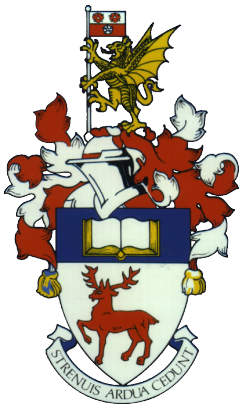


UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine
Human Development & Health

Genomic Data Analysis: populations, patients & pipelines

by

Reuben J. Pengelly

A thesis submitted for the degree of Doctor of Philosophy

Supervisory Team: Prof. Sarah Ennis, Dr. Jane Gibson & Prof. Andrew Collins

October 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Medicine
Human Development & Health

Doctor of Philosophy

Genomic Data Analysis: populations, patients & pipelines
by
Reuben John Pengelly MBiol

Methods for the ascertainment of genotype data have become more cost efficient by orders of magnitude with the use of high-density genotyping arrays and the advent of next generation sequencing (NGS). The resulting deluge of data has required ever advancing analytical approaches in order for the maximal information to be gleaned from these extensive data.

In this work, many application of NGS to clinical research are discussed. This includes the application of targeted gene sequencing to a cohort of 83 patients with chronic kidney disease, whole-exome investigations of eight families with cleft lip/palate phenotypes, as well as five cases where analytical lessons can be learned from exome sequenced cases harbouring pathogenic variants refractory to identification. Additionally, a novel QC tool for the unambiguous tracking of samples undergoing exome sequencing is presented.

Furthermore, work is presented investigating the linkage disequilibrium (LD) patterns in populations applying the Malécot-Morton model. We demonstrate that array genotyping is insufficient for the accurate determination of fine LD patterns in the human genome, with whole-genome sequencing providing more representative LD maps. Finally, we apply similar methods to *Gallus gallus*, generating the highest resolution maps of LD presented to date, showing that the patterns are highly discordant between commercial lines, and define features associated with recombination.

Overall, we highlight the diversity of ways in which genetic data can be utilised effectively in the age of genomic 'big data', and present tools which may be of benefit to other researchers utilising these technologies.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Reuben John Pengelly, 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.

Genomic Data Analysis: populations, patients & pipelines

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 detailed overleaf.

Signed:

Date: 19th December 2015

List of Publications

1. Pengelly, R. J., Gibson, J., Andreoletti, G., Collins, A., Mattocks, C. J. & Ennis, S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med* **5**, 89 (2013). DOI: 10.1186/gm492.

2. Collins, A., Arias, L., Pengelly, R., Martínes, 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* **1**, 10 (2013). DOI: 10.13172/2054-197X--1-987.

3. Coelho, T. A. F., Andreoletti, G., Ashton, J. J. *et al.* Immuno-genomic profiling of patients with inflammatory bowel disease: a systematic review of genetic and functional in vivo studies of implicated genes. *Inflamm Bowel Dis* **20**, 1813–19 (2014). DOI: 10.1097/MIB.0000000000000174.

4. Kadalayil, L., Rafiq, S., Rose-Zerelli, M. J. J. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* (2014). DOI: 10.1093/bib/bbu027.

5. Pengelly, R. J., Upstill-Goddard, R., Arias, L. *et al.* Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing. *Clin Genet* (2015). DOI: 10.1111/cge.12547.

6. Foulds, N., Pengelly, R. J., Hammans, S., Nicoll, J. A. R., Ellison, D. W., Ditchfield, A., Beck, S. & Ennis, S. Adult-Onset Leukoencephalopathy with Axonal Spheroids and Pigmented Glia Caused by a Novel R782G Mutation in CSF1R. *Sci Rep* **5**, 10042 (2015). DOI: 10.1038/srep10042.

7. Gast, C., Pengelly, R. J., Lyon, M., Bunyan, D. J., Seaby, E. G., Graham, N., Venkat-Raman, G. & Ennis, S. Collagen (*COL4A*) Mutations Are the Most Frequent Mutations Underlying Adult Focal Segmental Glomerulosclerosis. *Nephrol Dial Transplant* (2015). DOI: 10.1093/ndt/gfv325.

8. Pengelly, R. J., Tapper, W., Gibson, J., Knut, M., Tearle, R., Collins, A. & Ennis, S. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC Genomics* **16**, 666 (2015). DOI: 10.1186/s12864-015-1854-0.

9. Seaby, E. G., Pengelly, R. J. & Ennis, S. Exome sequencing explained: a practical guide to its clinical application. *Brief Funct Genomics* (2015). DOI: 10.1093/bfgp/elv054.

10. Seaby, E. G., Gilbert, R. D., Pengelly, R. J., Andreoletti, G., Clarke, A. & Ennis, S. Exome sequencing reveals the genetic cause of myoclonic epilepsy associated with Fanconi syndrome. *JRSM Open.* [Accepted] (2015).

Acknowledgements

Firstly, I would like to thank Profs. Sarah Ennis and Andrew Collins, and Drs. Jane Gibson & Will Tapper, for their patient guidance over the past three years. Of course, many thanks to all patients who have taken part in the studies detailed herein. Much of this work would not have been possible without open access to the data of several consortia, particularly that of the 1000 Genomes and UK10K Projects.

I appreciate the opportunity to work along side all those that I am in collaboration with. Without their clinical and laboratory investigations, and patients, much of this work would not have been possible. I would particularly like to thank Dr. Christine Gast for her collaboration performing the clinical aspects of the work presented in Chapter 9 and Dr. Ananth Ramakrishnan and Eleanor Seaby for collaboration on their respective cases presented in Chapter 7. Finally, thanks to Prof. Ignacio Briceño and his team for providing the invaluable clinical phenotyping for Chapter 8. Similarly, many thanks to Dr. Rick Tearle and Prof. Dave Burt for providing data and expertise for Chapters 10 & 11 respectively.

Though I am rarely in the lab, thanks to Nikki Graham for facilitation and assistance on the rare occasions, as well as for maintenance of the archival DNA storage. For when I am not in the lab, I am fortunate to have access to the excellent IRIDIS High Performance Computing Facility, with associated support, particularly from Dr. Elena Vataga, without which I would not have generated half as many results.

Finally, many thanks to all the members of the group, particularly Marcin Knut, Gaia Andreoletti, Enrico Mossotto and Eleanor Seaby, for making it an enjoyable three years, and putting up with basic questions on Perl at the start.

This studentship is part supported by the Faculty of Medicine Doctoral Training Fund, University of Southampton, with additional funding from the BBSRC for work regarding LD maps. Many thanks to all funders of this work.

List of Abbreviations

ABG	Array-based genotyping
ACMG	American College of Genetics and Genomics
AD	Alport disease
AF	Alternate-allele frequency
aHUS	Atypical haemolytic uremic syndrome
APDS	Activated PI3K-δ syndrome
BAF	B-allele frequency
BAM	Binary alignment/map
BEL	Brown egg layer
BRO	Broiler
C1-THF synthase	C-1-tetrahydrofolate synthase, cytoplasmic
CBS	Circular binary segmentation
CEU	CEPH (Utah residents with ancestry from northern and western Europe)
CHB	Han Chinese in Beijing
CHS	Southern Han Chinese
CLP	Cleft lip/palate
cM	Centimorgan
CNV	Copy number variation
DNA	Deoxyribonucleic acid
DOC	Depth of coverage
DTC	Direct to consumer
ESP	NHLBI Exome Sequencing Project
FDA	US Food and Drug Administration
FFPE	Formalin-fixed, paraffin embedded
FHx	Family history
FSGS	Focal segmental glomerulosclerosis
gDNA	Genomic DNA
GRC	Genome Reference Consortium
GWAS	Genome wide association study
HGMD	Human Gene Mutation Database

HGP	Human Genome Project
HLA	Human leukocyte antigen
HR	Homologous recombination
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state
IF	Incidental finding
IFN	Interferon
IL	Interleukin
IP	Incontinentia pigmenti
JMML	Juvenile myelomonocytic leukemia
JPT	Japanese in Tokyo, Japan
LD	Linkage disequilibrium
LDU	Linkage disequilibrium unit
lod	Logarithm of odds
LOH	Loss of heterozygosity
MAF	Minor-allele frequency
mBAF	Mirrored BAF
MDS	Multidimensional scaling
mtDNA	Mitochondrial DNA
NEMO	NF-$\varkappa$B essential modulator
NGS	Next-generation sequencing
NHLBI	National Heart, Lung, and Blood Institute
NSCLP	Non-syndromic cleft lip/palate
OMIM	Online Mendelian Inheritance in Man
PE	Paired end
PID	Primary immunodeficiency
PIP_2	phosphatidylinositol-4,5-biphosphate
PIP_3	phosphatidylinositol-3,4,5-triphosphate
QC	Quality control
RNA	Ribonucleic acid
SAM	Sequence alignment/map
SBS	Sequencing by synthesis
SCID	Severe combined immunodeficiency
SCLP	Syndromic cleft lip/palate
SNP	Single nucleotide polymorphism
SRNS	Steroid resistant nephrotic system
TSCA	TruSeq Custom Amplicon
TSS	Transcription start site

UKGTN	UK Genetic Testing Network
UPD	Uniparental disomy
VCF	Variant call file
WEL	White egg layer
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
YRI	Yoruba in Ibadan, Nigeria

List of Nomenclature

α	Significance level
α_{set}	Corrected significance level (per test)
C	Likelihood of collision within dataset
E	Error probability
ϵ	Rate of decline with distance of association between two markers
f	Number of shared founders
L	Component of $\hat{\rho}$ not due to linkage disequilibrium
L_{linked}	Observed likelihood of marker co-inheritance
$L_{unlinked}$	Likelihood of marker co-inheritance presuming independence
m	Number of matings
M	Anticipated linkage at 0 distance
n	Number of samples
n_{set}	Number of tests performed
O	Number of possible profiles
Φ	Identity by descent
q	Likelihood of collision between two samples
r	Probability of profile assignation
r^2	Coefficient of determination
ρ	Spearman's rank correlation coefficient
$\hat{\rho}$	Observed correlation between two markers

List of IUPAC Nomenclature

Nucleotide

A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
-	gap

Amino Acid

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

For John & Rowan, for getting me here,
and Alice, for keeping me going.

Part I

Introduction

Chapter 1

Foundations of Genetics

> "The science of genetics is in a transition period, becoming an exact science just as the chemistry in the times of [Antoine] Lavoisier, who made the balance an indispensable implement in chemical research."
>
> *Wilhelm Johannsen, 1911*[1]

1.1 A primer on molecular biology

Genetics, the study of the transfer of traits in discrete heritable units, largely stems from the works of Gregor Mendel in the mid 19th century on inheritance in *Pisum sativum* (the common pea). Mendel observed that traits passed down through generations of the pea in predictable patterns, abiding by ratios that stem from the biallelic inheritance of the traits[2]. It was not until the 1940s that the chemical basis of this inheritance was identified. Avery *et al.* investigated the transformation of benign *Streptococcus pneumoniae* to a pathogenic form through incubation of benign cells with cellular lysate of the pathogenic form. Following isolation of the 'transforming principle', chemical analyses determined it to be deoxyribonucleic acid (DNA)[3].

Further to the identification of DNA as the vehicle for inheritance, an appreciation of the properties of the molecule has allowed further advancements in molecular biology. Discoveries such as the elucidation of the semi-conservative nature of the process by which DNA replicates[4], along with the solving of the characteristic double helix crystal structure[5], have laid the groundwork for the burgeoning field. DNA is formed from a dictionary of four nucleotide bases (adenine, thymine, cytosine and guanine; A, T, C and G respectively), coding under the so-called 'central dogma' of molecular biology for proteins with a complement of 20 directly translated amino-acids monomers *via* trinucleotide codons (Figure 1.1)[6].

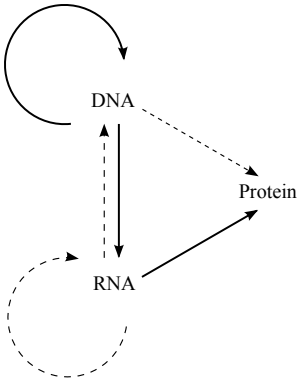


Figure 1.1: Information transfer paths available under the 'central dogma' of molecular biology. The major transfers of information in human molecular biology are shown in bold arrows: DNA, which will self-replicate, is transcribed into RNA, which may in turn be translated into polypeptides. Special cases of RNA self-replication and reverse transcription are also seen, though are not performed by integral human cellular mechanisms. Translation of DNA is a rare case, possible to perform experimentally. Adapted from Crick, 1970[7].

1.2 The Human genome

Humans have a diploid genome with a haploid size of ~3 Gbp, comprising 22 autosomal homologous chromosome pairs (1–22), and two allosomes (X & Y), totalling 46 chromosomes within somatic cells for a euploid individual (Figure 1.2). Being diploid, one of each chromosome is received from each haploid parental gamete upon fertilisation of the oocyte[8].

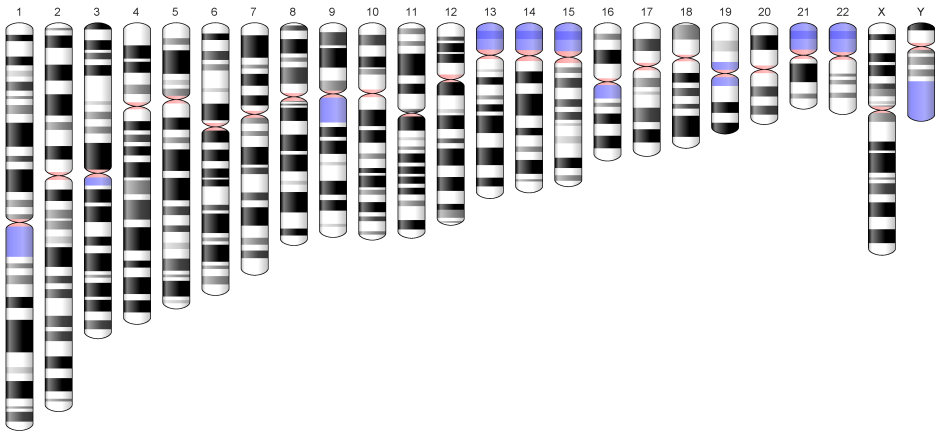


Figure 1.2: Ideogram showing representative human prometaphase chromosomes as observed following Giemsa staining. Pink regions indicate centromeric regions, while blue represent non-centromeric heterochromatin. Dark bands indicate AT-rich regions of chromosomes. Taken from `www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/`[9].

Within the genomic DNA (gDNA), there are 20,000–22,000 protein coding genes, in addition to functional ribonucleic acids (RNA), such as transfer, ribosomal and micro RNAs. From these ~20,000 protein coding genes, a large array of discrete mRNA transcripts can be generated by alternative splicing of the pre-mRNA, allowing for the

complexity of the human cellular processes[8]. Transcribed DNA is estimated to make up 1–2% of the human genome, with the majority of the remainder formed of repeat elements and other 'junk' DNA. Recent advances in our understanding of the function of many DNA elements however show that the vast majority of this 'junk' DNA is functional in some regard[10,11].

1.3 Mutation types

There is a wide gamut of mutation types that can occur. A range of small coding mutations that may occur in the exons of a gene are shown in Figure 1.3. These mutations will have diverse effect upon the translated protein, and thus also on any potential ultimate phenotypic effects. In addition to the small mutations shown, larger mutations also occur, including gross structural changes at a chromosomal level and nucleotide repeat expansions.

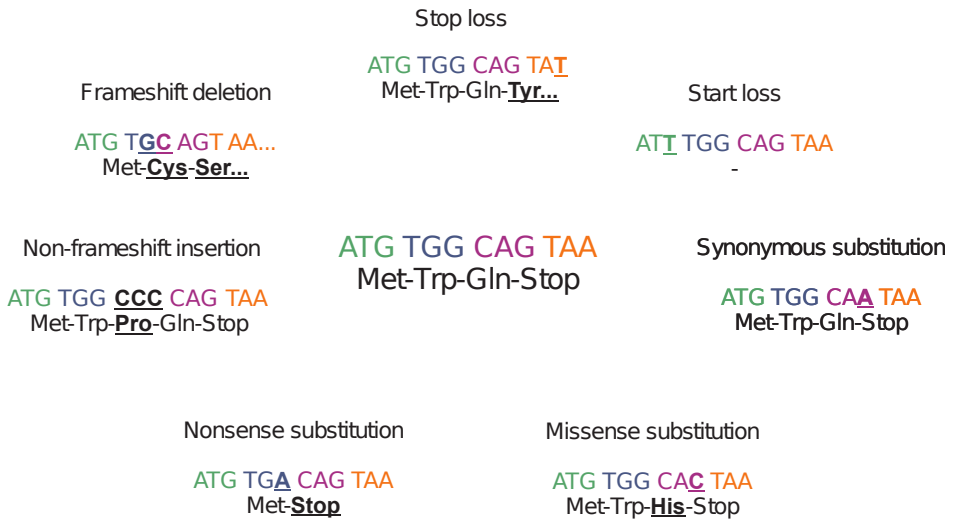


Figure 1.3: Summary of coding mutation types and effect on protein. The open reading frame for a hypothetical tripeptide is shown in the centre, with seven possible small mutation types shown surrounding this. Sequence changes are underlined. Note that frameshift and non frameshift variants may both be insertions or deletions, this has not been illustrated due to space constraints. Though synonymous substitutions are not expected to cause a protein change due to coding alterations, but may affect other factors, for instance altering a binding site motif. All other mutation types are expected to alter the primary sequence of the resultant protein. Frameshifts and stop loss mutations may result in the read through of a previous stop codon; therefore, translation may continue until an in-frame stop codon is reached. For start loss mutations, no translation is expected, unless there is a proximal alternative start codon, in which case this mutation would merely cause N-terminal truncation of the peptide.

1.4 Population structure

The genomes of humans are diverse, with several million deviations from the reference genome in any one individual. Many of these variants are highly common, with a high

alternate-allele frequency (AF), and some will be private to the individual. However, these variants are far from uniformly distributed across populations. A variant may be vanishingly rare in one population, or absent, and common in another[12]. If care is not taken, these population differences can hinder some studies.

1.5 Inheritance

As in peas, inherited traits in humans often follow predictable patterns in heredity. There are many possible modes of inheritance for a genetic trait (say, for simplicity a disease), even if we presume adherence to Mendelian monogenic inheritance of a trait. This predictability is due to the consistent passage of a proportion of DNA through generations.

With each separating meiosis between individuals, the proportion of alleles with identity by descent (IBD or Φ) is halved. The anticipated proportion of IBD between two relatives (denoted a and b for this example) can be calculated, presuming that all pedigree founders are unrelated for simplicity:

$$\Phi_{ab} = 0.5^m f \tag{1.1}$$

where m is the number of matings separating a and b *via* the nearest common founder, and f is the number of shared founders (example values are shown in Table 1.1; adapted from Lange, 1997[13])[14]. It should be noted that even for an entirely non-consanguineous pedigree, the identity by state (IBS) is expected to be greater than this calculated IBD due to the common alleles recurring within the pedigree derived from independent founders.

Table 1.1: Expected proportion of autosomal IBD for relatives within an outbred pedigree.

Relationship	f	m	Φ
Monozygotic twin	2	1	1
Parent	1	1	0.5
Sibling	2	2	0.5
Half-sibling	1	2	0.25
Grandparent	1	2	0.25
Aunt/Uncle	2	3	0.25
1st Cousin	2	4	0.125

Several modes of inheritance are discussed below, in the context of disease alleles for clarity, though it should be noted that this is not an exhaustive list.

1.5.1 Autosomal dominant

The inheritance of a single pathogenic allele will be sufficient for the manifestation of the disease. Molecular mechanisms underlying dominant conditions may include haploinsufficiency, where the half-dosage of the functional gene copy is insufficient for cellular processes, and dominant negative effects, for example as seen in proteins that form homodimers such as receptor tyrosine-kinases. Here, because the non-functional monomers still bind with the functional monomers, the homodimer is non-functional due to the required reciprocity of function between the monomers. An affected individual will have a 50% probability of passing the disease onto their child. Huntingdon's disease is a classical example of an autosomal dominant condition[8].

1.5.2 Autosomal recessive

Both inherited alleles of the disease locus are required to be pathogenic for the manifestation of the disease. Recessive conditions may be caused by the absence of a metabolic process: where half-dosage of function would have been sufficient, abrogation of function is pathogenic. For autosomal recessive conditions, there is a 100% chance that an affected individual will pass on a pathogenic allele, but where the partner is unaffected, the probability of them also passing on a disease allele will be dependant upon the carriage rate in the population and any family-history of the disease. It is of note that both pathogenic alleles in a gene are not required to be the *same* pathogenic allele; compound heterozygosity is often a more likely cause of autosomal recessive disease in non-consanguineous families. Cystic fibrosis is a canonical autosomal recessive disorder[8].

1.5.3 Sex-linked

Conditions can be either X-linked or Y-linked. In the case of an X-linked recessive condition, the same requirements for pathogenesis apply in females as with AR conditions. As males typically possess a single X-chromosome, there is not the allelic redundancy as with autosomes, so this hemizygosity for a pathogenic allele will be sufficient to cause disease. X-linked dominant conditions will manifest in both males and females; as in some autosomal dominant conditions, homozygosity for a pathogenic allele tends to be more severe, and may be lethal at some stage of development. As such, X-linked dominant conditions can tend to manifest more severely in males than heterozygous females, due to the obligate hemizygosity for the allele. Y-linked disease will manifest purely in males. Genes within the pseudoautosomal region, being homologous between the X and Y chromosomes, will exhibit an inheritance pattern more similar to autosomal loci. Allosome aneuploidies may interfere with the inference of mode of inheritance,

for instance, a male with Klinefelter's syndrome (karyotype 47,XXY) may carry an X-linked recessive allele without manifestation[8].

1.5.4 Alternative modes

In addition to the above Mendelian modes of inheritance, many diseases have alternative modes. For example, mitochondria contain a small genome (mtDNA) of ~16,500 bp, with a high coding density. As only the oocytic mitochondria are retained post-fertilisation, inheritance will only be apparent though the maternal lineage. Furthermore, due to the high copy-number of heterogeneous mtDNA in a cell, the resultant heteroplasmy may lead to variable penetrance in carriers of the variant[15]. In many cases a presumption of monogenic, completely penetrant inheritance of a trait is unfounded, and several genes may be involved in the disease processes, or require additional environmental triggers. The ultimate realisations of this concept, aptly named 'complex diseases', are those that require a complex interplay of factors for manifestation, with genetic variants merely predisposing an individual to the disease, and thus require different approaches for the identification of genes involved in Mendelian disease, as discussed in Chapter 2.

1.6 Genetics as aetiology

Aetiology in disease can be broadly considered to have several main classes, including deficiency, where disease is brought about by the lack of an essential nutrient (e.g. microcytic anaemia caused by iron deficiency), and pathogenic disease, caused by the uncontrolled presence of pathogenic micro-organisms, parasites or particles (e.g. hepatitis C virus). In addition to these exogenous aetiologies, endogenous factors can lead to disease. An inborn genetic defect in a metabolic or signalling pathway may manifest in a clinical phenotype, for example defects in the hedgehog signalling pathway may result in erroneous growth patterning during foetal development[16]. Furthermore, acquired somatic mutations may contribute to the development of malignant neoplasia.

In the vast majority of cases, these are not discrete factors and there will be some degree of interplay. For example, an individual may be born with cystic fibrosis, but the disease course is modified throughout life by events such as infection with respiratory pathogens (e.g. *Pseudomonas aeruginosa*). The focus of the research detailed herein is on congenital diseases which have a clear associated phenotype, regardless of the intervening factors, and are therefore expected to have a *strong* genetic cause. The study of these genetic diseases has enjoyed accelerating success as regards the identification of disease genes (Figure 1.4), driven largely by advances in associated technologies, as discussed in detail in Chapter 2.

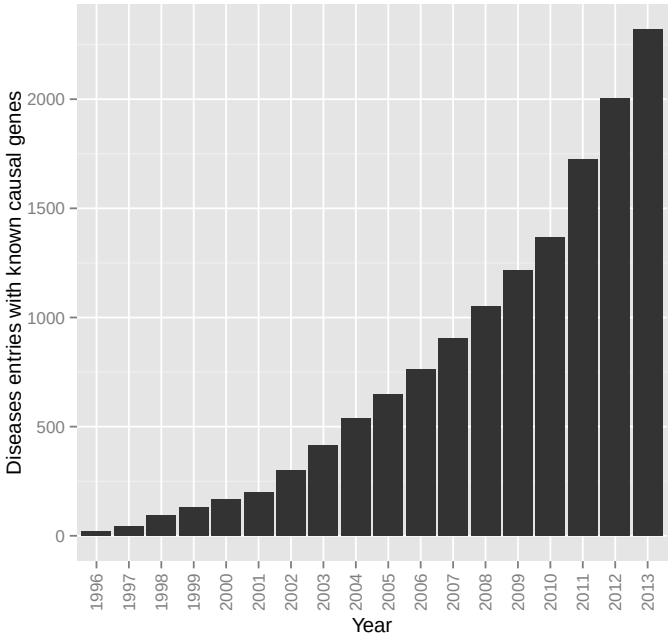


Figure 1.4: Accelerating rate of disease gene identification, 1996–2013. Cumulative count of the creation of additional disease entries in the Online Mendelian Inheritance in Man (OMIM) database[17] for which a causal gene is known, with 1995 as the baseline. Diseases for which only a locus is identified, as opposed to the specific gene, are not included in the count. Values based upon data-freeze downloaded 18th February, 2014.

Chapter 2

Medical Genetic Research as Driven by Emergent Technologies

The field of human medical genetics is a rapidly evolving, and accelerating field, with this continued progress being driven by the availability of new technologies for the determination and analysis of genetic data. Here I will discuss a few of the most significant methods that have been used for medical genetic research since the latter half of the 20th century, with a critical analysis of the methods.

2.1 Linkage mapping

One of the earliest approaches to the mapping of disease genes was linkage mapping. In linkage studies, related individuals exhibiting the disease are genotyped for a low density of markers. As technologies have progressed, so greater marker densities have been available to researchers, progressing from single point markers such as ABO blood-type, to several 1,000 independent microsatellite/single nucleotide polymorphism (SNP) markers. Statistical analyses are undertaken in order to determine which (if any) marker most closely cosegregates with disease (Figure 2.1). The seminal statistic for linkage analysis is the logarithm of odds (lod) score. Despite complex mathematics, the fundamental principle of the lod score can be expressed as:

$$\text{lod} = log_{10}\left(\frac{L_{observed}}{L_{unlinked}}\right) \tag{2.1}$$

where $L_{observed}$ is the likelihood, as empirically determined, of co-inheritance of the marker allele with the trait-defining locus, and $L_{unlinked}$ is the likelihood calculated presuming the marker and locus are independent (equal to 0.5 for the residual co-transmission in a fully-stochastic manner)[18]. Determination of the lod should be carried out in several independent families to allow for pooling of results and resultant increased certainty afforded due to the additive nature of lod scores.

