total of 475,141 SNPs with observed genotypes were available for meta-analysis in both the studies after excluding SNPs based on QC criteria. In stage-1 we used fixed-effects meta-analysis to pool hazard ratio estimates from the 536 POSH and 805 Helsinki breast cancer patients (Table 1). In the two study meta-analysis we found five SNPs which were associated at p-values lower than 9.9×10^{-6} (Table 2, Figure 1). The 25 most associated loci were selected for replication in POSH stage-2 patients. For loci with more than one SNP associated with survival, the most significant SNP and any other SNP(s) from the same locus which were not in high LD with the lead SNP ($r^2<0.6$) were selected for follow up in stage-2 (Table 3).

Replication testing in POSH stage-2 samples

A total of 18 SNPs with independent association signals were tested for replication in POSH stage-2 validation samples (n = 1523). One SNP demonstrated high duplicate error rate (> 8%) and was excluded from analysis. Of the 18 SNPs which were formally tested for replication, two demonstrated replication signals in the validation cohort. Nine of the eighteen SNPs which were tested for replication were observed to be associated in the same direction as in the POSH and HEBCS meta-analysis. In the stage-1 and stage-2 meta-analysis the strongest association signal was observed at rs421379. The minor allele of rs421379 is found to be associated with a higher risk of breast cancer related death (HR (95% CI) = 1.49 (1.27–1.75), $p = 1.1\times10^{-6}$) (Figure 2). The p-value for Cochran's heterogeneity test Q-statistic was not indicative of heterogeneity in meta-analysis estimate (p = 0.09). This variant was previously identified as the most significantly associated variant in a two stage GWAs for breast cancer survival in early onset cases from POSH. A weak replication signal in HEBCS allowed this SNP to be identified as the most strongly associated variant in this study too (Figure 2). The second most significant variant was located at 10p14, where the minor allele of rs12358475 was associated with protective effect on breast cancer mortality (HR (95% CI) = 0.75 (0.67–0.75), $p = 1.8\times10^{-6}$) (Figure 3). We observed another strong association with rs1728400 which is 113.6 kb upstream of the *FOXF1* locus (Table 3). In addition, the three SNPs (rs421379, rs12358475 and rs1728400) were also associated with distant disease free survival in the same direction as those observed for overall survival times, although none of these reached a genome wide level of significance (Table 4).

Genome wide imputation and meta-analysis

Following quality control of imputed data we had 7105428 SNPs available (maf ≥5%) in POSH stage-1 patients and 7353135 SNPs available (maf ≥5%) in the HEBCS study. In the two study meta-analysis we had close to 6.5 million SNPs available for meta-analysis. We did not identify any novel SNPs as associated with survival at p-values smaller than those observed using genotyped SNPs.

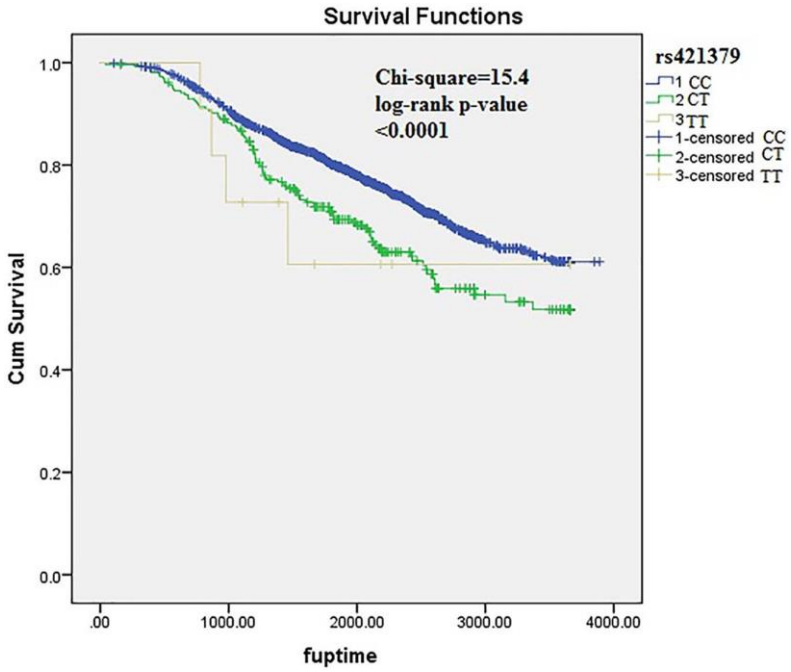


Figure 2. Kaplan-Meier plots depicting breast cancer related survival in response to rs421379 genotypes in pooled POSH stage-1, HEBCS and POSH stage-2 samples.
doi:10.1371/journal.pone.0101488.g002

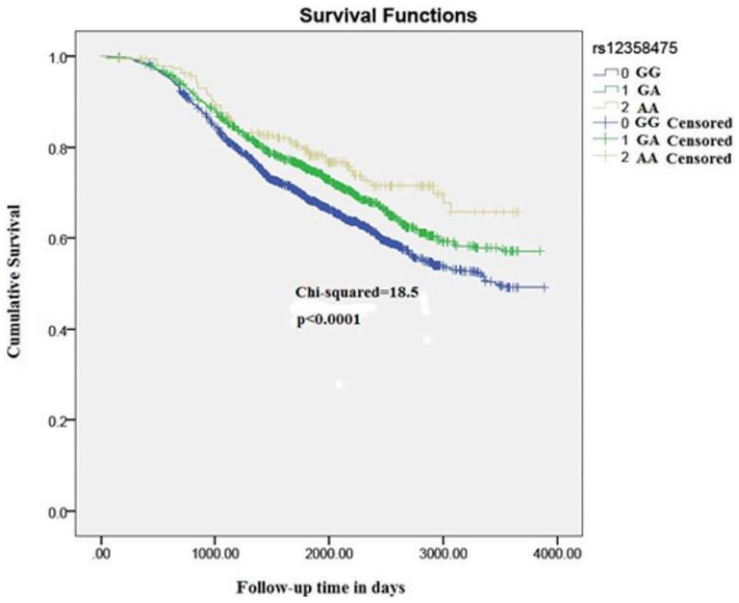


Figure 3. Kaplan-Meier plots depicting breast cancer related survival in response to rs12358475 genotypes in pooled POSH stage-1, HEBCS and POSH stage-2 sample.
doi:10.1371/journal.pone.0101488.g003

Table 5. Associations of SNPs with nominal replication signals with clinical characteristics associated with breast cancer in a pooled set of discovery and replication cohorts.

	N-Stage			M-Stage			Estrogen Receptor status (0 = positive, 1 = negative)			T-stage		
SNP	Odds Ratio	95% Confidence Interval	p-value	Odds Ratio	95% Confidence Interval	p-value	Odds Ratio	95% Confidence Interval	p-value	Odds Ratio	95% Confidence Interval	p-value
rs12358475	0.88	0.78–0.99	0.04	0.94	0.70–1.27	0.70	1.09	0.96–1.24	0.16	0.96	0.85–1.08	0.49
rs421379	1.06	0.83–1.35	0.65	1.25	0.93–1.67	0.14	1.09	0.89–1.34	0.39	0.98	0.77–1.25	0.88
rs1728400	1.09	0.96–1.22	0.17	1.30	0.99–1.70	0.05	1.07	0.9501.21	0.29	1.11	0.99–1.24	0.05

N-stage = metastasis to lymph node, M-stage = metastasis stage and T-stage = Tumour stage.

doi:10.1371/journal.pone.0101488.t005

Gene Expression variation by SNP in publically available database

We queried the Genevar 3.2.0 and SNP and CNV annotation database (scandb) to identify Cis or Tran's eQTL effects resulting from rs12358475, rs421379 and rs1728400. No associations of rs12358475 and rs1728400 with expression of any nearby genes were noticed in Genevar. In scandb too there were no strong trans-effect associations observed with rs12358475 and rs1728400. In Scandb we observed that rs421379 had impact on expression of *ABCD1* ($p = 1 \times 10–5$) and *RAB34* ($p = 9 \times 10–5$).

Univariate associations of most associated SNPs with N-stage, M-stage, T-stage and ER-status

In univariate analysis we did not observe any strong associations of rs12358475 with ER-status, N-stage, M-stage and T-stage. A nominally significant association with N-stage did not survive correction for multiple testing (Table 5). The SNP rs1728400 demonstrated weak associations with M-stage and T-stage (Table 5). No significant association of rs421379 with any of the clinical variables were observed.

Strength of association of SNPs most associated with survival in multivariate models

In pooled analysis involving the discovery and replication samples we observed a slight decrease in the strength of association at the rs421379 and rs12358475 variants. A prominent decline in association statistics at the rs1728400 variant was observed. The HR's for rs421379 and rs12358475 after adjusting for N-stage, M-stage, ER-status, and tumour size were 1.41 (1.15–1.72), $p = 0.001$ and 0.85 (0.75–0.97), $p = 0.01$. The observed HR for rs1728400 was 1.04 (0.94–1.15) $p = 0.46$.

Discussion

In this study we report a genome wide meta-analysis for identifying genetic variants associated with breast cancer related mortality. In combined meta-analysis involving 2864 individuals the strongest associations that we have identified locate to three SNPs at chromosomes 5, 10 and 16. We have previously discussed the potential biochemical pathways by which rs421379 could impact survival times [19]. It is important to note that the previous GWAs study that we had undertaken was performed exclusively in early onset cases alone. As such the findings from the current study are potentially important as these suggest a wider role for this variant in altering survival times in older breast cancer patients. We did not observe any significant effect of rs12358475 and rs421379 on clinical factors associated with breast cancer mortality suggesting that fluctuations in levels of clinical variables could be a by-product of disease rather than being driving factors.

rs12358475 is intergenic between *ECHDC3* (64 kb downstream) and C10orf47 (16 kb upstream), and 113 kb upstream of *UPF2*. *ECHDC3* encodes enoyl CoA hydratase domain containing 3 which has been described as a new inhibitor of mitochondrial fatty acid oxidation [34]. Although the clinical significance of this protein is not clear, it has been found to be differentially expressed in different breast cancer subtypes in mouse models [35]. *ECDHC3* has also been shown to be differentially expressed in acute coronary syndrome [36]. *UPG2* is involved in both mRNA nuclear export and mRNA surveillance and initiates nonsense-mediated mRNA decay (NMD) [37]. rs12358475 is predicted to disrupt a binding site for transcription factors *ETS1* and *NFAT*. ETS-1 is overexpressed in human breast cancer and this is indicative of poorer prognosis [38–40].

rs1728400 lies close to the *FOXF1* locus which is a putative tumour suppressor gene. This variant has previously been associated with oesophageal adenocarcinoma along with other SNPs close to rs1728400 which demonstrated even stronger associations [41]. As such if rs1728400t has a replicable impact on breast cancer prognosis then it could act via a different set of transcription factors than those activated in oesophageal carcinoma.

Although the study reported here is not the largest study undertaken for identifying common variants associated with breast cancer mortality [15,16], it has several methodological strengths. It is the first study to Meta-analyse associations of common genetic variants with breast cancer related mortality on a genome wide level across two independent prospective studies of breast cancer patients. Further both POSH and HEBCS are prospective studies of breast cancer patients who were recruited in similar clinical settings and both cohorts have relatively high breast cancer specific mortality. As such, heterogeneity between causes of death is reduced in the meta-analysis. With respect to potential tumour phenotypic heterogeneity both studies were not selected for specific breast tumour sub-types so this remains a potential methodological problem if the effect of SNPs relates to a particular tumour sub-type or a particular modality of treatment.

It was encouraging to note that 9 off the 18 SNPs which we had marked for replication testing were associated in the same direction as in the discovery set. Furthermore 4 of the 18 SNPs which were tested for replication had previously been identified as amongst the top 50 associations in GWAs of breast cancer mortality in early onset patients. rs11723068, rs11491815, rs421379 and rs1578790 were the first, fourth, eighteenth and 20th most strongly associated SNPs among the top 50 association [19].

Although previous studies have not described any SNPs as irrevocably associated with survival at genome wide levels of significance [15,18], we attempted to test associations of the most significant SNPs from these studies. None of the 10 SNPs which Azzato et al [15] tested for replication in the SEARCH study were associated at p-values$\leq$0.05 in the POSH and HEBCS meta-analysis results. The strongest replication signal we identified was with rs17299684 (HR = 1.15, p = 0.07). Similarly the two SNPs highlighted by Shu et al [18] as potentially associated with survival in the Chinese population, were not associated in our meta-analysis (rs3784099, HR = 0.94, P = 0.37 and rs9934948, HR = 1.09, P = 0.32). The association of SNP rs3803662 (*TOX3*), highlighted by Fasching et al [16], as potentially associated with breast cancer specific survival did not replicate in our meta-analysis (HR = 0.90, p = 0.09). The lone SNP highlighted by Azzato et al [17], as associated with survival in ER-negative patients was not available in the genome wide genotyped or imputed data, further no proxies at $r^2 \geq 0.6$ were identified based on HapMap phase 3 data. So unfortunately replication of this SNP could not be tested in our study.

Future studies with a similar ascertainment framework but with larger sample size, detailed tumour sub-type phenotyping and similar treatment modalities will be required to allow sub-type specific patient cohorts to be used for discovery and validation. A more detailed search for variants with MAF $<$0.05 may be necessary to fully comprehend the extent of intrinsic host genetic factors in determining breast cancer prognosis.

The main strengths of this study are the high genetic coverage achieved by using the Illumina 550 K and Illumin660 K chips in the Helsinki and POSH studies respectively. In addition we have also performed comprehensive imputation of common genetic variation (maf $>=$ 5%) based on the LD patterns in the 1000 genomes project. We had sufficient statistical power to detect genetic variants which were associated with survival at HR$\geq$1.23 while studying SNPs with maf $\geq$10%. Future studies using well annotated collaborative samples will be needed to perform sub-type specific analysis and replication to detect small effect sizes. Such a strategy has the potential to identify multiple genetic variants which are associated at HRs lower than 1.20. However a trade-off between the increases in effect sizes that may result from studying associations in specific homogeneous sub-groups may mitigate smaller sample sizes.

Supporting Information

Checklist S1 PRISMA Checklist.
(DOC)

Acknowledgments

We would like to thank Drs. Kirsimari Aaltonen, Dario Greco, Xiaofeng Dai, Päivi Heikkilä and Karl von Smitten, as well as Tuomas Heikkinen, M.Sc. for their help with the HEBCS patient samples and data, and research nurses Hanna Jäntti and Irja Erkkilä for their assistance in the HEBCS data collection and management. The authors thank Nikki Graham (DNA bank) and the staff of the Southampton CRUK Centre Tissue Bank.

Author Contributions

Conceived and designed the experiments: DE HN AC SR WT. Performed the experiments: SR SK RUG SG KA FJC JL. Analyzed the data: SR SK WT. Contributed reagents/materials/analysis tools: FJC JL DE HN SG KA CB. Wrote the paper: SR SK HN DE.

References

1. Office for national statistics website. Available: http://www.statistics.gov.uk/hub/index.html. Acessed 2014 Mar 17.
2. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
3. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, et al. (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. Journal of the National Cancer Institute 98: 1183–1192.
4. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826.
5. Sotiriou C and Pusztai L (2009) Gene-expression signatures in breast cancer. N Engl J Med 360: 790–800.
6. Lindstrom LS, Hall P, Hartman M, Wiklund F, Gronberg H, et al. (2007) Familial concordance in cancer survival: a Swedish population-based study. Lancet Oncol 8: 1001–1006.
7. Hartman M, Lindstrom L, Dickman PW, Adami HO, Hall P, et al. (2007) Is breast cancer prognosis inherited. Breast Cancer Res Tr 9.
8. Lakhani SR, Reis-Filho JS, Fulford L, Penault-Llorca F, van der Vijver M, et al. (2005) Prediction of BRCA1 status in patients with breast cancer using estrogen receptor and basal phenotype. Clin Cancer Res 11: 5175–5180.
9. Wilson JR, Bateman AC, Hanson H, An Q, Evans G, et al. (2010) A novel HER2-positive breast cancer phenotype arising from germline TP53 mutations. J Med Genet 47: 771–774.
10. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, et al. (2013) Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet 45: 392–398, 398e391–392.
11. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, et al. (2011) A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 43: 1210–1214.
12. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40: 703–706.
13. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, et al. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39: 865–869.

14. Chen W, Zhong R, Ming J, Zou L, Zhu B, et al. (2012) The SLC4A7 variant rs4973768 is associated with breast cancer risk: evidence from a case-control study and a meta-analysis. Breast Cancer Res Treat 136: 847–857.
15. Azzato EM, Pharoah PD, Harrington P, Easton DF, Greenberg D, et al. (2010) A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19: 1140–1143.
16. Fasching PA, Pharoah PD, Cox A, Nevanlinna H, Bojesen SE, et al. (2012) The role of genetic breast cancer susceptibility variants as prognostic factors. Hum Mol Genet 21: 3926–3939.
17. Azzato EM, Tyrer J, Fasching PA, Beckmann MW, Ekici AB, et al. (2010) Association between a germline OCA2 polymorphism at chromosome 15q13.1 and estrogen receptor-negative breast cancer survival. J Natl Cancer Inst 102: 650–662.
18. Shu XO, Long J, Lu W, Li C, Chen WY, et al. (2012) Novel genetic markers of breast cancer survival identified by a genome-wide association study. Cancer Res 72: 1182–1189.
19. Rafiq S, Tapper W, Collins A, Khan S, Politopoulos I, et al. (2013) Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. Cancer Res 73: 1883–1891.
20. Eccles D, Gerty S, Simmonds P, Hammond V, Ennis S, et al. (2007) Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH): study protocol. BMC Cancer 7: 160.
21. Fagerholm R, Hofstetter B, Tommiska J, Aaltonen K, Vrtel R, et al. (2008) NAD(P)H: quinone oxidoreductase 1 NQO1*2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer. Nature genetics 40: 844–853.
22. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nature genetics 42: 504–507.
23. Rafiq S, Tapper W, Collins A, Khan S, Politopoulos I, et al. (2013) Identification of inherited genetic variations influencing prognosis in early onset breast cancer. Cancer research.
24. Li J, Humphreys K, Heikkinen T, Aittomaki K, Blomqvist C, et al. (2011) A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat 126: 717–727.
25. (2003) The International HapMap Project. Nature 426: 789–796. genabel website. Available: http://www.genabel.org/. Accessed 2013 Dec 9.
26. Harris RJ, Bradburn MJ, Deeks JJ, Harbord RM, Altman DG, et al. (2008) metan: fixed- and random-effects meta-analysis. Stata Journal 8: 3–28.
27. Downey L, Livingston RB, Koehler M, Arbushites M, Williams L, et al. (2010) Chromosome 17 polysomy without human epidermal growth factor receptor 2 amplification does not predict response to lapatinib plus paclitaxel compared with paclitaxel in metastatic breast cancer. Clin Cancer Res 16: 1281–1288.
28. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337.
29. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS genetics 7: e1002003. scandb Website. Available: http://www.scandb.org/newinterface/about.html. Accessed 2013 Dec 9.
30. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, et al. (2010) SCAN: SNP and copy number annotation. Bioinformatics 26: 259–262.
31. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21: 2933–2942. Hashimoto T, Shindo Y, Souri M and Baldwin GS (1996) A new inhibitor of mitochondrial fatty acid oxidation. J Biochem 119: 1196–1201.
32. Herschkowitz JI, Zhao W, Zhang M, Usary J, Murrow G, et al. (2012) Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells. Proceedings of the National Academy of Sciences of the United States of America 109: 2778–2783.
33. Silbiger VN, Luchessi AD, Hirata RD, Lima-Neto LG, Cavichioli D, et al. (2013) Novel genes detected by transcriptional profiling from whole-blood cells in patients with early onset of acute coronary syndrome. Clin Chim Acta 421: 184–190.
34. Lykke-Andersen J, Shu MD, Steitz JA (2000) Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. Cell 103: 1121–1131. Lincoln DW, 2nd and Bove K (2005) The transcription factor Ets-1 in breast cancer. Front Biosci 10: 506–511.
35. Verschoor ML, Wilson LA, Singh G (2010) Mechanisms associated with mitochondrial-generated reactive oxygen species in cancer. Can J Physiol Pharmacol 88: 204–219.
36. Buggy Y, Maguire TM, McGreal G, McDermott E, Hill AD, et al. (2004) Overexpression of the Ets-1 transcription factor in human breast cancer. Br J Cancer 91: 1308–1315.
37. Levine DM, Ek WE, Zhang R, Liu X, Onstad L, et al. (2013) A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. Nat Genet 45: 1487–1493.
38. Lincoln DW 2nd, Bove K (2005) The transcription factor Ets-1 in breast cancer. Front Biosci 10: 506–511.
39. Verschoor ML, Wilson LA, Singh G (2010) Mechanisms associated with mitochondrial-generated reactive oxygen species in cancer. Can J Physiol Pharmacol 88: 204–219.
40. Buggy Y, Maguire TM, McGreal G, McDermott E, Hill AD, et al. (2004) Overexpression of the Ets-1 transcription factor in human breast cancer. Br J Cancer 91: 1308–1315.
41. Levine DM, Ek WE, Zhang R, Liu X, Onstad L, et al. (2013) A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. Nat Genet 45: 1487–1493.