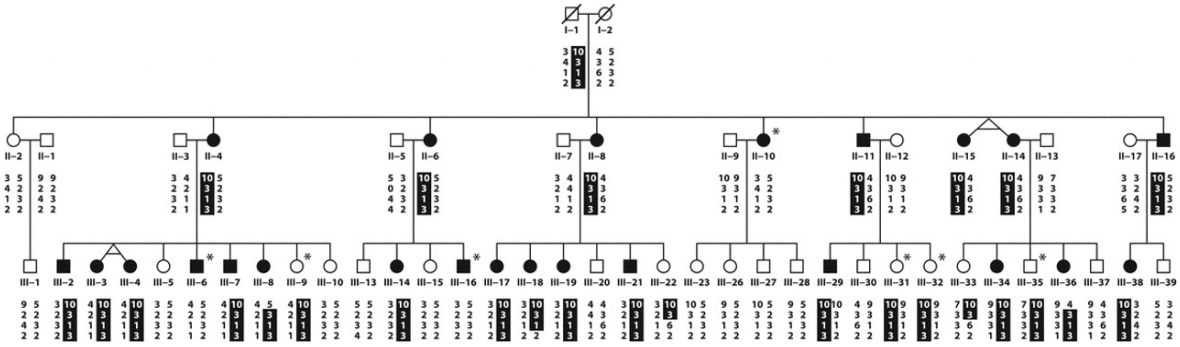


Figure 2.1: Principle of linkage mapping as applied to a pedigree exhibiting episodic ataxia. Haplotypes within 1q42, indicated below individuals, are shaded in black to denote the 10-3-1-3 putative risk haplotype; the risk halotype can be seen to segregate with disease in the majority of cases. 1q42 was previously found to be the most strongly linked region to the condition in a genome-wide linkage analysis prior to fine mapping, with a lod score of 3.65. Deviations are however seen from the expected pattern, marked with *, e.g. in individuals II-10 and III-9, this may be due to incomplete penetrance and phenocopy phenotypes. Taken from Cader *et al.*, 2005[19]. Reprinted by permission from Wolters Kluwer Health, © 2005.

Linkage studies, as with all methodologies, have many limitations; foremost of these is the sensitivity to errors in both genotyping and phenotyping which can greatly affect results given the small sample size within a family. Initial ascertainment of extended families with the disease of interest may also prove problematic, particularly where the disease has a strongly detrimental effect on fitness. Secondly, the parametric lod score is best suited for total-penetrance Mendelian traits, as deviation from this will reduce the power of locus detection, though can be compensated for[20]. Additionally, once a genomic region has been identified as linked, the fine mapping of the locus is non-trivial. Despite these issues, linkage mapping has had many successes in the identification of disease loci such as that for Huntingdon's disease[21]. Alternate non-parametric statistical methods for linkage mapping have also been used with some successes for complex diseases[22].

2.2 The Human Genome Project & reference genome

True 'genomics' could arguably be thought to have initiated with the advent of the publicly-funded Human Genome Project (HGP)[23–25]. The HGP stands out as one of the largest, most ambitious, non-military scientific endeavours so far completed, being particularly impressive for the sheer scale of international collaboration involved. The HGP had a broad array of goals in addition to the generation of a human genome reference sequence. These included educational initiatives to ensure maximum advantage could be taken of the completed genome, continuation of the development of technologies for genomic analysis and to produce similar reference resources for model organisms.

As with any newly emerging field, the early years of genomics have been marked by a lag between the rapidly advancing science and the ethico-legal framework within which it is be expected to function. This was by no means an unexpected issue, being another of the target areas of research within the HGP[24].

A draft reference sequence of the human genome was published in 2001[23], followed by the final release from the HGP in 2004[25]. Taking up the work of the HGP, the Genome Reference Consortium (GRC) now maintains the reference genomes of several species, releasing regular intermediate patches as required and major releases for the human reference currently approximately every 3 years[9]. An accurate reference genome is essential to facilitate modern genomic research, as discussed in subsection 4.3. There have been significant economic benefits resulting from the HGP; a report on the economic impact concluded in part that:

> "The federal government invested $3.8 billion [USD] in the HGP through its completion in 2003... generating the economic output of $796 billion, and thus shows a return on investment to the U.S. of 141 to 1....
>
> The HGP is arguably the single most influential investment to have been made in modern science and a foundation for progress in the biological sciences moving forward."
>
> *Simon Tripp & Martin Grueber, 2011*[26].

Since the initial HGP, significant improvements to the quality of the reference genome have been made with each release; two crude statistics are presented below by means of illustration (Table 2.1). The number of discrete contigs initially decreases from the draft sequence as adjacent contigs are successfully merged; also the N50 length, a measure of the length of contigs becomes greater. Other quality metrics for the releases tend to follow the same clear pattern of improvement. Note here the large increase in the number of contigs for the GRCh38 release; this increase is due to alternative assemblies being created for highly variable regions where we observe diverse haplotypes such as the human leukocyte antigen (HLA) region on chromosome 6[9]. The existence of alternative assemblies for these regions ensures that accurate alignment can still be obtained for individuals where the genome does not agree with the canonical reference sufficiently to allow for accurate alignment of short reads. The continued increase of the N50 for the contigs bear testament to the work assembling the reference.

Table 2.1: Contiguity statistics for major releases of the human reference genome.

	Draft (2001)	**NCBI35** (2004)	**NCBI36** (2006)	**GRCh37** (2009)	**GRCh38** (2013)
Contigs	87,757	390	388	461	1,385[b]
N50[a] (bp)	274,300	38,509,590	38,440,852	46,395,641	56,413,054

[a]The size at which contigs of length $\geq$ N50 comprise $\geq$ 50% of the total assembly length.
[b]GRCh38 contains a large increase in the number of alternative assemblies for highly variable regions, accounting for this increase.

2.3 Association studies

The investigation of association of alleles with disease has been a successful methodology for studying complex disease. The methodology of these studies is relatively simple (using the example of a binary trait). A large cohort of unrelated individuals containing cases (individuals affected with your trait of interest) and controls (individuals matched to the case cohort, particularly as regards ethnicity) are genotyped. Following this, standard statistical approaches are applied to see if an allele is significantly overrepresented in the cases *vs.* controls. Early examples of this methodology involved testing the association with a single locus, such as the ABO blood-type (albeit indirectly *via* phenotypic characterisation)[27] or HLA loci[28]. As genotyping technologies have progressed, the numbers of markers assayed has increased dramatically.

2.3.1 Genome-wide association studies

With the availability of high-density genotyping arrays, the concept of the genome wide association study (GWAS) was made feasible. In a GWAS a large number of markers, generally SNPs, are genotyped using these high-density genotyping arrays, followed by testing of SNPs for association[29]. Commonly utilised high-density genotyping platforms are the genome-wide human SNP array 6.0 (Affymetrix) as well as the BeadChip range (Illumina), with many allowing for simultaneous genotyping of ~1,000,000 SNPs.

The rationale of a GWAS is that, due to LD, the genotyped 'tag' SNPs can be utilised to identify genomic regions of significance when one allele of the tag SNP is over-represented in disease cases when compared to controls for example[29–31]. Tag SNPs used are considered surrogate markers for their encompassing haplotype (Figure 2.2). Elucidation of the exact pathogenic variant can be undertaken following identification of an associated haplotype. Initial selection of tag SNPs, as well as analysis and refinement of data from GWAS requires a catalogue of sites of genetic variation and alternate-allele

frequencies (AF); for this purpose the International HapMap (haplotype map) Project was initiated[32–35].

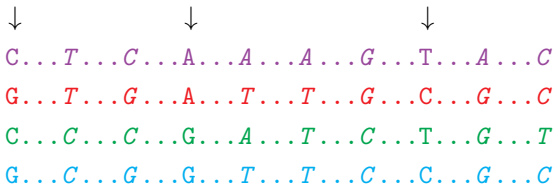


Figure 2.2: Rationale for the use of tag SNPs as surrogate markers for haplotypes. Three biallelic tag SNPs (indicated with arrows) are shown within the encompassing 4 haplotypes (coloured) present within the hypothetical population (as concatenates of variant sites). These three appropriate tag SNPs are sufficiently informative to uniquely identify each haplotype, and can thus be used as an efficient surrogate for haplotype determination within the population. Adapted from International HapMap Consortium, 2003[32].

One of the first successful GWAS studies was reported by Klein *et al.* in 2005[36]. This exemplar study utilised a cohort of 96 European patients with age-related macular degeneration, and an age matched control cohort of 50. Even with this small sample size, the authors were able to identify an associated SNP (rs380390) with a large effect size of a 7.4 fold increase in risk for individuals homozygous for the minor allele. Further to this, the pathogenic variant (rs1061170) was elucidated by sequencing of the encompassing gene in the cases and controls and validated based on experimental evidence.

Some authors have criticised the standard approaches of tag SNP selection as relying largely on unwarranted assumptions on the nature of LD patterns across the human genome[37,38]. Specifically, Terwilliger and Hiekkalinna published an 'utter refutation' of the core rationale behind the HapMap project, raising several concerns on largely statistical grounds regarding underpowered studies, and the assumption that pathogenicity of an allele would not affect LD patterns, which of course it would under strong negative selection pressures[37]. This in part explains the bias towards small effect sizes common in GWAS results; highly penetrant aetiological variants circulating within the population will tend to be present in only a minority of the founder haplotype in which it arose, hindering detection by GWAS[39]. It is of note that family-based linkage mapping will not suffer from this limitation.

Limitations of GWAS can be reduced with the infilling of intervening genotypes between tag SNPs by statistical inference, termed imputation. This imputation potentially allows for the prediction of rare non-genotyped intervening SNPs (AF < 0.05) with a stronger disease association that the tag SNPs, though with decreasing efficiency with decreasing AF, and does not allow for the detection of highly rare variation (AF < 0.003)[40]. More recently, bypassing this issue, arrays investigating rare variation

uncovered through large-scale resequencing projects have become available, with the intention to directly genotype functional coding variants within the cohort, already affording some success[41–43]. These rare-variant arrays however provide far less efficient imputation, and therefore allow for information on a much smaller proportion of the genome and are thus complementary, and not a viable replacement to, genome-wide arrays[44].

Despite the raising of some concerns, GWAS have been fairly successful in the identification of associated loci with complex disease. Large consortia such as the Wellcome Trust Case Control Consortium, which performed analyses with 3,000 common controls and 7 case cohorts of 2,000 individuals, were successful in the identification of associated variants, particularly for Crohn's disease and type I diabetes mellitus[45]. This one study (albeit a large one) identified 24 independent signals which achieved GWAS significance ($p < 5 \times 10^{-7}$) and a further 58 'suggestive' signals ($p < 5 \times 10^{-5}$) across the 7 diseases[45]. Many studies have had similar successes, with many significant regions being identified through the 'GWAS era'.

The vast majority of reported associations to date have an effect-size of much less than twofold; the challenge remains in the refinements of GWAS signals, and ultimately the clinical application of these associations. To illustrate this, the NHGRI GWAS catalogue[46] contains 9,947 GWAS significant ($p \leq 5 \times 10^{-8}$) records for all traits, with a median odds ratio of 1.064 (inter-quartile range 0.075–1.310). The odds ratios seen are highly variable by trait, for instance, for height, this is 1.044 (1.030–1.084) compared to 1.190 (1.129–1.380) for Crohn's disease.

2.3.2 Statistical considerations

There are two main considerations that hinder the identification of medically meaningful associated loci. Firstly, given the large number of statistical tests performed (most commonly one test per marker, so let us say for the sake of example 1,000,000), multiple testing correction must be applied to reduce the risks of false-positive findings[47]. There are many approaches for the limitation of false positive rates, the Bonferroni correction method is the simplest and can be informally presented as:

$$\alpha_{set} = \frac{\alpha}{n_{set}} \tag{2.2}$$

where α is the desired significance level for the set of tests (typically $\alpha = 0.05$), n_{set} is the number of tests being performed, and α_{set} is the corrected significance level that

must be used for each test in the set. Presuming the numbers mentioned above, this will give us $\alpha_{set} = 5 \times 10^{-8}$. Bonferroni corrections are often considered overly conservative as they presume that all tests are independent[47,48], which will not be the case in a GWAS due to the LD. The number of tests performed may be far higher in some cases, for instance where pairwise epistatic interations are investigated. Additionally, due to concerns over false-discovery, replication of results in an independent cohort is critical to validate initial results[49]. This correction for multiple testing means that large cohorts must be ascertained to provide the best possible power. The pressure for large cohorts may require the relaxation of criteria for inclusion; additional heterogeneity within the cohort may counter-productively reduce the power of the study.

The second consideration in GWAS interpretation is that the effect size of an associated haplotype is often modest. An odds-ratio of 4 would be considered substantial as an outcome for a GWAS[29,50], the issue becomes whether this readily translates into clinical utility. Whilst often it may not, the identification of these loci allows for the elucidation of pathways of importance and potential biological mechanisms for disease manifestation. This leads to the final challenge in GWAS interpretation, that of missing heritability. For most traits GWAS have characterised a small percentage of the observed heritability (measured by methods such as rates in siblings), with the remainder currently unexplained[50].

One striking example of this challenge of missing heritability is the analysis by Allen *et al.*[51] to identify genetic variation associated with height. The study utilised 180,000 individuals, identifying 180 associated loci; these identified loci collectively account for just 10% of the phenotypic heredity. There are several potential explanations for this missing heritability: the effects of rare variation (omitted by design from GWAS); other forms of genetic variation such as copy number variation; and also epistatic and epigenetic mechanisms. A further interesting possibility is that common variation as a whole contributes to the heritability of traits, as opposed to specific arbitrarily significant SNPs[50–53]. This hypothesis has profound implications for the potential translational application of genetics to complex disease. Further study of a broad range of hypotheses will hopefully help fill in this missing heritability[11,29,43,50,54,55].

2.4 Next-generation sequencing

Next-generation sequencing (NGS) is the massively parallel sequencing of DNA molecules, allowing for a sequencing throughput several orders of magnitude greater than Sanger sequencing; therefore the cost of sequencing a human genome has dropped by several orders of magnitude over the past 5 years (Figure 2.3). NGS has the capacity

for a far higher genotyping density than even the highest density array. NGS reads provide direct information for each nucleotide covered; as such, the loss of power due to recombination between the aetiological and tag variants seen in GWAS will not apply. Furthermore, no prior knowledge of potential variant sites is required, allowing the identification of rare and novel variants.

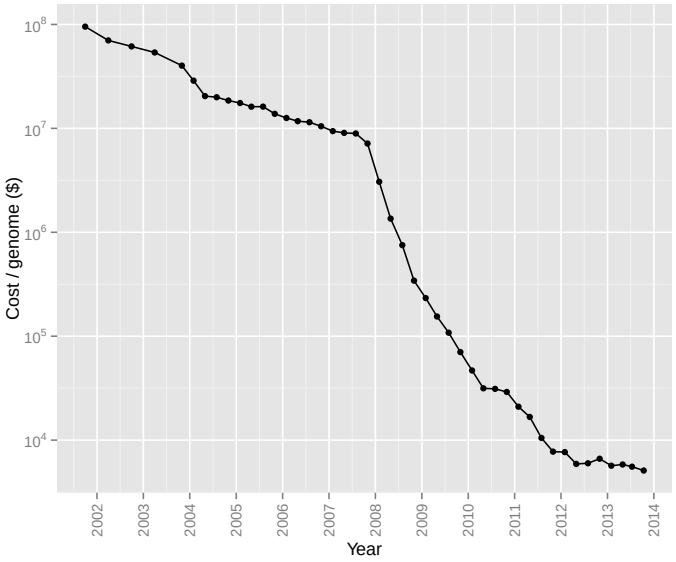


Figure 2.3: Cost of sequencing a human genome, 2001–2013. Costs (in USD) include all essential supplementary costs such as purchase of equipment and personnel costs. The sharp decline in cost beginning January 2008 is concomitant with the introduction of NGS. Note the logarithmic scale. The cost stands at $5,096/genome as of October, 2013. Data from `www.genome.gov/sequencingcosts`[56].

2.4.1 Applications of NGS

The development of high-throughput next-generation sequencing NGS platforms has allowed for rapid, cost-effective population re-sequencing projects in many prokaryotic and eukaryotic species including *Arabidopsis thaliana*[57], and of course *Homo sapiens*[12,58,59], as well as entire eukaryotic genera such as the *Saccharomyces*[60–62], allowing for comprehensive evolutionary analyses and comparative genomics[61,62]. Establishing a high-resolution catalogue of variation within population-specific cohorts provides researchers with a baseline of supposedly tolerated genetic variation; this baseline provides a hugely powerful filtering tool for the exclusion of common, and thus presumably relatively benign, variation when looking at genomic data derived from an individual sample[63]. Furthermore, projects to systematically catalogue the phenotypic effect of gene knockouts in mice allow a better understanding of gene function[64].

This distinction between benign and pathogenic variants is blurred when investigating complex diseases, contributory alleles for disease may be present at high frequency in a population, conferring a small increase in risk[30]. Now that NGS technologies have led

to vastly reduced sequencing costs compared to Sanger sequencing, it is now also possible to utilise re-sequencing to investigate specific traits, essentially with case-control designs, utilising existing variation databases as a shared control between studies[12,35,65,66]. This has the potential to identify some proportion of the missing heritability in complex diseases that remains following the GWAS-era.

Due to the increased power per sample, small sample-sizes can prove sufficient for identification of novel loci with Mendelian disease causality. Indeed, trio studies have been shown to have a high success rate for causal gene identification across diverse disorders with varying modes of inheritance, including dominant, recessive and *de novo* arising mutations, though alternative study designs are also effective, such as small cohorts and singletons[67]. The ability to identify causal variants from small cohorts allows investigation of very rare disorders, including those with incomplete penetrance, and unidentified biological causes. A recent study detailing the experience of clinical application of whole-exome sequencing (WES) reported a 25% rate of putative molecular diagnoses across large cohorts with diverse Mendelian disease[68,69].

Whole-genome sequencing (WGS) costs are currently prohibitively expensive for many research groups and clinical application. This, twinned with the computational challenges posed by the vast amounts of sequence data produced in WGS has led many investigators instead to currently utilise methods that target sequencing. Sequencing a desired small minority of the genome as opposed to the entirety further decreases costs and improves sequencing sample throughput[70]. With decreasing sequencing costs however, it is becoming increasingly viable to forego the exome enrichment phase and perform WGS.

Recent cohort studies have shown that WGS provides a greater diagnostic yield that WES, with a 34% diagnosis rate in Mendelian disease, increasing to 57% in trio analyses[71]. The greatly improved diagnostic rates when analysing trios highlights the major challenge in WGS, the interpretation of the vast amounts of data. In addition to the inclusion of non-exonic regions of the genome, WGS provide more complete coverage of the exome, providing greater variant detection sensitivity[72]. Beyond merely attaining sufficient coverage, WES sample processing leads to several other issues with variant detection. These are discussed in detail in Chapter 4. Due to the rapidly moving nature of the field, it is likely that there will be a move increasingly towards routine WGS in the next couple of years.

2.5 Implementation of genomics in healthcare

As genomics moves to the fore, including in the context of routine healthcare provisions, careful consideration must be given to potential uses arising issues to ensure that the introduction is maximally beneficial, while avoiding potential public backlash in the event of unexpected negative consequences, which could hinder the field.

2.5.1 Personalised medicine

Personalised medicine can be described as "an integrated, coordinated, and evidence-based approach for individualising patient care across the continuum from health to disease"[73]. Application of the tenets of personalised medicine requires identification and evaluation of biomarkers within the patient prior to decision-making. Genetic disease markers provide both the earliest indications of congenital disease-risk, as well as the least information on the dynamic progression of said risk, due to the intrinsic (mostly) stable nature of the genome (Figure 2.4)[73].

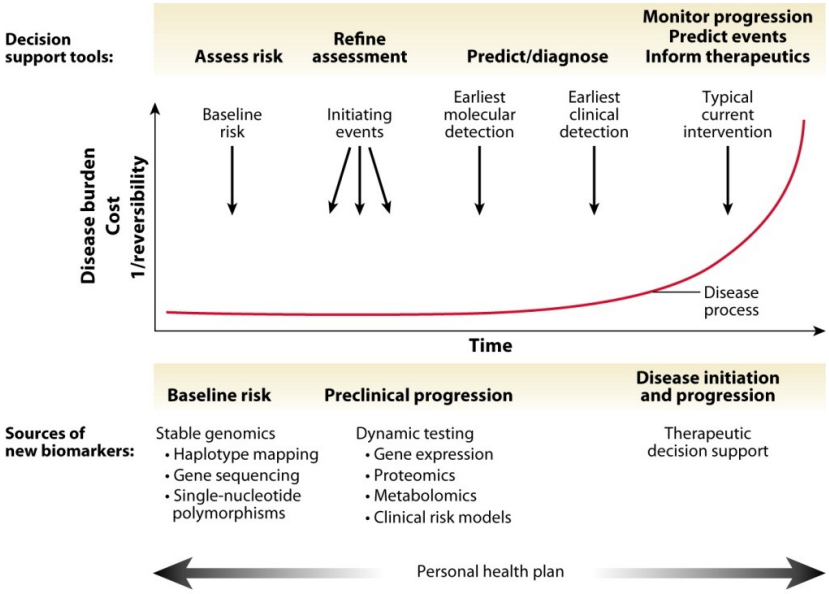


Figure 2.4: Potential roles for various biomarkers in disease risk prediction and diagnosis. Genomic information about an individual provide the earliest identifiers of disease risk, including even pre-fertilisation of the oocyte. However, genomic information alone will not provide information on risk progression, as dynamic biomarkers such as mRNA expression profiles might, (except where genetic instability is aetiological, as in oncogenesis). Interventions introduced following disease initiation may be less able to reverse disease progression compared to prophylactic interventions. Taken from Chan & Ginsburg, 2011[73].

For Mendelian disease traits, particularly those exhibiting dominant inheritance, disease risk can be approximated via evaluation of a thorough family history. Following this, where there is a perceived genetic risk, targeted genetic tests can be readily and cost-effectively undertaken to confirm genotype where a strong candidate aetiological locus for the disease of interest is known. Fulfilment of this caveat requires extensive

prior investment in genetic research to identify the locus. More than 2,500 monogenic diseases currently have available validated diagnostic genetic tests for clinical use[55].

Complex diseases pose a far greater challenge for both the elucidation of genetic risk markers, as well as clinical application of these markers. Identification of associated markers requires broadly-targeted genetic research, meaning ideally genome-wide methodologies. Despite the challenges, the field of large-scale genetic research continues to make huge advances in the identification of these markers[29,36,55,67,74–76]. Direct clinical application of these markers however remains a challenge, due principally to the relatively small increase in disease risk conferred by each variant identified[29,75].

In addition to the determination of personal risk for disease, genes and pathways identified by genomic research can also identify novel drug targets for rational drug design campaigns[77,78]. In addition to drug design, the identification of pharmacodynamically relevant markers allows for pre-emptive adjustment of drug regimens prior to administration, reducing the probabilities of adverse drug reactions[79]. Of particular interest for this approach are genes which are involved in modifying the adsorption, distribution, metabolism and excretion of the drug, such as transporters (e.g. *ABCB1*, encoding P-glycoprotein, an important cellular efflux pump)[80], as well as enzymes involved in metabolic processing, for both xenobiotic activation (e.g. *TMPT*, encoding thiopurine *S*-methyltransferase, activating the pro-drug azathioprine)[81], and degradation (e.g. *CYP3A4*, encoding cytochrome P450 3A4, catabolising ciclosporine and many others)[82].

2.5.2 Ethico-legal considerations

The unbiased nature of WES is also perceived as one of the major obstacles to routine clinical application of the technology. In all likelihood, within many individuals sequenced, healthy or otherwise, disease associated variants will be found, secondary to the reason for referral for sequencing, termed incidental findings (IFs). In some cases these may be low penetrance, entailing small increases in disease risk; in others however, variants will be highly penetrant for significantly detrimental phenotypes, such as certain *BRCA1/2* genotypes, associated mainly with breast and ovarian cancers[83,84]. There is a discrepancy in the attitudes towards disclosure of IFs between clinicians and lay persons. Lay persons were significantly more likely in one study to support the disclosure of IFs concerning themselves than the clinical geneticists who would be performing the disclosure[85].

There is a correlation of the willingness of clinicians to report IFs to patients with types of identified variant. For example, clinicians were more willing to report IFs which were

linked, with high levels of certainty, to serious treatable disease, than for untreatable serious disease or more dubious disease associations[85–87]. Resultant investigations and counselling for IFs will impose a significant cost to health services, though in some cases could also save costs thanks to avoidance of future acute interventions by implementation of cheaper prophylactic interventions[88].

The American College of Medical Genetics and Genomics (ACMG) have issued recommendations for the disclosure to patients undergoing medical sequencing, of all putatively deleterious variants observed in a curated list of genes. This list is selected so as to include genes in which deleterious variants predispose to serious, yet treatable conditions[89]. These genes are recommended to be actively screened where there is data. The apparent disregard for patient autonomy by not allowing patients to opt out of this disclosure, as well as the requirement to disclose information regarding minors, has been a source of much criticism, particularly in terms of the overly paternalistic nature of the recommendations[90]. In light of this criticism, the ACMG issued a clarification article, with no significant alteration to the stated position. More conservative recommendations have since been published by the European Society of Human Genetics[91].

It is worth noting that the issue of IFs is by no means unique to NGS studies; IFs are a major opposition to the utility of whole-body magnetic resonance imaging in diagnostics[92]. IFs do not purely entail disease risk factors; it is possible that information regarding false paternity and unknown consanguinity may be obtained. Some effort has been put into creating informatics approaches to the categorisation of variants within sequence data with pre-defined criteria[93]. This removes the human, time-consuming, and highly subjective aspect of the case-by-case decision making, and therefore may have a future role in the future simplification of the process for the end-user of the data.

While some have argued that the withholding of IFs pertaining to strong associations with treatable disease is ethically unjustifiable, regardless of consent[85], it would appear that thoroughly informed consent, detailing the patient's wishes as regards IF disclosure prior to data generation, and adherence to this agreement, would seem a reasonable path to take, and more in line with existing practices in medicine[86,87].

Informed consent, by its very nature requires clear communication with patients/-participants, which can be problematic. In a particularly extreme example of the issue of miscommunication, during a public health project involving genetics with Yup'ik Eskimos, researchers were unable to accurately convey the concept of genetics to some participants:

> "The formulation of "things that are passed through the blood from parents to children" can house real misunderstandings of the genetic basis of disease. One young [Yup'ik] man offered an example: "Parents who have HIV or AIDS will pass it on to their daughters or sons." An elder male also offered tuberculosis as such an example, as indeed it might appear to be, if the sick person has contact only with his or her family members. Such confusions highlight the need to clarify the differences between infectious and genetic mechanisms in discussions of hereditary traits, especially if using the "through the blood" descriptor."
>
> West *et al.*, 2013[94]

It is clear that if the broad concept of genetics can pose such difficulties then communication of the more complex implications of genetic testing may also prove problematic.

The problem of dealing with non-diagnostically relevant findings is enhanced by the availability of direct to consumer (DTC) testing. 23andMe, Inc. undertook a pilot programme offering consumers raw exome data for $999 USD (~£639 GBP as of 15th August, 2013) per individual[95]. That 23andMe offer only raw reads for the consumer to perform their own analyses, arguably absolves the company of responsibilities pertaining to causal variant identification, as these variants will have been obtained by the end consumer directly. The direct availability of WES data to the lay public will increase demand on already stretched genetic counselling services[96]. The ambiguity in the regulatory niche of DTC genetic testing has been clarified with the U.S. Food and Drug Administration (FDA) instructing 23andMe and other providers to cease providing medical interpretation of their DTC array genotyping results[97]. Regulatory vacuums are not uncommon for rapidly advancing medical technologies, for example this is also seen with some stem-cell treatments[98].

One of the intrinsic features of genetic information is that by definition, in most cases, the information acquired does not pertain solely to the proband. Ownership of information is a troublesome issue, requiring the balance of autonomy of the patient with a duty to potentially affected relatives, particularly where there is a high risk of a disease for which there are effective interventions[99]. The decisions by the patient in these cases is highly influenced by societal factors, and the cohesiveness of the family[100]. The issue of ownership of genetic information has been highlighted recently with the legal challenges of the family of Henrietta Lacks over the publication of the complete haplotype-resolved genome sequence derived from the HeLa cell line[101–103]. IFs may also directly affect relatives of the proband; in the case of IFs it increases the cost

of providing potentially unnecessary counselling and tests, whilst also amplifying the possible benefits of successful prophylactic intervention[88].

Chapter 3

Linkage Disequilibrium

3.1 Introduction

Many medical genetic approaches utilise the phenomenon of linkage disequilibrium (LD, also known as allelic association). LD relies on the property of chromosomes as continuous molecules; without further interference, alleles on the same chromosome would always be inherited together. However, this LD is degraded along a chromosome primarily through homologous recombination (HR; Figure 3.1)—the formation of chiasmata and resultant reciprocal exchange of DNA between sister-chromatids—pertinently during meiosis for these purposes, though HR is also important in DNA repair[104].

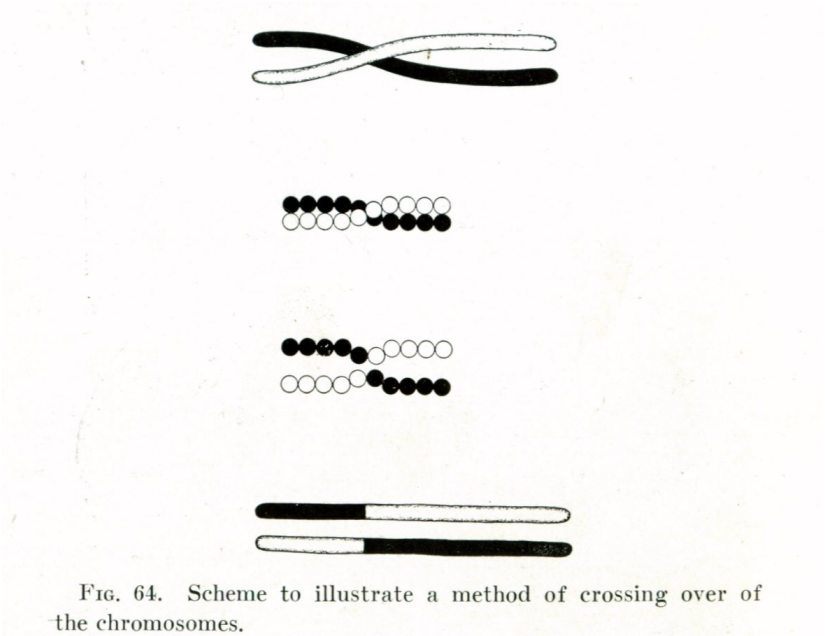


Fig. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Figure 3.1: Early illustration of the concept of homologous recombination by Thomas Hunt Morgan. It can be seen the the two sister chromatids crossover, forming a Holliday junction, which, on resolution, may result in the switching of chromosome regions between the pair. Taken from Morgan[105].

Due to HR, we find that, on average, proximal markers are more likely to be co-inherited than distal markers as HR is less likely to occur between closely spaced markers through the generations. On a population level, LD patterns across the genome are influenced by factors aside from recombination, though this is the primary architect. In addition, regions with a higher mutation rate, such as that seen in the HLA region, will show a greater breakdown in LD, and will also be more influenced by evolutionary selection[106].

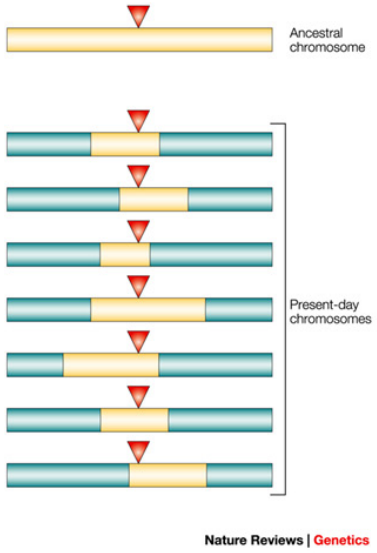


Figure 3.2: Degradation of LD from ancestral chromosome (yellow). A mutation arising on the ancestral chromosome (red triangle) will remain associated with the surrounding genetic background, with recombination events introducing new stretches (blue). Regions near the mutation are less likely to be interrupted by recombination. Take from Ardlie *et al.*[107]. Reprinted by permission from Nature Publishing Group, © 2002

3.2 Applications of LD

The principle of LD is exploited research such as linkage and GWAS studies, as well as population genetics[108]. A few examples of the applications of LD are discussed here; this list is not intended to be exhaustive, merely to illustrate the range of possibilities.

3.2.1 GWAS refinement

Arrays used for GWAS studies are optimised to provide as complete coverage of the genome through tag SNPs as possible. These tag SNPs act as surrogate markers for the encompassing haplotype (Figure 2.2)[35,46]. Where a tag SNP is identified as significantly associated with a trait, further work must be performed in order to identify the functional variation in LD with the tag SNP[36]. High resolution appreciation of LD in the region of the tag SNP enables definition definition of a region of interest flanking the tag SNP[109].

Information regarding LD can also be used in the initial stages of an association study. For instance, one approach implemented in *CHROMSCAN*[110] utilises an LD map[106] in order to best incorporate LD information into association mapping. The authors find that this algorithm provides a 5% improvement in statistical power; furthermore, they report a 46% improvement in the accuracy of the localisation of the causal SNP on the physical map compared to alternative methods. It is noteworthy that this study was performed using data with ~100,000 SNPs across the genome, it is therefore reasonable to assume that further increases in resolution would be obtained with higher marker densities[109,110]. Other groups have also had success using LD maps for GWAS refinement[111].

3.2.2 Selection

A genomic region under purifying selection will have reduced haplotypic diversity within a population, as variation arising through mutation will be removed from the population over generations. As such, there will be significantly increased LD across a region in a population where there is selection *vs.* a population where no selection pressure is applied. Similarly, there are disctinctive patterns of LD where a locus is under selection, without requiring differential selection between studied populations[112].

3.2.3 Recombination mapping

The predominant architect of LD patterns in the genome is meiotic recombination[104,109,113–115]. Because of this, patterns of LD can be used to identify recombination hotspots, ~2 kb regions with high recombination intensity. Using LD structure, Myers *et al.*[114] identified a short motif underlying ~40% of recombination hotspots, indentified to be a binding site for PRDM9.

3.3 Visualisation of LD

LD patterns are shaped by multiple factors, namely recombination, mutation rates, drift, selection and population history. As such, they are often highly complex; conceptualisation, particularly visually, therefore poses a real challenge for researchers utilising LD. Many software packages are available to help tackle this problem, with a wide array of approaches. One example of the methods for visualising LD is *Haploview*[116], which is the most commonly used (based upon citations). However, there are many alternative visualisation approaches, with diverse rationales; some examples of this are shown in Figure 3.3. It can be seen that the pairwise visualisations (i.e. Figure 3.3a,c,d) become cluttered and confusing with a large number of markers, whereas model based plots (i.e. Figure 3.3b) facilitate the identification of the signal in the noise for clear display.

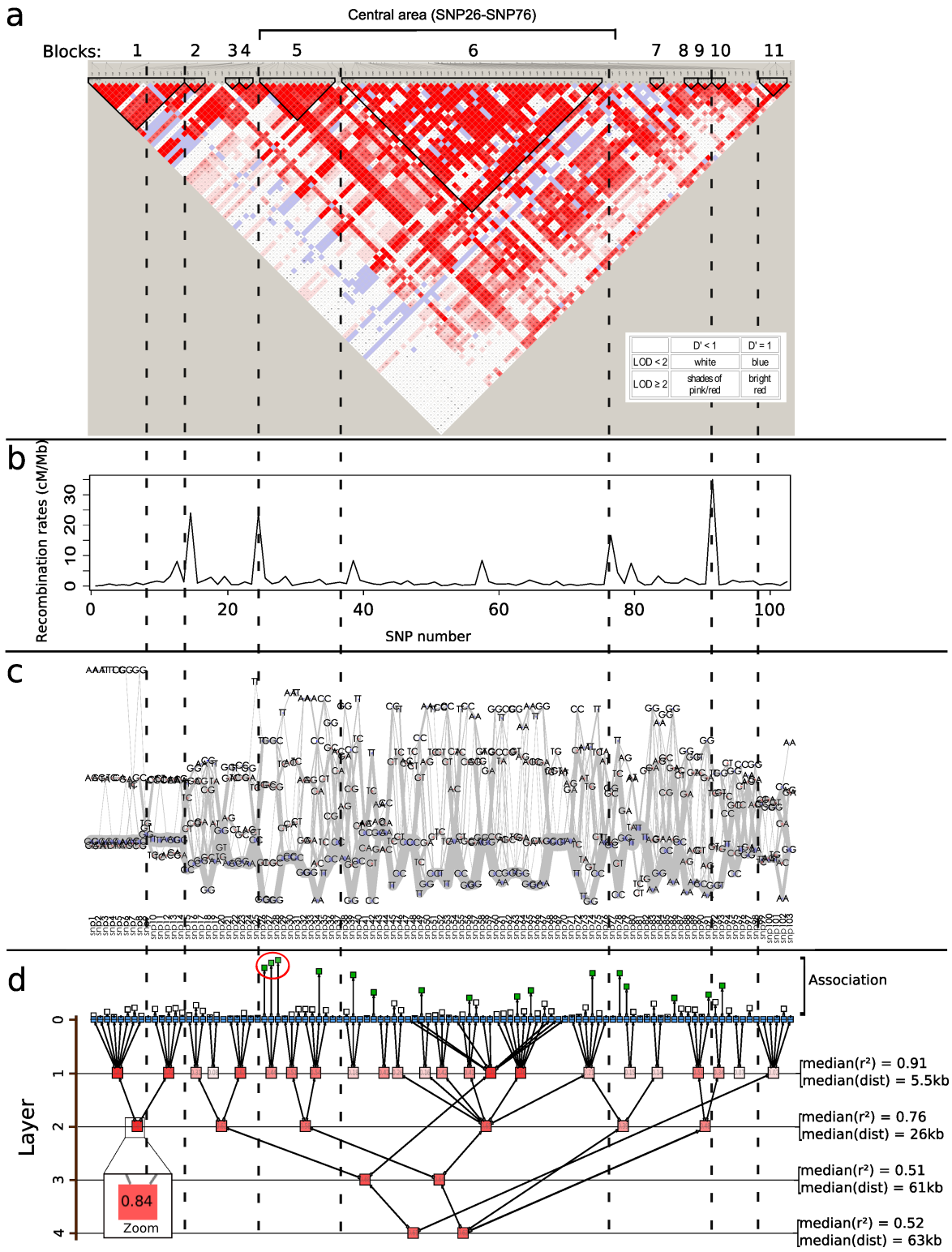


Figure 3.3: Comparison of LD visualisation software. Software used: a) *Haploview*[116], a triangular heat-map based visualisation; b) *PHASE* v2.1[117], a coalescent model based metric; c) a Textile Plot[118] and d) *Tulip*[119], a latent forest based method. Common regions of LD breakdown are indicated by vertical dashed lines. Figure taken from Mourad *et al.*[119] under the Creative Commons V2 Attribution License.

3.4 Measures of LD

Given the importance of LD in genetics, a plethora of measures are utilised by researchers in order to quantify LD. A range of metrics used in LD utilising studies are discussed below.

3.4.1 Pairwise metrics

Pairwise LD metrics are the most commonly used, being relatively free of biological assumptions. For considering pairwise LD metrics, it is informative to first construct a 2×2 contingency table for possible haplotypes (Table 3.1, adapted from Mueller, 2004[120]).

Table 3.1: 2×2 contingency table for possible haplotypes of biallelic loci A and B.

		B_1	B_2	
A_1		A_1B_1	A_1B_2	
	Actual	pA_1B_1	pA_1B_2	
	Expected	pA_1pB_1	pA_1pB_2	pA_1
A_2		A_2B_1	A_2B_2	
	Actual	pA_2B_1	pA_2B_2	
	Expected	pA_2pB_1	pA_2pB_2	pA_2
		pB_1	pB_2	1

There are many metrics available for the quantification of pairwise LD between markers, which can be defing using the nomenclature in Table 3.1 (Table 3.2), each with their own advantages and disadvantages[120]. The two most commonly used metrics for pairwise LD are r^2 and D'. r^2 is the rate at which one allele successfully predicts the other allele, rendering r^2 metric sensitive to AF. D' however utilises the D_{max} parameter to correct for AF, allowing for normalised comparison between marker pairs with differing AF. The difference between these two commonly utilised pairwise metrics is illustrated in Figure 3.4

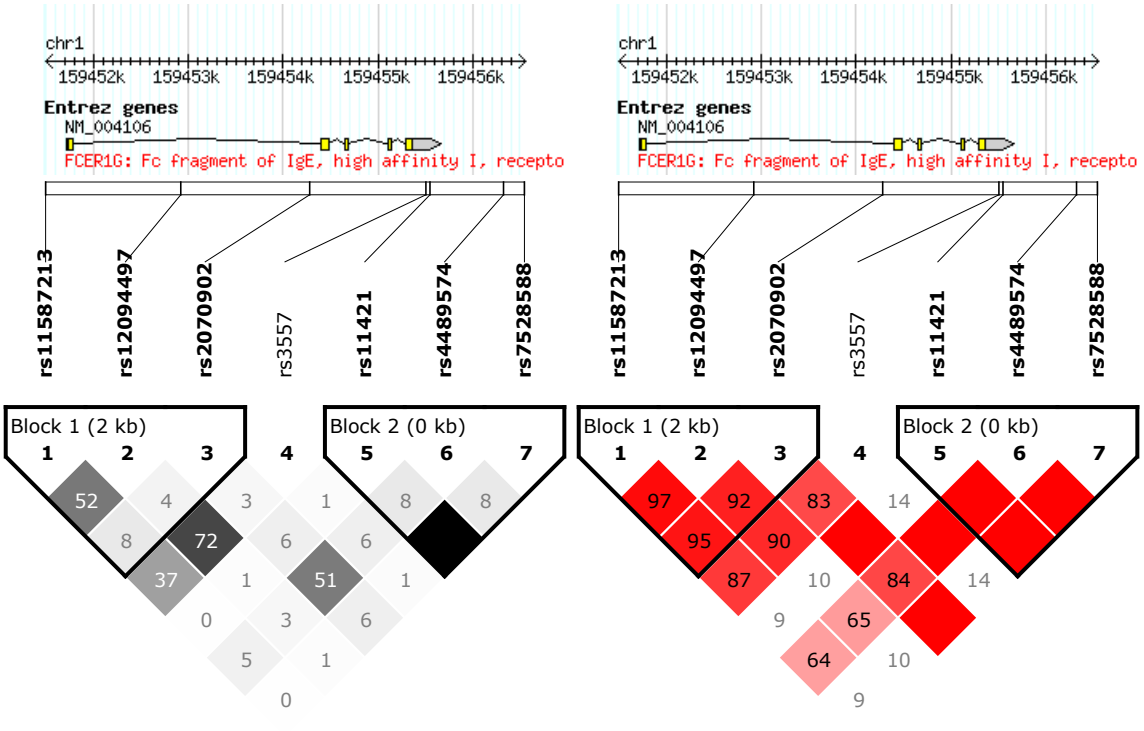
Table 3.2: Available metrics for pairwise LD quantification and their properties.

Metric	Definition
D	$pA_1B_1 - pA_1pB_1$
D'	D/D_{max} [a]
r	$D/(pA_1pA_2pB_1pB_2)^{\frac{1}{2}}$
r^2	$D/pA_1pA_2pB_1pB_2$
$\hat{\rho}$ [c]	$\lvert D' \rvert$
Δ	$pA_1B_1 + pA_1/B_1 - 2pA_1B_1$ [b]

[a] $D_{max} = \begin{cases} \min(pA_1pB_1,\ pA_2pB_2) \text{ when } D < 0 \\ \min(pA_1pB_2,\ pA_2pB_1) \text{ when } D > 0 \end{cases}$

[b] pA_1/pB_1 is the frequency of alleles A_1/B_1 being inherited in *trans*.

[c] Note that $\hat{\rho}$ is termed ρ in the literature, we use this variant symbol herein to prevent confusion with the Spearman's correlation, ρ.



(a) r^2 **(b)** D'

Figure 3.4: Comparison of r^2 and D' for SNPs in the *FCER1G* gene. The *Haploview* display[116] shows a triangular heatmap for 7 SNPs in the region, highlighting the differences in the two metrics. Values for D' between markers are consistently greater and more stable than r^2, largely due to compensation for marker AF though the D_{max} component. Both metrics have utility dependant upon the information desired.

3.4.2 Multi-locus measure of LD

While the pairwise measure of LD discussed are suitable for some purposes, where more complex studies of LD are desired, particularly for downstream analyses, alternative

methods are required. Two methods, the Malécot-Morton and coalescent models for quantifying LD will be discussed.

3.4.2.1 Malécot-Morton model

The Malécot-Morton model was developed by Newton Morton during his tenure as the head of the Genetic Epidemiology group at the University of Southampton. The model is based upon the Malécot model of isolation by distance, itself initially derived for application to separation of populations by geographic distance[121]. The final Malécot-Morton model is defined as:

$$\hat{\rho} = (1 - L)Me^{-\epsilon d} + L \tag{3.1}$$

where $\hat{\rho}$ is the association between SNPs, the asymptote L is the 'background' association between unlinked markers which is increased in small sample sizes and with residual population structure, M reflects association at zero distance, with values of 1 consistent with monophyletic origin and < 1 with polyphyletic inheritance, ϵ is the rate of LD decline, and d is the physical distance in kb between SNPs[122].

The variable $\hat{\rho}$ has been shown to be the most efficient representation of LD in a region, as well as being highly insensitive to AF, and intuitive to interpret[123]. The software *LDMAP* iteratively fits the Malécot-Morton model for values of $\hat{\rho}$ between multiple markers to identify the values of L, M and ϵ which provide the closest fit for the observed data. The software ultimately produces a map in linkage disequilibrium units (LDU), which equal ϵd (Figure 3.5).

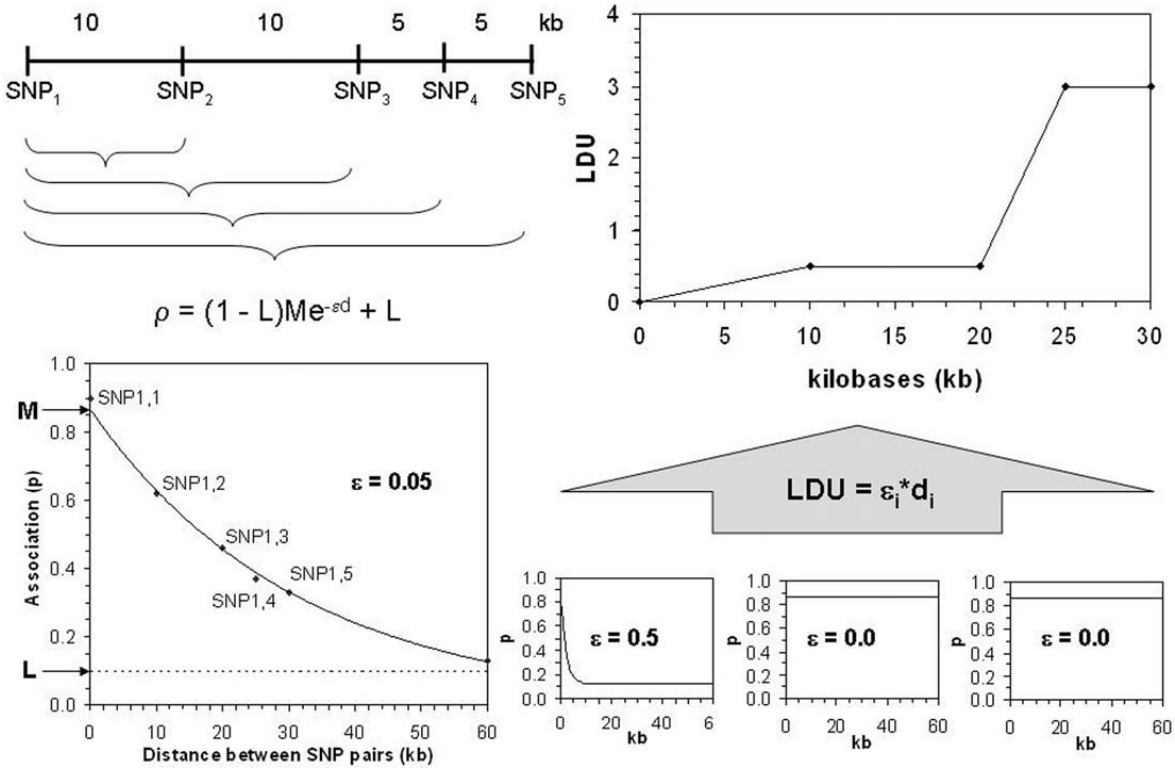


Figure 3.5: Illustration of *LDMAP* algorithm. Firstly, *LDMAP* calculates pairwise $\hat{\rho}$ between markers, then using these to estimate ϵ, as well as M and L accross the region. Once ϵ is estimated, the final LD map is constructed in LDU (i.e. ϵd). Taken from Tapper[124]. Reprinted by permission from Springer, © 2007.

3.4.2.2 Coalescent models

The coalescent model of evolution is based upon the principle that evolutionary processes in a population can be represented as a Markov chain of events[125]. These approaches are utilised for the simulation of population genetic data, as illustrated in Figure 3.6[126].

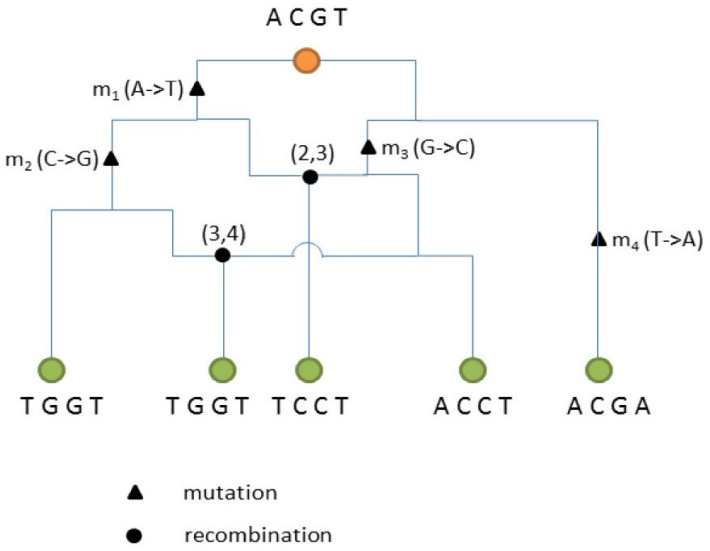


Figure 3.6: Illustration of the coalescent model. After four mutations (denoted by m_n) and two recombination events [indicated on nodes], the sequence of the last common ancestor has evolved into five (four distinct) present-day sequences. Figure taken from Yang *et al.*[126] under the Creative Commons V2 Attribution License.

Software such as *LDhat* takes empirical genotype data from a population and derives estimates of recombination. Specifically, the `rhomap` function implemented in *LDhat* provides a value of ρ, defined here specifically as:

$$\rho = 4N_e r \tag{3.2}$$

where N_e is the effective population size and r is the sex averaged recombination rate in the population[127]. In a comparison of *LDMAP* and *LDhat*, Tapper *et al.*[128] showed that LDU maps have a greater correlation with empirical linkage maps than their coalescent counterpart ($R^2 = 0.37$ and 0.32 respectively).

Chapter 4

Experimental & Analytical Methodologies Utilising NGS

> "On two occasions I have been asked,—"Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?"... I am not able rightly to apprehend the confusion of ideas that could provoke such a question."
>
> *Charles Babbage, 1864* [129]

As discussed in Chapter 2, NGS and associated methods are a powerful approach for clinical molecular diagnostics. Translation of NGS into clinical science however requires improvement and validation of the quality of final data.

NGS has proven to be a disruptive technology in the field of genomics. The improvements in technologies have required a concomitant increase in our analytical capabilities, not purely in terms of computing power and storage, but also in intelligent methodologies for efficient analyses. In this section I will discuss the practical processes for NGS analysis of DNA, from patient selection for sequencing through to aetiological candidate identification in Mendelian disease.

4.1 Sample selection and acquisition

4.1.1 Patient selection

It is essential that appropriate selection of individuals is undertaken to ensure that sufficient power is available for the identification of candidate aetiological variants. Firstly, the pedigree should be formally recorded in as much detail as practicable to allow for evaluation of the mode of inheritance of the disease (e.g. Figure 4.1). For the purposes of this work, primarily focused on Mendelian disease, the pedigree should be

evaluated to ensure that the disease is likely to have a *strong* genetic component, as otherwise the standard methodologies of analysis utilised are unlikely to result in useful conclusions.

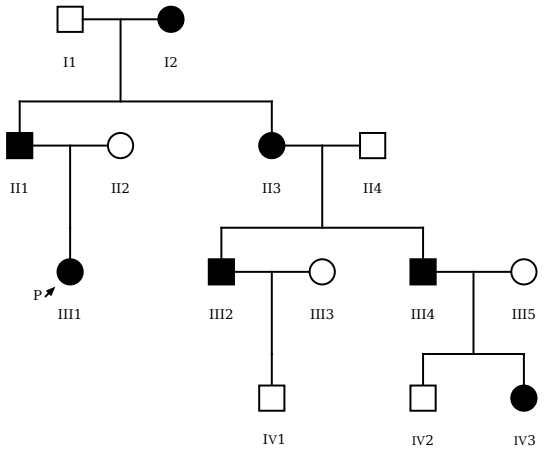


Figure 4.1: Illustrative pedigree showing inheritance of autosomal dominant disease across 4 generations. P indicates the pedigree proband (III1). Of the affected individuals sequencing the proband and individual IV3 would provide the best segregation filtering power.

Where it is deemed suitable (and genetic material is available) to sequence multiple members of a pedigree, it must be ensured that the affected members selected are maximally informative by minimising the probability that alleles are shared by chance between members, improving filtering power on variants based on segregation. Individuals selected from a pedigree for sequencing may vary dependent upon the presumed mode of inheritance. A recessive condition will require more members of the pedigree to be sequenced than a dominant condition. In severe autosomal dominant disease, the causal variant is expected to be a strong outlier in terms of conservation at the site whereas in recessive conditions it is possible that the aetiological variants circulate within the population at an appreciable frequency, and are thus more problematic to identify by comparison with databases.

Based upon Table 1.1 it can be seen that to minimise the IBD between sequenced affected individuals, 1st cousins would be optimal out of the above options. Conversely, to sequence an unaffected member of the proband's family then a sibling or parent is ideal. Caution must be used with distant relatives; it is possible that 'affected' individuals may possess a dissimilar phenotype or even an alternative aetiology for the disease (dependent upon disease frequency). It must be ensured that the intervening pedigree information is consistent with continuous carriage of the disease through the pedigree. In Figure 4.1, between the proband and IV3 it is expected that there is excellent segregation filtering power ($\Phi = 0.0625$). At this stage also it cannot be overly stressed the importance of accurate phenotyping; erroneous assignment of affected/unaffected status can nullify the value of segregation filtering of variants. In

consanguineous pedigrees, autozygosity mapping for regions of IBD that are homozygous in affected individuals can be a powerful approach[130].

4.1.2 DNA isolation

Following selection of pedigree members to be sequenced, DNA must be appropriately sourced. In some cases a specific source tissue may be used e.g where somatic mosaicism is anticipated as in cancer. Where there is no anticipated tissue specificity then the most commonly used sources of DNA are peripheral whole-blood and saliva where this is not feasible. NGS of a large proportion of the genome requires a large amount of high-quality gDNA (ideally in the order of µg, though smaller quantities are viable). The quantity and quality of DNA required for NGS can be problematic for some studies, particularly where non-fresh sources such as formalin-fixed, paraffin embedded (FFPE) samples are used. FFPE fixation produces cross links between the protein and DNA in the sample, and chemical modification and fragmentation of the DNA. However, specific approaches can be undertaken in order to maximise the likelihood of obtaining high-quality DNA from samples such as these[131,132].