ORIGINAL ARTICLE

Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes

Katja Christodoulou,[1] Anthony E Wiskin,[2] Jane Gibson,[1] William Tapper,[1] Claire Willis,[2] Nadeem A Afzal,[3] Rosanna Upstill-Goddard,[1] John W Holloway,[4] Michael A Simpson,[5] R Mark Beattie,[3] Andrew Collins,[1] Sarah Ennis[1]

► Additional materials are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2011-301833).

For numbered affiliations see end of article.

Correspondence to
Dr Sarah Ennis, Genetic Epidemiology and Genomic Informatics Group, Human Genetics, Faculty of Medicine, University of Southampton, Duthie Building (Mailpoint 808), Southampton General Hospital, Southampton SO16 6YD, UK; s.ennis@soton.ac.uk

KC and AEW contributed equally to this study.

Revised 27 March 2012
Accepted 1 April 2012
Published Online First 28 April 2012

ABSTRACT

Background Multiple genes have been implicated by association studies in altering inflammatory bowel disease (IBD) predisposition. Paediatric patients often manifest more extensive disease and a particularly severe disease course. It is likely that genetic predisposition plays a more substantial role in this group.

Objective To identify the spectrum of rare and novel variation in known IBD susceptibility genes using exome sequencing analysis in eight individual cases of childhood onset severe disease.

Design DNA samples from the eight patients underwent targeted exome capture and sequencing. Data were processed through an analytical pipeline to align sequence reads, conduct quality checks, and identify and annotate variants where patient sequence differed from the reference sequence. For each patient, the entire complement of rare variation within strongly associated candidate genes was catalogued.

Results Across the panel of 169 known IBD susceptibility genes, approximately 300 variants in 104 genes were found. Excluding splicing and *HLA*-class variants, 58 variants across 39 of these genes were classified as rare, with an alternative allele frequency of <5%, of which 17 were novel. Only two patients with early onset Crohn's disease exhibited rare deleterious variations within *NOD2*: the previously described R702W variant was the sole *NOD2* variant in one patient, while the second patient also carried the L1007 frameshift insertion. Both patients harboured other potentially damaging mutations in the *GSDMB*, *ERAP2* and *SEC16A* genes. The two patients severely affected with ulcerative colitis exhibited a distinct profile: both carried potentially detrimental variation in the *BACH2* and *IL10* genes not seen in other patients.

Conclusion For each of the eight individuals studied, all non-synonymous, truncating and frameshift mutations across all known IBD genes were identified. A unique profile of rare and potentially damaging variants was evident for each patient with this complex disease.

Significance of this study

What is already known on this subject?

- Genome-wide association studies have implicated numerous candidate genes for inflammatory bowel disease (IBD), but evidence of causality for specific variants is largely absent. Furthermore, by design, genome-wide association studies are limited to the study of common variants and overlook the functionally detrimental variation imposed by rare/novel mutation.
- Exome analysis is fully informative for the spectrum of variation within the protein coding sequence of genes. It has been used to successfully identify disease causing variants in Mendelian disorders, but its potential to identify the missing heritability in complex diseases such as paediatric IBD has not yet been realised.

What are the new findings?

- This study examines genetic variants from the perspective of the patient rather than the gene—for each paediatric case a profile of deleterious variation is determined across a comprehensive panel of known IBD genes.
- Paediatric IBD patients carry a wide spectrum of low frequency variants within candidate IBD genes.
- In silico analyses indicate a substantial proportion of these mutations are potentially deleterious.
- Consistent with complex inheritance, this small subset of patients with severe IBD exhibit a varied profile of mutation with limited sharing of specific variants across the set of eight exomes.

INTRODUCTION

Ulcerative colitis (UC) and Crohn's disease (CD) are the two main clinical phenotypes of inflammatory bowel disease (IBD), both resulting in chronic and relapsing inflammation. The incidence of IBD in the paediatric population of the UK is 5.2 per 100 000 children per year, with breakdown figures of 3.1 for CD, 1.4 for UC and 0.6 for IBD unclassified (IBDU).[1] While the precise aetiology and pathogenesis is complex and incompletely understood, it is widely accepted that IBD occurs as the result of a dysregulated mucosal immune response to commensal gut flora in the genetically susceptible host.[2] Familial aggregation of disease implies a strong genetic component,[3] although

Significance of this study

How might it impact on clinical practice in the foreseeable future?

- Functional studies are required to confirm in silico assessment of variation impact on biology.
- Even mutations confirmed to confer susceptibility must be considered among the full profile of disease predisposing variation present in any individual.
- As the cost of next generation sequencing falls and the number of mutation profiles increases, there is clear potential for genetic characterisation of IBD phenotypic sub-types facilitating targeted therapeutic intervention/personalised medicine.

environmental factors may play a greater role in ulcerative colitis.[4]

Over recent years, genome-wide association studies (GWAS) have been applied with huge success to identify common genes involved in both CD and UC. Genes with replicated evidence for strong association suggest that pathways involving disruption of the innate and adaptive immune system, compromised epithelial barrier function and impaired autophagy play a significant role in disease.[2] However, despite the identification of over one hundred unique genes in IBD susceptibility, these common variants in combination account for less than a quarter of the genetic risk.[5–7] The source of this missing heritability is the subject of much debate with various explanations: overestimates of original heritability statistics; underpowered GWAS studies (in terms of sample size and single nucleotide polymorphism (SNP) coverage) to detect common variants associated with decreasing effect sizes; poorly investigated epistatic and gene–environment interactions; and rare variation.[8]

Rare variants form the group of infrequent mutations that occur in <5% of the population. A large proportion of variants in this class occur at a much lower frequency (<0.1%), and many thousands are likely to be specific to ethnic groups, isolates, families or even individuals. Nevertheless, this class of variation harbours multiple penetrant disease mutations conferring medium to high risk. Rare variants escape detection by GWAS. *BRCA1* and *BRCA2* are examples of familial breast cancer genes that harbour many high risk variants but go undetected by GWAS. This is consequent to each of the disease causing mutations being shared by only a fraction of the patient group and so no common SNP can act as a proxy or 'tag' to flag the gene as causal. It is entirely plausible that a proportion of IBD and other complex disease heritability unaccounted for by common variation lies within higher risk rare variants. Furthermore, many of these mutations may lie within genes already implicated by association studies.

Exome sequencing determines each letter of the genetic code at nearly all coding regions or exons in the genome (the 'exome'), thereby generating the complete profile of coding variation. It has already proved its success in identifying causal mutations in an ever growing list of both recessive and dominant rare Mendelian disorders whereby sequencing of a small number of unrelated cases has been used to identify disease causing variants.[9] One such case reported exome sequencing undertaken in a male child presenting at 15 months with intractable IBD; exome sequencing was used to successfully identify a causal mutation in the *XIAP* gene (X-linked inhibition of apoptosis gene) for which the child was hemizygous. After haematopoietic progenitor cell transplant treatment, as recommended for XIAP deficiency, the IBD resolved, suggesting that the Crohn's-like illness seen in this patient was driven by this single mutation.[10]

As next generation sequencing technology advances, it becomes increasingly affordable. Nevertheless, while costs remain in the region of several hundred pounds per sample, targeted analyses of those patient groups most likely to yield positive results is prudent. Prioritisation of cases with strong family history and/or patients representing the phenotypic 'extreme' of common traits is a useful strategy.[11] One such example of an 'extreme' phenotype is paediatric disease in which onset is particularly early. Genetic susceptibility is thought to play a more important role in the aetiology of early-onset IBD than in late-onset IBD.[12] This is supported by a higher rate of positive family history of IBD in patients with a younger age at diagnosis compared to the older age group, suggesting that an earlier presentation may be due to a higher burden of disease-causing mutations in the genomes of these affected children compared to those in whom disease manifests later in life.[13] In addition, environmental confounding factors such as smoking are less likely to be exerting an influence on disease in paediatric cohorts. It has also been suggested that early-onset disease may in itself be a more aggressive phenotype; in CD, earlier age at diagnosis is associated with a greater need for surgery and increased small bowel disease.[12–14]

Two of the most comprehensive association studies investigating IBD have used adult cohorts, but a recent GWAS of 3246 early-onset IBD cases successfully identified five new loci associated with childhood susceptibility as well as replicating loci previously implicated in adult-onset disease.[15] Early-onset disease genes have also been located using linkage analysis and candidate gene sequencing approaches undertaken in two unrelated consanguineous families.[16] Despite distinct clinical and histopathological features of the CD and UC phenotypes, an estimated 30% of IBD-related loci are shared between both phenotypes.[2] It is likely that further study of rare variation across implicated genes may uncover more commonality.

The application of exome sequencing to complex diseases is fraught with analytical difficulty; finding disease causing variants among the many innocent variants present in the genome has been likened to finding 'needles in stacks of needles'.[17] Targeting analyses to subsets of genes in patients with extreme phenotype is a practical approach to examining genetic influence in disease. In this study we apply next generation sequence technology to paediatric IBD (PIBD). The study is focused on a small cohort of eight paediatric patients with markedly early onset/severe disease. Patients are representative of the spectrum of IBD presentation, and limiting the study to this modest number makes data interpretable on a case-by-case basis. We focus on a comprehensive panel of known causal genes and for each patient describe their individual burden of rare and novel damaging variation.

MATERIALS AND METHODS

Recruitment of paediatric IBD cohort of patients

Children included in this study were selected from the 'Genetics of Paediatric IBD' cohort between October 2010 and October 2011. This cohort was recruited through tertiary referral paediatric IBD clinics at the University Hospital Southampton Foundation Trust. This hospital is the regional centre for paediatric gastroenterology, providing a tertiary paediatric

Gut 2013;**62**:977–984. doi:10.1136/gutjnl-2011-301833

gastroenterology and endoscopy service for the Wessex region, and draws on a patient population of 3.5 million. The service has a rolling database of over 300 paediatric IBD cases and approximately 50–70 patients are diagnosed each year. All children had a diagnosis of IBD and were aged between 5 and 18 years at time of recruitment, although their diagnosis may have been made at an earlier age. Diagnosis was established using the Porto criteria[18]; all children had compatible history, examination and laboratory investigation results, and infectious causes excluded. All were investigated with upper gastrointestinal endoscopy and ileo-colonoscopy. Written informed consent was obtained from the attending parent of all children, and the child where appropriate. In the initial recruitment interview, clinical data and venous blood samples (10 ml for DNA extraction and 8 ml for plasma extraction) were collected. Additional comprehensive clinical data were extracted from patient records. For each patient we gathered information on gender, dates of birth and initial diagnosis, disease extent currently and at diagnosis using the Paris classification,[19] disease activity score at diagnosis (using the paediatric CD activity index (PCDAI) and the paediatric ulcerative colitis activity index (PUCAI)), height and weight currently and at first diagnosis, time to and date of first relapse, treatment history (use of steroids, immunomodulators, biological therapies, surgery), history of potential aetiological and modifying conditions such as smoking, gastrointestinal infection and other autoimmune disease, and family history.

Ethics statement

This study was approved by the Southampton and South West Hampshire Research Ethics Committee (REC) (09/H0504/125) and University Hospital Southampton Foundation Trust Research & Development (RHM CHI0497).

Selection of samples

Eight patient samples from our PIBD cohort as previously described were selected for exome sequencing for this study. These eight patients were selected based on age of diagnosis, disease severity or positive family history in a first degree relative. Selection criteria and patient phenotypic characteristics are summarised in table 1.

DNA and plasma extraction

Genomic DNA was extracted from EDTA anticoagulated peripheral venous blood samples using the salting out method. Plasma was isolated from lithium–heparin anticoagulated peripheral venous blood samples using standard methods.

Exome sequencing

Targeted exome capture was performed using the SureSelect Human All Exon 50Mb kit (Agilent). The Illumina HiSeq system was used to generate sequence data. These steps were conducted at the Wellcome Trust Centre for Human Genetics at Oxford University. The resultant paired end sequencing data were aligned against the human genome reference sequence 18 (hg18) using the Novoalign software (2.06.09MT, Novocraft Technologies, Selangor, Malaysia). Duplicate reads, resulting from PCR clonality or optical duplicates, and reads mapping to multiple locations were excluded from downstream analysis. Depth and breadth of sequence coverage was calculated with custom scripts and the BedTools package.[20] Single nucleotide substitutions and small insertion deletions were identified and quality filtered within the SamTools software package[21] and in-house software tools. Variants were annotated with respect to genes and transcripts with the Annovar tool.[22] Summary statistics for exome sequencing, mapping and coverage are shown in supplementary table 1 (available online only). Data from the 1000 Genomes Project (1KG) phase I (2010 November release) were utilised using LiftOver (University of California Santa Cruz Genome Browser, http://genome.ucsc.edu/cgi-bin/hgLiftOver) for the conversion of 2010 November coordinates to hg18. Variants were characterised as novel if they were previously unreported in the dbSNP129, dbSNP132, 1KG data and our 22 in-house reference exomes (supplementary table 2). Southampton reference exomes for evaluating the burden of mutation comprised independent DNA samples from unrelated individuals who were exome sequenced on the same platform at the same time as part of other local projects. Each reference exome was from a patient with a distinct clinical diagnosis but no history of gastrointestinal or autoimmune disease. The clinical phenotypes of the 22 reference exomes included 10 with leukaemia, 5 with lymphoma, 4 with Beckwith–Wiedemann syndrome and 3 with macrocephaly malformation syndrome.

The National Heart Lung and Blood Institute Exome Sequencing Project Exome Variant Server (http://evs.gs.washington.edu/EVS/) (Feb 2012) was used as a reference dataset for rare variant allele frequency in a European American population (table 2). This project contains exome data from approximately 3500 European American individuals taken from 12 disease cohorts with a range of heart, lung or blood disorders.

Selection of a panel of known IBD genes

We constructed a panel of high priority genes previously shown to be strongly associated with IBD. Our aim was to include all

Table 1 Summary of patient phenotypes and characteristics (specific selection criteria are in bold)

Sample ID	Age at diagnosis (years)	Sex	Disease	Phenotype description and selection criteria	Ethnicity	Family history
Proband 1	11	Male	CD	**Severe disease requiring surgery**/ Stricturing ileo-colonic disease requiring right hemicolectomy within 6 months of diagnosis.	White British	–
Proband 2	7	Female	CD	**Early age of onset**/non-stricturing, non-penetrating mild to moderate pancolitis, disease resistant to treatment.	White British	–
Proband 3	6	Male	CD	**Early age of onset**/non-stricturing, non-penetrating granulomatous colitis and duodenitis. **Mother diagnosed CD aged 21 years.**	White British	+
Proband 4	6	Female	CD	**Early age of onset**/non-penetrating pancolitis with possible ileo-caecal stricture.	White British	–
Proband 5	13	Male	CD	Non-stricturing, non-penetrating, colitis. **Family history** including maternal CD and maternal grandparental UC.	White British	+
Proband 6	9	Male	UC	**Severe** left sided colitis, also with oral pemphigus.	White British	–
Proband 7	2	Male	UC	**Early age of onset**/mild to moderate pancolitis.	White British	–
Proband 8	3.5	Male	IBDU	**Early age of onset**/left sided colitis.	Iraqi	–

CD, Crohn's disease; IBDU, inflammatory bowel disease unclassified; UC, ulcerative colitis.