4.2 *In vitro* technologies for NGS

4.2.1 NGS sequencing platforms

Since the advent of mainstream NGS with the release of the 454 sequencing platform in 2005, several platforms have been made available, based upon diverse chemistries and rationales. Comparisons of several NGS platform have been made (Table 4.1, adapted from Liu *et al.*, 2012[133]), in which the Illumina HiSeq 2000 system appears to be the frontrunner, all factors considered[133,134]. These data were collated in 2012; given the rapid progression of technologies in the field all platforms have been improved, the table however still proves informative. This is reflected by the fact that Illumina is also the clear market leader in NGS, with a market share of 56% in 2012[135]. Furthermore, in Q4, 2013 Illumina, Inc. received approval from the FDA for its MiSeqDx diagnostic sequencing system and associated targeted gene sequencing panels[136], the first such NGS technology to receive FDA approval.

Table 4.1: Comparison of considerations for 3 NGS platforms.

	454 GS FLX	**HiSeq 2000**	**SOLiDv4**
Methodology	Pyrosequencing	Synthesis	Ligation
Read length (bp)	700	150 PE	50 PE
Runtime (Days)	1	7	14
Accuracy rate (%)	99.9	98[a]	99.94
Cost/Mbp ($)	10	0.07	0.13

PE - paired end reads

[a]As quoted in Liu *et al.*, 2012[133], 99.74% is the quoted accuracy in Quail *et al.*, 2012[134], this large discrepancy may be due to different metrics being used, as error rates will be dependant upon sequence context and nature of the errors counted.

The Illumina sequencing by synthesis (SBS) technology published in 2008[137] could produce 35 bp paired end (PE) reads, with the best cost-efficiency per Mbp of the 3 platforms compared. The maximum read-length is now far longer, with 300 bp PE reads possible on the MiSeq platform due to process improvements. The rationale of SBS is that of reversible chain termination, as opposed to the irreversible chain termination utilised in Sanger sequencing. On each reagent-cycle within the microfluidic flowcell, fluorescently labelled, terminated nucleotides are passed over immobilised single-stranded, primed DNA, allowing the progression of the complementary strand synthesis by a single base. Each base is terminated with a different fluorophore, with non-overlapping emissions maxima; following removal of unbound nucleotides, the base incorporated can be identified by imaging of the flowcell following laser excitation. The cycle is ended by chemical cleavage of the terminating fluorophore, leaving the 3'-OH open to nucleotide addition in the new cycle.

This process is repeated for a defined number of cycles, which may be followed by equivalent sequencing from the opposing end of the template DNA fragment, which is typically ~300 bp in length[137]. This use of PE reads allows for more long-range information to be gleaned from the data, particularly useful for alignment, as well as investigating structural rearrangements (as discussed in subsection 4.3) and local phasing. The length of the short reads however is limiting in their use for alignment in certain regions of the genome, and can be problematic for *de novo* assembly of genomes and transcriptomes[138].

In addition to the above platforms available for purchase, Complete Genomics, Inc. specialise in providing service WGS, using their proprietary DNA nanoball sequencing by ligation methodology[139]. The method has also been adapted to allow for long range phasing of genotypes to produce haplotype contigs of N50 $>$ 500 kbp[140], further enhancing the utility of the technology.

NGS technologies are constantly evolving, with regular updates to sequencing chemistries and control software, improving sequencing and analysis accuracy, particularly for *de novo* assemblies. Also, new platforms and methodologies emerge, with approaches providing long single-molecule reads becoming more mainstream[141–143].

4.2.2 Genomic subset enrichment

Due to the continued relative expense of WGS as a single test there are several approaches available for the selection of regions of gDNA of interest, theoretically enhancing the efficiency in terms of variants of interest observed per unit of sequence data acquired. There are two major classes of enrichment: whole-exome sequencing (WES) and more limited in scope targeted enrichment, which may be custom designed to cover tens of genes. An increasing move toward more selective panels can be seen recently, particularly as NGS diagnostics becomes more routine. This further increases efficiency where the genomic regions of interest have been identified in previous studies, while also greatly simplifying the analytical and ethical issues due to the narrower scope of the investigations, resulting in less data to analyse per patient.

4.2.2.1 Whole-exome sequencing

The human exome (complement of protein-coding regions of the gDNA) is oft-quoted as bearing 85% of aetiological variants, despite constituting 1–2% of the genome[144], though the provenance of this statistic is unclear. The utility of WES was first demonstrated by Ng *et al.* in 2009[145] on 12 individuals, and has since been demonstrated to be an exceptionally useful tool in the geneticists toolbox[54,67,70,76,144,146–152]

Sample preparation in WES is more complex than WGS, due to the requirement for this pre-enrichment of gDNA for exonic regions (Figure 4.2), with implicit additional costs for this stage of processing. However, this additional preparation cost is offset currently by savings in required sequence data for suitable data, as well as downstream *in silico* processing. Sample preparation utilises sequence-specific hybridisation: 'baits' of oligonucleotides complementary to exomic regions of the genome are incubated with fragmented gDNA. Subsequent retrieval of baits will provide a pool of enriched DNA for downstream processing. As with NGS, there are several exome enrichment platforms available, each with their own defined target regions[153].

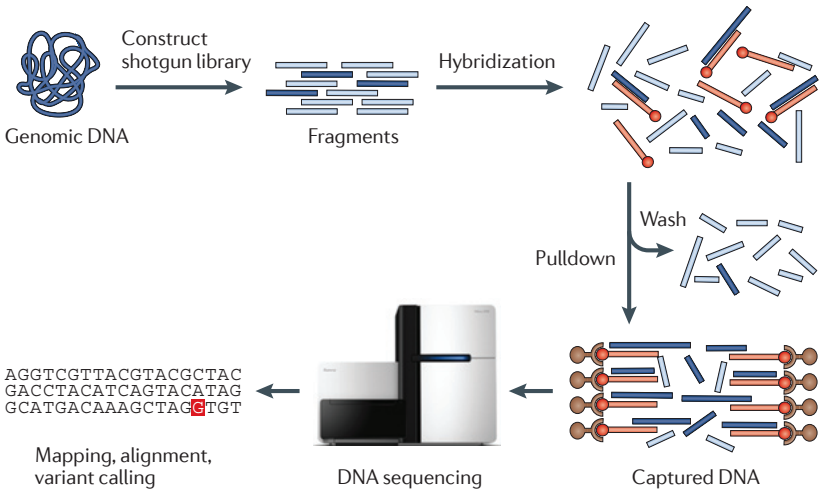


Figure 4.2: Workflow of exome sequencing sample processing. Extracted genomic DNA is fragmented by ultra-sonication, which may be followed by ligation of identifying barcodes to the fragmented DNA. DNA fragments are then hybridised to targeted baits (red) of an exome capture kit such as Agilent SureSelect. These baits are designed to be complementary to protein coding regions, thus the pulldown of the baits will provide a DNA sample enriched for exonic DNA, ~1% of the genome. This is then sequenced and analysed downstream *in silico*. Taken from Bamshad *et al.*, 2011[70]. Reprinted by permission from Nature Publishing Group, © 2011.

One of the main strengths of WES over targeted candidate gene sequencing lies in the relatively unbiased nature of the data acquisition. Excessive masking of data acquisition too early in the study would limit the answers to this initial subjective area of interest[154]. This will limit the utility of the experiment in cases where the cause lies within a non-candidate gene for the disorder, or where the disease is mischaracterised or uncharacterisable. However, once the unbiased data are acquired, a initial tiered interrogation of the data may still be performed to reduce the analytical burden where strong candidate loci are known.

There are several weaknesses within the WES methodology. Due to the requirement for sequence identity between the gDNA and WES capture kit, there can be biases in capture of alleles. This may for example be caused by a deletion preventing correct alignment of bases for the bait annealing stage[155]. Furthermore, the core principle of WES in only sequencing exonic regions of the genome can result in the non-identification of non-exonic variants or CNVs and large-scale structural rearrangements due to the low level of information on 99% of the genome. These factors could result in the absence of data on potentially clinically relevant alleles.

4.2.2.2 Targeted capture

While WES provides an attractive cost-effective alternative to WGS currently, for some purposes a smaller genomic subset may be desired. In these cases, smaller panels of gDNA enrichment are available, and are also amenable to custom design. There are two

approaches used for this: hybridisation based, similarly to in WES pre-enrichment, and amplicon based, where the enrichment is by means of multiplexed PCR amplification of regions of interest[156]. The small proportion of the genome that can be captured using these approaches allows for much larger throughput of samples than WES, and also allows use of low-throughput sequencing platforms such as the Illumina MiSeq. A further effect of the small region of enrichment is an increased depth of coverage for the regions captured. This is particularly useful for cancer resequencing, where the tumour DNA purity is likely to be low due to stromal contamination. High depth allows for increased detection of somatically acquired variants[157], including developmental mosaics; the cost of very high-depth sequencing for a whole-exome would be excessively high for many purposes.

4.3 *In silico* analytical processing of NGS data

The nature of NGS data requires several processing stages for the gleaning of biologically interpretable data (Figure 4.3). The three phases will be discussed below, as well as the quality control (QC) that should accompany analysis. The storage requirements for NGS data and analysis files can be substantial, requiring ~10 GB for WES data, as well as additional capacity for working files and back-ups as required. This is unlikely to pose a challenge for individual patients, however, as NGS becomes more commonplace, an appropriate storage infrastructure will become essential. Computational power is also a consideration, with WES alignment taking many hours on a modern desktop computer. Again, single samples do not pose a challenge, but analysis parallelisation becomes essential as throughput increases, ultimately requiring high-performance computing facilities, either locally or cloud-based, though this poses additional challenges of data security.

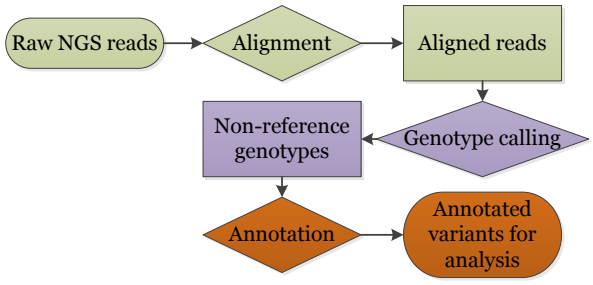


Figure 4.3: Generalised analysis workflow for NGS data analysis. The three major stages of NGS analysis are shown, namely alignment, variant calling and annotation.

4.3.1 Alignment of NGS short-reads

Alignment of NGS short-reads is the first, and most computationally intensive stage of the *in silico* analysis in resequencing following data generation. Alignment entails

defining the position of each read in relation to the reference genome based upon finding the position in the reference genome with which the read has least mismatches, be they SNPs, indels or sequencing errors. A commonly utilised aligner for WES is *BWA* (Burrows-Wheeler Aligner[158]). *BWA* applies the Burrows-Wheeler transformation[159] to efficiently hold the reference genome in memory for comparison with read sequences, outputting a sequence alignment/map (SAM) file containing read sequences along with best alignment positions; following SAM file generation, data can be more efficiently stored in binary alignment/map (BAM). The SAM/BAM formats are the *de facto* standard for alternative alignment software. Other popular aligners such as *Bowtie*[160] and Novoalign[161] are based on an alternative implementation of the same principle. For PE reads, there is an extra layer of complexity. In *BWA* both reads are independently mapped to the reference genome and viable positions are then compared to select the pair most closely collocated in the correct orientation[158].

Certain aligners will be preferable in different circumstances. *Bowtie* for instance is able to more rapidly process reads compared to *BWA*, but at the expense of tolerance to errors in the reads[162]. Additionally, regions in which there is a high sequence diversity require alternative approaches for accurate alignment, and thus downstream calling. The HLA region is a region towards which considerable attention has been directed due to the region's biological import. As such tools are available for the specific alignment to all known haplotypes and subsequent calling, e.g. *Omixon Target*[163].

4.3.2 Variant calling from aligned reads

There are many tools for the calling of genetic variants from aligned sequence data, some of which are focused on calling of particular classes of variants; an illustrative selection of tools is discussed below. The variant call file (VCF) has become the *de facto* standard for genotype calls of all variant classes[164].

4.3.2.1 SNP calling

SNP genotypes are the most readily called class of genetic variation. The software *SAMtools* is currently the most highly cited tool suitable for this purpose[165], followed closely by *GATK*[166]. *SAMtools* initially produces a raw 'pileup' of sequence data at a position from an alignment, then applying a Bayesian probabilistic framework to determine the most likely genotype at the position, as well as assigning a phred-scaled score to the genotype call indicating the quality of the call (see subsubsection 4.4.1). As well as the calling of singleton samples, *SAMtools* and *GATK* can be used for calling multiple samples in an analysis; here several pileups are analysed concurrently

with prior probabilities for genotypes for each sample being dependant upon the allele frequencies within the pool, improving calling of shared genotypes.

4.3.2.2 Indel calling

Short insertion/deletion variation (indels) can often be called using standard SNP calling software such as *SAMtools*, however increased accuracy can be obtained with dedicated software for various sizes of indel. Two types of read-level evidence can be utilised for indel discovery (Figure 4.4). Firstly, split reads, where portions of a single read map discontiguously to the chromosome, and split pairs, where a read pair maps with an insert size between the pair significantly greater than that expected given the distribution of insert sizes in the sample. *Pindel* takes advantage of both sources of information, affording the ability to call medium–large indels of 10 kb from 36 bp PE reads with base-pair precision[167]. Similarly, *SoftSearch* takes advantage of this information, as well as 'soft-clipping' of aligned reads, where an end of the read has low mapping quality due to a gap in the alignment that has not been opened, resulting in multiple mis-matches which would normally be ignored in downstream analysis[168].

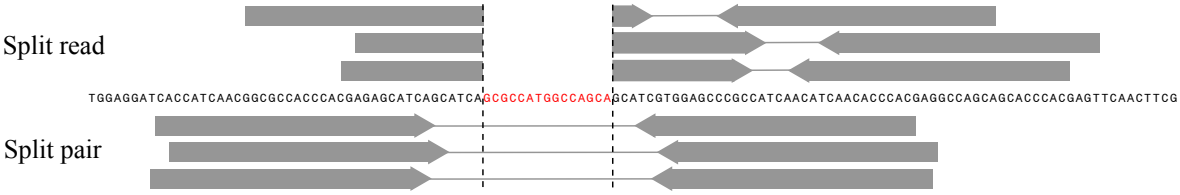


Figure 4.4: Informative features in NGS reads for the detection of indels. Reads (thick arrows) may be split directly by a deletion as compared to the reference sequence (red sequence), allowing for identification of the indel with base-pair resolution. PE reads may also have an increased aligned insert size (thin joining line) where the insert spans the indel event. Multiple pieces of such evidence will be required to confidently call the genotype.

4.3.2.3 Copy number variation calling

Copy number variations (CNVs) are a form of large-scale structural variation of the genome which entails the duplication/deletion of a region. Here there are three main forms of evidence for these events in NGS data. Since CNVs can be considered large scale indels, the information described in Figure 4.4 still proves informative. In addition, depth of coverage (DOC) of a region will be proportional to the genomic dosage of that sequence, i.e. where a heterozygous deletion is present the expected DOC for that region is halved.

CNV detection methods required vary between WGS and WES studies. In WGS we expect a relatively uniform DOC in most regions, whilst in WES DOC is highly heterogeneous due to variable capture efficiency at the exome enrichment stage. WGS by definition has reads mapping to the majority of the genome, making read-level

information of more utility than in WES, where this information is available for a minority of the genome; off-target reads, where the region has some data despite not being a target of the exome enrichment can prove informative in these situations (e.g. see subsection 7.3.2). There is a plethora of CNV calling software available, particularly for WES where more complex approaches are required, and can be broadly categorised as split-read/pair based (such as *Pindel*[167]), DOC based (such as *XHMM*[169]), or a hybrid methodology (such as *SoftSearch*[168])[170].

4.3.2.4 Somatic variants

Variants present at an individual's conception will be present with an allelic dose of one or two alleles in diploid cells, and will therefore be expected to be in a high proportion of the reads from a sequencing experiment (see also subsubsection 4.4.4), allowing detection with the software such as *SAMtools*[165] and *GATK*[166]. Variants may arise during foetal development, resulting in mosaicism. Alternatively, variants may also arise later in life, for example, though nor purely, as associated with cancer. In these cases, we expect a lower proportion of reads to be derived from the variant DNA, as well as greater variation dependant upon the exact DNA source; here more sensitive methods of detection are required than for germline variation.The power of detection of somatic variants in inexorably linked to the read depth of the sequencing experiment. If a variant if present at an allelic proportion of 1%, it is highly unlikely that a sequencing experiment delivering 20 X would detect this variant. Where low level variant detection is a priority, read depths $\gg$ 1,000 X may be used.

To allow for greater variant detection sensitivity *in silico*, for example as required in cancer genomics, there are two main approaches. The first is simply the lowering of thresholds for variant calling when sequencing tumour derived DNA. This approach however will perform poorly in terms of specificity. Alternatively, paired sequence derived from both tumour and germline material allow for comparative calling between the samples in order to discriminate between germline and somatically acquired genotypes, including small variants and CNVs. Both of these approaches require alternative software, such as *VarScan 2*[157,171].

The properties of somatic variants in terms of allelic ratios presents opportunities as well as challenges. Loss of heterozygosity (LOH) post-conception, such as that resulting from a chromosome arm deletion, will result in a tract of variants exhibiting skewed allelic ratios (as measured by the B allele frequency (BAF), i.e. the proportion of reads harbouring the alternate allele). These regions can be identified using specialist tools such as *BAFsegmentation*[172]. In this, the BAF is transformed to the mirrored BAF

(mBAF) using:

$$\text{mBAF} = |\text{BAF} - 0.5| + 0.5 \tag{4.1}$$

Regions with recurrent deviation from mBAF = 0.5 beyond the pre-defined cutoff are ultimately segmented using circular binary segmentation (CBS) to identify continuous regions which are likely to have LOH[172,173].

4.3.3 Annotation of called variants in WES data

There are several annotations that are required for the downstream interrogation of genotypic data. At the most basic, annotation as regards the genes and transcripts in which the variant is situated, as well as resultant changes to the gene product are essential. However, several other information sources of information are also required for downstream analysis.

4.3.3.1 Allele frequencies

A key annotation type is the AF of a variant. These data are derived from the large scale sequencing/genotyping consortia discussed in Chapter 2, such as the 1000 Genomes[12] and HapMap[35] Projects. Where Mendelian disease is being investigated, particularly with a severe phenotype, a high AF for a variant will support its exclusion as a aetiological candidate. AFs will vary between populations dependant upon the level of historical isolation between populations. As such, it is essential that an ethnically matched data-source is used for association analyses, for Mendelian disease however it is worth having a broader panel of comparison. Similarly, some apparent genotypes, particularly erroneous calls, may be sequencing/analysis/batch specific, and thus a database of in-house samples will allow us to recognise systematic artefactual genotypes, at both a platform and batch level.

4.3.3.2 Conservation metrics

As an extension of the AF in human populations, one can use conservation across multiple species to investigate the possible deleteriousness of a variant, effectively utilising information from a far longer evolutionary history, enhancing the power of discrimination between variants. Many scores are available for this, including PhyloP[174] and GERP++[175]. However, all of these tools are imperfect in their predictive capacity, and a consensus approach is required for reliable prioritisation.

4.3.3.3 Physicochemical properties

A change of amino acid within a protein can have vastly differing consequences dependent upon the nature of the change. For instance, the change from a glycine (R-group: -H)

to a cysteine (R-group: -CH_2SH) will be likely to have a far greater effect on protein function compared than to an alanine (R-group: -CH_3). Additionally, proline, with its cyclic structure making it the only secondary amine amino acid, imposes severe constraints on the φ bond angle at that position, affecting protein folding[176]. Scores are available, based upon the physicochemical factors alone (such as the Grantham score[177]) or integrating this with sequence conservation (such as the SIFT[178] and PolyPhen-2[179] scores).

4.3.4 Filtering of genotypes for the identification of aetiological candidates

There are many possible stages for the filtering of genotypes. Firstly, where multiple related individuals have been sequenced, genotypes can be filtered based upon patterns of segregation, where phenotyping is clear. In the majority of cases variants will be expected to be present in all affected individuals, but not in the unaffected individuals. Furthermore, variants known to be seen at an appreciable frequency in population datasets will be excluded. For highly penetrant, severe, dominant Mendelian disease variants would be expected to be present a rate of $< 1\%$ in a healthy population, though this will also vary slightly with mode of inheritance. Prioritisation of variants by effect type will be useful; a rare frameshift or stopgain variant is more likely to cause disease than a synonymous variant. Finally, where candidate genes are known, these should be interrogated first, reducing both the analytical burden if the causal variant is seen within this subset, as well as reducing the likelihood of discovering clinical variation unrelated to the primary diagnosis of referral.

4.4 Quality metrics & QC of NGS data

There are several key quality considerations in the use of NGS data. Two common metrics (mean DOC and phred) are detailed below, as well as QC approaches that should be undertaken in the processing of NGS data throughout all stages of the analysis.

4.4.1 Phred

Within a read, bases will be of different qualities, due to various factors such as those inherent in the sequencing chemistries and starting DNA quality. Downstream analyses are required to consider this variable quality to allow for weighting in the determination of a consensus between reads for instance. During the HGP, *Phred* became the standard software for sequence determination from Sanger reads. On base calling, *Phred* assigns a quality score to each base, based upon factors such as the amplitude and resolution

of peaks in the electrophoretogram trace[180]. The quality score (itself termed phred) directly relates to the probability of a base call being erroneous:

$$\text{phred} = -10 \times log_{10}(E) \tag{4.2}$$

where E represents the probability of the base call being in error. Phred scores are a standard method of reporting error probabilities; while the background method for determining E will vary between platforms, the integer scores represent the same concept. A phred score of > 20 is considered a standard for 'good' quality of a read, corresponding to a 1% error rate (Table 4.2). It is worth noting that with NGS, one hopes to have multiple reads covering the same position, and thus the confidence in a consensus call for a position will be increased with increased coverage, allowing for cumulative phred scores for a position $\gg 100$.

Table 4.2: Error rates for a range of phred scores.

phred	Error rate (%)
3	50
10	10
20	1
30	0.1
40	0.01
100	0.0000000001

4.4.2 Depth of coverage

NGS reads have a higher error rate than Sanger reads. Due to the massively parallel nature, a multiplicity of reads spanning a region of interest can be readily produced. The number of reads aligning to a site in the reference genome is a key consideration during experimental design. Required DOC varies greatly dependent on the intended use of the data. For example, the 1000 Genomes Project utilises low mean DOC (2–6 X) WGS data in order to determine genotype calls for individuals by using a pooled approach that considers external genotyping data from both genotyping arrays and WES for the individual, as well as prior knowledge of AFs[12].

Due to the cumulative nature of the evidence for each read at the position, far greater DOC is required when dealing with individuals, particularly where sequencing is for clinical purposes, where accuracy is paramount[181]. Profiling of the mutational spectrum of cancer can require extremely high DOC due to the heterogenous nature of the polyclonal sample from which DNA would be sourced, including both cancerous cells and stromal cells, in varying proportions depending on the cancer type and stage. Reads will cover the various alleles present in sub-clones in proportion to the stoichiometry

present in the DNA sample, high DOC is therefore required to ensure that the rarer alleles at a position are observed, and distinguishable from sequencing errors[157].

Another related factor is the uniformity of coverage. Sequencing to a lower mean DOC is more likely to be viable knowing that there is a narrow distribution of DOC across the exome. This allows us confidence that the majority of regions of interest will have sufficient DOC to provide useful data. As such, the proportion of the genome, or defined subset of interest, that is covered to a certain threshold DOC is also often a more useful metric than the mean.

4.4.3 Confirmation of identity

Prior to the interrogation of data, it must be validated that the correct data-set is being interrogated for the correct individual. Approaches such as validation of gender based upon X-chromosome calls—with a significant deficiency in heterozygous genotypes in males due to their monosomy—and ethnicity utilising principal-components analysis, allowing comparison with individuals of known ethnicity, are two low-resolution methods that can prove informative[182]. Where multiple family members are sequenced, pairwise IBS should be checked to ensure that it is consistent with the reported relationship (see Table 1.1); this will also allow for the discovery of issues such as false-paternity, which would hinder variant filtering by segregation.

The ultimate validation of identity will be by comparison of the NGS data with external data such as SNP genotypes, and there are approaches available to do so[183,184]. A key advantage of this approach is that it allows comparison of samples at all stages though processing due to the intrisic nature of the markers, and also allows comparison with fresh blood from the individual in question to avoid all ambiguity if required (see Chapter 6).

4.4.4 Contamination checks

Even where the identity of the sample has been successfully validated, the inclusion of sequence data from exogenous DNA may affect the results. Possible sources may include from cross contamination between concurrently processed samples, as well as environmental contamination such as from bacterial DNA. Contamination can be assessed through interrogation of the alternate allele read-counts across variant loci. We would expect a trimodal distribution centred on 0%, 50% and 100% corresponding to homozygous reference, heterozygous and homozygous alternate genotypes respectively. Significant deviation from this pattern may indicate the presence of DNA from another individual, or somatically acquired variants. For contamination with non-human

sequence, exclusionary pre-alignments of raw data can be performed to remove reads that map to non-human sequence and not the human reference[185].

Genomics, and particularly NGS has the capacity to be of great utility in genetic research and clinical medicine. However, in reference to the quote from Wilhelm Johannsen with which this part was started, it is clear that genetics is still in this transition period. Where the stoichiometric balance of chemistry is now well defined and understood, the problem of missing heritability is still a challenge to be met in medical genetics, limiting the translational application currently. NGS should help meet this challenge, allowing for the identification of novel causes of disease at a greater rate.

Chapter 5

Aims

The overarching purpose of the work detailed herein is to present examples of the research utility of large scale genetic data. This is particularly relevant given the current rapid advances in NGS technologies, affording even greater data resources in the very near future. I will present a summary of the specific intentions of each primary research chapter below.

5.1 Part II - Application of NGS to Diagnostics

5.1.1 Chapter 6 - Sample tracking in WES studies

In this chapter, a novel tool for the tracking of DNA samples from an individual throughout an exome sequencing workflow is presented, given the need for increased robustness in clinical WES. The major intended properties of this tool were that it would be: robust, even with large numbers of samples sequenced; cost efficient and; be effective across populations. This tool, in the format of a SNP fingerprinting panel, also required stringent validation in both existing NGS data, as well as in theoretical simulations to allow for the inevitable larger future number of samples sequenced.

This Chapter was predominantly my own work, with significant input from Gaia Andreoletti, Chris Mattocks and Prof. Sarah Ennis.

5.1.2 Chapter 7 - Identification of cryptic variants

Chapter 7 is intended to illustrate a selection of cases, which were interrogated in partnership with clinical colleagues, in which the identification of the aetiological variants in patients has been particularly challenging. The hope is that these cases illustrate broader paradigms in terms of the challenges facing those who wish to best interrogate clinically applied exomes. The cases detailed will also highlight some of the

current deficiencies in WES/NGS technologies which will need to be resolved in order for NGS to have the best possible application to human samples.

For this work, my contribution was mainly data analysis and interpretation, with clinical phenotyping, recruitment and interpretation performed by clinical colleagues, including Dr. Ananth Ramakrishnan, Eleanor Seaby, Dr. Rodney Gilbert and Prof. Ignacio Briceño.

5.1.3 Chapter 8 - Cleft lip WES

Chapter 8 reports the application of WES to 10 families with cleft lip/palate phenotypes from the Bogotá region of Colombia. These patients display a mix of syndromic and non-sydromic presentations. Evaluation of the WES data is with a view to identifying the aetiology in the ten families, whilst also highlighting the genetic differences between syndromic and non-syndromic presentations.

For this work, my contribution was mainly data analysis and interpretation, with clinical phenotyping, recruitment and interpretation performed by clinical colleagues led by Prof. Ignacio Briceño, and Prof. Andrew Collins contributing to variant interpretation.