Inflammatory bowel disease

Table 2 Characterisation of non-synonymous, stopgain and indel variants with an alternative allele frequency of <0.05 or not reported in 1000 genomes across 39 known IBD genes

Gene	Chromosome	Exon	Variant	Functionally implicated in pathway	Base pair location in hg18	rs ID number in dbSNP 132	Nucleotide change	Protein change	Frequency in 1000 genome	Frequency in NHLBI ESP	Sift score	Grantham Score	Grantham	Polyphen2	Observed n/8	No. homozygote	No. heterozygote	proband 1	proband 2	proband 3	proband 4	proband 5	proband 6	proband 7	proband 8
BACH2	6	7	ns	B-cell regulation	90,717,215	NR	C1331T	S444L	NR	NR	0	145	MR	B	1	0	1						◊		
BACH2	6	7	ns	B-cell regulation	90,717,675	rs61754114	C871G	L291V	0.010	0.030	0	32	C	PrD	1	0	1							◊	
BSN	3	5	ns	Presynaptic cytoskeletal support	49,665,631	rs35762866	G3638A	G1213D	0.035	0.108	0.13	94	MC	PrD	2	0	2			.	.				
BSN	3	5	ns	Presynaptic cytoskeletal support	49,667,511	NR	G5518A	E1840K	NR	NR	0.07	56	MC	PrD	1	0	1							.	
BTNL2	6	6	ns	T-cell negative regulation	32,470,681	rs41521946	C1178A	P393Q	NR	0.003	0.75	76	MC	B	3	0	3		.	.		.			
BTNL2	6	5	ns	T-cell negative regulation	32,471,871	rs28362679	C1001T	S334L	0.014	0.020	0	145	MR	PrD	1	0	1								◊
C1orf93	1	5	ns	Prostaglandin processing	2,509,900	NR	G526A	G176R	NR	NR	0	125	MR	PrD	1	0	1						◊		
CD19	16	13	ns	B-cell receptor signalling	28,857,552	rs34763945	G1544A	R515H	0.025	0.066	0.56	29	C	PrD	2	0	2	.		.					
CDKAL1	6	8	ns	Methylthiotransferase family	20,889,397	NR	G560A	R187K	NR	NR	0.40	26	C	B	1	0	1					.			
CXCR1	2	2	ns	Chemokine receptor	218,737,177	rs16858808	C1003T	R335C	0.018	0.030	0.09	180	R	PrD	2	0	2					.	.		
CXCR1	2	2	ns	Chemokine receptor	218,738,088	rs16858811	T92G	M31R	0.040	0.032	0.60	91	MC	B	2	0	2					.	.		
ERAP2	5	6	ns	Antigen presentation	96,253,828	rs75263594	C1040T	T347M	0.013	0.033	0.01	81	MC	PrD	1	0	1			◊					
ERRFI1	1	4	ns	Epithelial barrier function	7,996,921	rs34781518	G325A	D109N	0.005	0.015	0.14	23	C	B	1	0	1								
FUT2	19	2	ns	Blood group antigen synthesis	53,898,797	rs602662	G772A	G258S	NR	0.515	0.06	56	MC	PrD	5	2	3	.			.	.	.	.	
GMPPB	3	5	ns	Catalyses mannose processing	49,735,146	NR	G448C	E150Q	NR	NR	0.01	29	C	B	1	0	1						◊		
GSDMB	17	7	ns	Unknown	35,316,029	rs35104165	A710G	D237G	0.012	0.036	0	94	MC	B	1	1	0	◊							
HORMAD2	22	2	ns	Unknown	28,819,945	rs34150968	G4A	A2T	0.004	0.012	0	58	MC	PoD	1	0	1					◊			
ICAM1	19	5	ns	Leukocyte adhesion ligand	10,256,141	-	G988A	V330M	0.001	0.003	0	21	C	PrD	1	0	1								◊
ICAM1	19	5	ns	Leukocyte adhesion ligand	10,256,252	-	C1099T	R367C	0.004	0.000	0	180	R	PrD	1	0	1								◊
IL10	1	2	ns	Innate immune recognition	205,011,338	-	C211A	L71M	NR	NR	0.07	15	C	PrD	1	0	1						.		
IL10	1	1	ns	Innate immune recognition	205,012,361	-	G43A	G15R	NR	0.002	0.04	125	MR	PoD	1	0	1							◊	
IL18RAP	2	11	ns	Enhances IL18 binding	102,433,811	-	C1282A	L428M †	NR	0.001	0.15	15	C	PrD	1	0	1	.							
IL18RAP	2	11	ns	Enhances IL18 binding	102,433,812	-	T1283A	L428Q †	NR	0.001	0.15	113	MR	PrD	1	0	1	.							
IL1RL1	2	11	ns	T-helper cell function	102,334,643	rs10192036	C1501A	Q501K †	NR	0.019	1.00	53	MC	B	4	2	2			.	.	.		.	
IL1RL1	2	11	ns	T-helper cell function	102,334,644	rs10204137	A1502G	Q501R †	0.018	0.032	0.61	43	C	B	4	2	2			.	.	.		.	
IL1RL2	2	11	ns	Interleukin receptor	102,217,903	-	C1412T	A471V	NR	NR	0	64	MC	PrD	1	0	1							◊	
JAK2	9	9	ns	Th17-cell differentiation	5,055,003	rs2230723	C1177G	L393V	0.016	0.006	0.38	32	C	B	1	0	1		.						
KIF21B	1	33	ns	Microtubule-binding protein	199,210,518	-	C4722A	D1574E	NR	NR	0.1	45	C	PrD	1	0	1				.				
LRRK2	12	18	ns	Autophagy	38,958,256	rs10878307	A2167G	I723V	0.046	0.070	0.52	29	C	B	2	1	1	.					.		
LTA	6	3	ns	Cytokine receptor interaction	31,648,736	rs2229092	A152C	H51P	0.039	0.072	0.3	77	MC	B	2	0	2					.		.	
MST1	3	17	sg	Apoptosis	49,696,816	-	C1951T	R651X	0.004	0.013	0.14	-	-	-	1	0	1				.				
MST1	3	13	ns	Apoptosis	49,697,765	rs62262682	G1478T	R493L	0.015	0.058	0.09	102	MR	B	1	0	1					.			
MTMR3	22	17	ns	Lipid phosphatase	28,745,983	rs61737780	C2335T	L779F	0.005	0.012	0.21	22	C	PoD	1	0	1					.			
MTMR3	22	17	ns	Lipid phosphatase	28,746,527	rs41278853	A2879G	N960S	0.041	0.086	0.08	46	C	B	1	0	1				.				
NOD2	16	4	ns	Autophagy	49,303,427	rs2066844	C2104T	R702W	0.029	0.047	0	101	MR	PrD	2	0	2	◊		◊					
NOD2	16	9	ns	Autophagy	49,314,777	rs5743291	G2863A	V955I	0.044	0.095	0.46	29	C	B	1	0	1				.				
NOD2	16	11	fi	Autophagy	49,321,282	-	3019_3020insC	L1007fs	NR‡	NR	-	-	-	-	1	0	1			.					
PARK7	1	5	ns	Autophagy	7,953,581	rs71653619	G293A	R98Q	0.003	0.012	0.50	43	C	B	2	0	2				.		.		
PNMT	17	3	sg	Adrenaline processing	35,080,063	-	C744A	Y248X	NR	0.088	0.18	-	-	-	1	0	1						.		
PTGER4	5	3	ns	Epithelial barrier function	40,727,650	rs111866313	G880A	V294I	0.009	0.027	0.51	29	C	B	1	0	1		.						
RTEL1	20	24	ns	DNA repair	61,791,572	rs35640778	G2051A	R684Q	0.003	0.018	0.58	43	C	B	1	0	1							.	
SEC16A	9	23	ns	Endoplasmic reticulum traffic	138,465,668	rs45519739	C6173T	T2058M	NR	0.015	0.01	81	MC	-	1	0	1			◊					
SEC16A	9	3	ns	Endoplasmic reticulum traffic	138,490,409	-	G1480C	G494R	NR	NR	0	125	MR	-	1	0	1						◊		
SEC16A	9	3	ns	Endoplasmic reticulum traffic	138,490,852	-	G1037A	R346H	NR	0.001	0.13	29	C	-	1	0	1		.						
SEC16A	9	3	ns	Endoplasmic reticulum traffic	138,490,870	-	G1019A	G340E	NR	0.002	0.07	98	MC	-	1	0	1	.							
SH2B1	16	1	ns	Adaptor for TYK receptors	28,785,470	-	T554A	L185Q	NR	NR	0.04	113	MR	PrD	1	0	1								◊
SH2B1	16	5	ns	Adaptor for TYK receptors	28,790,787	-	A1495G	T499A	NR	NR	0.19	58	MC	PoD	1	0	1		.						
SMAD3	15	3	ns	TGF-β signalling pathway	65,244,752	rs35874463	A376G	I126V	0.018	0.053	0.65	29	C	B	2	0	2	.	.						
SNAPC4	9	17	ns	RNA polymerase transcription	138,396,228	rs34569521	G2186A	R729Q	0.046	0.088	0	43	C	PrD	1	0	1							◊	
SNAPC4	9	10	ns	RNA polymerase transcription	138,402,740	-	G1100C	G367A	NR‡	0.003	0	60	MC	PrD	1	0	1							◊	
SP140	2	24	ns	Nuclear body protein	230,883,793	rs62192163	T2266C	C756R	0.048	0.245	0.44	180	R	B	1	1	0	.							
SULT1A2	16	2	ns	Sulphate conjugation	28,514,733	rs1136703	T20C	I7T	0.014	0.059	0.12	89	MC	B	1	0	1	.							
TAGAP	6	6	ns	T cell regulation	159,382,412	rs41267765	G439A	E147K	0.014	0.020	0.64	56	MC	B	1	0	1	.							
TAGAP	6	5	ns	T cell regulation	159,383,130	-	G283A	G95S	NR	NR	0.58	56	MC	B	1	0	1	.							
THADA	2	28	nd	Apoptosis	43,508,785 *	-	4014_4016del	1338_1339del	NR	NR	-	-	-	-	1	0	1	.							
TYK2	19	20	ns	Th17-cell differentiation	10,325,843	rs35018800	C2783T	A928V	0.003	0.008	0	64	MC	PrD	1	0	1				◊				
TYK2	19	8	ns	Th17-cell differentiation	10,336,649	rs2304255	G1087A	G363S	0.035	0.080	0.54	56	MC	B	1	0	1								.
ZNF365	10	3	ns	Zinc finger	64,084,667	-	C97A	L33I	0.006	0.000	0	5	C	U	1	0	1	◊							

Novel variants are shown in grey.
Where a specific variant is present in a proband, this is indicated by a dot (.).
Where a specific variant is present in a proband and has a SIFT score of <0.05 this is indicated by a ◇.
*Indicates the first bp location of a 3-bp deletion.
†Indicates a dinucleotide variant (that for *IL18RAP* results in a codon change from CTG>AAG, resulting in p.L428K amino acid change).
‡NR indicates variants that despite not being reported in dbSNP132 or 1000 genomes, are reported in dbSNP129 or seen in our in-house reference exomes and are therefore not characterised as novel.
B, benign; C, conservative; fi, frameshift insertion; MC, moderately conservative; MR, moderately radical; nd, non-frameshift deletion; NR, not reported; ns, non-synonymous; PoD, possibly damaging; PrD, probably damaging; R, radical; sg, stopgain; U, unknown.

genes with convincing evidence for disease causality in previous studies. Selection was based on the findings of two genome-wide meta-analyses of IBD,[5 6] one genome-wide association study of early-onset IBD,[15] and one linkage study in consanguineous families with early-onset IBD.[16] Gene names were cross referenced with the Human Genome Nomenclature Committee to ensure that the most up-to-date versions of gene names were applied (http://www.genenames.org/). Our consolidated panel represented 169 genes (supplementary table 3).

Evaluation of spectrum of mutation and predicted functional impact

Exome data from our eight patients were cross-referenced against our gene panel described above. Synonymous variants

were excluded from analysis due to their decreased likelihood of functional effect on protein. SIFT ('sorting intolerant from tolerant') scores[23] were annotated using Annovar, or where scores were missing, were derived indirectly using the database of non-synonymous functional prediction.[24] A small number of additional missing scores were obtained from the SIFT server at http://sift.jcvi.org. SIFT is a sequence homology-based tool that predicts whether an amino acid substitution is likely to affect protein function. Variants with SIFT scores of <0.05 are considered 'deleterious', and SIFT therefore allows prioritisation of amino acid changes by ranking according to score.

We examined in silico predictions from the Polyphen2 (Polymorphism Phenotyping v2) server at http://genetics.bwh.harvard.edu/pph2/bgi.shtml.[25] Polyphen2 uses a probability model to generate thresholds and classify polymorphisms as benign, possibly damaging or probably damaging, based on 11 predictive features relating to sequence, phylogenetic and structural information which characterise the substitution. Additional functional predictions of the result of each amino acid change were derived from Grantham scores,[26] which predict the effect of amino acid substitutions according to chemical properties including polarity and molecular volume. The Grantham distance, d, between two amino acids is classified as conservative ($0<d\leq50$), moderately conservative ($50<d\leq100$), moderately radical ($100<d\leq150$) or radical ($d>150$).[27] Radical changes predicted by these scores are linked to clinical phenotypes.[28]

Burden of mutation

Using only novel variants or variants with an alternative allele frequency of <0.05 in the 1000 genomes data, a χ^2 contingency test was performed to test for an excess of rare potentially deleterious variants (non-synonymous and frameshift indels) compared to neutral synonymous variants, within the panel of known IBD genes in our eight cases compared to 22 reference exome samples from non-IBD patients.

RESULTS

Exome sequencing

On average, each PIBD exome had 78% of mappable bases of the Gencode defined exome represented by coverage of at least 20 reads (supplementary table 1). For each patient approximately 23 000 variants were found. After exclusion of synonymous variants, approaching 13 000 variants were found per patient, of which approximately 300 were novel (supplementary table 2).

Characterisation of mutations in genes known to be associated with IBD

Across all eight exomes, we found 332 variants (excluding synonymous) among 104 of our panel of 169 genes (supplementary table 4). Of these, approximately 40% (122) were found in HLA class genes. Seventeen were novel variants not previously reported in public databases or our own in-house database of non-IBD patient reference exomes.

Table 2 describes the set of variants remaining after removal of splicing, common (where the alternative allele frequency in 1000 genomes is reported as >0.05) and *HLA* variants. Fifty-eight variants within 39 genes remain, of which 17 are novel.

The χ^2 analysis to test for an excess of deleterious rare variants in known and candidate IBD genes in IBD cases listed in table 2 compared to 22 reference exomes did not reach statistical significance (supplementary table 5).

Crohn's disease patient profiles

Only two patients with early onset CD exhibit rare potentially deleterious variations within *NOD2*.

Proband 1 was diagnosed with CD aged 11 years and required a right hemicolectomy for extensive ileo-caecal stricturing. He is a heterozygote carrier of the *NOD2* R702W variant that is associated with a twofold increase in odds ratio of CD.[29] In addition he harbours potentially damaging mutations in *GSDMB* and *ZNF365* and a dinucleotide variant of undetermined functionality on one chromosomal copy of the *IL18RAP* gene. The presence of ileal disease and a stenotic phenotype in this patient is also consistent with his *NOD2* variant profile.[29]

Proband 2 carries a novel variant in each of the *SEC16A* and *SH2B1* genes. This patient also has a rare variant in *JAK2*; however, SIFT scoring suggests none of these mutations are likely to be particularly deleterious.

Proband 3 is the second patient with *NOD2* variation and carries both the R702W variant and the L1007 frameshift insertion. Carriage of two or more high risk alleles in *NOD2* confers a 17-fold increased risk of IBD.[29] Exome analysis cannot determine if both variants have been co-inherited on the same chromosome. Proband 3 additionally possesses potentially deleterious variants in *ERAP2* and *SEC16A*.

Proband 4 presented with severe disease aged 6 years. She carries the *NOD2* V955I variant, but this is predicted to be innocuous as is her private variant in *KIF21B*. She is a heterozygote for a number of previously seen variants with borderline (~0.05) SIFT scores (*FUT2*, *MTMR3*). The most distinct rare (frequency of 0.003) and potentially deleterious variant observed in this patient is the A928V variant in the *TYK2* gene.

Proband 5 possesses one variant in the *GMPBB* gene and another in *HORMAD2*, both estimated by SIFT to be harmful. The former is ascertained as novel to this individual, whereas the latter occurs in <0.5% of chromosomes studied in the thousand genomes project, but in just over 1% of the 3500 exomes tested in Exome Variant Server.

UC and IBDU patient profiles

Proband 6 has a histological diagnosis of UC and carries novel deleterious mutations in the *BACH2*, *C1orf93* and *SEC16A* genes. A fourth novel variant in the *IL10* gene also has a low SIFT score.

Proband 7 is a boy, diagnosed aged 2, and similar to our other UC patient, exhibits a potentially functionally detrimental mutation in *BACH2* and a second very rare and possibly damaging mutation in *IL10*. The *IL1RL2* and *SNAPC4* genes are also apparently compromised in this individual.

Proband 8 was diagnosed at a young age with IBDU, and possesses two possibly harmful variants in *ICAM1*, one in *BTNL2* and a novel deleterious variant in *SH2B1*.

Predicted functional impact

Figure 1 illustrates relationships between SIFT, Grantham and Polyphen2 scores for all non-synonymous variants in table 2. There is particularly close agreement between SIFT and Polyphen2 scores as noted previously.[30] Agreement with Grantham scores is less clear, but there is striking concordance between the vast majority of variants with a SIFT score >0.2 (benign) being independently designated benign by Polyphen2 and conservative by Grantham. Notably, two variants are classified as radical by Grantham and probably damaging by SIFT and/or Polyphen2—*CXCR1* (R335C) and *ICAM1* (R367C)—with the latter being classified as radical/damaging by all three criteria.

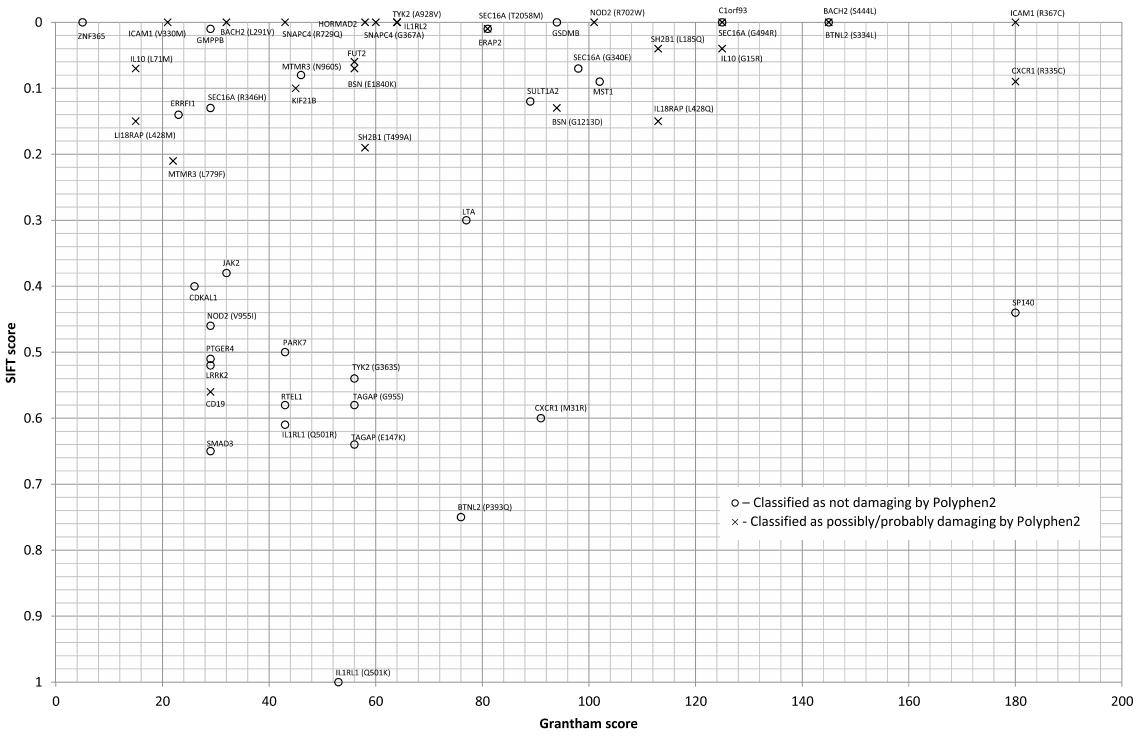


Figure 1 In silico functional predictions.

DISCUSSION

In this study we have applied exome sequencing, which allows the screening of the complete spectrum of variation within protein coding genes. There is abundant evidence that such regions are likely to be highly enriched for disease causing variation.[31] We have focused on the identification of rare and novel variation within genes known to contain causal variants or identified as candidate genes for IBD. Excluding *HLA* variants and considering only rare non-synonymous, stop-gain mutations and indels, we uncovered 58 variants across 39 genes, of which 17 were not previously reported. Of these, 35% (20 variants) have SIFT scores under 0.05, 12 of these are also classified as probably damaging by Polyphen2 and five of these (*BTNL2*: S334L; *C1orf93*: G176R; *ICAM1*: R367C; *NOD2*: R702W and *SH2B1*: L185Q) are also classified as moderately radical or radical by Grantham score. One variant, *CXCR1* (R335C), has a borderline SIFT score of 0.09 and is classified as probably damaging by Polyphen2 and radical by Grantham score. These variants may compromise protein function and contribute to the PIBD phenotype in these patients.

Our study included five patients with childhood onset CD. The variant profiles show that four of these patients carry potentially deleterious mutations in one or more IBD candidate genes. One child had a 17-fold increased risk of IBD on the basis of his *NOD2* profile alone. Others in this group bear variants with likely impact on antigen presentation (*ERAP2*), endoplasmic reticulum trafficking (*SEC16A*) and T-helper cell differentiation. A variant in the *IL18RAP* gene was recently reported by Rivas *et al*[7] to carry a threefold OR for CD, and variants in the same gene have also been implicated in coeliac disease.[32] We identify a rare, non-synonymous, two-base pair mutation in this gene in one of our severely affected early onset CD cases. Our study examined only two patients with a clear diagnosis of UC and intriguingly we observe unique, potentially deleterious variation in both the B-cell regulatory gene *BACH2* and *IL10* genes in both patients. Interestingly, defective *IL10* functioning is already recognised in UC pathogenesis,[33 34] whereas although other components of B-cell signalling (*IL7R* and *IRF5*) have shown previous association with UC,[6] variation in *BACH2* has shown previous association with CD only. Our patient with undetermined IBD is the only patient with rare *ICAM1* variants. This gene, in which our IBDU patient carries two functionally damaging variants, plays a role in cell-mediated inflammation and has been identified as a therapeutic target in IBD.[35]

Assessing our results obtained for each individual in our cohort with IBD, we can see clearly that it is possible to generate an individualised variant profile for each patient. Individualised profiles are already being usefully applied to refine disease diagnosis. For example, Franke *et al*[36] reported recently on a whole genome sequencing undertaken on a 47-year-old patient diagnosed with CD in her 20s. Her case was particularly severe, as she had failed standard treatments including anti-TNF, had undergone multiple bowel resections, and required intermittent parenteral nutrition. Sequencing in this patient revealed multiple 'hits' in the autophagy pathways. This prompted in-depth mycobacterial diagnostics and ultimately resulted in a diagnosis of chronic active *Mycobacterium avium* infection.

Although suggestive and interesting mutation profiles have emerged from our small panel, it is clear that our picture is far from complete. Proband 3 displays rare variation across many genes, but not one of these appears to have potential functional consequence. Furthermore, in 65 genes previously linked to IBD, we identified no variants in our eight probands. It is possible

that these genes do not contribute to disease in this small group, consistent with a high degree of genetic heterogeneity in this complex disease. It is also possible that limitations of sequencing technology or the analytical pipeline could have resulted in failure to call true variants. By focusing our analysis on exomes, we rely on the fact that many of the non-coding SNP variants previously implicated by GWAS simply flag coding variants in the genomic vicinity. Protein-coding genes harbour about 85% of the mutations with large effects for disease-related traits,[37] but it is entirely possible that restriction of the exome capture to coding regions might have overlooked non-coding variants with significant impact on protein expression. By tabulating rare and novel variants, we are focusing attention on those variants hypothesised to have larger effect sizes on the assumption that such variants confer significant genetic contribution to childhood severe and/familial disease.[38] However, for any complex disease, multiple common susceptibility variants, each contributing very modest effect sizes, should not be ignored.

SIFT, Polyphen2 and Grantham scores provide an indication of potential causality but they must be interpreted with caution, particularly for complex traits. Kumar *et al*[39] describe in silico prediction such as SIFT as effective for monogenic disease, but consider such tools to be less effective for lower penetrance variants associated with complex diseases. Furthermore, one study compiled in silico prediction scores and found pairwise agreement between all methods to be in the range 60–70%, implying fairly substantial disagreement.[24] These and other studies underpin the difficulty in ascribing functional evidence and translational importance of genetic variants, and the particular difficulty in heterogeneous complex disease. However, it is notable that published evidence demonstrates a clear functional impact for two of the six variants listed above as having an overall deleterious score by two or more of the in silico measures. The *CXCR1* gene R335C variant has been previously implicated in chronic obstructive pulmonary disease and asthma.[40] The two *CXCR1* mutations listed in table 2 (R335C and M31R) are in tight linkage disequilibrium and both are known to alter the structure and charge of the protein at the respective positions. The N-terminus of CXCR1 protein has been identified as potentially important for receptor–ligand binding, leading to the suggestion that the M31R variant may affect this interaction. This led to the hypothesis that both polymorphisms could impact receptor function through alterations in structure.[41] The upper right quadrant of figure 1 indicates those variants where all three *in silico* prediction tools are concordant in ascribing detrimental effects of the variant. Mutations such as the rare R376C *ICAM1* variant may modify the function of the encoded glycoprotein expressed on immune and endothelial cells and should be prioritised for functional assessment. Another non-synonymous variant highlighted by the in silico scores is the *NOD2* R702W variant which, together with the *NOD2* L1007fs variant, has been found to impair the activation of the NF-κB pathway in response to muramyl dipeptide (MDP), a bacterial wall component, with the L1007fs mutant unable to respond.[42] NOD2 is localised to the cell membrane but the L1007fs polymorphism disrupts this association and thus the protein has cytoplasmic distribution. Forcing the L1007fs mutant protein to associate with the plasma membrane does not lead to activation of the NF-κB pathway in response to MDP; thus it is not the localisation of the NOD2 mutant, but rather an inability to respond to MDP, that affects induction of the NF-κB pathway. The L1007fs mutation has been shown to produce a truncated protein with impaired function.[43] The *NOD2* R702W variant occurred in four of the 22 non-IBD reference exomes, representing a higher than expected frequency. Although the reference exomes were composed of germline DNA from patients with diverse diagnoses (various lymphomas, leukaemias and congenital growth disorders), all four of these IBD negative controls had a diagnosis of chronic lymphocytic leukaemia. Interestingly, a population based cohort study of 47 679 Swedish patients with CD or UC, reported a 20% increased risk of haematopoietic cancers in these patients.[44] However, the role of *NOD2* polymorphisms has been further investigated in a variety of cancers, with most finding no association.[45] Recently, however, Sivakumaran *et al*[46] found abundant evidence for pleiotropy in complex disease, defined as one gene having an effect on multiple phenotypes. The authors identified many genes harbouring variants associated with CD and other immune-mediated phenotypes. These associations include a CD association with chronic lymphocytic leukaemia, through the *SP140* gene (within which a rare variant is listed in table 2). Other gene/disease associations linked with CD include *BACH2* with type 1 diabetes and coeliac disease, *IL18RAP* in coeliac disease, *IL1RL1* with eosinophil count and coeliac disease, *MST1* with UC and primary sclerosing cholangitis, *ZNF365* with breast cancer, and *NOD2* with leprosy, among many others.[47] All of these genes contain rare variants listed in table 2 within the eight patients we have exome sequenced.

The abundance of potentially damaging variants arising from next generation sequencing renders interpretation of the potential impact of disease challenging. However, focusing on early onset and other forms of 'severe' phenotype, including familial cases, coupled with our ability to filter variants identified with increasingly large and reliable databases of apparently neutral variants, offers the prospect of identifying important rare variants involved in complex traits such as IBD. This is the first study whereby a cohort of patients have been exome sequenced with the specific aim of generating a unique and personalised profile of rare variants across known disease genes for each patient. The rare variant profiles presented here provide a relatively small number of potential causal variants and include many mutations classed as deleterious by in silico prediction, a number of potential compound heterozygotes and a number of variants for which there is established functional evidence of roles in disease. These data, assessed from the perspective of individual patients, provide one of the first glimpses of personal mutation profiles and establish a foundation to elucidate the disease significance of these variants in future next-generation sequencing analyses of PIBD patients.

Author affiliations
[1]Genetic Epidemiology and Genomic Informatics Group, Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, Duthie Building (Mailpoint 808), University Hospital Southampton NHS Foundation Trust, Southampton, UK
[2]NIHR Biomedical Research Unit (Nutrition, Diet & Lifestyle), University Hospital Southampton NHS Foundation Trust, Mailpoint 218, Southampton General Hospital, Tremona Road, Southampton, UK
[3]Paediatric Medical Unit, University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Tremona Road, Southampton, UK
[4]Human Genetics & Genomic Medicine, Human Genetics, Faculty of Medicine, University of Southampton Duthie Building (Mailpoint 808), University Hospital Southampton NHS Foundation Trust, Southampton, SO16 6YD, UK
[5]Division of Genetics and Molecular Medicine, King's College London School of Medicine, Guy's Hospital, London, UK

Acknowledgements The authors would like to thank Nikki J Graham from the DNA laboratory in Human Genetics & Genomic Medicine, University of Southampton; and David Buck and Lorna Gregory from the Wellcome Trust Centre for Human Genetics, Oxford University.

Contributors KC was responsible for analysis, and with AEW, interpretation of data, drafting of the manuscript, critical revision of article and final approval. RMB, NA and

CW were responsible for acquisition of data, critical revision and final approval of article. AC, JG, RU-G, WT, JWH and MS were responsible for interpretation of data, critical revisions and final approval. SE was responsible for conception, design, acquisition of data, analysis and interpretation of data, drafting, revision and approval of the final manuscript.

Funding This project was supported by: NIHR Biomedical Research Unit (Nutrition, Diet & Lifestyle), University Hospital Southampton NHS Foundation Trust with specific thanks to Liz Blake, Senior Paediatric Research Sister, and Rachel Haggarty, Senior Children's Research Nurse; University Hospital Southampton Foundation Trust R&D; and the Crohn's in Childhood Research Association (CICRA).

Competing interests None.

Patient consent Obtained.

Ethics approval This study was approved by the Southampton & South West Hampshire Research Ethics Committee (REC) (09/H0504/125).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Sawczenko A,** Sandhu BK, Logan RF, *et al*. Prospective survey of childhood inflammatory bowel disease in the British Isles. *Lancet* 2001;**357**:1093—4.
2. **Khor B,** Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011;**474**:307—17.
3. **Bengtson MB,** Solberg C, Aamodt G, *et al*. Familial aggregation in Crohn's disease and ulcerative colitis in a Norwegian population-based cohort followed for ten years. *J Crohns Colitis* 2009;**3**:92—9.
4. **Spehlmann ME,** Begun AZ, Burghardt J, *et al*. Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. *Inflamm Bowel Dis* 2008;**14**:968—76.
5. **Franke A,** McGovern DP, Barrett JC, *et al*. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;**42**:1118—25.
6. **Anderson CA,** Boucher G, Lees CW, *et al*. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011;**43**:246—52.
7. **Rivas MA,** Beaudoin M, Gardet A, *et al*. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011;**43**:1066—73.
8. **Bodmer W,** Tomlinson I. Rare genetic variants and the risk of cancer. *Curr Opin Genet Dev* 2010;**20**:262—7.
9. **Gilissen C,** Hoischen A, Brunner HG, *et al*. Unlocking Mendelian disease using exome sequencing. *Genome Biol* 2011;**12**:228.
10. **Worthey EA,** Mayer AN, Syverson GD, *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;**13**:255—62.
11. **Day-Williams AG,** Zeggini E. The effect of next-generation sequencing technology on complex trait research. *Eur J Clin Invest* 2011;**41**:561—7.
12. **de Ridder L,** Weersma RK, Dijkstra G, *et al*. Genetic susceptibility has a more important role in pediatric-onset Crohn's disease than in adult-onset Crohn's disease. *Inflamm Bowel Dis* 2007;**13**:1083—92.
13. **Biank V,** Broeckel U, Kugathasan S. Pediatric inflammatory bowel disease: clinical and molecular genetics. *Inflamm Bowel Dis* 2007;**13**:1430—8.
14. **Lacher M,** Kappler R, Berkholz S, *et al*. Association of a CXCL9 polymorphism with pediatric Crohn's disease. *Biochem Biophys Res Commun* 2007;**363**:701—7.
15. **Imielinski M,** Baldassano RN, Griffiths A, *et al*. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* 2009;**41**:1335—40.
16. **Glocker EO,** Kotlarz D, Boztug K, *et al*. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med* 2009;**361**:2033—45.
17. **Cooper GM,** Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**:628—40.
18. **IBD Working Group of the European Society for Paediatric Gastroenterology, Hepatology and Nutrition.** Inflammatory bowel disease in children and adolescents: recommendations for diagnosis—the Porto criteria. *J Pediatr Gastroenterol Nutr* 2005;**41**:1—7.
19. **Levine A,** Griffiths A, Markowitz J, *et al*. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm Bowel Dis* 2011;**17**:1314—21.
20. **Quinlan AR,** Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841—2.
21. **Li H,** Handsaker B, Wysoker A, *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078—9.
22. **Wang K,** Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.
23. **Ng PC,** Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812—14.
24. **Liu X,** Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;**32**:894—9.
25. **Adzhubei IA,** Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248—9.
26. **Grantham R.** Amino acid difference formula to help explain protein evolution. *Science* 1974;**185**:862—4.
27. **Li WH,** Wu CI, Luo CC. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* 1984;**21**:58—71.
28. **Botstein D,** Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;(33 Suppl):228—37.
29. **Economou M,** Trikalinos TA, Loizou KT, *et al*. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* 2004;**99**:2393—404.
30. **Rudd MF,** Williams RD, Webb EL, *et al*. The predicted impact of coding single nucleotide polymorphisms database. *Cancer Epidemiol Biomark Prev* 2005;**14**:2598—604.
31. **Lehne B,** Lewis CM, Schlitt T. Exome localization of complex disease association signals. *BMC Genomics* 2011;**12**:92.
32. **Dubois PC,** Trynka G, Franke L, *et al*. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;**42**:295—302.
33. **Franke A,** Balschun T, Karlsen TH, *et al*. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* 2008;**40**:1319—23.
34. **Festen EA,** Stokkers PC, van Diemen CC, *et al*. Genetic analysis in a Dutch study sample identifies more ulcerative colitis susceptibility loci and shows their additive role in disease risk. *Am J Gastroenterol* 2010;**105**:395—402.
35. **Philpott JR,** Miner PB Jr. Antisense inhibition of ICAM-1 expression as therapy provides insight into basic inflammatory pathways through early experiences in IBD. *Expert Opin Biol Ther* 2008;**8**:1627—32.
36. **Franke A,** Kuehbacher T, Nikolaus S, *et al*. The complete individual genome of a Female Crohn's disease patient—What can you Learn? *Gastroenterol* 2011;**140** (5 Suppl 1):S-90.
37. **Majewski J,** Schwartzentruber J, Lalonde E, *et al*. What can exome sequencing do for you? *J Med Genet* 2011;**48**:580—9.
38. **Bodmer W,** Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;**40**:695—701.
39. **Kumar S,** Dudley JT, Filipski A, *et al*. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet* 2011;**27**:377—86.
40. **Stemmler S,** Arinir U, Klein W, *et al*. Association of interleukin-8 receptor alpha polymorphisms with chronic obstructive pulmonary disease and asthma. *Genes Immun* 2005;**6**:225—30.
41. **Vasilescu A,** Terashima Y, Enomoto M, *et al*. A haplotype of the human CXCR1 gene protective against rapid disease progression in HIV-1+ patients. *Proc Natl Acad Sci U S A* 2007;**104**:3354—9.
42. **Lecine P,** Esmiol S, Metais JY, *et al*. The NOD2-RICK complex signals from the plasma membrane. *J Biol Chem* 2007;**282**:15197—207.
43. **Ogura Y,** Bonen DK, Inohara N, *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;**411**:603—6.
44. **Askling J,** Brandt L, Lapidus A, *et al*. Risk of haematopoietic cancer in patients with inflammatory bowel disease. *Gut* 2005;**54**:617—22.
45. **Yazdanyar S,** Nordestgaard BG. NOD2/CARD15 genotype, cardiovascular disease and cancer in 43,600 individuals from the general population. *J Intern Med* 2010;**268**:162—70.
46. **Sivakumaran S,** Agakov F, Theodoratou E, *et al*. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* 2011;**89**:607—18.
47. **Lees CW,** Barrett JC, Parkes M, *et al*. New IBD genetics: common pathways with other diseases. *Gut* 2011;**60**:1739—53.

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 2. 251–260
Advance Access published on 18 May 2012
doi:10.1093/bib/bbs024

Machine learning approaches for the discovery of gene–gene interactions in disease data

Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege and Andrew Collins

Submitted: 14th March 2012; Received (in revised form): 20th April 2012

Abstract

Because of the complexity of gene–phenotype relationships machine learning approaches have considerable appeal as a strategy for modelling interactions. A number of such methods have been developed and applied in recent years with some modest success. Progress is hampered by the challenges presented by the complexity of the disease genetic data, including phenotypic and genetic heterogeneity, polygenic forms of inheritance and variable penetrance, combined with the analytical and computational issues arising from the enormous number of potential interactions. We review here recent and current approaches focusing, wherever possible, on applications to real data (particularly in the context of genome-wide association studies) and looking ahead to the further challenges posed by next generation sequencing data.

Keywords: *machine learning; gene–gene interaction; random forest; support vector machines; multifactor-dimensionality reduction; genome-wide association study*

INTRODUCTION

Genes influence all human diseases and yet much of the genetic landscape of many common diseases is still uncharacterized. Genome-wide association studies (GWAS) using single nucleotide polymorphisms (SNPs) have been extensively used to uncover genetic architecture [1] by testing variants individually for association with particular diseases or traits [2, 3]. However, GWAS have explained only a small proportion of the genetic variation underlying disease [1, 4]. For common diseases the effect of an individual SNP on disease susceptibility is generally small and emerging evidence suggests that many low-penetrance variants interact multiplicatively [5] with increasing numbers of risk alleles contributing to significantly elevated disease risks [6]. Therefore, it is likely that much of the genetic variation underlying common diseases arises through interactions between many genes and environmental factors; a form of epistasis [7]. Thus the identification of individual disease-related SNPs may be less useful for disease prediction than the identification of the epistatic relationships underlying genetic disease.