5.1.4 Chapter 9 - Gene panels in kidney disease

In Chapter 9, targeted NGS sequencing using a custom gene panel is applied to a cohort of 83 patients in the Wessex region with focal segmental glomerulosclerosis. The aim of this work was to evaluate the mutational spectrum in these patients, and further interrogate the presence of genotype/phenotype correlations in subsets of the cohort.

For this work, my contribution was mainly data analysis and interpretation, with clinical phenotyping, recruitment and interpretation performed by Dr. Christine Gast.

5.2 Part III - Mapping of Linkage Disequilibrium

5.2.1 Chapter 10 - Characterisation of WGS LD maps

Chapter 10 details the work done assessing the utility of WGS data for LD map generation. The aim was to validate that WGS is a viable source of genotypic data for LD map generation, as well as being computationally feasible. Furthermore, an assessment of gains attributable to the increase in genotype density, and, the corollary of this, the specific deficiencies of array-based genotyping data for LD map generation.

This work was predominantly my own, with significant input from Prof. Andrew Collins, Prof. Sarah Ennis and Dr. Rick Tearle.

5.2.2 Chapter 11 - LD in commercial chickens

The work in Chapter 11 details the generation of LD maps from array-based genotyping of several lines of commercial chickens. This was with a view to quantify the degree of concordance in LD patterns between distinct populations, as well as investigating the features underlying patterns of LD, i.e. primarily sequence features associated with recombination hotspots.

This Chapter was predominantly my own work, with significant input from Prof. Andrew Collins, Prof. Sarah Ennis, Dr. Almas Gheyas and Prof. David Burt.

5.3 A note on terminologies

As in all fields, geneticists utilise some terms with scant consistency, particularly between specialities. A few cases will be discussed here briefly in order to ensure clarity in the later chapters. Some situations require the use of less standard terminologies to ensure that there is no ambiguity due to the broad scope of this work. Choice of terminology in this work is not intended to suggest that these terms should be used across genetics, it is merely a choice for clarity.

5.3.1 Allele frequencies

The minor allele frequency (MAF) of a variant is defined as the second most common allele observed across the sample that the MAF is being defined in. In some situations, this can lead to confusing and unintuitive presentation of results. For instance, for some loci, the minor allele is the reference allele, as the reference genome does not accurately represent the variation across populations, though advancements in this are being made[186]. Further to this, the minor allele may be different for populations.

An alternative to the MAF metric is the alternate-allele frequency (AF)[164]. This can be defined as the most common non-reference allele at a position. The advantage of the AF is that the reference allele is constant between populations, removing some ambiguity. It should be noted that the most common alternative allele may still be different between populations, though this will mostly be the case in the case of highly rare variants[12].

For the purposes of this work, the term AF will be used in preference to MAF. The exception to this is where it is truly the MAF that we require for analyses, for instance in Part III. The verbose term allele frequency will be used when referring to a specific alternative allele, which will be specified.

Another usage of allele frequencies is in describing the ratio of alleles within a sample where technologies which sample multiple copies of DNA are used, such as NGS; this is discussed in more detail in Chapter 4. For these situations, we will borrow the term B allele frequency (BAF) from microarray analysis[172]. This will refer to the frequency of the most common non-reference allele within the read data of an individual at a position.

5.3.2 Genetic variants

In medical genetics particularly, there is a confusion of terms regarding the description of deviations from the reference genome. In the medical literature, these are often referred to as mutations, particularly where they are expected to cause disease. This is at odds with population genetics, where the term variant is utilised. In this work we will use the term variant in preference, aside from where the acquisition of the variant by the process of mutation has been observed. In cases such as somatically acquired and *de novo* variants, these will be referred to as mutations.

Part II

Application of NGS to Diagnostics

Chapter 6

Post Hoc Sample Tracking in Whole-exome Sequencing Studies

6.1 Background

The high start-up investment required for in-house WES is currently prohibitive to many groups so sample preparation and/or sequencing is commonly outsourced. This transference of sample custody, combined with the complex sample preparation workflow, makes sample mix-ups possible, and difficult to detect. In both clinical and research contexts, ensuring provenance of data is essential to allow the accurate assignment of clinical details to sequence data. It is possible that samples may be misidentified at any stage of the analytical process, both *in vitro* and *in silico*. Therefore, sample tracking must be contiguous throughout both data generation and analysis. Consequent to sample mix-ups in a research setting, erroneous data and sample matching may result in a loss of power for identification of causal variants[187]. In a clinical setting, these mix-ups may instead lead to delayed or inaccurate reporting of results to patients. Whilst good practice in the handling of samples and increased laboratory automation minimises potential for error, additional checkpoints are still required to support QC[188]. A method for the *post hoc* confirmation of sample identity is therefore highly desirable.

Genetic sample identification methods have an advantage over alternative sample management systems in that the genetic 'label' is intrinsic to the biological sample itself, removing the possibility of manual labelling errors. SNPs are increasingly utilised for DNA-based identification of human samples, with several benefits compared to standard forensic methods, such as amenability for highly degraded samples[189–191]. Existing SNP panels for human forensic identification and commercial SNP panels for sample identification, such as the iPLEX Sample ID Plus panel (Sequenom, San Diego, CA, USA), utilise pan-genome SNPs, the majority of which are non-exonic, and are therefore not useful for WES studies, as the majority of markers will not lie within

the enriched regions of the genome. In addition to existing SNP panels, short tandem repeat markers, as used in standard forensic identification procedures, can be used for genetic sample tracking. However markers applied are again frequently outside exomic regions and, if captured, will be prone to erroneous NGS genotyping using standard pipelines due to the repetitive nature of the markers[190,192].

Several methods for genetic tracking of human biological samples have been previously described, some of which are application specific, such as for transcriptome microarray studies[187,193,194]. Although software for the validation of NGS (including WES) sample identity, such as *verifyBamID* is available, for the detection of sample misidentifications external array-based genotypes of the samples are required, without which only contamination of the samples can be assessed[183].

Due to the lack of an existing tool for the identification of sample mix-ups without the availability of array genotypes, we aimed to formulate a cost-effective panel of a small number of SNPs. Here we describe an optimised panel of SNPs for which WES data are typically informative, the genotypic profile of which can be utilised to extract intrinsic identifiers from human genomic DNA. These SNP profiles have high discriminatory power, even in large datasets. The profile derived from this panel can be compared to an independently genotyped profile for the same individual, allowing accurate validation of data and sample pairings, at a modest cost per sample.

6.2 Methods

6.2.1 Panel selection

6.2.1.1 Candidate SNP identification

Regions of overlap between three current commonly used whole-exome enrichment kits, (namely Agilent SureSelect Human All Exon V4, Illumina TruSeq Exome Enrichment and Nimblegen SeqCap EZ Human Exome Library V3.0 kits), and common SNPs (as contained in dbSNP 137[65]), were established using BEDTools[195]. SNPs were further filtered for inclusion based upon their presence in genes targeted by the Illumina TruSight Exome kit, which targets only genes of clinical interest.

Primary candidate selection criteria required SNPs to:

1. Represent bi-allelic substitutions, excluding substitutions of complementary bases, that is, A↔T and G↔C transversions;

2. Be technically amenable to both accurate WES and orthogonal genotyping, that is, not present in large-scale genomic repeats[196], or homopolymeric tracts of ≥ 5 bp, GC content for the flanking 250 bp was restricted to a range of between 40% and 55% and no other variant within 50 bp with an AF ≥ 0.01 was permitted;

3. Conform to desirable HapMap Phase 3 AFs across several populations, explicitly AFs of between 0.2 and 0.8 in: CEPH (Utah residents with ancestry from northern and western Europe; CEU), Japanese in Tokyo, Japan (JPT), Han Chinese in Beijing, China (CHB) and Yoruba in Ibidan, Nigeria (YRI)[35] and;

4. Not alter the primary sequence of the encoded protein or have an associated OMIM record[17].

6.2.1.2 Candidate SNP selection

Following primary candidate identification steps, SNPs were further optimised by the following requirements:

1. Be located at least 10 bp from intron–exon boundaries to minimise the likelihood if involvement in splicing processes;

2. Not be situated in regions with a high sequence similarity to non-target regions, that is, no non-target BLAT score > 100[197], as this could result in nonspecific genotyping and;

3. Be outside of linkage disequilibrium with all other selected SNPs.

Finally, candidate SNPs were prioritised for inclusion in the panel by proximity of the AFs to 0.5, across HapMap populations, in order to maximise discriminatory power.

6.2.2 Validation & application

6.2.2.1 WES coverage

A set of 91 in-house exome samples was evaluated for depth of sequence coverage for the candidate SNPs, with a requirement that no samples had < 10 reads covering the SNP. Exome capture was performed using Agilent SureSelect Human All Exon V3 ($n = 22$) and V4 ($n = 55$), Illumina TruSeq Exome Enrichment ($n = 9$) and Nimblegen SeqCap EZ Human Exome Library V3.0 ($n = 5$). Exome enrichment, sequencing and *in silico* analysis of samples was performed as previously described[146,198].

6.2.2.2 Publicly available data

The power of sample resolution for the panel was validated using NGS derived genotype data from phase 1 of the 1000 Genomes Project (n = 1,092 WES samples)[12] and the UK10K project (n = 2,688; 256 of which are WES samples, the remaining 2,432 are low coverage imputed whole-genome data)[59]. Genotypes were extracted from VCF files using custom scripts and *Tabix*[199]. Quantification of mismatches between samples was performed using *MEGA5*[200].

6.2.2.3 Simulated data

Estimates for the true probability of repeat profiles were determined using a Monte Carlo simulation approach. Simulated datasets were generated by taking the individual population AF for each SNP as input and defining numeric boundaries in accordance with the expected proportions of genotypes under Hardy-Weinberg equilibrium. A pseudo-random number generating function was then used to assign a genotypic state for each SNP within these boundaries, outputting a concatenate of genotypes, and repeating until the desired dataset size is populated. This was implemented in the custom Perl script `generate_fingerprint.pl` (Appendix A.1), with the output passed to a wrapping shell script, outputting a count of the unique genotype concatenates within the dataset. We performed 20,000 bootstrapping pseudoreplicates of dataset generation in all cases.

6.2.2.4 Calculation of power

Due to the computational intensity and non-empirical nature of the Monte Carlo simulation, especially for large simulated datasets, a mathematical method for approximation was attempted. To perform an approximation allowing for variable likelihoods for each genotype profile, we used:

$$C \approx \frac{qn^2}{2} \tag{6.1}$$

where C is the likelihood of a collision within the dataset, and n is the number of samples within the dataset and q is the probability of a collision between two samples. The value for q is calculated as:

$$q = \sum_{i=1}^{O} r_i^2 \tag{6.2}$$

where i refers to a possible profile, and O the number of possible profiles (3^{24}, equal to 282,429,536,481 for the described panel), and r is the probability of a sample being assigned profile i; the probability r can be readily calculated from the AF data. Calculation of q is required once for each population for the panel, and can then be

utilised for all sample sizes of interest. Implementation of the calculation of q was attempted in custom Perl scripts.

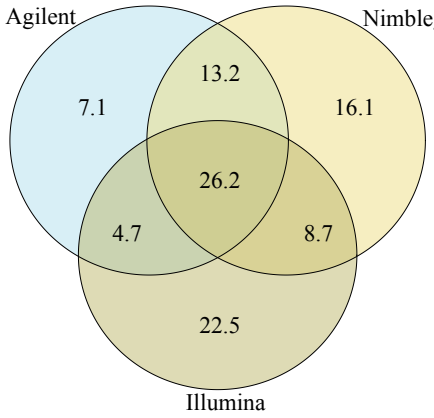
6.2.2.5 Application of panel

We applied the panel to a batch of 48 samples exome sequenced by an external service provider, for which orthogonal genotypes were obtained concurrently through an independent genotyping provider using KASP genotyping (LGC Genomics, Hoddeston, UK). Following plating of DNA samples for dispatch, a replica plate was made directly from the primary plate, to be dispatched for the orthogonal genotyping. Genotypes derived from exome data and orthogonal genotyping assays were compared using *PLINK* v1.07[182] and custom Perl and shell scripts.

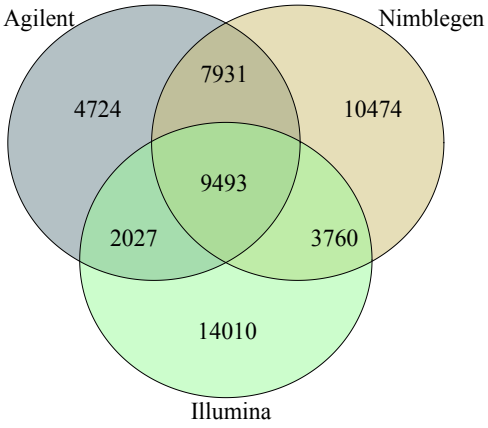
6.3 Results

6.3.1 Panel selection

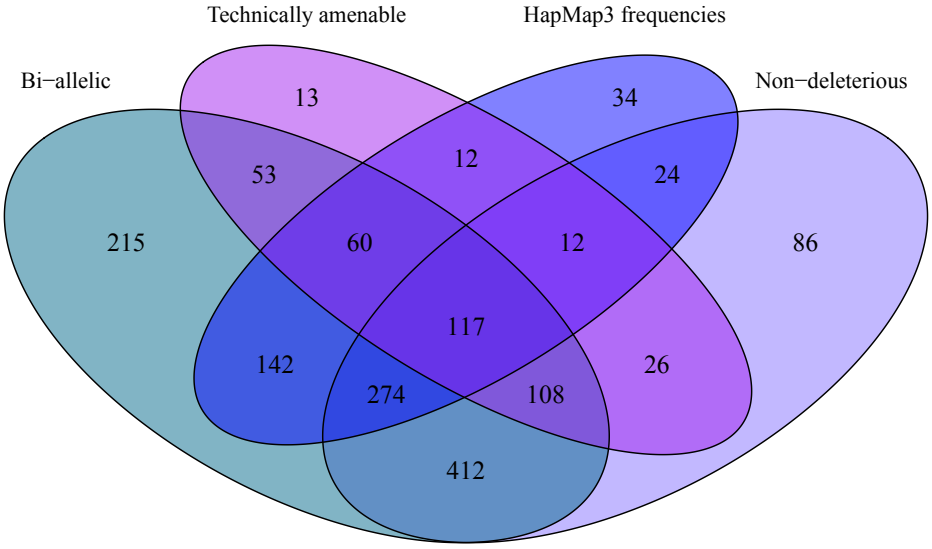
In total, 26.2 Mbp of genome sequence was found to overlap all three commonly applied whole exome capture kits, containing 9,493 common SNPs. Of these, 1,662 SNPs are additionally covered by the Illumina TruSight Exome kit. Within this subset, following the filtering for all primary candidate criteria, 117 candidate SNPs were identified (Figure 6.1; Table B.1).



(a) Mbp overlap between kits



(b) SNPs within kit overlap



(c) Properties of 1,662 SNPs covered by all kits

Figure 6.1: Venn diagrams showing commonality of targeting between capture kits (a,b) and properties of encompassed SNPs (c). Overlap between exome capture kits is presented in Mbp (a) and number of SNPs with an AF ≥ 0.3 (b). Agilent - SureSelect Human All Exon V4; Illumina - TruSeq Exome Enrichment; Nimblegen - SeqCap EZ Human Exome Library V3.0. For a subset of SNPs present in both the intersection of the three kits shown, and the Illumina TruSight Exome kit, a breakdown of fulfilment of the four classes of candidate filtering criteria is shown (c) (see the main text for details of filtering criteria). 117 SNPs exhibited all desired characteristics; 74 SNPs exhibited none of the desired characteristics.

From the 117 available SNPs, an optimised panel of 24 SNPs was selected (Table 6.1). Within the set of 91 in-house WES samples, all 24 SNPs were sequenced at sufficient read-depth for accurate genotype calling, across all capture kits.

Table 6.1: Optimised panel of identifying SNPs.

					HapMap Phase 3 AF			
Chr	**Position**[a]	**rsID**	**Gene**	**Alleles**	**CEU**	**CHB**	**JPT**	**YRI**
1	179520506	rs1410592	*NPHS2*	A/G	0.59	0.62	0.54	0.53
1	67861520	rs2229546	*IL12RB2*	A/C	0.64	0.36	0.44	0.58
2	169789016	rs497692	*ABCB11*	A/G[b]	0.55	0.65	0.51	0.22
2	227896976	rs10203363	*COL4A4*	C/T	0.46	0.44	0.36	0.57
3	4403767	rs2819561	*SUMF1*	C/T[b]	0.56	0.73	0.73	0.72
4	5749904	rs4688963	*EVC*	A/G[b]	0.33	0.65	0.67	0.52
5	82834630	rs309557	*VCAN*	A/G[b]	0.49	0.34	0.52	0.50
6	146755140	rs2942	*GRM1*	A/G	0.54	0.49	0.55	0.47
7	48450157	rs17548783	*ABCA13*	C/T	0.46	0.72	0.53	0.48
8	94935937	rs4735258	*PDP1*	C/T	0.40	0.64	0.66	0.46
9	100190780	rs1381532	*TDRD7*	C/T[b]	0.48	0.59	0.50	0.58
10	100219314	rs10883099	*HPSE2*	A/G	0.52	0.52	0.53	0.62
11	16133413	rs4617548	*SOX6*	A/G	0.52	0.65	0.61	0.51
12	993930	rs7300444	*WNK1*	C/T	0.46	0.55	0.48	0.28
13	39433606	rs9532292	*FREM2*	A/G	0.29	0.41	0.44	0.54
14	50769717	rs2297995	*L2HGDH*	A/G	0.55	0.65	0.67	0.59
15	34528948	rs4577050	*SLC12A6*	A/G	0.68	0.75	0.63	0.32
16	70303580	rs2070203	*AARS*	C/T[b]	0.53	0.28	0.51	0.49
17	71197748	rs1037256	*COG1*	A/G	0.50	0.67	0.65	0.56
18	21413869	rs9962023	*LAMA3*	C/T	0.67	0.81[c]	0.75	0.51
19	10267077	rs2228611	*DNMT1*	A/G[b]	0.47	0.73	0.56	0.48
20	6100088	rs10373	*FERMT1*	C/T[b]	0.54	0.31	0.35	0.58
21	44323590	rs4148973	*NDUFV3*	G/T	0.65	0.33	0.38	0.73
22	21141300	rs4675	*SERPIND1*	C/T	0.46	0.62	0.51	0.57

[a]Position as defined in genome reference assembly GRCh37 (hg19).

[b]SNP is defined on the negative strand.

[c]AF marginally outside target range for candidate selection. Selected due to paucity of candidates on chromosome 18.

6.3.2 Validation & application

6.3.2.1 Publicly available data

1000 Genomes Project

The 24 biallelic SNPs afford 48 points of allelic comparison. Testing the optimised panel in the 1000 Genomes Project data ($n = 1{,}092$)[12], an average of 18.0 (SD = 3.3) allelic differences between all pairwise combinations was observed, with a range of 3–34. As such, there will be, on average, 18 differential alleles between any two samples, enabling discrimination.

UK10K Project

On addition of the UK10K data (n = 2,688) to the 1000 Genomes Project data (n_Σ = 3,780), there remained an average of 17.8 allele mismatches across the profiles. Eighteen UK10K sample pairs produced duplicate profiles. On investigation of these pairs, they were found to share > 98% genotypic concordance across an extended panel of 1,662 SNPs in all cases compared to an average of 42%, with a range of 27–77% for all 18 sample pairs with unique SNP profiles (Figure 6.2). As such, these pairs represent extreme outliers, and are presumed to be derived from genetically identical biological samples, either from the same individual or monozygotic twins, and were therefore excluded from the mismatch average. In several cases, sample data producing concordant profiles bore consecutive sample designations.

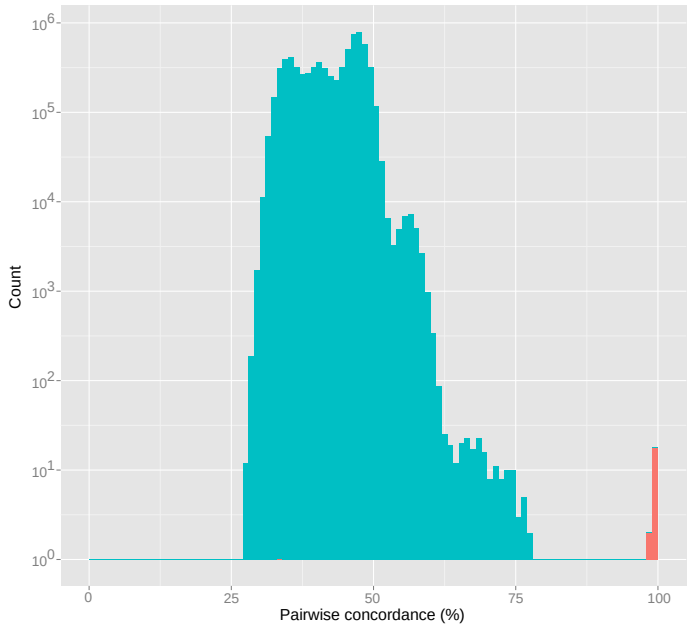


Figure 6.2: Distribution of pairwise genotype concordance between samples. Pairs resulting in duplicate SNP profiles (n = 18; red) and pairs between samples with unique SNP profiles (n = 7,142,293; blue) within the combined dataset of 3,780 samples are shown. Concordance across the 1,662 SNPs detailed in Figure 6.1c was evaluated. All pairs resulting in duplicate profiles have > 98% concordance, well separated from the distribution of samples with unique profiles. Note the logarithmic scale.

6.3.2.2 Simulated data

The discriminatory power of the panel was evaluated by Monte Carlo simulation. First, we evaluated the time taken for the analyses to run for a range of dataset sizes to confirm that the approach was computationally feasible. 50 pseudoreplicates of CEU 1000 Genomes Phase 1 AF data based simulation for a range of dataset sizes was performed and the CPU run-time was recorded, allowing for extrapolation to estimate the time required to obtain the desired 20,000 pseudoreplicates (Table 6.2). Monte Carlo simulation runtime increases approximately linearly with increased simulated

dataset size. Based upon these data we parallelised the simulations across multiple CPUs, allowing for the large CPU-time required to be completed in real-time inversely proportional to the number of CPUs applied. We also limited our simulated dataset sizes to 102,400 to maintain reasonable computational run-times.

Table 6.2: Time taken for simulation of collision frequency for varying dataset sizes

Size	50 replicates (s)	20,000 replicates (m)
100	0.45	3.02
200	0.62	4.15
400	0.99	6.62
800	1.73	11.51
1,600	3.23	21.52
3,200	6.28	41.89
6,400	12.52	83.43
12,800	25.11	167.37
25,600	50.96	339.72
51,200	102.88	685.85
102,400	210.15	1,401.02
204,800	429.04	2,860.24

We simulated datasets of 10,000 individuals, that conformed to AF distributions for investigated HapMap populations (CEU, CHB, JPT and YRI), 1000 Genomes Project pilot average[58], as well as for a hypothetical perfect allele distribution (AF = 0.5 for all SNPs) (Table 6.3). In all simulated populations, $< 2.5\%$ of simulated datasets of 10,000 contained any repeat SNP profiles (henceforth termed 'collisions'). This translates approximately into less than 1 in every 40 independent datasets of 10,000 individuals containing a single matching pair of profiles.

Table 6.3: Profile collisions per simulated dataset of 10,000 individuals with population AFs.

AF Source	Average collisions per dataset (± SD)
1000 Genomes average	0.0039 (0.062)
HapMap Phase 3:	
CEU	0.0064 (0.079)
CHB	0.0239 (0.154)
JPT	0.0082 (0.086)
YRI	0.0076 (0.086)
Theoretical perfect[a]	0.0031 (0.056)

[a]All 24 SNPs assigned an AF of 0.5, which will give the most even trifurcation per SNP, and thus discriminatory power.

The effect of dataset size on the frequency of collisions was investigated for populations present in 1000 Genomes Project Phase 1 data[12]. An exponential increase in the

frequency of collisions was observed with increasing dataset size, though the panel continued to have high power for the discrimination of samples (Figure 6.3). For instance, were we to have 85,000 unrelated Southern Han Chinese (CHS) samples, (the worst performing 1000 Genomes population evaluated, due to the AF distribution for SNPs within this panel), we would expect the dataset to contain, on average, a single duplicate SNP profile.

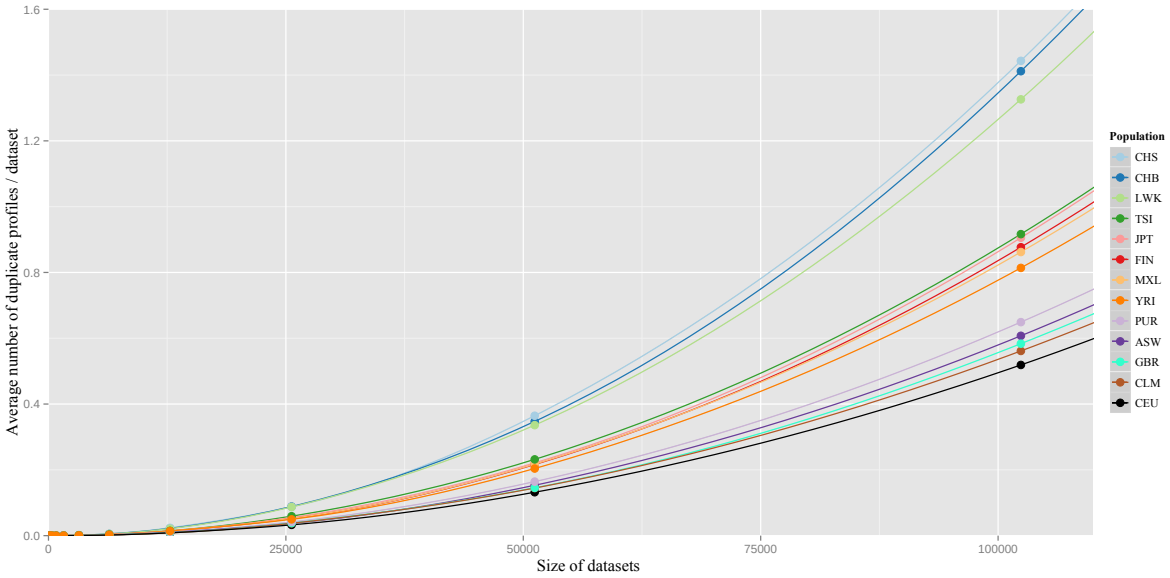


Figure 6.3: Relationship between sample size and incidence of repeat SNP profiles for 13 populations. Collision rate was simulated for multiple populations using custom scripts. An exponential increase in the probability of non-unique SNP profiles is observed with increase in sample sizes. In the case of the worst performing population, an average of 1 repeat profile per dataset of 85,000 would be expected. Allele frequencies are based on samples from the 1000 Genomes Phase 1 dataset[12]. Additional populations are Americans of African ancestry in Southwest USA (ASW), Columbians from Medellin, Colombia (CLM), Finnish in Finland (FIN), British in England and Scotland (GBR), Luhya in Webuye, Kenya (LWK), Mexican ancestry from Los Angeles, USA (MXL), Puerto Ricans from Puerto Rico (PUR) and Toscany in Italia (TSI).

In addition, total SNP absence, for example through technical failure of orthogonal genotyping, was modelled. For each SNP that entirely failed to provide data, a less than three-fold drop in discriminatory power was observed in all cases (data not shown). This suggests that our approach is robust against technical failure.

6.3.2.3 Calculation of power

Several Perl implementations for the calculation of q were attempted, in an effort to attain computational feasibility. Provisional testing indicated that calculation of q for a single population utilising a single 2 GHz CPU core would require ~42 days of continuous processing. Given the iterative nature of the calculation, parallelisation is programmatically challenging, and without parallelisation calculation is not practicable

within a reasonable time-frame. As such, attempts to empirically calculate C were abandoned at this stage.