The term epistasis has been used to refer to at least two phenomena which may be related in complex ways. Biological epistasis, which occurs at the cellular level, corresponds to the physical interactions amongst biomolecules in gene regulatory networks and pathways that impact on phenotype. Hence, the impact of a gene on an individual's phenotype depends on one or more additional genes. Alternatively, statistical epistasis reflects differences

Corresponding author. Andrew Collins. Genetic Epidemiology and Bioinformatics, Faculty of Medicine, University of Southampton, Duthie Building (808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD. Tel: 44 (0) 2380796939; Fax: 44 (0) 2380794264. E-mail: arc@soton.ac.uk

Rosanna Upstill-Goddard is studying for a Breast Cancer Campaign funded PhD at the University of Southampton focusing on machine learning approaches applied to early-onset breast cancer data.

Diana Eccles is Professor of Cancer Genetics at the University of Southampton and is undertaking research into early onset breast cancer using the 'Prospective study of Outcomes in Sporadic versus Hereditary breast cancer' (POSH) cohort.

Joerg Fliege is head of the Operational Research Group in the School of Mathematics at the University of Southampton and is involved in developing new mathematical tools and techniques in optimization and machine learning.

Andrew Collins is head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in association and sequencing studies for a number of diseases for which machine learning methods show promise.

in biological epistasis among a population of individuals: the deviation from additivity within a statistical model of the relationship between multiple genotypes and phenotype(s) at a population level [1, 8]. Moore and Williams [8] present conceptual relationships between biological and statistical epistasis diagrammatically in their Figure 2. Phillips [9] has suggested that epistasis can be split into three categories: compositional epistasis, functional epistasis and statistical epistasis. Compositional epistasis is introduced to represent the traditional definition of epistasis as the blocking of the effect of an allele by an allele at another locus. However defined, the relationships between biological and statistical forms of epistasis are complex and statistical interaction does not necessarily reflect interaction on a biological level [10].

One of the major problems associated with uncovering epistatic interactions is the volume of data to be analysed; as the number of SNPs increases the number of potential interactions increases exponentially [7], known as the 'curse of dimensionality'. The potential complexity of such interactions supports the use of machine learning and data mining techniques. Machine learning (ML) approaches employ algorithms to 'learn' from training data sets to solve problems and enable predictions about outcomes in other data based on patterns and rules learned. There are several issues that need to be considered when developing ML methods for the identification of epistasis including: genetic heterogeneity (which may be common in complex diseases[11]), the presence (or absence) of main effects, and the number of SNPs involved in the interactions (which is usually unknown in advance) [11].

EARLY ML APPROACHES

A range of ML methods have been developed over the past 10–15 years with the aim of uncovering gene–gene interactions implicated in common complex diseases. Here we discuss some approaches that have been used to detect epistasis, namely multifactor-dimensionality reduction (MDR), neural networks (NNs), random forest (RF), and support vector machines (SVMs).

Multifactor-dimensionality reduction

MDR was one of the first ML methods developed to detect and characterize gene–gene interactions [12, 13]. In the first stage of MDR, *n* genetic factors (e.g. SNPs) are selected from the entire set of factors. All possible multifactor (SNP genotype) combinations are represented in cells in *n*-dimensional space and each cell is assigned a case-control ratio. Multilocus genotypic predictors are thus reduced from *n* dimensions to one dimension by classifying each cell as either low-risk or high-risk, based on a threshold value of cases-to-controls [12, 14]. Following classification cross-validation is carried out to estimate the prediction error of each model by splitting the data into a training set consisting of 90% of the data and a testing set of the remaining 10%. A model is developed based on the classification of genotypes in the training set which is used to predict disease status of genotypes in the test set. The cross-validation process is repeated 10 times and the prediction error is averaged [12]. MDR modelling can thus be applied to real disease data to search for epistasis and any predictors designated as 'high-risk' are, therefore, potentially disease-related. This approach was evaluated using a sporadic breast cancer data set [12]. A statistically significant high-order interaction was detected amongst four polymorphisms in the absence of any significant main effects, one of the earliest reports of such an interaction associated with a common multifactorial disease. The power of MDR was found to be robust to the presence of 5% genotyping error, 5% missing data and a combination of the two for a number of different two-locus epistasis models. Additional advantages of using MDR for the discovery of epistasis include:

(i) The model-free approach, invaluable for diseases such as sporadic breast cancer for which the mode of inheritance is unknown and likely to be complex.
(ii) The capability of MDR for detecting and characterizing multiple genetic loci simultaneously and, through the use of cross-validation, minimizing the false-positive rate.
(iii) The number of interaction terms does not grow exponentially as each new variable is added [12].

However, some disadvantages associated with this method impact upon its reliability as a predictor of disease–genotype interactions. In the presence of a high (50%) phenocopy–genetic heterogeneity rate, power is greatly compromised [14] supporting the need for refinements to effectively deal with genetic heterogeneity in complex trait data. The resulting

models can be difficult to interpret [12]—although genotypes are classified as 'high-risk' or 'low-risk' there is no quantitative assessment of how high- or low-risk they are, thus it is difficult to determine which of the putative interactions are most likely to be disease-related and warrant further investigation. MDR (and extensions to MDR) have only been successful when applied to a small number of SNPs in certain genes of (known) interest [12, 13, 15, 16]. The MDR approach alone is not directly applicable to GWAS data, given the huge number of interactions to be assessed; however, using a filter algorithm to isolate a subset of potentially interesting SNPs for MDR analysis can overcome this limitation. Finally, MDR has a high false positive and negative error rate when the case and control ratio in a genotype combination is closely similar to that in the whole data set[15].

Neural networks

NNs were originally developed to model neurons but are now regularly used for data mining in a wide range of fields [17, 18] with 'feed-forward/back-propagation' networks being the most common [19]. They have excellent power for performing pattern recognition and classification [11] and are capable of dealing with voluminous data [18]. A NN resembling a directed graph where the nodes represent genetic elements (SNPs), and the arcs are the connections (interactions) between the elements, has been developed for genetic applications [18]. The nodes are arranged into layers. One or more nodes reside in the input layer and receive the information to be processed by the NN. The input layer links to multiple nodes in a hidden layer (of which there may be several) via arcs. Finally, there is an output node. Each arc is assigned a weight which, initially, are chosen randomly, but through training the network on test data, weights are adjusted to minimize the error rate [19]. The target of the NN is the recognition of corresponding patterns in real data, based on patterns observed in test data, and for predictions about patterns not seen before through recognizing sub-patterns and correlations in the data [19]. To uncover genetic loci potentially involved in epistatic interactions NNs are trained using known genotypes as inputs and known phenotypes as outputs and the development of the internal weighting structure is of particular importance. The internal weight structure of the network can be analysed after training to determine the effect of each locus on the resulting phenotype [19].

NN applications to disease data have shown variable success. Motsinger-Reif *et al.* [18] suggest that this may be due to the use of sub-optimal NN architecture. Exhaustive search of all possible architectures to find the optimal structure is infeasible and so one solution is to optimize architecture with ML algorithms. Examples of such algorithms include the Genetic Programming optimized NN (GPNN) [20, 21] and Grammatical Evolution NN (GENN) [18] [using genetic programming (GP) or grammatical evolution (GE) respectively to optimize a NN]. GP aims to 'evolve' computer programs to solve complex problems [22]. First, an initial population of randomly generated computer programs is produced. Each program is run on a problem and assigned a fitness value based on its performance. The best programs are chosen to go forward for 'reproduction' following the 'survival of the fittest' principle. Some programs are taken into the next generation unaltered, while others undergo 'crossover' in which new programs are created from combinations of components of the original programs. This procedure is repeated for a number of generations to find the optimal program [23]. GE is a variation of, and improvement on, GP, with more flexibility[17]. GE uses populations consisting of linear genomes which constitute individuals. Each genome is divided into codons which are translated into phenotypes (the NN) by the grammar [17]. In a similar way to GP, the resulting phenotypes can be tested for fitness and subsequent generations produced to find the optimal model. GPNN has higher power to detect gene–gene interactions in the presence of non-functional SNPs than the more traditional Back Propagation NN (BPNN) [23] while power comparisons have shown that GENN consistently outperforms GPNN [17, 18]. NNs can screen out loci that do not affect the phenotype, thus reducing the number of genetic locus combinations to be tested [19]. Network approaches can also be used to identify genetic interactions through exhaustive enumeration of all possible pairwise interactions; however, this approach only searches for SNPs with strong pairwise interactions so may overlook SNPs with higher order interactions [24].

Genetic heterogeneity, polygenic inheritance, high phenocopy rates, and incomplete penetrance are problematic in the search for epistasis. Some of the characteristics of NN methods render them capable of addressing these difficulties; pattern

recognition is well suited to address genetic heterogeneity and polygenic inheritance while signal filtering addresses high phenocopy rates and incomplete penetrance [19].

Random forest

RF are a type of high-dimensional non-parametric predictive model composed of a collection of classification or regression trees [25] generated from random vectors [26]. Each tree of a RF is grown from a training set (or bootstrap sample) from the original data using random feature selection and trees are grown to their full extent without pruning. The bootstrap sample of size *n* is produced from the original sample, also size *n*, with variables chosen with replacement. Thus some variables will be chosen multiple times while others will not be chosen at all [25].The best split at each node in each tree is chosen from a random subset of the predictor variables [27]. The so-called 'out-of-bag' (OOB) estimates of prediction error are then generated from the observations that are not chosen in the bootstrap sample (often up to one third of cases are not included). The RF algorithm is an effective prediction tool with the potential to uncover interactions among genes that do not exhibit strong main effects [22], however, it has been suggested that their ability to detect interactions actually depends on the presence of main effects, no matter how weak [28]. Thus, this approach may lack power to uncover those interactions that occur in the absence of any main effects.

A recent study used the RF approach to uncover interacting SNPs contributing to rheumatoid arthritis, but no significant interactions were found that could be replicated in a follow-up cohort [29]. Power calculations have further indicated that this method will only detect those interactions with a large effect size [29]. However, an advantage of RFs is that they do not 'overfit' the data, and, as the number of trees in the RF increases, the prediction error converges to a limiting value [26]. An importance score is provided for each variable in a RF [27] rendering it capable of identifying SNPs predictive of a phenotype. This has prompted suggestions that RFs could be used to highlight significant SNPs for analysis with other methods [25]. However, this would conflict with the suggestion that RFs are useful tools to uncover genetic epistasis since the detection of interactions between variables is more important than the effect of single SNPs on disease status. A further downside of the RF method is that, although it has shown considerable promise in low-dimensional data (~100 SNPs and 10,000 observations), it has not been successfully applied to GWAS data [28].

Support vector machines

SVMs are classification techniques which are potentially as powerful as NNs [30]. In the development of a supervised learning approach the actual outcome of the (training) data is given and similar patterns are searched for during testing [31]. In its simplest form, a SVM is focused on identifying a linear separator to divide data points of two classes and is thus a non-probabilistic binary linear classifier. Furthermore, using kernel functions, non-linear separators can be established by modifying the input space. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. SVMs have shown excellent power to detect epistasis in both simulated and real datasets [11, 31]; Listgarten *et al.* [31] identified variants in a number of genes associated with breast cancer risk. A quadratic kernel was used and the authors showed that multiple SNP sites from several genes at distant parts of the genome were better at identifying breast cancer patients than single SNPs. When compared to MDR this approach provides more interpretable output; however, unlike MDR, SVMs cannot cope well with missing data [11]. Chen *et al.* [11] employed an SVM approach which was combined with search algorithms to produce four different models to detect epistasis, in the absence of genetic heterogeneity. Sparse SVMs [32] have been developed to select variables for inclusion in the model as a pre-processing step. This technique aims to reduce the instabilities in SVM results that arise from small changes in training/validation data. Such an approach might be usefully applied to the study of epistasis.

MORE RECENT MODELS AND APPLICATIONS

Recently a number of new ML methods have been proposed which utilize different approaches to detect epistasis [13, 15, 16, 27, 28, 33–35], some based on the methods already discussed while others introduce new approaches.

Extensions to MDR

MDR has proven a popular method for identifying epistasis. Recently, new MDR models have been developed to address some of the limitations of the original methods, some of which have been successfully applied to a range of genetic association studies [13, 15, 16, 34, 35]. It was previously demonstrated that MDR cannot effectively model epistasis when genotype combinations have case-control ratios similar to those in the full data set. However, robust MDR (RMDR) was proposed to deal with this limitation. In traditional MDR if the case-control ratio of a cell equals that of the whole data set, or a cell is empty, then it is randomly assigned high- or low-risk status. In robust MDR genotypes are pooled into three groups: high-risk, low-risk, or unknown-risk—based on the statistical significance of association of each multilocus cell with case-control status. If a cell has case-control ratio equal (or close) to that of the whole data set, then it is labelled unknown-risk and is excluded from the model. Results are thus simplified and easier to interpret than traditional MDR. The RMDR approach was evaluated with a bladder cancer dataset and confirmed findings of the MDR approach but with a simpler, easier-to-interpret model which was more computationally efficient [13].

Quantitative classification of SNP–SNP interactions is absent from traditional MDR, making it difficult for researchers to establish which interactions are potentially the most important. Two recently proposed methods designed to produce more definitive results are OR-MDR and MB-MDR. Chung *et al.* [15] proposed Odds Ratio-based MDR (OR-MDR) which uses the same method as MDR to categorize genotypes as high-risk or low-risk but includes the odds ratio for each genotype combination. A high odds ratio indicates an interaction that is potentially more high-risk compared to an interaction with a low odds ratio, thus the most high-risk interactions can be easily identified with such an approach. Model-based MDR (MB-MDR) [34, 35] assigns genotypes into three categories: high-risk, low-risk, or no evidence and all three are tested for association with the disease/trait. This test is designed to reduce the risk of missing of important interactions and the inability to deal effectively with main effects and confounding factors. A two-stage approach is utilized to uncover interactions: a synergy measure among potentially interacting genes is used in the first stage and MB-MDR is used in the second stage. This method aims to produce more accurate identification of which interactions are high-risk and results show that, when compared to MDR, it identifies far fewer interactions as high- or low-risk. Many interactions that are classified as high-risk by MDR are assigned to the 'no evidence' group in MB-MDR. Evidence suggests that this is a smaller set of more reliable results and some of the genotypes assigned to the 'no evidence' group are in fact disease-related and should not be excluded from the model. The inclusion of this group seems to be particularly important in scenarios where minor allele frequencies (MAFs) are low, where genetic heterogeneity is present and when there is less power to make reliable statements about the risk status of the genotypes.

To address MDR's inability to take into account covariates such as age, sex and ethnicity, and inability to deal with continuous data, Generalized MDR (GMDR) has been proposed. GMDR is similar to MDR in that it splits genotypes into the two classifications of risk; however, it uses a score based on the maximum-likelihood estimates for each variable, rather than the case-control ratio. The score is calculated with the inclusion of terms for covariates and dichotomous/continuous data. GMDR significantly increases the accuracy of risk prediction when the data contains covariates and is applicable to balanced case-control or random data [16].

Extensions to RF

SNPInterForest [27] is based on the RF approach but is more successful at uncovering disease-associated SNPs and has the capacity to simultaneously identify multiple interactions. SNPInter Forest is more sensitive to SNPs with limited marginal effect, something the original RF algorithm performs poorly with. A modification to the RF method is introduced to prevent the importance scores of SNPs without marginal effects from being underestimated. Multiple-SNP selection occurs at each node, in contrast to the original RF algorithm in which only a single SNP is used. This significantly improves the ability of this approach to detect SNPs associated with disease. It can be challenging to extract useful biological information from RF analyses with respect to biological interactions. SNPInter Forest addresses this issue by evaluating the interaction strength of SNP combinations. Each branch of a tree represents a possible SNP interaction on that branch and, if a certain SNP combination appears

more often on a branch, then those SNPs are likely to interact more strongly. Interaction strength is calculated from the number of times each SNP combination appears in each branch of each tree. Normalisation is applied in order to identify the weaker interactions and those interactions due to single SNPs with strong associations. SNPInterForest also demonstrates high recall rates and low false positive rates, however, it is very computationally demanding [27].

SNPInterForest has been identified as outperforming other methods such as 'BOolean Operation-based Screening and Testing' (BOOST). BOOST is a computationally efficient two-stage statistical method applied to analyse all pairwise interactions in genome-wide data [36] and in simulated data with weak marginal effects. In the absence of marginal effects, BOOST has been shown to produce many more false positive results. The ability of SNPInterForest to detect high order epistatic interactions between more than two SNPs was also assessed in simulated data. Five datasets were produced using a model of three SNPs, two of which are moderately associated with the disease by a pure epistatic interaction and a third SNP with a weaker effect that amplifies the interactive effect. SNPInterForest successfully identified the interactions in all five datasets.

RFs are often used for the selection of a subset of variables [22] rendering them useful for identifying potentially interesting SNPs in a two-stage approach [33]. For example, the TRM method [33] uses RF to identify and select important variants and Multivariate Adaptive Regression Splines (MARS), a nonparametric regression method, to detect interactions. RFcouple [37] on the other hand has been suggested as a pre-screening method for MDR. The advantage of two-stage approaches is that a subset of potentially significant SNPs is selected by a filter algorithm and a ML approach is employed to search for potential interactions; thus a smaller set of potentially interesting SNPs can be exhaustively searched for epistasis [3, 38].

Lin *et al.* [33] combined both RF and MARS in the TRM approach because neither method alone was considered optimal for selecting the optimal genotype combination for predicting phenotypes in studies with large numbers of SNPs. Individually RFs can have difficulties revealing underlying interaction patterns and MARS can have difficulty coping with numerous non-functional SNPs. In this study three approaches were compared: TRM_{OOB}, TRM_{IS}, and MARS alone. TRM_{OOB} is a version of TRM that uses RF_{OOB} and MARS while TRM_{IS} uses RF_{IS} and MARS. RF_{OOB} is based on the unused 'OOB' observations and RF_{IS} is based on the importance spectrum of the original data compared with that of the permuted data. TRM_{OOB} demonstrated higher true positive rates and lower false positive rates than the other two approaches in a simulation study with 100 SNPs. TRM has not yet been applied to a large dataset, such as a GWAS, nor has it been compared to other more established methods with proven success in the past. The study has, however, demonstrated a two-stage approach for screening and testing SNPs as capable of uncovering potential interactions.

De Lobel *et al.* [37] propose using RFcouple as a pre-screening method for MDR. RFcouple is based on RFs but uses information from the ratio of cases to controls for each genotype to define a new variable for each SNP pair. Thus the data set contains a variable for each SNP pair rather than for each individual SNP. An RF is constructed based on these data and SNP pairs are chosen based on Z-scores, which are related to prediction error and standard error of the RF in the permuted data. The single SNPs that make up these pairs are retained and then analysed by a method such as MDR. Power when RFcouple is used before MDR is always comparable to or greater than using MDR alone.

Random Jungle (RJ) is an implementation of the RF method which is aimed at analysing data on a genome-wide scale, i.e. 1000s of SNPs [28]. Application of RJ to Crohn's disease GWAS data confirmed previous GWAS findings as well as uncovering new interactions between Crohn's-associated genes. RJ is much more computationally efficient than other RF implementations allowing for feasible analysis of high-dimensional GWAS data in a realistic time frame. However, RJ differs from many methods in that it tests association allowing for interaction, rather than testing directly for interaction [36]. In line with the traditional RF approach, RJ has difficulty in detecting interactions when SNPs only have weak main effects; the trees are constructed based on the main effects of SNPs [36], thus such an approach is not useful in the absence of main effects. Table 1 provides an overview of some of the main ML approaches used to detect gene–gene interactions and some of their strengths and limitations.

Table 1: Strengths and limitations of some machine learning approaches

Method	Strengths	Limitations
MDR	Detects multiple genetic loci simultaneously, keeping false-positive rate low. Model-free—important when mode of inheritance is unknown. Nonparametric—the number of interaction terms does not grow exponentially as each new variable is added. Power remains high with 5% genotyping error and/or 5% missing data for various two-locus epistasis models. Cross-validation minimizes false positive rate.	Power significantly reduced with high (50%) phenocopy/genetic heterogeneity. No quantitative assessment of each model to determine which is the most high-risk—models difficult to interpret. Can be computationally intensive, particularly when the number of SNPs to be evaluated exceeds 10. False positive/negative error rates high when case-control ratio in test data is close to that in the whole dataset. May identify totally different models influenced by missing values in the data.
RMDR	Produces easier to interpret models than MDR—classifies genotypes with a case-control ratio close to the whole data set as 'unknown risk' (excluded from model).	Significant computational burden. Takes longer to evaluate one-way, two-way and three-way models than MDR.
OR-MDR	Like MDR but with odds ratio for each genotype combination—a quantitative measure of disease risk. Provides confidence interval for each genotype combination.	Cannot classify an empty cell. Computationally expensive, particularly when the number of SNPs exceeds 10. Impact of genetic heterogeneity/phenocopy unknown.
MB-MDR	Genotypes classified as low-risk/high-risk/no evidence—reducing number of interactions classified as high-risk. Genotypes in the 'no evidence' group potentially disease related and considered in the model. Improved power and false positive rate compared to MDR.	Phenocopy and genetic heterogeneity significantly reduce power.
GMDR	Uses score based on maximum-likelihood (ML) rather than case-control ratio. ML score includes covariates-significantly increasing accuracy of risk prediction.	Like MDR genotypes only assigned to two risk groups with no quantitative assessment. Can be computationally intensive, as above.
Neural Networks (NNs)	Excellent power for pattern recognition/classification Capable of dealing with large volumes of data. Accommodates genetic heterogeneity/polygenic inheritance/high phenocopy rates/incomplete penetrance.	It is impossible to enumerate all possible NN architectures and altering the architecture can change results of data analyses. Thus there is no way to be certain that the architecture being used is optimal.
GPNN	GP optimizes the architecture of NN. High power to detect interactions in the presence of non-functional SNPs. Preferable when functional SNPs are unknown and variable selection as well as model fitting required. Does not 'overfit' data. High power in epistasis model with weak marginal effect. Modelling flexibility—no need to select optimal inputs, weights, connections, or hidden layers.	High false positive rates in three locus models. Requires a parallel processing environment. The output is a binary expression tree which can be large (up to 500 nodes) and difficult to interpret.
GENN	GE optimizes the NN architecture. Consistently outperforms GPNN—optimizes NN more efficiently in fewer generations than GP. High power to detect risk loci in complex disease.	
RF	May uncover interactions among genes that do not exhibit strong main effects. Does not 'overfit' the data and prediction error converges to a limiting value. Identifies SNPs predictive of a phenotype.	Ability to detect interactions depends on main effects, no matter how weak. No demonstrated success in GWAS data. Sometimes underestimates importance scores of SNPs without marginal effects. Can be challenging to extract useful biological information. Only detects interactions with large effect size.
SNPInterForest	Simultaneously identifies multiple interactions. Does not underestimate importance scores of SNPs without marginal effects. Multiple SNP selection at each node improves ability to detect disease SNPs even when marginal effects absent. Evaluates interaction strength of SNP combinations. Demonstrates high recall/low false-positive rates. Interactions found in presence of genetic heterogeneity.	Very computationally intensive.

(continued)

Table 1 Continued

Method	Strengths	Limitations
TRM	A subset of potentially interesting SNPs can be searched exhaustively for interactions.	Has only been applied to small data sets (100 SNPs). Has not been compared to more established methods.
RJ	Designed to analyse data on a genome-wide scale. More computationally efficient than RF implementations—analysis of high-dimensional GWAS data feasible.	Tests association allowing for interaction rather than testing directly for interactions. May fail to detect interactions when only weak main effects.
SVM	More interpretable output compared to MDR. Readily generalized to new data structures. No user-defined decisions required for classification.	May not cope well with missing data. Power reduced in the presence of genetic heterogeneity.

LIMITATIONS OF CURRENT MODELS AND FUTURE DIRECTIONS

There are many difficulties associated with the detection of epistasis in GWAS related to both the data to be analysed and the capabilities of the ML methods being used. Firstly, there is the complexity of the disease data which includes allelic/locus heterogeneity, phenocopies, trait heterogeneity, phenotypic variability [38] and incomplete penetrance [10]. Some of the models discussed here have been developed to deal with such limitations. For example, it has been suggested that RF methods may be successful at dealing with certain types of heterogeneity [22, 27, 28], while some of the characteristics of NNs render them capable of addressing genetic heterogeneity, polygenic inheritance, high phenocopy rates and incomplete penetrance [17, 19].

Secondly, the computational burden associated with the search for gene–gene (SNP–SNP) interactions is potentially huge [39], particularly when searching for interactions between two or more SNPs within a GWAS [3]. The majority of methods discussed have demonstrated success in simulated data containing, at most, a few hundred SNPs. While such results are promising, it is currently unclear how successful some of these methods will be at dealing with up to 500 000 SNPs in a GWAS. Aside from the computational burden, the outputs may present serious challenges for biological interpretation. The power of many methods is significantly reduced when attempting to uncover higher order interactions. The issue of designing 'sufficiently powerful' studies has received relatively little attention to date. Although an exhaustive search of pairwise interactions in GWAS data might become computationally feasible extensive validation of candidate interactions in independent samples is essential, as in GWAS generally, to confirm or refute discoveries. More sophisticated approaches capable of modelling higher order interactions need to be developed but may require the use of expert knowledge of biological and biochemical pathways to choose SNPs likely to be associated with a particular disease [22]. It may also be necessary, and more powerful, to employ a two-stage model, in which filter algorithms select a subset of SNPs and a ML method exhaustively searches for interactions [3, 38, 40]. This approach may be less time-consuming and produce easier-to-interpret models [40]. However, some argue it is likely that SNPs with strong epistasis but weak main effects will be filtered out [36], so these methods will not necessarily find the optimal solution. Moreover, it is often the case that individual SNPs are assessed for disease association based on an importance score that does not take into account interactions with other SNPs. Clearly, when searching for epistasis, it is the interactions between SNPs that are of importance. Thus, a SNP with a high importance score but no involvement in SNP–SNP interactions is clearly not useful in this context.

There may be little similarity between biological and statistical epistasis; biological epistasis occurs at the cellular level within an individual while statistical epistasis addresses genetic variation on a population scale [1]. Most methods, however, test statistical, rather than biological, epistasis [36].

Finally, all approaches discussed are successful at uncovering epistasis in simulated data, with some also being successful in application to disease data of varying volume and complexity (Table 2). However, most studies have focussed on validating previously produced results with few actually uncovering new disease related interactions. While it is important that methods are tested on real data

Table 2: Some applications of machine learning approaches to genetic data

Method	Successful scenario	Reference
MDR	Applied to 10 polymorphisms in five genes related to oestrogen metabolism in breast tissue. A four-locus interaction associated with risk for sporadic breast cancer identified.	[12]
RMDR	Study examined the relationship between DNA repair gene SNPs, smoking and bladder cancer. Seven SNPs in five genes involved in DNA repair tested. Verified results from an MDR study using the same data but provided a much clearer model of high risk interactions.	[13]
OR-MDR	Applied to 42 SNPs in 10 genes related to chronic fatigue syndrome. Both MDR and OR-MDR applied to all possible SNP combinations up to the fourth order.	[15]
MB-MDR	Applied to 282 SNPs in 108 genes of the inflammation pathway of bladder cancer. Eight second-order interactions and 14 third-order interactions were identified.	[34]
GMDR	Applied to 23 SNPs in four genes to identify susceptibility genes for nicotine dependence. GMDR and MDR identified the same interactions.	[16]
GPNN	Applied to 22 SNPs in nuclear-coded mitochondrial complex I genes in a Parkinson's disease cohort. A two-locus interaction between the DLST gene and sex was detected.	[41]
GENN	Applied to 35 SNPs in five genes that encode proteins involved in IL-2/IL-15 signalling. Replicated findings from analysis using MDR.	[18]
RF	Applied to 42 SNPs from the asthma-related ADAM33 gene.	[25]
SNPInterForest	Applied to GWAS data of rheumatoid arthritis from the Wellcome Trust Case Control Consortium (~500 000 SNPs).Two novel interactions identified.	[27]
TRM	Applied to 106 SNPs in six oestrogen receptor-related genes from prostate cancer patients. Interactions identified between SNPs in two genes previously linked to prostate cancer risk and premature ovarian failure.	[33]
RJ	Applied to GWAS data of Crohn's disease containing ~275 000 SNPs. Results validated findings from other GWAS and identified new interactions.	[28]
SVM	Applied to 57 SNPs in 18 genes in a prostate cancer study. Identified high-order interactions between up to five SNPs in line with results from MDR on the same data.	[11]

for which there are already known interactions, there is a pressing need for ML applications that uncover important new interactions in common diseases. No single method has been particularly successful in this respect as yet.

Given the increasingly voluminous genetic data now being produced by next generation sequencing studies, and the emerging evidence that very large numbers of individually low risk variants underlie common diseases, the need for powerful ML models is more pressing than ever. It is evident that current methods require further development before successful application to these enormous data sets can be claimed and their outputs enhance understanding of the genetic epidemiology of disease or become useful in a clinical disease risk predictive setting.

Key Points

- ML methods including multifactor dimensionality reduction, RFs, NNs and SVMs have been developed and applied with some success to detect gene–gene interactions in simulated and disease data.
- Extensions to the original models have facilitated application to large data sets for some approaches but computational and problems of biological interpretation remain.
- Recent two-stage methods which initially reduce the number of potentially interesting SNPs in which to search for interaction may be promising but no single method fully addresses the complexity of the problem.
- Suitable models and applications for the even more complex and voluminous next generation sequencing data are currently lacking.

FUNDING

This work was supported by the Breast Cancer Campaign.

References

1. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet* 2009;**85**:309–20.
2. Hirschhorn JN. Genomewide association studies—illuminating biologic pathways. *N Eng J Med* 2009;**360**:1699701.
3. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
4. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
5. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet* 2008;**40**:17–22.

6. Harlid S, Ivarsson MIL, Butt S, *et al.* Combined effect of low-penetrant SNPs on breast cancer risk. *Br J Cancer* 2012; **106**:389–96.
7. Moore JHP, Ritchie MDP. The challenges of whole-genome approaches to common diseases. *JAMA* 2004;**291**: 1642–3.
8. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* 2005;**27**:637–46.
9. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 2008;**9**:855–67.
10. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002;**11**:2463–8.
11. Chen S-H, Sun J, Dimitrov L, *et al.* A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol* 2008;**32**:152–67.
12. Ritchie MD, Hahn LW, Roodi N, *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001;**69**:138–47.
13. Gui J, Andrew AS, Andrews P, *et al.* A robust multifactor dimensionality reduction method for detecting gene–gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann Hum Genet* 2011;**75**:20–8.
14. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003;**24**:150–7.
15. Chung Y, Lee SY, Elston RC, *et al.* Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics* 2007;**23**:71–6.
16. Lou X-Y, Chen G-B, Yan L, *et al.* A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am J Hum Genet* 2007;**80**:1125–37.
17. Motsinger A, Dudek S, Hahn L, *et al.* Comparison of neural network optimization approaches for studies of human genetics. *Lect Notes Comp Sci* 2006;**3907**:103–14.
18. Motsinger-Reif AA, Dudek SM, Hahn LW, *et al.* Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol* 2008;**32**:325–40.
19. Lucek PR, Ott J. Neural network analysis of complex traits. *Genet Epidemiol* 1997;**14**:1101–6.
20. Ritchie MD, Motsinger AA, Bush WS, *et al.* Genetic programming neural networks: a powerful bioinformatics tool for human genetics. *Appl Soft Comput* 2007;**7**:471–9.
21. Koza JR, Rice JP. 'Genetic generation of both the weights and architecture for a neural network'. *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference* 1991, Vol. 392, pp. 397–404.
22. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010;**26**:445–55.
23. Ritchie M, White B, Parker J, *et al.* Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics* 2003;**4**:28.
24. Hu T, Sinnott-Armstrong N, Kiralis J, *et al.* Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 2011;**12**:364.
25. Bureau A, Dupuis J, Falls K, *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005;**28**:171–82.
26. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
27. Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinformatics* 2011;**12**: 469.
28. Schwarz DF, Konig IR, Ziegler A. On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* 2010;**26**:1752–8.
29. Liu C, Ackerman H, Carulli J. A genome-wide screen of gene–gene interactions for rheumatoid arthritis susceptibility. *Human Genetics* 2011;**129**:473–85.
30. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
31. Listgarten J, Damaraju S, Poulin B, *et al.* Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clin Cancer Res* 2004;**10**:2725–37.
32. Bi J, Bennett K, Embrechts M, *et al.* Dimensionality reduction via sparse support vector machines. *J Mach Learn Res* 2003;**3**:1229–43.
33. Lin H-Y, Ann Chen Y, Tsai Y-Y, *et al.* TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. *Ann Hum Genet* 2012;**76**:53–62.
34. Calle ML, Urrea V, Vellalta G, *et al.* Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat Med* 2008;**27**:6532–46.
35. Cattaert T, Calle ML, Dudek SM, *et al.* Model-based multifactor dimensionality reduction for detecting epistasis in case–control data in the presence of noise. *Ann Hum Genet* 2011;**75**:78–89.
36. Wan X, Yang C, Yang Q, *et al.* BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010;**87**:325–40.
37. De Lobel L, Geurts P, Baele G, *et al.* A screening methodology based on Random Forests to improve the detection of gene–gene interactions. *Eur J Hum Genet* 2010;**18**: 1127–32.
38. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet* 2004;**20**:640–7.
39. Wang Y, Liu G, Feng M, *et al.* An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 2011;**27**:2936–43.
40. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–7.
41. Motsinger A, Lee S, Mellick G, *et al.* GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics* 2006;**7**:39.

OPEN ACCESS Freely available online

PLOS ONE

Support Vector Machine Classifier for Estrogen Receptor Positive and Negative Early-Onset Breast Cancer

Rosanna Upstill-Goddard[1], Diana Eccles[1], Sarah Ennis[1], Sajjad Rafiq[1], William Tapper[1], Joerg Fliege[2], Andrew Collins[1]*

1 Human Genetics and Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom, **2** Centre for Operational Research, Management Science and Information Systems, University of Southampton, Southampton, United Kingdom

Abstract

Two major breast cancer sub-types are defined by the expression of estrogen receptors on tumour cells. Cancers with large numbers of receptors are termed estrogen receptor positive and those with few are estrogen receptor negative. Using genome-wide single nucleotide polymorphism genotype data for a sample of early-onset breast cancer patients we developed a Support Vector Machine (SVM) classifier from 200 germline variants associated with estrogen receptor status ($p<0.0005$). Using a linear kernel Support Vector Machine, we achieved classification accuracy exceeding 93%. The model indicates that polygenic variation in more than 100 genes is likely to underlie the estrogen receptor phenotype in early-onset breast cancer. Functional classification of the genes involved identifies enrichment of functions linked to the immune system, which is consistent with the current understanding of the biological role of estrogen receptors in breast cancer.

Citation: Upstill-Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, et al. (2013) Support Vector Machine Classifier for Estrogen Receptor Positive and Negative Early-Onset Breast Cancer. PLoS ONE 8(7): e68606. doi:10.1371/journal.pone.0068606

Editor: Syed A. Aziz, Health Canada and University of Ottawa, Canada

Received April 11, 2013; **Accepted** May 30, 2013; **Published** July 19, 2013

Funding: This work was supported by the Breast Cancer Campaign. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: arc@soton.ac.uk

Introduction

Breast cancer sub-types may be classified according to the number of estrogen receptors present on the tumour. Tumours expressing large numbers of receptors are termed estrogen receptor positive (ER+) and, conversely, estrogen receptor negative (ER−) for few or no receptors. ER status is extremely important since ER+ cancers grow under the influence of estrogen, and may therefore respond well to hormone suppression treatments, while the proliferation of ER− cancers is not driven by estrogen and does not respond to estrogen modulation. Deroo and Korach [1] describe the "classical" (or genomic) pathway of estrogen action: an estrogen molecule binds to a receptor which induces receptor phosphorylation and dimerization to form a nuclear estrogen-ER complex [1,2]. The transcription of target estrogen responsive genes is regulated through the binding of the estrogen-ER complex to specific estrogen response elements (EREs) located in the gene promoter region [3]. The target genes of this pathway are many and varied; the majority are crucial for normal cell physiology, growth and differentiation and can promote the growth of breast tumours under certain conditions [2,4].

Two hypotheses seek to explain the relationship between estrogen and breast cancer. The first considers the proliferation of mammary cells stimulated by the binding of estrogen to the ER leading to an increase in the number of target cells and associated elevated risk for replication errors and acquisition of deleterious mutations during cell division and DNA replication. A second hypothesis identifies genotoxic by-products of estrogen metabolism which may lead to DNA damage and, subsequently, cancer. Evidence exists to support both hypotheses as mechanisms to initiate and promote tumour development [1]. Estrogen is necessary for breast tumour formation regardless of the receptor status of the cells and the tumour-promoting effects of estrogen are not limited to ER+ cells alone [5]. While estrogen influences the growth of ER+ tumour cells through binding receptors it is suggested that the growth of ER− tumour cells is the result of estrogen acting on cells of the tumour microenvironment which enhances angiogenesis, stromal cell recruitment and thus, tumour development and progression [5,6].