6.3.2.4 Application of panel

Application of the SNP panel to our batch of 48 samples revealed a discrepancy between exome and orthogonal genotypes for two samples dispatched in adjacent wells, suggesting a reciprocal transposition (Figure 6.4). The occurrence of this error in the exome data was also supported by interrogation of X-chromosome heterozygosity to confirm sample gender. In addition to the identification of the switch, the panel allowed for expeditious resolution of the error, permitting the continued use of the data in downstream analyses.

Sample		rs1410592	rs2229546	rs497692	rs10203363	rs2819561	rs4688963	rs309557	rs2942	rs17548783	rs4735258	rs1381532	rs10883099	rs4617548	rs7300444	rs9532292	rs2297995	rs4577050	rs2070203	rs1037256	rs9962023	rs2228611	rs1103	rs4148973	rs4675
1	Exome	A A	G A	C C	C T	G G	T T	T C	G A	T C	C T	A A	A A	A G	C T	G A	G A	A A	G A	G G	T C	C C	G G	T G	T T
	Geno	N N	G A	C C	C T	G G	T T	T C	G A	T C	C T	A A	A A	A G	C T	G A	G A	A A	G A	G G	T C	C C	G G	T G	T T
2	Exome	C A	A A	C C	C T	A G	T T	T C	G A	T C	C C	A A	G G	G G	C T	A A	G G	G A	G A	G G	C C	C C	G G	G G	T T
	Geno	C A	G A	T C	C T	G G	T T	T C	A A	C C	C C	G G	A G	A G	C C	A A	A A	G A	A G	A C	C T	T C	A G	G G	C T
3	Exome	C A	G A	T C	C T	G G	T T	T C	A A	C C	C C	G G	A G	A G	C C	A A	A A	G A	A G	A C	C T	T C	A G	G G	C T
	Geno	C A	A A	C C	C T	A G	T T	T C	G A	T C	C C	A A	G G	G G	C T	A A	G G	G A	G A	G G	C C	C C	G G	G G	T T
4	Exome	C A	A A	T T	C T	G G	T T	T C	G A	T C	C T	A A	G G	A A	C T	A A	A A	A A	G A	G A	C C	T C	A A	G G	C T
	Geno	C A	A A	T T	C T	G G	T T	T C	G A	T C	C T	A A	G G	A A	C T	A A	A A	A A	G A	G A	C C	T C	A A	G G	C T

Figure 6.4: Exome derived and orthogonal genotypes for four samples, showing a sample-switch between 2 and 3. Informative markers for the resolution of this switch are highlighted in yellow.

6.4 Discussion

Validation of sample identity is essential in order to ensure data integrity and validity of conclusions drawn from data. We have described a powerful tool for the identification and validation of data provenance throughout the workflow of WES data collection and analysis. The power of discrimination, that is, the precision with which samples can be uniquely identifiable, is sufficient and robust for most projects on the current scale of up to 10,000 samples, with inbuilt redundancy of SNPs to protect against technical failures. In WES, the exome enrichment process provides the limiting step for the availability of data on SNPs for use in sample identification. As such, this panel will also be of utility for whole-genome sequencing data, where there is no such limitation on SNP coverage. This will be beneficial where there are mixed datasets of both whole-genome sequence and WES data.

NGS is now developing as the diagnostic methodology of choice across a range of applications, including mutation scanning in targeted gene panels and WES for congenital disorders, as well as high depth analysis for tumour profiling. Whilst the service model for delivery of these tests is not fully resolved at this stage, there will certainly be economic arguments for centralising certain tests. This will have the effect

of increasing the throughput requirements as well as physically moving samples between labs. Both of these factors will increase the opportunity for sample misidentification.

Even for testing within a single lab, the use of inherent sample and data identification methods, as described in this study, seems a robust approach to fulfil the regulatory requirement for providing a full audit trail and ensuring data provenance[181,201]. The SNP panel presented here is immediately usable across all commonly used exome capture kits, and would be equally applicable to any gene panel by incorporating, or 'spiking', the SNP regions into the custom capture kit at the design stage.

We have shown our panel to have a high discriminatory power across a diverse range of populations. The discriminatory power of the panel may be reduced for various reasons, such as geographically localised variation in AFs, and degradation of DNA samples, resulting in incomplete data. Additionally, the discriminatory power will be marginally reduced where many relatives are sequenced. In the case of highly consanguineous families, sample tracking methods such as barcoding will afford optimal certainty in these particular cases. Should concerns over insufficient discriminatory power arise, additional SNPs may be added to the panel from the existing list of candidates, also allowing the tailoring of an enhanced panel to the population(s) of interest, should this be desired. Nevertheless, we have demonstrated our panel to be sufficiently robust to withstand power reductions without loss of utility for most purposes. Simulation of panel power as implemented in `generate_fingerprint.pl` relies on the Perl `rand` function for pseudo-random number generation; there are known issues with this function providing poor randomisation performance[202], though this is unlikely to have influenced these results due to the low resolution binning of the random numbers, i.e. from 32-bit accuracy numbers to trinary genotype categories. While more random alternatives to this algorithm exist, it was decided to utilise the stock `rand` function due to its low computational intensity, a requirement to facilitate a large number of pseudoreplicates.

We have also presented a recent case in which use of this panel has allowed us to identify, confirm, and resolve a sample switch, highlighting the importance of using such a tool. Monetary cost will vary with the technology used for orthogonal genotyping and sample throughput. We have intentionally designed the panel to be platform nonspecific, allowing for the establishment of in-house assays using preferred genotyping methodology or outsourced where required. Our own chosen methodology has a list-price of approximately £10 GBP per sample, representing a small fraction of the cost of exome data generation; this will of course vary dependant upon chosen method and throughput.

Since the publication of our panel, there has been one further publication by Hu *et al.* elucidating a similar method, and SNP panel[203]. Hu *et al.* utilised information theory to identify the optimal set of SNPs for sample tagging. They also provide a tool, *SNPtagger*, for the generation of custom SNP panels where specific requirements exist. Ultimately, they describe their 30 SNP panel as having an average mismatch distance between simulated samples of 18, comparable to the average observed mismatch distance in our study of 17.8 across 3,780 actual samples.

The demand for the development of effective tools for bioinformatic analysis, data compression, mutation effect prediction and quality control is high. As such, we have formulated this panel of SNPs for the discrimination of human biological samples on the basis of data intrinsic to WES data derived from samples processed using common capture kits. Since the panels inception we have utilised it routinely in our routine sample analysis pipeline.

Following publication of the described final panel, the panel is now offered as a genotyping service by LGC Genomics, allowing for use of the panel by groups without laboratory facilities, as well as a pre-validated genotyping kit, to allow for ready incorporation into existing laboratory workflows, without the requirements of assay design and validation[184,204].

Chapter 7

Lessons Learned in the Identification of Cryptic Aetiological Variants in Whole Exome Sequencing

7.1 Background

Whole-exome sequencing (WES) has proven to be a powerful tool for the identification of aetiological variance, providing a cost effective means of leveraging the clinical diagnostic power of NGS[68–70]. As discussed in Chapter 4, WES involves the pre-selection of coding regions of gDNA by hybridisation with complementary baits, followed by NGS sequencing, providing a high, cost-efficient, diagnostic yield.

I have worked closely with several local clinicians, identifying patients for which WES may prove useful, and performing data analysis. In many cases, this has resulted in a successful molecular diagnosis for the patients in question, informing appropriate treatment and allowing for genetic counselling where this is desired by the patient. In some cases however, the apparent aetiological variants have been refractory to identification.

A genetic variant may be refractory to identification using WES methodologies (i.e. a 'cryptic variant') for a multitude of reasons. In this chapter I discuss several cases where variants which are thought to contribute to disease pathogenesis have been identified, though requiring (sometimes extensive) further analysis than purely the default pipeline detailed below. This chapter focusses on cryptic variants as other identified aetiological variants, while their identification is of clear benefit to the patient and may further understanding in the relevant clinical field, are nonetheless of limited interest from a

bioinformatic data analysis perspective. Categorisation of cryptic variants however should allow for the identification of weaknesses in standard analytical processes. Key weaknesses in the local pipeline used are identified, and approaches to resolving these issues are discussed in the conclusion. These represent key areas where continued research and method development will facilitate a diagnostic uplift upon the routine clinical application of NGS in healthcare.

7.2 Methods

7.2.1 *In vitro* sample processing

gDNA was isolated from either whole-blood by the salting-out method or spin-column preparation, or from stabilised saliva according to manufacturer protocol (Oragene Discover 250 kit, DNA Genotech, Ontario, Canada). Downstream sample processing steps from this stage were outsourced to an external service provider as detailed herein. In brief, isolated DNA was fragmented by ultrasonication and size selected to give a mean fragment size of 200 bp; whole-exome enrichment was performed using either the SureSelect Human All Exon V4 or V5 kit (Agilent, Santa Clara, CA, USA) according to manufacturer instructions prior to sequencing for 100 bp PE reads on either the HiSeq 2000 or 2500 platform (Illumina, San Diego, CA, USA). Where necessary, Sanger sequencing was performed following PCR of gDNA with standard methods, using the forward amplification primers for sequencing; primer design was performed using *Primer3Plus*[205].

7.2.2 *In silico* data processing

Data was analysed using the in-house Soton Mendelian V3.0 or V3.1 pipeline (collectively referred to as V3.x)[146,184,198,206]. Raw FASTQ reads were aligned to the reference genome GRCh37 (hg19) using *Novoalign MPI* v2.08.02[i][161] (see Figure 7.1 for pipeline overview). Following primary alignment, duplicate reads—reads which align originating at the same genomic position, and are thus presumed to be technical artefacts as opposed to true independent reads—were flagged using *Picard* v1.108[207]. Variant sites were called using the *SAMtools* v0.1.18 `mpileup` command on individual samples[ii][165]. All standard statistical analyses were performed in *R* v3.0.1 unless otherwise stated. Pedigrees were drawn using *Madeline 2.0*[208].

[i]Non-default parameters used were a presumed mean fragment length of 200 bp, SD 30 bp, gap-opening penalty of 65 and gap-extension penalty of 7 (these gap-penalties are the weighting against the opening and extension of gapped alignments as compared to the reference genome assembly, as seen in indels).

[ii]For calling, non-default parameters were to: only consider reads with a mapping quality of ≥ 20; perform extended base alignment quality computation; skip indel calling where DOC > 2000 and; require $\geq 5\%$ of reads to support an indel call.

Where it was desired to investigate somatic LOH in samples, *BAFsegmentation*[172] was applied to the exome data. *BAFsegmentation* assesses the deviation in the allelic ratios in heterozygotic loci (assessed by the mBAF), with contiguous regions segmented using CBS. VCF files were converted to *BAFsegmentation* input format using custom scripts; only variants with a read depth of ≥ 20 (this depth required in both samples for pairwise analysis) were considered in order to minimise stochastic noise in the BAF at very low read depths. Regions consistently exhibiting an mBAF of ≥ 0.6 were considered to be regions of LOH.

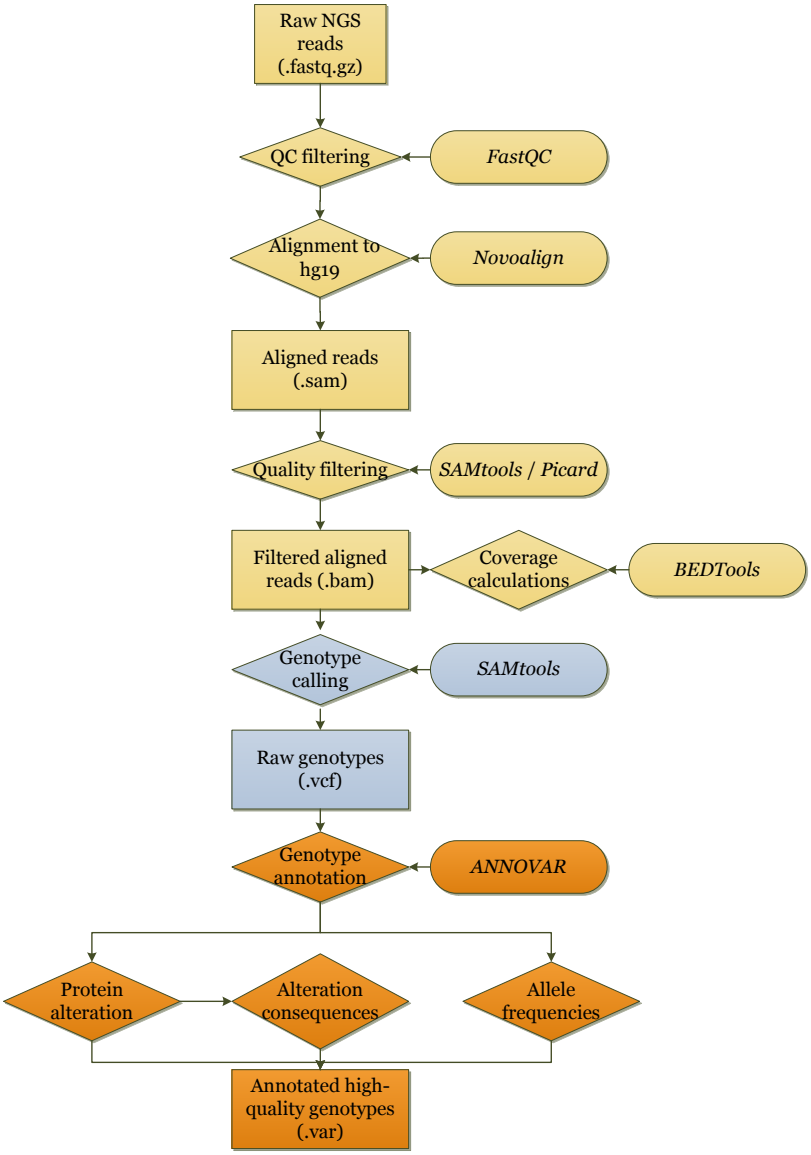


Figure 7.1: Overview of the Soton Mendelian V3.x analysis pipeline. Flowchart embodies the final section of Figure 4.2 as performed in-house. The analysis is aggregated into three major segments, read alignment (beige), genotype calling (blue) and annotation (orange). Early stages are computationally intensive, with hardware requirements and file sizes decreasing throughout the process due to the aggregation and filtering of data. Rectangular objects represent output files (file formats in parentheses), diamonds denote processes and ovoid objects, software used.

7.2.3 Annotation of called variants

Called variants were annotated using *ANNOVAR*[209] with: the AF in the 1000 Genomes Project Phase 1 dataset[12]; 5,400 individuals from the NHLBI Exome Sequencing Project (ESP)[210] and; ~250 individuals WES analysed in-house with the Soton Mendelian V3.x pipeline. Protein alterations were annotated based upon the RefSeq transcript databases, and with predictions of deleteriousness including GERP++[175], PolyPhen-2[179] and SIFT[178] where the variant is non-synonymous. Variants within 10 bp of intron–exon boundaries were annotated as having a putative involvement in splicing processes, and assessed using *MaxEntScan*[211] as required. A ΔMaxEnt score of ≥ 2.5 was considered indicative of a variant likely to alter splicing processes. All chromosomal positions are defined as in GRCh37 (hg19), and all SNP rsIDs are as contained in dbSNP build 139[65].

7.2.4 Filtering of annotated variants

Filtering of variants was informed by prior information, and as such is not consistent across all cases. As a flexible framework, variants were filtered out using the following exclusion criteria, specifically being:

1. Synonymous, with the exception of exonic variants located within 10 bp of an intron–exon boundary;
2. Present outside of coding regions (aside from if within 10 bp of an exon) of defined candidate genes, where this information was available;
3. Common in the ESP and/or 1000 Genomes Project datasets (AF ≥ 0.01 or ≥ 0.05 dependant upon disease frequency/severity/mode of inheritance);
4. Present at a zygosity inconsistent with the expected mode of inheritance, and;
5. Segregating inappropriately within family members, where multiple members have WES data available.

Furthermore, remaining variants were prioritised for investigation if they were: known clinical variants; novel; predicted to be protein truncating or missense and predicted to be deleterious.

7.2.5 Quality control

Raw FASTQ reads were subjected to standard quality checks using *FastQC* v0.10[212]. Following alignment, DOC statistics were compiled using *BEDTools* v2.17[195] and evaluated for mean DOC and the percentage of target regions covered to 1, 5, 10 and

20 X. The proportion of reads mapping both to the genome, and to the exome target was also calculated. Following genotype calling, the X-chromosome and autosomal heterozygosity was calculated, to confirm gender and check for evidence of contamination (see subection 4.4). An aliquot of DNA was contemporaneously dispatched for orthogonal genotyping for the validation of the identity of the final data (see Chapter 6)[184]. IBS was calculated between all samples dispatched together to validate relatedness of samples where this is expected, and to highlight cross-contamination between samples. *VerifyBamID*[183] was applied to the BAM files in order to assess whether the data harboured consistent deviations in BAF from the expectation, indicative of contamination with exogenous DNA.

7.3 Indels

7.3.1 Family A - Nager syndrome

7.3.1.1 Clinical presentation

The female proband of Family A received a putative diagnosis of Nager syndrome (MIM 154400) at age 9, with no remarkable family history known, though an extended family history was not available (Figure 7.2). Presentation included micrognathia (an undersized jaw), hypoplasia of the ear-canal and absent index fingers. Further detail is given in subsubsection 8.3.1.1. Given the sporadic nature of the case, it is likely that the aetiological variant arose *de novo*, though it is also possible that it is recessive if distantly related parents have formed the union. Nager syndrome has been previously reported to occur in similar sporadic cases, as well as in familial cases with both dominant and recessive modes of inheritance—including resulting from compound heterozygosity[213]. The cause of the majority of Nager syndrome cases has been recently identified to be mutations in *SF3B4*[214], with the aetiology of the non-*SF3B4* cases currently unresolved.

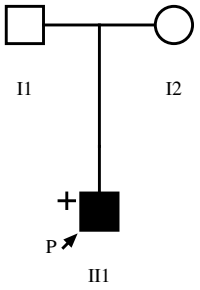
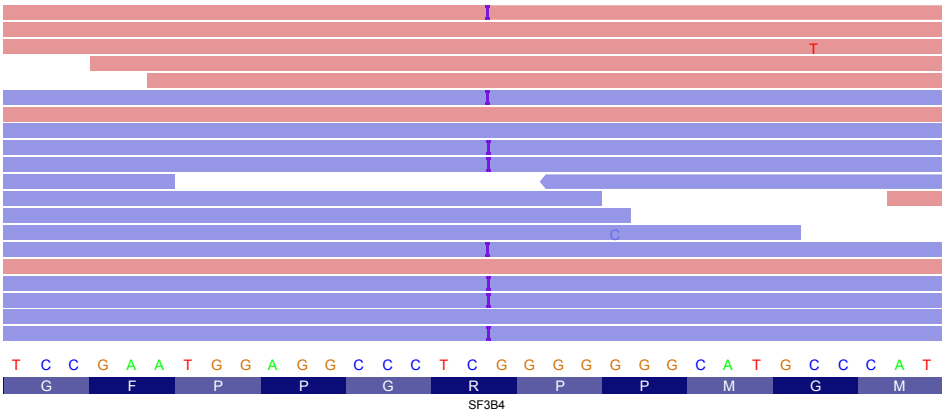


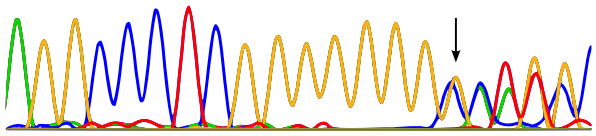
Figure 7.2: Pedigree showing inheritance of Nager syndrome in Family A. WES analysed individuals indicated by '+'.

7.3.1.2 Genetic analysis

The singleton proband was chosen for WES; the trio was not sequenced due to cost constraints. Returned WES data were of a good quality, passing standard checks (Table B.2). A mean DOC of 56.9 X was attained, with 25,139 variants called by our pipeline. Given the putative diagnosis of Nager syndrome, we first queried the variant calls for the candidate gene *SF3B4*[213,215]; no variants were called by our Soton Mendelian v3.1 pipeline. Given the strong candidacy, raw read data mapping to *SF3B4* was manually investigated in *IGV*[216], identifying a c.1060_1061insC:p.R354fs variant (transcript NM_005850, transcribed from the reverse strand). This variant was supported by $\frac{8}{19}$ reads at the position and was subsequently confirmed as *de novo* by myself using Sanger sequencing. It appears the indel was not called by our pipeline as only $\frac{1}{8}$ variant reads were mapped to the forward strand (Figure 7.3), leading to exclusion of the variant due to quality filtering at the variant calling stage. This effect has also been observed by other members of the group in other cases.



(a) NGS alignment



(b) Sanger electrophoretogram

Figure 7.3: *SF3B4*:p.R354fs as seen in alignment data (a) and Sanger electrophoretogram (b) for the proband of Family A. In (a), reads mapping to the forward stand are coloured in red, and mapping to the reverse strand in blue; mononucleotide insertions in reads are represented by purple bars. Note the bias in reads harbouring insertions toward reads mapping to the reverse strand. Colour coding in (b) is consistent with sequence colouring in (a), frameshift indicated with arrow. Primers used for amplification were F: 5'-TTCTCTTTCAGCCCTTGCCC-3' and R: 5'-ATGCTAAACTTCCTCCCCGC-3'. Figure (a) produced using *IGV*[216].

7.3.1.3 Discussion

SF3B4:p.R354fs has been previously reported to be a dominant cause of Nager syndrome, with patients testing positive for deleterious variants in this gene in 32 of 53 cases across two WES studies[213,215]. The remaining patients negative for *SF3B4* variants

may comprise a subset of Nager syndrome with a distinct aetiology, or alternatively testing may be negative due to limitations inherent in the WES approach[214]. The negative cases may carry cryptic loss of function mutations for example in transcription binding motifs or intronic splicing regulatory regions. *SF3B4* encodes for a component of the pre-mRNA splicing machinery; given the recent identification of aetiological variants in the gene, the mechanism of pathogenesis is not yet known.

The Soton Mendelian pipeline should be optimised to call this variant and other similar examples, as it would currently appear to be overly conservative in the calling of small indels where there is a significant strand bias. Variants exhibiting a strong strand bias are not called as these are more likely to result in false positive calls due to technical artefacts. This is a particulat issue for indel variants, and thus *SAMtools* is more stringent with these variants. This may be in the form of optimisation of parameters for the *SAMtools* `mpileup` command currently used for the calling of SNPs and small indels, or the incorporation of other more specialised software into the pipeline. Care must be taken however to minimise the increase in false positives. A known limitation of NGS data is a weakness in the calling of indels, with lower sensitivity than for the detection of SNPs, as well as higher false positive rate and greater between-software heterogeneity[217].

This case highlights the need for prior hypotheses in the interrogation of WES data. The presence of a strong prior hypothesis will allow for appropriate expenditure of time investigating genes, proportional to the perceived likelihood of that gene being clinically relevant. Where interrogation efforts are evenly distributed across the entire WES data-set, an increased type II error rate is likely to be seen; additionally, care must be taken in efforts to improve exome-wide sensitivity, as an increase in type I errors which may accompany this will hinder meaningful interrogation of the data. For a further example of the necessity of a targeted curatorial approach, see subsection 7.3.2.

7.3.2 Family B - Severe combined immunodeficiency with megaloblastic anaemia

7.3.2.1 Clinical presentation

Primary immunodeficiencies (PIDs) are a diverse class of disorders which are characterised by the lack of an effective immune response to pathogens, including opportunistic pathogens. PIDs are further defined by the cause of this immunodeficiency being endogenous, as opposed to being exogenous causes such as infection and chemotherapy[218]. Individual II2 (see Figure 7.4) presented at 4 months with *Pneumocystis jirovecii* pneumonia, and responded well to appropriate antifungal treatment. *P. jirovecii*, like

most fungi, is an opportunistic pathogen, infection with which is indicative of underlying immunodeficiency[219]. Haematological testing indicated significant lymphopoenia across all sub-sets; as such, a putative diagnosis of severe combined immunodeficiency (SCID) was made. Whilst further tests were ongoing, II2 was listed for a bone-marrow transplant, though continued to respond well to prophylactic antifungals.

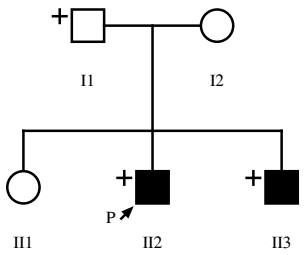


Figure 7.4: Pedigree showing inheritance of severe combined immunodeficiency in Family B. WES analysed individuals indicated by '+'. Note that the mother, I2, is of Asian descent.

Individual II3 was born following diagnosis of II2, and was therefore diagnosed at birth with a similar condition on the basis of haematology, going on to suffer septic arthritis of the hip at 9 months. Following the birth of II3, II2 developed megaloblastic anaemia, initially thought to be treatment related. Supplementation with folinic acid (a vitamer of folic acid, vitamin B_9) was successful in treating the anaemia. On folinic acid supplementation an increase in lymphocyte counts was observed, affording a partially reconstituted immune system in the brothers, obtaining low-normal lymphocyte counts. It became apparent that the megaloblastic anaemia was not treatment related, but likely due to a congenital metabolic deficiency.

7.3.2.2 Genetic analysis

Prior to these WES investigations, Sanger sequencing for several SCID candidate genes was carried out in a clinical genetics laboratory (namely *IL2RG*, *IL7R*, *JAK3*, *ADA*, *PNP*, and *RAG1/2*); results were negative for pathogenic variants in all cases. As such, it was decided to utilise WES to broaden the search for an aetiological candidate. The unaffected father and two affected brothers were exome sequenced. Returned WES data were of a good quality, passing most standard checks (Table B.2). A mean DOC of 70.6, 67.0 and 59.7 X was attained, with 23,488, 24,577 and 23,886 variants called by our pipeline for I1, II2 and II3 respectively. Data for II2 and II3 did however exhibit a significant excess of autosomal heterozygosity, as has been previously been observed with substantial contamination. This excess was determined to be due to the mixed-ethnicity of the brothers; no other evidence for contamination was observed. Given the strongly positive response to folinic acid supplementation primary candidate genes interrogated were those involved in the 'folic acid metabolic process' (GO:0046655[220]). 15 variants were called in these 14 genes.

Filtering of the 15 variants in folic acid metabolism genes, for those present in both brothers, excluding synonymous variants and those present with an AF ≥ 0.05 in the 1000 Genomes Project dataset, resulted in a single candidate remaining, a heterozygous novel *MTHFD1*:c.T152C:p.L51P substitution (transcript NM_005956). Analysis of segregation of p.L51P in Family B showed that the variant was inherited from the mother, I2. *MTHFD1* was clearly a strong aetiological candidate gene, though the single p.L51P SNP, also present in the healthy mother, would be unlikely to be pathogenic in isolation.

MTHFD1 was further investigated for variation in DOC across the gene, with a view to identifying any exonic deletions. In order to do this, the number of reads mapping to each exon of the gene was first enumerated for the three family members, as well as 13 unrelated controls sequenced contemporaneously. Given the maternally inherited p.L51P variant, it was hypothesised that a paternally inherited deletion may be the *trans* aetiological counterpart in the brothers. The raw count of reads was then normalised for each sample based upon the total number of reads aligned to the *MTHFD1* gene, and subsequently this was normalised by the mean normalised read count across the 13 control samples. A large deletion spanning the entire gene could be ruled out due to the heterozygous nature of the p.L51P call. A significant deficiency of coverage ($p = 0.00019$) within Family B across exon 13 of *MTHFD1* was observed (Figure 7.5), being indicative of a deletion of the exon.

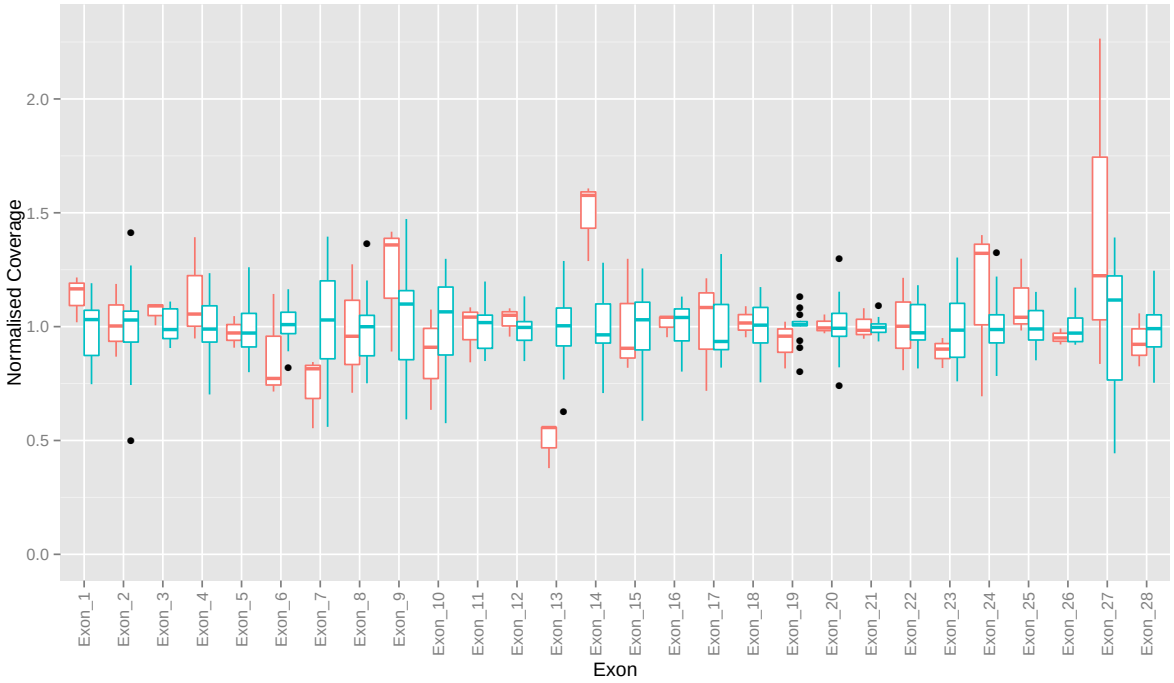


Figure 7.5: Normalised coverage across the *MTHFD1* gene in Family B (red) compared to 13 controls from the same sequencing batch (blue). A significant lack of coverage of exon 13 if observed for all WES analysed members of Family B ($p = 0.00019$, one-tailed t-test). In addition exon 14 appears to have an excess of reads in Family B samples ($p = 0.011$), though this significance level does not withstand Bonferroni correction.

As an aside, there is a clear inverse correlation between the standard deviation of the normalised coverage and the average coverage of an exon ($\rho = -0.90$, $p = 7.28 \times 10^{-7}$, Spearman's rank). This correlation highlights the requirement for high DOC data where CNVs are to be interrogated as a priority, in order to improve the power of detection; similarly, a greater number of reference samples is invaluable for maximal power of CNV detection. Out of several CNV and indel calling software applied to these data (namely *Pindel*, *SoftSearch* and *XHMM*[167–169]) with fully relaxed criteria, none have called this variant. However, the *ExomeDepth*[221] `TestCNV` function was able to call the deletion, with a Bayes factor of ~9. The `TestCNV` differs from the other software applied in that it allows the user to specify the region to test, limiting concerns overs false positives due to the greater *a priori* probability of the region harbouring a deletion; the standard exome-wide `CallCNVs` function stil fails to identify the deletion.

In addition to the reduced DOC for exon 13 in Family B, a split-pair of reads was observed across the region in the data derived from I1 and II3 (Figure 7.6). This pair was used to approximate breakpoint location, informing the primer design for confirmation. PCR and sequencing primers were placed outside this read-pair, ensuring that both primers would flank the deletion. Sanger sequencing confirmation successfully identified a 1,745 bp deletion in all WES analysed members of Family B, and confirmed segregation of the two *MTHFD1* variants with disease (Table 7.1).

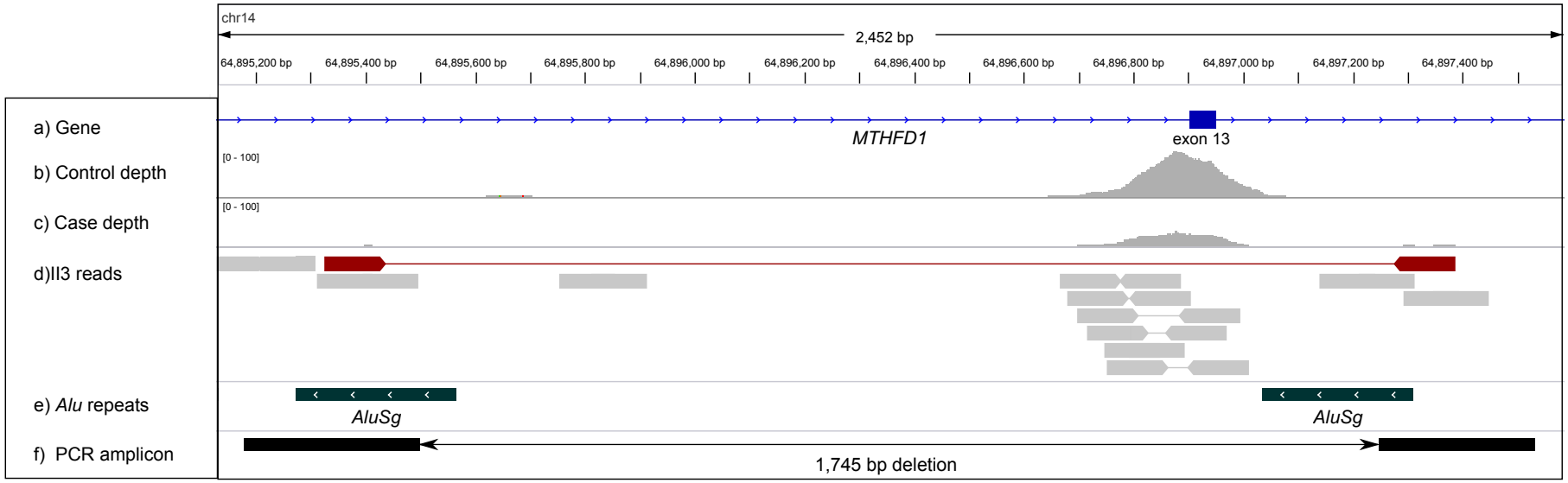


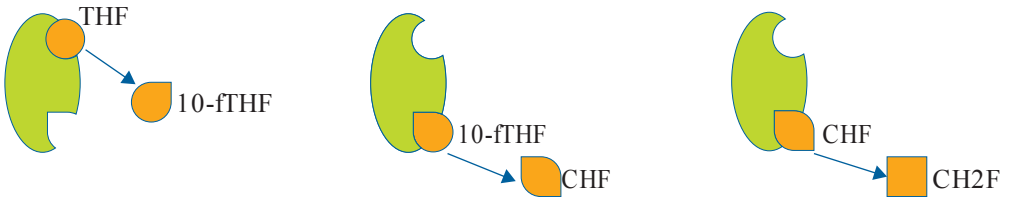
Figure 7.6: Supporting evidence for deletion of exon 13 of *MTHFD1* in Family B. DOC histograms are shown for a representative control (b) and case (II3; c) sample under the chromosomal coordinates and exonic structure of the gene. The pileup of reads is also shown for II3 (d), with an anomalously diverged read-pair evident (red). This is further indication of the exon 13 deletion (see Figure 4.4 for further explanation). Breakpoints of the deletion were confirmed by Sanger sequencing (f), and are shown to lie within the anomalous read-pair. The breakpoints lie within homologous positions in two *AluSg* repeats, suggesting a mutational mechanism (e). Figure modified from *IGV*[216] output.

Table 7.1: Segregation of *MTHFD1* variants with SCID in members of Family B.

		MTHFD1	
Individual	**SCID**	**L51P**	**Δ exon 13**
I1	-	-	+
I2	-	+	-
II2	+	+	+
II3	+	+	+

7.3.2.3 Discussion

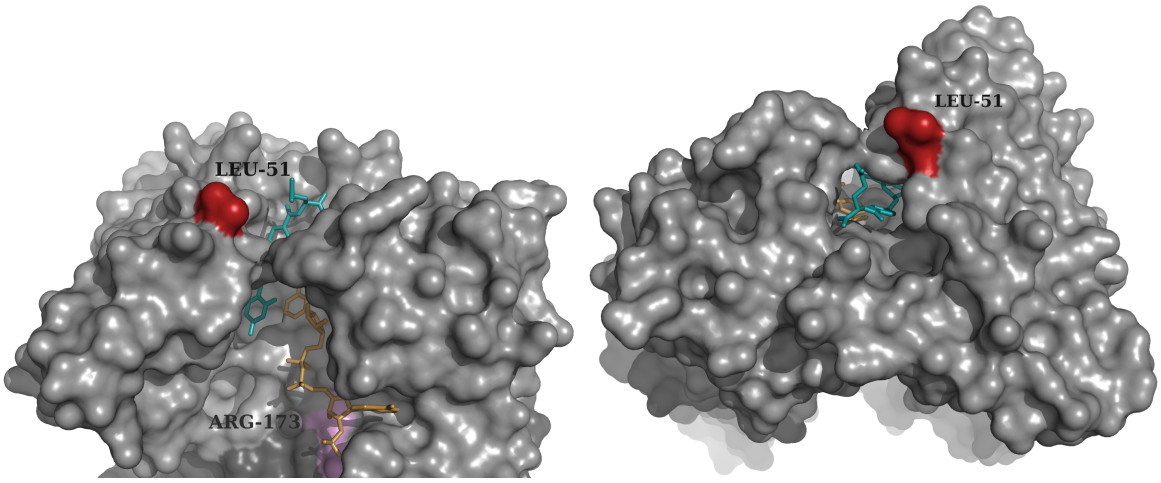
MTHFD1 encodes C-1-tetrahydrofolate synthase, cytoplasmic (C1-THF synthase). C1-THF synthase is a trifunctional enzyme with dehydrogenase (EC 1.5.1.5) and cyclohydrolase (EC 3.5.4.9) activities in one active site, and synthase activity (EC 6.3.4.3) in a second (Figure 7.7). These three activities occur in sequence and are required for the shuttling of folate metabolites between several key metabolic cycles in the mitochondria, nucleus and cytosol, including purine biosynthesis[222]. A review of the literature revealed that a similar paediatric SCID case was reported in 2011, also in an admixed pedigree, with compound heterozygosity for deleterious variants in *MTHFD1* and similar response to folate supplementation[150].



(a) Synthase **(b)** Cyclohydrolase **(c)** Dehydrogenase

Figure 7.7: Activities of the trifunctional C1-THF synthase enzyme. THF - tetrahydrofolate, 10-fTHF - 10-formylTHF, CHF - 5,10-methenylTHF, CH2F - 5,10-methyleneTHF. All reactions are reversible, with the second and third reactions requiring NADP cofactor. Taken from Scotti *et al.*, 2013[222]. Reprinted by permission from Wiley Periodicals Inc., © 2013.

The substitution of c.T152C is assigned a GERP++ score of 3.2, indicating a high level of phylogenetic sequence conservation, in agreement with the PolyPhen-2 score of 0.99. Due to the substitution of a proline residue within an α-helix, this is expected to disrupt the helix, deforming the proximal region[176]. Leu-51 is located near to the bifunctional active site of C1-THF synthase (Figure 7.8).



(a) C1-THF synthase top view **(b)** C1-THF synthase end view

Figure 7.8: Structure of C1-THF synthase showing substrate binding and residues mutated in Family B and previous case. Leu-51 (red), Pro-51 in our cases, can be seen to lie next to the binding pocket in the bifunctional dehydrogenase/cyclohydrolase active site for the substrate (teal). The previously reported p.R173C (pink) lies within the binding cleft for the NADP cofactor[150]. Both of these variants would as such be expected to alter the kinetics of this enzyme, most likely through the reduction of substrate/cofactor affinity. Data from 1.5 Å crystal structure, PDB ID: 1DIA[223].

It appears likely, as the breakpoints of the indel lie within two nearby homologous *Alu* repeats in equivalent positions (Figure 7.6), that the high sequence similarity led to the excision of the indel. A similar *Alu*-mediated excision has been previously reported as an aetiological mutational mechanism, possibly due to replication slippage[224]. The excision of exon 13 from the *MTHFD1* mRNA results in a premature stop codon due to a frameshift. As such, it would be indicated to undergo nonsense-mediated decay; this was confirmed to be the case using reverse-transcriptase PCR, revealing the prevalence of the c.T152C allele in the mRNA population *vs.* the wild-type allele at this site as present on the deletion allele. It is of note that the previously reported case also carried a compound heterozygote comprising c.[C517T];[727+1G>A], with the splice site variant again expected to induce nonsense-mediated decay of the transcript[150,225].

It must be noted that, despite the similarities in clinical phenotype between Family B and the previously reported case, there are some discrepancies, with the reported case having a more severe phenotype, including renal and neurological issues. The reported intellectual disability and seizures were not respondent to folate supplementation, though neurological issues are known in conditions where homocysteine (a metabolite that requires the products of C1-THF synthase for further processing) levels are increased[226]. Family B exhibits no signs of neurological defects, and did not have abnormally increased levels of homocysteine when off supplementation. There are two possible, non-mutually exclusive reasons for this: the p.L51P may not abrogate enzymatic activity to the same degree as p.R173C, and dietary levels of folate may have been higher in Family B than in

the previous case report. In an *Mthfd1*$^{+/-}$ mouse model, where it is the monofunctional synthase active site that is fully disrupted, a significant increase in homocysteine levels *vs. Mthfd1*$^{+/+}$ was only observed when the mice were given a folate deficient controlled diet. Furthermore, *Mthfd1*$^{-/-}$ was developmentally lethal, making it highly likely that these SNPs permit some residual activity[227,228].

Overall, it is clear, based on both a single previous case report and evidence from a murine model that the *MTHFD1* compound heterozygote is aetiological for the SCID and megaloblastic anaemia phenotypes in Family B; confirmation of this allows for the confident continued use of folinic acid supplementation in the affected brothers, with consideration no longer being given to bone marrow transplantation.

7.4 Loss of heterozygosity

7.4.1 Family C - Juvenile myelomonocytic leukaemia

7.4.1.1 Clinical presentation

The patient first presented at four months with a stroke, secondary to moyamoya, the constriction of arteries in the brain, and had no relevant family history (Figure 7.9). The patient was also developmentally delayed, initially attributed to the stroke. Aged two, he was referred to nephrology with marked thrombocytopaenia, proteinuria and hypertension. Low serum complement 3 suggested a perturbation in the alternative complement pathway consistent with atypical haemolytic uremic syndrome (aHUS). Renal electron microscopy confirmed endothelial cell separation from the glomerular basement membrane and he commenced eculizumab therapy for aHUS.

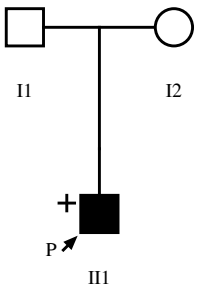


Figure 7.9: Pedigree showing inheritance of apparent aHUS syndrome in Family C. WES analysed individuals indicated by '+'.

Following the first tier genetic interrogation which is detailed below, it was not felt that these variants sufficiently explained the patient's phenotype so further phenotyping was performed. At this point it was determined that the patient had splenomegaly, dysmorphology (Figure 7.10) and continuing thromobocytopaenia, a symptom of aHUS

which would be expected to be resolved by eculizumab therapy. An additional 44 genes identified based upon these features were therefore interrogated[229].

Figure 7.10: Facial dysmorphology apparent in the patient from Family C. The photo, taken at age four shows low set ears, microcephaly and broad neck. The patient is in the bottom percentile for height and weight, and is developmentally delayed.

7.4.1.2 Genetic analysis

WES was performed for the proband with the intention of resolving the cause of the aHUS. Returned WES data were of a good quality, passing all standard checks (Table B.2). A mean DOC of 58.5 X was attained, with 24,955 variants called by our pipeline. 540 genes associated with aHUS in HGMD were interrogated as tier one[229]. Two potentially pathogenic variants were identified at this point, *CFH*:p.Q950H and *VWF*:p.R1339H. As it was not felt that these variants sufficiently explained the patient's phenotype, further phenotyping was performed. At this point it was determined that the patient had splenomegaly, dysmorphology (Figure 7.10) and continuing thromobocytopaenia, a symptom of aHUS which would be expected to be resolved by eculizumab therapy. An additional 44 genes identified in HGMD based upon these features were therefore interrogated[229].

In the second tier interrogation, a *CBL*:c.1096-1G>T was identified. *CBL* germline mutations are known to cause Noonan-like syndrome, which includes moyamoya and dysmorphology[230]. The *CBL* variant in the patient was determined to have arisen *de novo* by Sanger sequencing of the parents. This *CBL* syndrome is also known to commonly progress to juvenile myelomonocytic leukaemia (JMML)[230]. The progression is known to be initiated by the acquisition of a somatic uniparental disomy (UPD) for 11q, within which the *CBL* gene is located, resulting in LOH, and absence of functional CBL in the cell[231].

In order to evaluate the potential progression to JMML in the patient, evidence LOH genome-wide was investigated, applying *BAFsegmentation*[172]. The application of *BAFsegmentation* to the WES data for the patient identified an 11q LOH (Figure 7.11). There was no evidence of a reduction in the depth of coverage across this region, indicating that this is a balanced UPD.

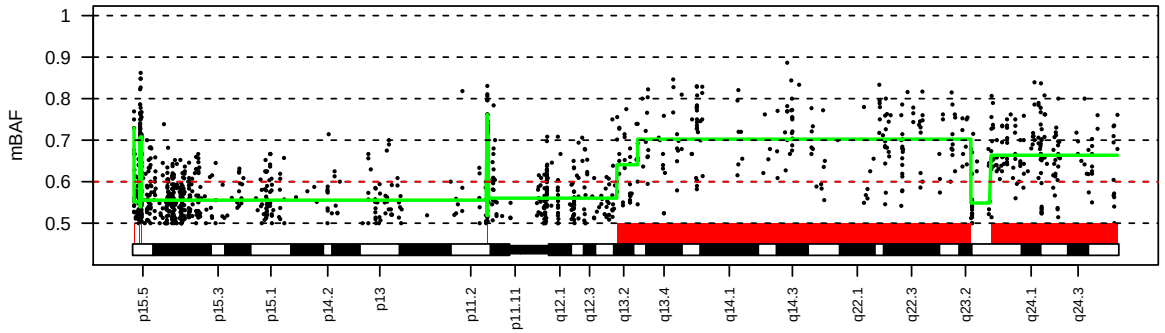


Figure 7.11: *BAFsegmentation* output showing 11q LOH in the patient from Family C. The mBAF of genotypes is shown in black points at the appropriate position on the karyogram, green indicated the segmentation results; regions indicated in red are segmented regions with an mBAF > 0.6. Note that the apparent LOH at p15.5 and p11.2 are likely due to alignment issues in these small regions, no additional large regions were deemed to have LOH by the software.

7.4.1.3 Discussion

The identification of the 11q UPD indicated that the patient is progressing to JMML[232]; however the patient has not yet shown clinical symptoms of the disease, aside from splenomegaly. As such, this finding in the WES data allow for the monitoring of the patient's burden of leukocytes carrying the UPD, as well as clinical manifestations, in order to respond with appropriate treatments if/when required. A small number of *CBL* germline mutation carriers do not progress to clinical JMML, and thus the treatment of a bone marrow transplant would be unnecessary[233].

Though the initial referring diagnosis of aHUS did not lead to resolution of the case, the flexibility of WES allowed for the interrogation of further genes when new information was available. Furthermore, it allowed new approaches to the data analysis to be applied where allelic imbalance was of interest, as well as depth of coverage.

7.4.2 Patient D - Actinic keratosis

7.4.2.1 Clinical presentation

The Patient D, an 83 year old female, presented with an actinic keratosis lesion on the left middle finger, which required surgical removal in 2013 (Figure 7.12). Actinic keratoses form due to mutations arising though prolonged, repeated exposure to the sun[234]. These lesions appear to share common UV-induced mutation profiles (including

frequent *TP53* mutations and acquired CNVs) with skin cancers, and there is a risk of progression of a lesion diagnosed as actinic keratosis to carcinoma[235]. Extensive genomic instability in even apparently benign lesions has been reported[236].

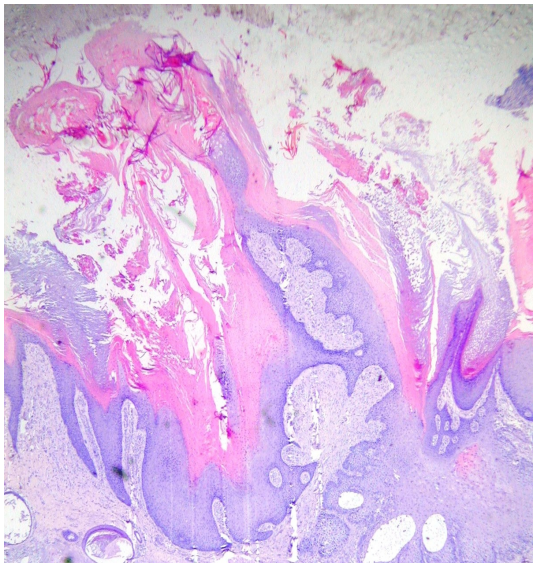


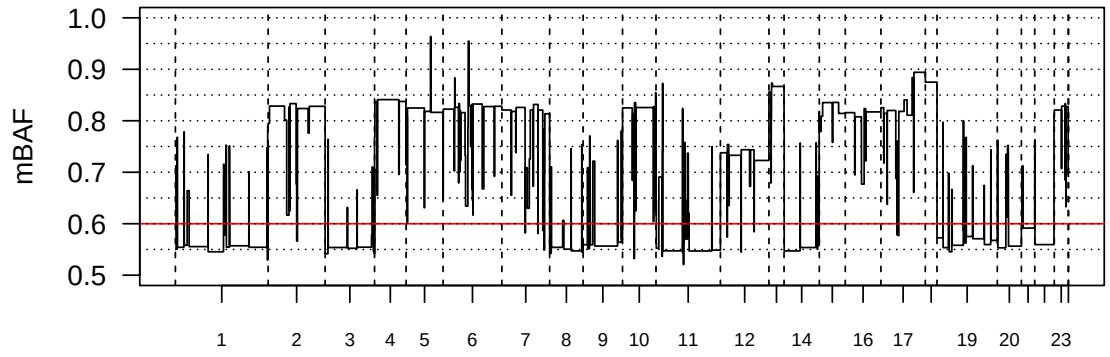
Figure 7.12: Histology of lesion from Patient D. Slide prepared using Haematoxylin and eosin staining

7.4.2.2 Genetic Analysis

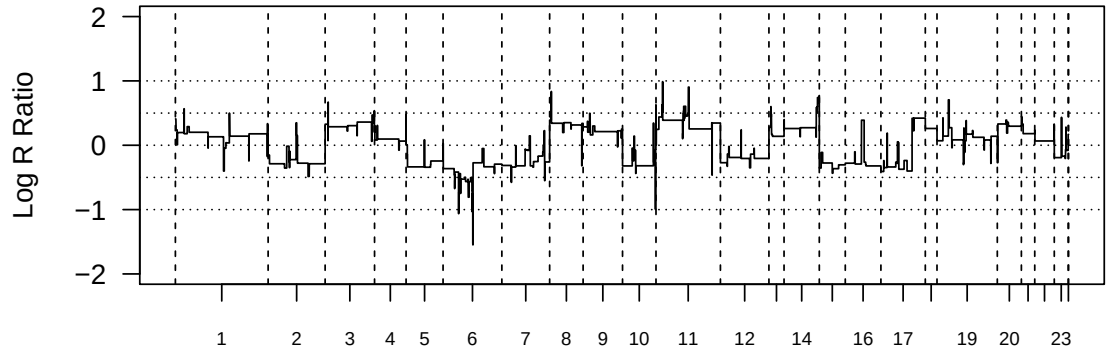
Archival FFPE embedded tissue was laser-capture microdissected to isolate the lesion at a high purity and appropriate normal tissue to allow for paired analysis. Returned WES data for both samples were of a good quality, passing all standard checks (Table B.2). A mean DOC of 119.4 X and 63.6 X was attained for the lesion and normal tissues respectively, with an intentional increased number of reads for the lesion. As actinic keratosis is expected to harbour somatic mutations, an alternative somatic pipeline was utilised for genotyping using a *VarScan 2*[157] paired analysis approach between the lesion and normal tissue in place of the *SAMtools*[165] based calling used in the Mendelian pipeline. The *Varscan 2* `copycaller` function was also used to identify acquired CNVs. Comparative CNV calling using the `copycaller` function utilises the log_2 ratio for the read depth between two paired samples (normalised for total read count), and applies CBS to identify regions of consistent deviation from $log_2 R = 0$[173].

When the `copycaller` function was applied to the paired samples, several regions of whole chromosome amplification were identified (Figure 7.13b), as well as many deletions. In order to evaluate if these were likely to be true amplifications, *BAFsegmentation*[172] was applied to assess evidence for allelic imbalance (Figure 7.13a); this was paradoxically not the case for the apparently amplified chromosomes, but was for all other chromosomes.

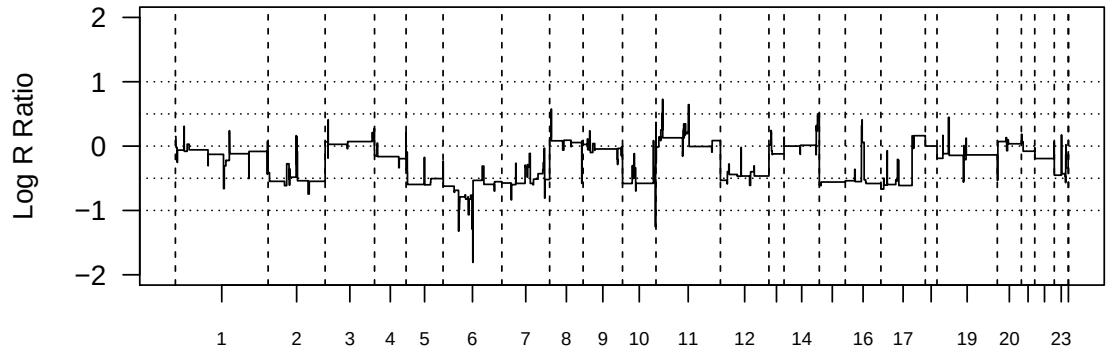
It was at this stage noted that it appeared that the majority of chromosomes showed LOH, consistent with acquired CNVs (Figure 7.13a), and that this may skew the normalisation procedure, given that appropriate normalisation relies upon the majority of the genome remaining euploid in both samples; if the average scenario for a chromosome is a deletion, then this explains the *apparent* amplification of the minority of chromosomes. As such, a normalisation using only reads aligning to the chromosomes exhibiting minimal evidence of LOH was applied (Figure 7.13c). Using this curated normalisation approach, chromosomes exhibiting minimal evidence of LOH also have $log_2 R \approx 0$, consistent with sustained euploidy between the two samples for these chromosomes, this is consistent with initial expectations.