The estrogen receptor has two forms, α and β, which are encoded by the *ESR1* and *ESR2* genes respectively. The two forms have distinct roles in breast tissue; ERα promotes cell proliferation in response to estrogen while ERβ inhibits proliferation and tumour formation [7,8]. Single nucleotide polymorphisms (SNPs) in the *ESR1* gene have been associated with increased susceptibility to breast cancer, however they are fairly rare [9–11]. Variation in the *ESR2* gene may also be important in disease susceptibility however, no SNPs demonstrating a strong association with breast cancer risk have been identified [1,12,13]. A number of SNPs have been identified through genome wide association studies (GWAS) as being breast cancer risk SNPs. In many cases these SNPs relate to the risk of developing a particular subtype of disease, often the ER+ type [14]. Overall, the genetic basis of the estrogen receptor cancer sub-types is not well understood and worthy of further analysis [1].

We hypothesized that patients who develop ER+ and ER− tumours would show distinct constitutional genetic profiles the exploration of which could yield new insights into the biological effect of the host genomic environment on the emergence of these forms of breast cancer. We developed machine learning (ML)

Table 1. Weka kernels and classification results using 200 SNPs with the strongest ER+/− association.

Kernel type	Percentage correctly classified	True positive rate	False positive rate	True negative rate	False negative rate	Area under ROC
Linear	93.28±3.07	0.88±0.07	0.04±0.03	0.96±0.03	0.12±0.07	0.92±0.04
Normalized quadratic polynomial	95.69±2.69	0.89±0.08	0.01±0.02	0.99±0.02	0.11±0.08	0.94±0.04
Quadratic Polynomial	93.89±3.06	0.89±0.07	0.04±0.03	0.96±0.03	0.11±0.07	0.93±0.04
Cubic Polynomial	94.54±2.94	0.89±0.07	0.03±0.03	0.97±0.03	0.11±0.07	0.93±0.04
RBF	95.95±2.61	0.89±0.07	0.01±0.02	0.99±0.02	0.11±0.07	0.94±0.04

doi:10.1371/journal.pone.0068606.t001

classifiers to explore the distinction between profiles in well characterised breast cancer cases. ML is used extensively in many scientific fields for classification purposes. ML methods have been used in genetic studies to explore the underlying genetic profile of disease and build models capable of (i) detecting gene-gene interactions; (ii) predicting disease susceptibility; (iii) predicting cancer recurrence; and (iv) predicting cancer survivability [15]. Genetic SNP data can be used to build such classification models, with high accuracy observed in many cases. Support Vector Machines (SVMs) have been shown to have excellent power and the ability to establish binary classification based on multiple features [16]. The aim of the SVM approach is to separate the data points from the two classification groups using a decision surface, called a hyperplane. The simplest classifier is a linear hyperplane but, for more complex datasets, it is necessary to map the input features into high-dimensional space using a non-linear mapping function, called a kernel function [16]. The placing of the separating hyperplane depends on maximising the margin between the hyperplane and the data points of two classes. If the input data are not cleanly separable by a hyperplane (a non-separable case, [17]), it is desirable to separate the data by the smallest sum of all classification errors: the 'soft margin hyperplane'. In the case of genetic data linear models may be sufficient in the absence of, for example, complex underlying gene-gene interactions whilst kernel functions are most applicable otherwise. We develop here a SVM classifier which discriminates ER+ and ER− breast cancer cases which provides new insights into the biological nature of the ER+/ER− breast cancer sub-division.

Results

SVM classification accuracy

The overall classification accuracy of a ML classifier is a measure of how successful the method is at assigning samples to the correct class. In this study the highest classification accuracy was achieved using 200 SNPs fully genotyped in all 542 study samples (Table S1) and individually associated with the ER-negative phenotype ($p<0.0005$). Five kernel models were produced, all with classification accuracy exceeding 93% (Table 1). Classification accuracy was reduced when the highest ranked 50 (<86%) and 100 (<93%) SNPs were considered (Table S2). The highest classification accuracy was achieved using the radial basis function (RBF) kernel and normalized quadratic polynomial kernel: 95.95% and 95.69% respectively. In both cases 99% of the ER− cases and 89% of the ER+ cases were classified correctly. The true positive rate (number of ER+ cases correctly classified) was equal in all five models, demonstrating that they are equally successful at recognising and classifying ER+ cases in the test data. The true negative rate always exceeds 0.95, indicating that at least 95% of ER− cases are classified correctly in each model. All models are superior at classifying ER− cases compared to ER+ cases.

Classifier performance was further evaluated using the receiver operating characteristic (ROC) area under curve (AUC) values which indicate these models have excellent accuracy: all exceed 0.9 (Table 1). ROC curves were produced for the linear model and RBF kernel model for both ER+ and ER− cases (Figure 1) based on true and false positive/negative values. Figure 2 shows the relationship between chi-squares for individual SNPs derived from PLINK [18,19] and weights from the linear classification model. Variants with the largest (absolute value) weights are the most discriminating in the classifier. The input chi-squares used in feature selection (see methods) are uncorrelated with the linear SVM model weights ($r = -0.026$).

SVM classifiers were produced for two additional subsets of SNP features to further investigate classification accuracy. A set of 200 SNPs showing no individual association for the ER+/ER− distinction and a subset of 200 randomly selected SNPs were used to produce classification models (Table S3). Accuracy is low for both subsets (<69%) as are true positive rates in both cases (<33%). Area under ROC curve values are also very low at 0.51 or less, indicating that these models perform no better than 'random' which achieves an AUC of 0.5.

DAVID functional annotation

To identify biological terms and pathways that are particularly enriched for genes represented in the classifier (Table S1) we used the DAVID annotation tool [20–22] DAVID identified four gene annotation clusters, three enriched pathways and 36 term annotation clusters. Of these, two gene annotation clusters and 9 term annotation clusters are particularly enriched (enrichment score ≥1.00) relative to the whole genome background. The cluster with the highest enrichment score contains genes related to the inflammatory response (Table 2) and the next highest (Table 3) shows enrichment of genes in specific pathways related to axon guidance and signalling.

DAVID analysis was also performed for the 100 SNPs with the highest absolute classifier weights (Table S1) from the linear SVM kernel model. Similar annotation clusters were identified (data not shown) with functions relating to immune cell activation again being particularly enriched in the gene set.

Discussion

Machine learning techniques have an important role to play in disease classification and the discovery of underlying disease mechanisms, including gene-gene interactions or signalling path-

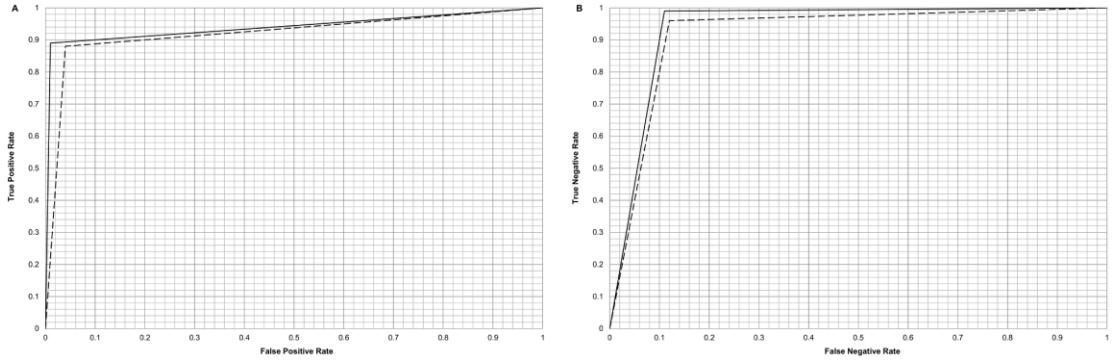


Figure 1. ROC curves for ER+ and ER− classification using linear and RBF kernels. ROC curves and area under ROC curve (AUC) values can be used as more robust measures of classifier accuracy beyond overall classification accuracy. (A) ROC curves for ER+ classification. (B) ROC curves for ER− classification. In both cases the linear model is represented by a dashed line and the RBF kernel model is represented by a solid line. The point on each curve corresponds to the true positive/negative and false positive/negative values obtained from 100 iterations of 10-fold cross-validation carried out on 542 samples with 200 SNP features. The ROC curve for any meaningful classifier needs to lie above the $y = x$ line; the case where equal proportions of cases would be classified correctly and incorrectly, as would occur if class values were assigned at random.
doi:10.1371/journal.pone.0068606.g001

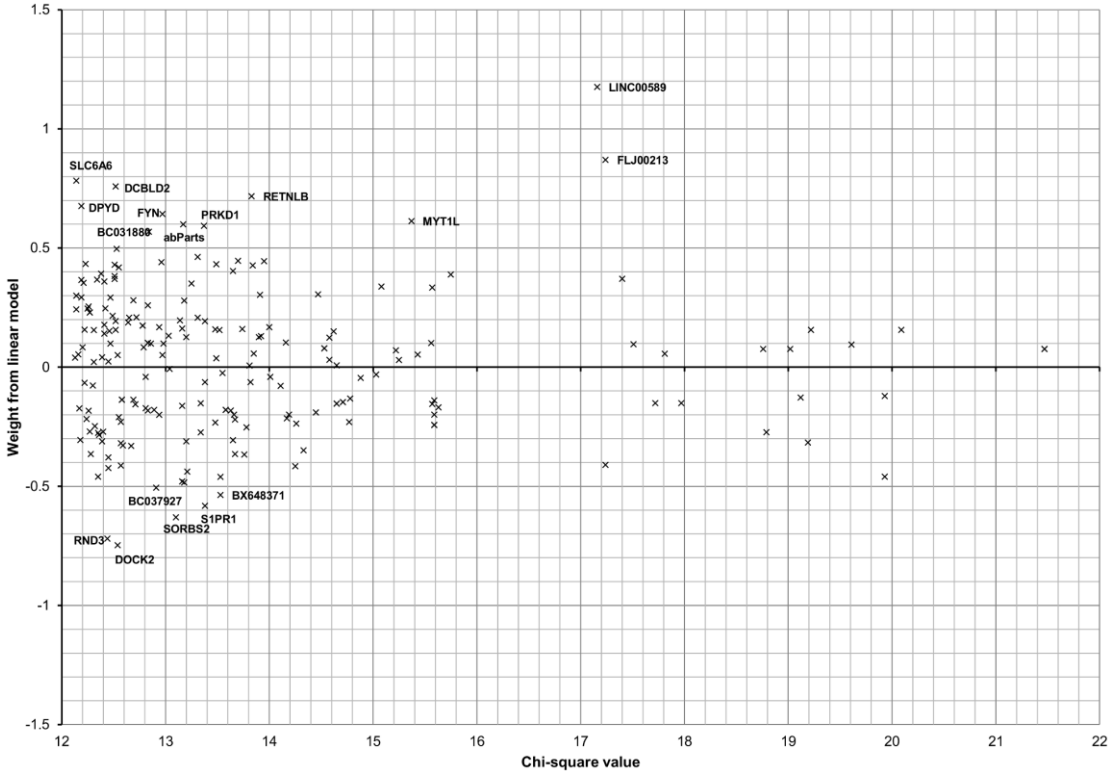


Figure 2. Relationship between weights under a linear classifier and chi-square values used in feature selection. SVM models were constructed on 542 study samples with genotype data for a subset of 200 SNPs chosen based on ER+/− association, determined from the chi-square statistic. SNP feature weights were obtained from the linear SVM model and used as an indicator of the importance of each feature for classification; SNPs with the largest absolute weight values are the most important for classification. Chi-square values used in feature selection and SVM classifier weight values are uncorrelated; Pearson's correlation coefficient $r = -0.026$. SNPs with absolute weight values > 0.5 are annotated with the name of the gene in which they reside or are in closest proximity to.
doi:10.1371/journal.pone.0068606.g002

Table 2. DAVID Annotation Clusters: Enriched gene ontology (GO) terms from the ER+/− classification.

Cluster 1: Enrichment Score: 1.97 (GO: Biological Process)				
Term	**No. genes**	**% genes**	**P value**	**Fold Enrichment**
calcium ion transport	7	6.03	0.00018	8.34
T cell proliferation	4	3.45	0.00051	25.05
di-, tri-valent inorganic cation transport	7	6.03	0.00056	6.73
T cell activation	6	5.17	0.00084	8.05
lymphocyte proliferation	4	3.45	0.00187	16.10
leukocyte proliferation	4	3.45	0.00214	15.37
mononuclear cell proliferation	4	3.45	0.00214	15.37
lymphocyte activation	6	5.17	0.00613	5.09
positive regulation of immune system process	6	5.17	0.01269	4.26
leukocyte activation	6	5.17	0.01356	4.19
cell proliferation	8	6.90	0.01360	3.10
response to abiotic stimulus	7	6.03	0.02048	3.22
cell activation	6	5.17	0.02619	3.54
Cluster 2: Enrichment Score: 1.93 (GO: Cellular Component)				
Term	**No. genes**	**% genes**	**P value**	**Fold Enrichment**
synapse	10	6.90	0.00024	4.50
cell junction	11	9.48	0.00128	3.39

doi:10.1371/journal.pone.0068606.t002

way enrichment which influences disease. Support vector machines in particular are state-of-the-art classifiers [23] with documented success at building accurate classifiers for disease versus control populations based on genetic data [24–26]. As discussed here, SVMs are useful for the analysis of disease sub-types given that many diseases (breast cancer included) comprise distinct sub-types with individual biology. The best resultant SVM model for ER+/ER− status in early-onset breast cancer cases successfully classifies cases into sub-types with ~96% accuracy (Table 1), with accuracy exceeding 93% with all kernels.

Although SVM classification accuracy is an important indicator of success it can be misleading, particularly in the case of an unbalanced data set (unequal numbers of cases in the two groups), as in this study [27]. Other indicators, such as the number of cases correctly classified into each group and the area under ROC curve (AUC) values should also be considered. A ML algorithm may produce a majority-class classifier when presented with an unbalanced dataset [27]. In this situation all cases are classified as members of the majority class, making the classifier appear more accurate than in reality. For example, using this dataset, an accuracy of 68.6% can be achieved by simply classifying all 542 samples into the majority group: ER−, giving a misleading impression that the classifier is correctly identifying a reasonable proportion of the samples. However, the true positive and true negative rates, 0.00 and 1.00 respectively, identify the model as invalid. The true positive (number of true ER+ cases classified as ER+; TPR ~89%) and true negative (number of true ER− cases classified as ER−; TNR ~99%) results from the models in this study (Table 1) indicate that a substantial proportion of the ER+ and ER− cases are correctly classified. Therefore it is reasonable to conclude that the SVM models produced are successful as ER+/ER− classifiers and any one of the five models is suitable as a classifier for unseen data. It is evident however that the ER− cases are classified more accurately than ER+ cases, which is likely to reflect the unbalanced data (372 ER− cases versus 170 ER+ cases). The greater difficulty in classifying ER+ cases arises from the more limited variation in the SNP profile given the smaller number of cases available to the classifier.

Table 3. Significant enrichment of genes in KEGG pathway identified by DAVID.

Pathway	**Genes**	**P value**
Axon guidance	EPHA4, FYN, NRP1, NTN4, PPP3CA	0.007
T cell receptor signalling pathway	FYN, IL5, PPP3CA, PTPRC	0.027
Fc epsilon RI signalling pathway	FYN, IL5, MAP2K4	0.081

doi:10.1371/journal.pone.0068606.t003

Classifier performance can be further evaluated using receiver operator characteristic (ROC) curves which are based on the true positive and true negative rates at several different thresholds. One of the major advantages of the ROC curve is that it is unaffected by unbalanced datasets [28]. The area under ROC curve (AUC) measure [29] takes values between 0.00 and 1.00 with values closer to 1.00 indicating good performance. A random classification would produce an AUC of 0.5, the AUC values for the ER+/ER− classifier (Table 1) are in the range 0.92–0.94, suggesting excellent classification ability.

Feature selection is an important component of building a ML classifier. Much of the SNP data in these samples will not be useful for building an accurate model (Table S3) so it was necessary to select a subset of SNP features from which to build a classifier. For a review on feature selection methods available for ML algorithms see [30]. Feature selection prior to SVM implementation is essential to avoid the 'curse of dimensionality', which tends to arise

from training of too few examples with too many variables [15]. Therefore, it is suggested that the sample-to-feature ratio should ideally exceed 5:1, which is clearly not achievable with unselected genome-wide SNP data. Machine learning theory considers the concept of VC-dimension [31]. The VC-dimension quantifies a learning machine's *capacity* describing how complex a model can be: learning machine functions with high capacity may generate lower training error rates but require larger training sets than simpler, low capacity models. The best theoretical performance guarantee is achieved through the right balance between the accuracy attained for a given training set and the model capacity. Because analysis of genomic disease data considers potentially very large number of features (SNPs) evaluated on relatively small numbers of samples (genomes) feature selection strategies aim to reduce overfitting. Alternatives to the approach to reduce feature complexity adopted here include Recursive Feature Elimination (RFE) applied to linear SVMs using the ranked SVM weights to recursively eliminate features [32]. Such an approach has been used extensively for DNA micro-array gene expression data but has received less attention thus far for GWAS disease data.

The underlying biological nature of the genes identified as discriminators of ER+/ER− breast cancer was of particular interest in this study. To identify gene enrichment in gene groups and pathways we used the DAVID toolset. Analysis of the 139 genes that the classifier SNPs reside in, or are closest to, identified gene groups, pathways and annotation terms that were particularly enriched (Tables 2 and 3). Of the two annotation clusters with the highest enrichment scores (Table 2) it is notable that cluster 1 contains genes relating to immune/inflammatory cell activation, differentiation and proliferation. This suggests one of the distinctions between ER+ and ER− tumours relates to genetic variation in immune system pathways. The role of the immune/inflammatory response in influencing tumourigenesis and tumour progression, through the formation of an inflammatory microenvironment at the tumour site, is well characterised [33–36]. It has been suggested that as much as 50% of breast tumour volume comprises cells of the immune system, in particular, tumour-associated macrophages (TAMs) and tumour-infiltrating lymphocytes (TILs) [37] that establish the tumour microenvironment. Infiltrating immune cells are likely to be a major source of pro-tumourigenic factors at the tumour site because they have the capacity to release cytokines, chemokines, metalloproteases, reactive oxygen species and a number of bioactive mediators into the stroma. Furthermore, infiltrating immune cells regulate a number of processes, including enhanced cell survival, angiogenesis and suppression of anti-tumour immune responses [38] suggesting a role in both tumour development and progression. In particular, TAMs have been implicated as a source of mitogenic signals for tumour cells through cytokine secretion [39] potentially enhancing cell division and tumour growth.

The role of estrogen and estrogen receptors as regulators of proliferation and differentiation in breast tissue is well-established and is crucially important for disease progression in many cases [34,40]. It has been suggested that infiltrating leukocytes are a major source of estrogen expression in breast tumours [41] which could contribute to disease development and progression.

The estrogen receptor status of breast cancer patients has long been recognised as a strong prognostic factor that influences patient treatment options and survival. Patients with ER− forms of the disease tend to show decreased survival rates in the first few years after diagnosis and present with more aggressive tumours [42–45]. However, after 10 years of disease-free survival a relapse is more likely to occur in a patient who originally presented with ER+ disease [45]. A number of other factors influence breast cancer patient survival, one of which is the infiltrating immune system cells. There is a suggested strong correlation between the infiltration of lymphocytic cells and patient survival, particularly in patients with disease onset before the age of 40 years [33]. The number of $CD8^+$ T lymphocytes present at the tumour site influences patient survival, with higher numbers being associated with better survival rates. This effect is more evident in patients presenting with ER− tumours compared to ER+ tumours [46]. In contrast, TAM levels in breast tumours appear to positively correlate with aggressiveness of disease and poor prognosis [47,48].

DAVID analysis of the gene set also identified five genes implicated in the 'axon guidance' pathway (Table 3). Axon guidance molecules are important in the mammary gland for maintaining normal cell proliferation and adhesion during tissue development [49] and the proximity of nerves and blood vessels in a number of tissues suggests that there may be molecular cross-talk and common cues between these structures [50]. Dysregulation of these guidance molecules in the mammary gland has been linked to breast cancer initiation and progression [49].

Genome wide association studies have identified risk-related SNPs for many diseases. Thirty-five SNPs, which lie in or near to 36 genes, are identified as breast cancer risk SNPs in the Catalog of Published Genome-Wide Association Studies [51]. From the SNPs used in the ER+/ER− classifier none of the 35 risk SNPs is present in this list nor are any of the classifier SNPs in or near the 36 catalogued genes. Thus, the SNPs identified in this study represent a set of genes not previously linked to breast cancer risk although some of the genes have been linked to roles in prognosis. Our analysis finds that variation in, or near, at least 139 genes defines the genetic background on which different estrogen receptor tumour phenotypes are most likely to arise in early onset breast cancer patients. The polygenic nature of complex phenotypes has become an emerging theme from the numerous genome-wide association studies which have identified a large number of causal variants with minor impacts on risk. A polygenic model seems appropriate to define the distinction between breast cancer sub-types such as ER+/ER−, which are likely to represent distinct forms of disease. The evidence that this distinction relates in part to genetic variation in highly complex immune system pathways reinforces the emerging concept that the presenting cancer phenotype is shaped not only by a random series of acquired somatic gene mutations but also by the stable genetic background of the individual in whom the cancer arises. Understanding interactions between the host genome and the process of oncogenesis will be an important contribution to the development of more individualised treatment and prevention approaches in the future.

Materials and Methods

Breast cancer samples

542 early-onset breast cancer patients were selected from the '**P**rospective study of **O**utcome in **S**poradic versus **H**ereditary breast cancer' (POSH) cohort [52] of ~3000 patients with disease onset before the age of 40 years. Germline DNA samples were genotyped for 490,732 SNPs spanning chromosomes 1 to 22. Tumours from all cases were classified for estrogen receptor status with 170 identified as ER+ and 372 as ER−. The POSH study received approval from the South and West Multi-centre Research Ethics Committee (MREC 00/6/69). Written consent was given by the patients for their information to be stored in the hospital database and used for research.

SNP genotyping

Genotyping of the breast cancer samples was conducted using the Illumina 660-Quad SNP array. Genotyping was conducted at the Mayo Clinic, Rochester, Minnesota, USA (261 samples), and the Genome Institute of Singapore, National University of Singapore (281 samples) [53]. To ensure complete harmonisation of genotype calling, the intensity data available from both locations, in form of .idat files, were combined and used to generate genotypes using the algorithm in the genotyping module of Illumina's Genome Studio software. A GenCall threshold of 0.15 was selected and the HumanHap660 annotation file was used. SNPs were excluded from further analysis if they had a sample minor allele frequency (MAF) below 0.01, a genotyping call rate <95% or showed significant deviation from Hardy-Weinberg equilibrium (HWE, P-value <0.0001). We used the pairwise Identity-By-State (IBS) and multidimensional scaling, implemented in PLINK v1.07 [18,19], to confirm that patients were ethnically homogeneous. A proportion of the SNPs had missing genotypes and we used the MACH 1.0 program [54–56] to impute missing genotypes, where possible, based on genotype and haplotype phase data specific for CEU population available from HapMap phase 2 project. Genotype imputation was used to establish a set of SNPs with complete genotypes for testing as features in the models. However, imputation failed to resolve all genotypes for 27 SNPs with high chi-squares and these were removed from further consideration in the SVM models and replaced with the next most associated and fully genotyped SNPs in the ranked list.

SNP feature selection

SNPs showing significant association with ER− cases were identified from the additive chi-squared association test implemented in the PLINK toolset in which ER+ samples were labelled as 'controls' and ER− samples were labelled as 'cases'. Based on results from the chi-squared test all SNPs were ranked in terms of association with the ER+/− classification. Subsets of SNPs were selected as features for SVM models from the ranked list of SNPs and models were produced from subsets of 50, 100 and 200 SNPs to test utility as discriminatory factors for ER+/ER− breast cancer.

SVM model input

The three genotypes at each SNP were converted into numeric values following [24] and [25]. Major and minor allele frequencies for each SNP were determined from all genotypes in the sample. Heterozygous genotypes were labelled 0, homozygotes for the major allele were labelled 1, and homozygotes for the minor allele labelled −1. The two classes of samples in the models were ER+ cases and ER− cases.

Building a support vector machine classifier

Support vector machines are supervised machine learning algorithms which build models based on 'training' data and search for similar patterns in 'test' data [16]. The training set is often a subset of all samples complete with all class and feature values and the resultant model is then applied to the remaining test data. Novel data can be presented to the model and classified according to the position of the data point relative to the hyperplane constructed from the training set. The robustness and reliability of the SVM classifier can be tested using cross-validation, where the data is split into n equally sized sets testing n models. We used 10-fold cross-validation: data were divided into 10 approximately equal-sized sets and a classifier built based on the data in 9/10 of these sets. The remaining 10% of data was used as a test set to determine the accuracy of the classifier. This process was repeated 10 times with each set representing the test data once and average classification accuracy determined. We further explored 10-fold cross-validation using 100 replicates and mean accuracy from 1000 resultant models was obtained for alternative kernel models.

The SVM classification model was produced using the Weka data mining software [57,58]. The Sequential Minimal Optimization (SMO) algorithm for training a SVM classifier was applied to the data. Five kernel models were evaluated; linear, normalized quadratic polynomial, quadratic polynomial, cubic polynomial, and radial basis function (RBF).

Gene annotation

Annotation of sets of SNPs used in the classification models was undertaken using the ANNOVAR software [59,60]. Gene-based annotation was carried out using the UCSC 'Known Gene' database. SNPs were annotated as intergenic, exonic, intronic, downstream, ncRNA intronic, ncRNA exonic, upstream, UTR3, or UTR5. For SNPs situated outside genes the closest gene was identified and gene names were taken from the HUGO Gene Nomenclature Committee database [61,62]. A total of 139 unique gene names were linked to the set of 200 SNPs used in the final classifier (Table S1).

Functional gene classification

Functional gene annotation clusters were identified using the 'Gene Functional Classification' tool in DAVID (Database for Annotation, Visualization and Integrated Discovery) [20–22]. DAVID determines significant enrichment of function within a submitted gene name list by contrasting with a 'whole genome' background. Annotation clusters were identified from the 139 genes using the 'Functional Annotation Clustering' tool and five annotation categories: disease, functional categories, gene ontology, pathways and protein domains. Enriched pathways were identified using only the 'Pathways' annotation category with BBID, BIOCARTA, and KEGG selected.

Supporting Information

Table S1 200 SNPs which most strongly discriminate ER+ and ER− breast cancers used in the classification models. The weights are taken from a linear model built using one iteration of 10-fold cross-validation in the WEKA Explorer. Classification accuracy for this model was 92.4%. The magnitude of the absolute values of the SNP weights indicates importance of the SNP for classifying cases. Positive SNP weights relate to classifying ER+ cases while negative SNP weights relate to classifying ER− cases. For those SNPs that are not located within a gene the nearest gene is given and the distance of the SNP from this gene is indicated by dist = .
(DOCX)

Table S2 Weka kernels and classification results for 100 and 50 SNPs with highest chi-squares. Comparison of classifiers built with 100 and 50 highest ranked SNPs from PLINK chi-square test.
(DOCX)

Table S3 Weka kernels and classification results for bottom 200 and random 200 SNPs. Comparison of classifiers built with 200 random SNPs and the lowest ranked by PLINK chi-square test.
(DOCX)

Author Contributions

Conceived and designed the experiments: RU-G DE SE SR WT JF AC. Performed the experiments: RU-G SR WT AC. Analyzed the data: RU-G SR SE WT JF AC. Contributed reagents/materials/analysis tools: DE JF AC. Wrote the paper: RU-G DE SR SE WT JF AC.

References

1. Deroo BJ, Korach KS (2006) Estrogen receptors and human disease. J Clin Invest 116: 561–570.
2. Osborne CK, Shou J, Massarweh S, Schiff R (2005) Crosstalk between Estrogen Receptor and Growth Factor Receptor Pathways as a Cause for Endocrine Therapy Resistance in Breast Cancer. Clinical Cancer Research 11: 865s–870s.
3. Björnström L, Sjöberg M (2005) Mechanisms of Estrogen Receptor Signaling: Convergence of Genomic and Nongenomic Actions on Target Genes. Molecular Endocrinology 19: 833–842.
4. Clemons M, Goss P (2001) Estrogen and the Risk of Breast Cancer. New England Journal of Medicine 344: 276–285.
5. Gupta PB, Proia D, Cingoz O, Weremowicz J, Naber SP, et al. (2007) Systemic Stromal Effects of Estrogen Promote the Growth of Estrogen Receptor–Negative Cancers. Cancer Res 67: 2062–2071.
6. Péqueux C, Raymond-Letron I, Blacher S, Boudou F, Adlanmerini M, et al. (2012) Stromal Estrogen Receptor-α Promotes Tumor Growth by Normalizing an Increased Angiogenesis. Cancer Res 72: 3010–3019.
7. Paruthiyil S, Parmar H, Kerekatte V, Cunha GR, Firestone GL, et al. (2004) Estrogen Receptor β Inhibits Human Breast Cancer Cell Proliferation and Tumor Formation by Causing a G2 Cell Cycle Arrest. Cancer Res 64: 423–428.
8. Ström A, Hartman J, Foster JS, Kietz S, Wimalasena J, et al. (2004) Estrogen receptor β inhibits 17β-estradiol-stimulated proliferation of the breast cancer cell line T47D. Proc Natl Acad Sci U S A 101: 1566–1571.
9. Andersen TI, Heimdal KR, Skrede M, Tveit K, Berg K, et al. (1994) Oestrogen receptor (ESR) polymorphisms and breast cancer susceptibility. Human Genetics 94: 665–670.
10. Roodi N, Bailey LR, Kao W-Y, Verrier CS, Yee CJ, et al. (1995) Estrogen Receptor Gene Analysis in Estrogen Receptor-Positive and Receptor-Negative Primary Breast Cancer. J Natl Cancer Inst 87: 446–451.
11. Iwase H, Greenman JM, Barnes DM, Hodgson S, Bobrow L, et al. (1996) Sequence variants of the estrogen receptor (ER) gene found in breast cancer patients with ER negative and progesterone receptor positive tumors. Cancer Lett 108: 179–184.
12. Maguire P, Margolin S, Skoglund J, Sun X-F, Gustafsson J-Å, et al. (2005) Estrogen receptor beta (ESR2) polymorphisms in familial and sporadic breast cancer. Breast Cancer Res Treat 94: 145–152.
13. Yu K-D, Rao N-Y, Chen A-X, Fan L, Yang C, et al. (2011) A systematic review of the relationship between polymorphic sites in the estrogen receptor-beta (ESR2) gene and breast cancer risk. Breast Cancer Res Treat 126: 37–45.
14. Easton DF, Eeles RA (2008) Genome-wide association studies in cancer. Hum Mol Genet 17: R109–R115.
15. Cruz JA, Wishart DS (2006) Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform 2: 59–77.
16. Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Learning 20: 273–297.
17. Burges CC (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2: 121–167.
18. Purcell S PLINK v1.07. Available: http://pngu.mgh.harvard.edu/purcell/plink/. Accessed 2012 Apr.
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81: 559–575.
20. DAVID 6.7. Avaliable: http://david.abcc.ncifcrf.gov/home.jsp. Accessed 2013 Jan.
21. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4: 44–57.
22. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37: 1–13.
23. Chen S-H, Sun J, Dimitrov L, Turner AR, Adams TS, et al. (2008) A Support Vector Machine Approach for Detecting Gene-Gene Interaction. Genetic Epidemiology 32: 152–167.
24. Waddell M, Page D, Shaughnessy Jr J (2005) Predicting Cancer Susceptibility from Single Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma. ACM. 21–28.
25. Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, et al. (2004) Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clinical Cancer Research 10: 2725–2737.
26. Ban HJ, Heo JY, Oh KS, Park KJ (2010) Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. BMC Genet 11: 26.
27. Ben-Hur A, Weston J (2010) A User's Guide to Support Vector Machines. In: Carugo O, Eisenhaber F, editors. Data Mining Techniques for the Life Sciences: Humana Press. 223–239.
28. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861–874.
29. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30: 1145–1159.
30. Guyon I, Elisseeff A (2003) An Introduction to Variable and Feature Selection. J Mach Learn Res 3: 1157–1182.
31. Vapnik V, Levin E, Cun YL (1994) Measuring the VC-dimension of a learning machine. Neural Comput 6: 851–876.
32. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 46: 389–422.
33. Ménard S, Tomasic G, Casalini P, Balsari A, Pilotti S, et al. (1997) Lymphoid Infiltration as a Prognostic Variable for Early-Onset Breast Carcinomas. Clinical Cancer Research 3: 817–819.
34. Chen R, Alvero AB, Silasi D-A, Mor G (2007) Inflammation, Cancer and Chemoresistance: Taking Advantage of the Toll-Like Receptor Signaling Pathway. American Journal of Reproductive Immunology 57: 93–107.
35. Mantovani A, Allavena P, Sica A, Balkwill F (2008) Cancer-related inflammation. Nature 454: 436–444.
36. Grivennikov SI, Greten FR, Karin M (2010) Immunity, Inflammation, and Cancer. Cell 140: 883–899.
37. Reed MJ, Purohit A (1997) Breast Cancer and the Role of Cytokines in Regulating Estrogen Synthesis: An Emerging Hypothesis. Endocrine Reviews 18: 701–715.
38. DeNardo D, Coussens L (2007) Inflammation and breast cancer. Balancing immune response: crosstalk between adaptive and innate immune cells during breast cancer progression. Breast Cancer Research 9: 212.
39. Ch'ng E, Tuan Sharif S, Jaafar H (2013) In human invasive breast ductal carcinoma, tumor stromal macrophages and tumor nest macrophages have distinct relationships with clinicopathological parameters and tumor angiogenesis. Virchows Archiv: 1–11.
40. Lee WJ, Monteith GR, Roberts-Thomson SJ (2006) Calcium transport and signaling in the mammary gland: Targets for breast cancer. Biochimica et Biophysica Acta (BBA) – Reviews on Cancer 1765: 235–255.
41. Mor G, Yue W, Santen RJ, Gutierrez L, Eliza M, et al. (1998) Macrophages, Estrogen and the Microenvironment of Breast Cancer. J Steroid Biochem Mol Biol 67: 403–411.
42. Anderson WF, Chu KC, Chatterjee N, Brawley O, Brinton LA (2001) Tumor Variants by Hormone Receptor Expression in White Patients With Node-Negative Breast Cancer From the Surveillance, Epidemiology, and End Results Database. Journal of Clinical Oncology 19: 18–27.
43. Dunnwald L, Rossing M, Li C (2007) Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Research 9: R6.
44. Hähnel R, Spilsbury K (2004) Oestrogen receptors revisited: long-term follow up of over five thousand breast cancer patients. ANZ Journal of Surgery 74: 957–960.
45. Foulkes WD, Smith IE, Reis-Filho JS (2010) Triple-Negative Breast Cancer. New England Journal of Medicine 363: 1938–1948.
46. Mahmoud SMA, Paish EC, Powe DG, Macmillan RD, Grainge MJ, et al. (2011) Tumor-Infiltrating CD8+ Lymphocytes Predict Clinical Outcome in Breast Cancer. Journal of Clinical Oncology 29: 1949–1955.
47. Leek RD, Lewis CE, Whitehouse R, Greenall M, Clarke J, et al. (1996) Association of Macrophage Infiltration with Angiogenesis and Prognosis in Invasive Breast Carcinoma. Cancer Res 56: 4625–4629.
48. Solinas G, Germano G, Mantovani A, Allavena P (2009) Tumor-associated macrophages (TAM) as major players of the cancer-related inflammation. Journal of Leukocyte Biology 86: 1065–1073.
49. Harburg G, Hinck L (2011) Navigating Breast Cancer: Axon Guidance Molecules as Breast Cancer Tumor Suppressors and Oncogenes. Journal of Mammary Gland Biology and Neoplasia 16: 257–270.
50. Klagsbrun M, Eichmann A (2005) A role for axon guidance receptors and ligands in blood vessel development and tumor angiogenesis. Cytokine Growth Factor Rev 16: 535–548.
51. Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, et al. A Catalog of Published Genome-Wide Association Studies. Available: www.genome.gov/gwastudies. Accessed 2013 Feb.
52. Eccles D, Gerty S, Simmonds P, Hammond V, Ennis S, et al. (2007) Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH): study protocol. BMC Cancer 7: 160.
53. Rafiq S, Tapper W, Collins A, Khan S, Politopoulos I, et al. (2013) Identification of inherited genetic variations influencing prognosis in early onset breast cancer. Cancer Res.
54. MACH 1.0. Available: http://www.sph.umich.edu/csg/abecasis/MACH/index.html. Accessed 2013 Jan.
55. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype Imputation. Annual review of genomics and human genetics 10: 387–406.

56. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34: 816–834.
57. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explorations 11: 10–18.
58. Weka 3.6.8. Available: http://www.cs.waikato.ac.nz/ml/weka/. Accessed 2012 Aug.
59. ANNOVAR Available: http://www.openbioinformatics.org/annovar/.Ac-Accessed 2012 Oct.
60. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38: e164.
61. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA (2011) genenames.org: the HGNC resources in 2011. Nucleic Acids Research 39: D514–D519.
62. HGNC Database, HUGO Gene Nomenclature Committee (HGNC), EMBL Outstation – Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. Available: www.genenames.org. Accessed 2012 Nov.