(a) mBAF



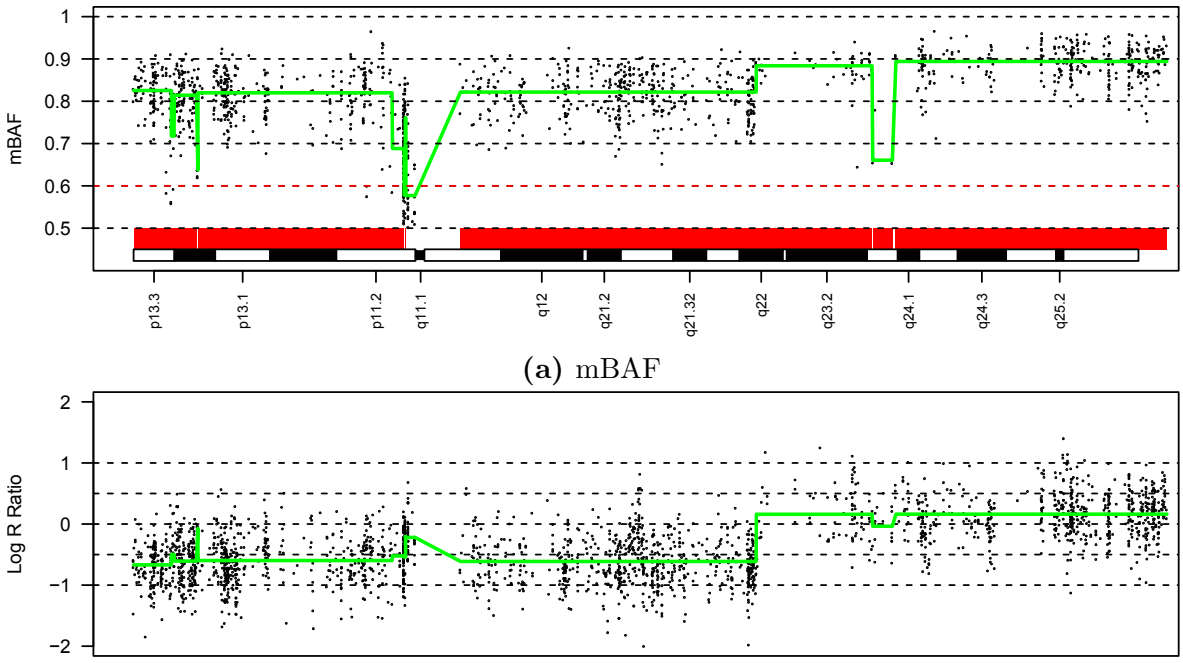
(b) log_2R whole-genome normalised



(c) log_2R non-LOH normalised

Figure 7.13: Genome-wide comparison of normalisation approaches for detecting copy number changes. Shown is the mBAF plot for the sample (a), with 14 chromosomes (e.g. chr2) showing significant deviation from the expected value of 0.5, and nine chromosomes (e.g. chr1) appearing to be have negligible deviation.
Where read-count normalisation for log_2R sample comparison is performed on a genome-wide level (b), it can be seen that chromosomes without LOH also appear to have undergone amplification compared to the normal samples, a result which is clearly counter intuitive.
However, where read-count normalisation only utilises the nine chromosomes without mBAF deviations (c), the more expected pattern of $log_2R \approx 0$ for chromosomes without apparent LOH, and $log_2R < 0$ for those showing LOH is seen. Figure generated using *BAFsegmentation*[172].

On assessment of the mBAF plot for the whole genome (Figure 7.13a), it was noted that the greatest sustained deviation in mBAF, with mBAF ≈ 0.9, was observed in 17q. As 17q UPD is a recurrent mutation in cancers, it was decided to investigate this in greater detail (Figure 7.14).



(a) mBAF

(b) log_2R non-LOH normalised

Figure 7.14: LOH and apparent copy-number change across chr17. Shown is the mBAF plot for the chromosome (a), and the log_2R where read-count normalisation for log_2R sample comparison is performed on chromosome without apparent LOH (see Figure 7.13) (b). Whilst the entire chromosome exhibits LOH, this is most pronounced downstream of 17q22. Interestingly, there is also a corresponding change in the log_2R in this region, being ≈ 0.5 upstream and ≈ 0 downstream. Figure generated using *BAFsegmentation*[172].

Given that the region upstream of 17q22 exhibits the greatest mBAF deviation, this would indicate that this 17q LOH was an early event in the progression of the lesion, and has therefore is present in the greater number of cells compared to the other mutations, with the 17p LOH occurring later in the progression of the lesion. 17q LOH is very common in cancers[237], and as such it is reasonable to assume that this is a common driver mutation. It is interesting to note that 17q appears to be a copy-neutral LOH as $log_2R \approx 0$, whereas the 17p LOH is accompanied by a marked reduction in read depth ($log_2R \approx -0.5$), indicating that this is the result of a later deletion, as opposed to the isodisomy seen in 17q. However, this leads us to a highly unintuitive conclusion. Given the high degree of skewing of the BAF (mBAF > 0.8 across the chromosome), it would appear that the majority of cells harbour LOH for the two segments of the chromosome. However, as the 17p deletion spans the centromere, it would be impossible for the 17q region to remain diploid during successive cell divisions. Further investigation would be required to elucidate the mechanism by which these mutations occurred, requiring alternative approaches such as single cell for phasing or WGS for breakpoint detection to resolve this apparent paradox.

7.4.2.3 Discussion

If paired copy number calling was applied in this case only using the total count of aligned reads, then incorrect conclusions could be drawn from the CNV calls produced. Whilst it would naively appear reasonable to presume that the majority of the genome remains euploid, with acquired CNVs being the exception, this is clearly not the case in this sample. Using a combination of the BAF and read depth information, it is possible to elucidate a more complete picture of the progression of successive structural alterations. To generalise the lessons learned while analysing this exome, it is clear that appropriate choice of controls and normalisation procedures is essential in order to generate meaningful results.

7.5 Clinical phenotyping

7.5.1 Family E - Activated PI3K-δ syndrome

7.5.1.1 Clinical presentation

Proband II1 (see Figure 7.15) presented with disseminated pneumonococcal infection at age 5 years, pneumonia and non-clonal lymphoproliferation leading to splenomegaly requiring surgical intervention, and haemolytic anaemia. In addition there was a deficiency in polysaccharide targeting antibodies and lymphopoenia of mature $CD4^+$ T-cells, as well as further perturbation in immunoglobulin levels. II1 responded well to immunoglobulin infusion. Parents had no apparent PID conditions, though the father had been treated some years prior for leukaemia, with no reported PID phenotypes in the extended family. It was noted by the clinical team that the phenotype resembled autoimmune lymphoproliferative syndrome, though without the expected increase in $CD4^-CD8^-$ T-cells.

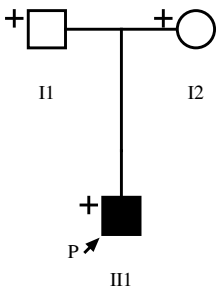


Figure 7.15: Pedigree showing inheritance of activated PI3K-δ syndrome in Family E. WES analysed individuals indicated by '+'.

7.5.1.2 Genetic analysis

The simplex trio of proband and parents was sequenced. Returned WES data were of a good quality, passing all standard checks (Table B.2). A mean DOC of 63.9, 68.0 and

61.6 X was attained, with 23,055, 23,797 and 23,297 variants called by our pipeline for I1, I2 and II1 respectively. Given the apparent recessive/*de novo* mode of inheritance, variants were filtered for being discordant between the proband and parents, with a higher allelic dosage in the proband. Initially, genes known to be involved in apoptosis were interrogated[238]. No discordant variants with 1000 Genomes Project AF ≤ 0.01 were called in the 63 apoptosis genes queried. As such, 248 genes known to be involved in PID disorders were interrogated[239]. No variants with an AF ≤ 0.01 were present within these genes with increased dosage in II1.

In November 2013, a report of 35 patients with a common PID phenotype was published, and highlighted by our clinical colleagues due to the comparable phenotype to Family E[240]. This case series identified a recurrent *PIK3CD*:c.G3061A:p.E1021K substitution in a high proportion of the cases. Given the phenotypic similarities, the *PIK3CD* gene was interrogated, identifying the p.E1021K variant in the affected proband and father. As such, II1 is affected with the newly described activated PI3K-δ syndrome (APDS; MIM 615513)[240].

7.5.1.3 Discussion

Given the phenotypic similarities, it is clear that the *PIK3CD*:p.E1021K is pathogenic in II1. However, what is not apparent is the reason for the lack of a PID phenotype in the father, I1. *PIK3CD* encodes phosphatidylinositol-4,5-biphosphate 3-kinase, catalytic subunit δ (PI3K-δ; EC 2.7.1.153). PI3K is a plasmalemma-bound complex which catalyses the phosphorylation of plasmalemma-bound phosphatidylinositol-4,5-biphosphate (PIP_2) to phosphatidylinositol-3,4,5-triphosphate (PIP_3); PIP_3 is an active signalling molecule, for which ~200 proteins contain complementary binding sites. Specifically, the PI3K-AKT-BAD signalling pathway is involved in the suppression of apoptosis[6]. The PI3K-δ subunit is leukocyte specific[241].

The p.E1021K substitution was shown to be activating, leading to increased PIP_3 levels and thus increased AKT activity and activation of BAD, and thus inhibition of apoptosis[240]. This inhibition of apoptosis would explain the lymphoproliferative phenotype, additionally, this substitution has been observed in B-cell lymphoma. As such, this germline variant may be contributory to the development of malignancies, as has been reported in a case of a female developing B-cell lymphoma at age 19[242,243]. Given this overlap with haematological malignancies, it was suggested that chemotherapeutic inhibitors of PI3K-δ may be clinically useful in the treatment of APDS[240]. Rapamycin (an inhibitor of mTOR, itself a coactivator of AKT) has already shown some clinical utility[241].

The segregation of p.E1021K in Family E highlights the importance of accurate phenotyping and presumption of mode of inheritance and penetrance when filtering based on phenotypic segregation is desired. *PIK3CD* was in the secondary list of PID candidate genes, however the p.E1021K was excluded due to its presence in the non-PID affected father. Immunology research is particularly susceptible to apparent incomplete penetrance, as in this case, due to the complexity of the immune system, it is challenging to phenotype the required characteristics where the characteristic of interest is not known.

7.6 Discussion

In this chapter, five cases have been discussed, wherein the identified aetiology was refractory to identification. In some of these cases, the identification of variant(s) is hindered due to deficiencies in current software tools and sequencing technologies. Particularly, the indels in Families A & B could not be identified due to limitations self imposed by software for the minimisation of false positives. This is challenge to be partly overcome statistically and programmatically, but improvements in long-read technologies and greater uniformity in genome coverage will both support these efforts.

For the proband of Family C, the lessons are twofold. Firstly, purely focussing on the subjective referral phenotype may result in the missing of important findings in a patient. A holistic approach to data interrogation is therefore required, on a multidisciplinary interface between diverse clinicians and informaticians. Furthermore, LOH detection proved to be informative in data for which only constitutional mutations were initially considered relevant. This highlights the malleability of NGS data, to be leveraged to answer an evolving question without additional data generation. In Family D, the necessity for appropriate curated controls was highlighted, beyond merely selecting appropriate individuals.

Finally, in Family E, experience in the challenges associated with incomplete penetrance and diverse presenting phenotypes were discussed. This can present a massive challenge when utilising segregation information for filtering. It is noteworthy that the *PIK3CD*:p.E1021K would likely have been identified sooner had familial data been unavailable.

Analysis has revealed some specific weaknesses in the Soton V3.x pipelines, particularly with regards to indel detection. *SAMtools* is known to be comparably weak for the detection of indels, and therefore an alternative should be sought. Callers such as *GATK*[166] and *Platypus*[244] use more encompassing approaches to variant calling,

with features such as local realignment around indels greatly improving sensitivity. Active pipeline development for the Soton V4.0 pipeline is currently underway, and will utilise control DNA for comparison to a gold standard. Specifically, NA12878 (NIST RM8398) has been sequenced, and is compared to genotype data produced by the Genome in a Bottle Consortium[245]. As a provisional assessment, the use of the *GATK*[166] `HaplotypeCaller` provided an uplift in indel sensitivity of 17.7% compared to *SAMtools*[165] (96.8% *vs.* 79.1% respectively) when applied to 200 X data captured using the Agilent SureSelect Focused Exome.

Overall, it is clear that WES will prove to be clinically valuable, particularly as large cohorts of data are amassed, allowing greater power to detect more subtle aetiological signals. However, targeted approaches will also play a valuable role, particularly in disorders where much of the heredity is understood. It is clear from the cases presented here that a proportion of the ~75% of WES investigations that remain unresolved will require a more targeted, customised interrogation approach[68,69]. The routine interrogation of WES data will likely become a relatively 'push-button' approach in the near future, requiring minimal human involvement in the data interrogation stages, this will allow bioinformaticians more time to better interrogate cryptic exomes, and develop novel analytical tools.

As exome sequencing rapidly approaches clinical practice, the increasing amounts of data will facilitate bioinformatic method development, as well as more powerful large cohort studies allowing the identification of further novel aetiological genes, possibly also genes with smaller pathogenic contributions.

Chapter 8

Application of Whole-exome sequencing to Cleft lip/palate phenotypes in Colombia

8.1 Background

Correct establishment of growth patterning during foetal development is key to ensure correct morphogenesis, including that of craniofacial features. This growth patterning can be perturbed by many factors, including environmental and genetic. Isolated, single feature disorders, such as a cleft lip/palate tend to be caused by the interaction between genetic predispositions and environmental factors, such as excessive consumption of alcohol, smoking and other teratogens, as well as prolonged developmental exposure to altitude. In contrast, familial syndromic phenotypes are more likely to have an underlying Mendelian genetic aetiology[206,246,247]. Our group have investigated several affected families in collaboration with Prof. Ignacio Briceño, based at the University of La Sabana, Bogotá, Colombia.

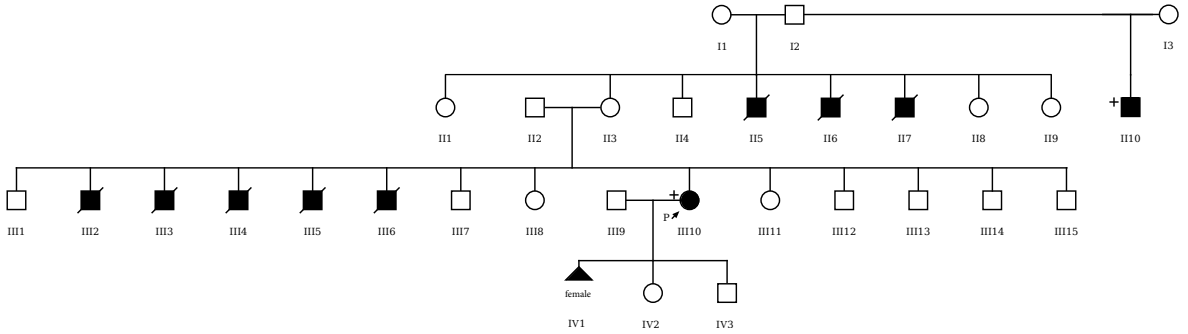
8.2 Methods

Ascertainment of the individuals detailed herein was at the Operation Smile Multidisciplinary Centre in the Bogotá region of Colombia, established for the treatment of individuals affected by orofacial clefting. Exome sequencing and data analysis was performed as described in Chapter 7. Given the diversity of cases analysed, these will be analysed in two sub-cohorts of non-syndromic (designated as sample IDs beginning NSCLP) and syndromic, (designated with SCLP).

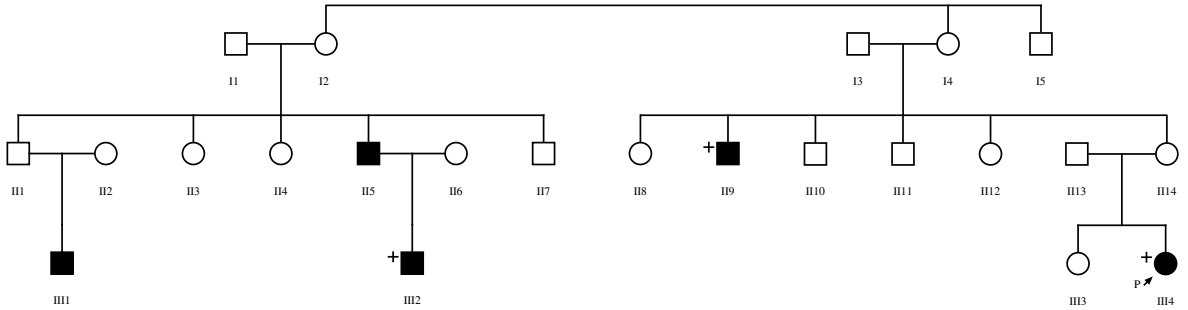
For filtering of identified variants, we established a comprehensive list of genes previously implicated in any form of CLP phenotype including search terms related to the clinical diagnoses made for the patients. First, we queried the Human Gene Mutation Database (HGMD professional)[229] in July 2014, 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 (accessed July 2014)[17], and a small number of additional CLP-related genes from the review were also included by Collins *et al.*[247]. 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[12]. More frequent variants were excluded from further consideration as unlikely causes of rare syndromic disease.

For NS variants, we used the scaled predictive scores from dbNSFP v2[248] 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)[174]; SIFT (scores < 0.05 are predicted to affect protein function)[178]; PolyPhen-2 HumVar (scores ≤ 0.446 considered 'benign'; scores between 0.447 and 0.908 considered 'possibly damaging'; scores ≤ 0.909 considered 'probably damaging')[179,249]; 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[250]; MutationTaster (variants with scores > 0.95 considered damaging)[251] 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)[175]. 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)[177]. 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)[252]; Logit (the conditional probability that a variant is Mendelian disease-causing given prediction scores from 13 programs, including SIFT, PolyPhen-2, LRT, MutationTaster, PhyloP, GERP++ and CADD, under a logistic regression model)[253]. We also produced a combined rank for variants with PhylopP, GERP++, CADD and Logit scores based on the summed ranks across all four scores.

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 consistently harbour erroneous NGS genotype calls, were removed from further consideration[254]. All identified variants were cross-referenced with an in-house database of exome-sequenced samples and variants present in any of these exomes. The 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. Finally, where multiple members of a pedigree were sequenced, a variant was required to be observed in all affected and be absent in all unaffected members in segregation analysis. 10 families were analysed in total, three syndromic (Figure 8.1) and seven non-syndromic (Figure 8.2), with a total of 15 individuals sequenced.



(a) SCLP2



(b) SCLP3

Figure 8.1: Pedigrees of families effected with syndromic CLP. Family SCLP1 is not shown as the proband has no relevant family history; the proband has history of swallowing disorder due to retrognathia; bilateral dacryostenosis; micrognathia; atresia of the right external auditory canal; agenesis of 1st finger (bilateral); normal external genitalia.
Phenotypes for affected members of SCLP2 (a): 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) postnatal death at 8–15 days and facial clefting; III6 (female) postnatal death at 8 days and cleft lip and palate; IV1 prenatal death and facial clefting; III10 (proband, exome sequenced), unilateral (left side) cleft lip and palate, clubbing, nail hyperpigmentation, cutaneous syndactyly.
Phenotypes of affected members of SCLP3 (b): 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.
Exome sequenced individuals are indicated with a '+'

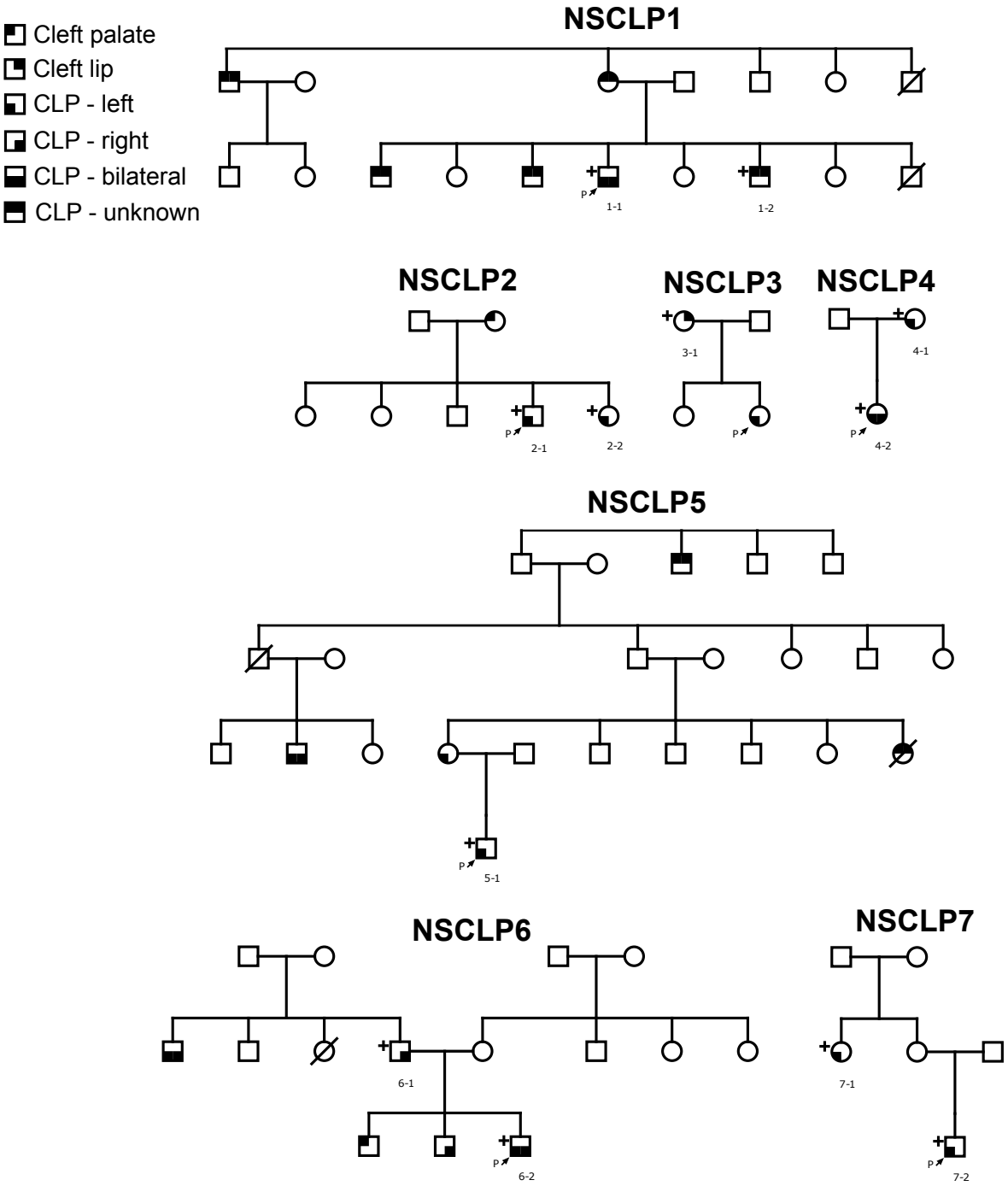


Figure 8.2: Pedigrees of families effected with non-syndromic CLP. All individuals have an isolated CLP phenotype. In family NSCLP3 the mother has submucous cleft palate and the child has global developmental delay, suggesting the possibility of an undiagnosed syndromic condition; the presentation is however not recognisable as a known CLP syndrome, and thus these comorbidities may be incidental.
Exome sequenced individuals are indicated with a '+'

8.3 Results

8.3.1 Syndromic CLP

WES data of a good quality was returned, with > 50 X coverage for all samples (Table B.3). Table 8.1 lists 35 variants which met the filtering criteria across the six exome-sequenced individuals, of which 30 are nonsynonymous SNPs, 3 are splicing variants and there are single stop gain and frameshift insertions. Analysis on each family suggests causal variation in each case, as described below.

8.3.1.1 Family SCLP1

The proband was diagnosed as a potential Nager syndrome patient. Nager syndrome is extremely rare, and fewer than 100 cases have been reported[213–215]. Nager syndrome belongs to a group of conditions displaying acrofacial dysostosis, characterized by association between craniofacial and limb malformations[213]. The patient phenotype (Figure 8.1) 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 8.1), *ERCC2* (rank 7) and *PROKR2* genes (rank 10). More significantly, sequencing also identified the known c.1060_1061insC:p.R354fs frameshift mutation in exon 5 of the *SF3B4* gene. This variant was confirmed as present by Sanger sequencing. Exome sequencing has previously established mutations in the *SF3B4* gene (splicing factor 3B, subunit 4) as responsible for autosomal dominant Nager syndrome[213]. *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[213]; 5 of 14 families tested[214] and 5 of 12 families[215]. Most patients who are negative for *SF3B4* mutations are phenotypically identical, indicating genetic heterogeneity.

Table 8.1: Deleterious variants in syndromic CLP cases

Gene	Nucleotide	Protein	AF	SIFT	PP-2	LRT	MT	GS	PhyloP	GERP	CADD	Logit	SCLP: Rank	1 1	2 III10	2 II10	3 III2	3 II9	3 III4
SF3B4	1060_1061insC	R354fs	.	.	.	.	.	.	.	.	.	.	.	◇					
TFR2	1483-7A>C	.	.	.	.	.	.	.	.	.	.	.	.			◇			
TNNT3	82+7C>T	.	0.001	.	.	.	.	.	.	.	.	.	.		◇				
COG1	743-10C>G	.	0.001	.	.	.	.	.	.	.	.	.	.		◇				
IFT140	G2569A	G857S	.	0.07	0.995	0	1	56	7.661	5.22	34	0.275	1		◇				
RPGRIP1L	G724T	E242X	.	0.14	.	0	1	.	4.463	5.87	36	0.164	2		◇	◇			
IRF6	G604A	V202I	.	.	0.916	0	1	29	7.311	6.17	27.1	0.119	3				◇	◇	◇
IFT172	G3604T	V1202L	.	.	.	.	1	32	4.362	5.7	28.3	0.148	4	◇					
IDUA	T965A	V322E	0.002	0	0.999	0	1	121	5.962	5.15	23.3	0.151	5		◇				
SH3PXD2B	C2288T	P763L	.	.	0.997	0	1	98	7.565	5.29	19.54	0.116	6			◇			
ERCC2	A1900G	K634E	.	0	0.925	0	1	56	5.182	5.13	28.3	0.123	7	◇					
IKBKG	G169A	E57K	.	0.16	0.997	0	0.99	56	5.105	5.6	21.9	0.094	8		◇				
NKX3-2	G493C	D165H	.	0.17	0.419	0.022	1	81	3.045	5.31	21.4	0.212	9		◇	◇			
PROKR2	C719T	T240I	.	0.23	0.841	0	1	89	5.246	5.16	23.1	0.096	10	◇					
COL1A2	C3226T	P1076S	.	0.02	0.063	0	1	74	3.858	5.32	16.87	0.147	11		◇				
PGM1	C143T	A48V	.	0.13	0.025	0	1	64	7.651	5.13	19.05	0.072	12		◇				
PKLR	C92T	A31V	.	0.03	0.935	0	0.74	64	1.654	4.74	26.1	0.962	13		◇	◇			
NOTCH2	T7223A	L2408H	0.001	0	0.969	0.006	0.74	99	2.431	5.35	15.05	0.1	14		◇				
SEC23A	A2116G	I706V	.	0.55	0.042	0	1	29	5.022	5.75	12.73	0.043	15		◇				
GRIN2A	A662G	K221R	0.001	.	0.027	0	1	26	6.107	5.09	12.59	0.057	16		◇				
TUBB2B	C743T	A248V	.	0	0.082		1	64	9.506	4.18	7.09	0.098	17	◇					
ECEL1	A1516G	M506V	.	0.01	0.76	0	1	21	3.251	5.36	14.54	0.048	18			◇			
SRCAP	A3859G	T1287A	.	0.47	0.091	.	1	58	2.494	5.17	8.04	0.062	19			◇			
IFT122	C496T	R166W	0.001	0.02	0.88	0.001	1	101	3.254	4.8	15.93	0.041	20	◇					
UBE3B	C136T	R46W	.	0	1	0	1	101	1.832	4.43	16.02	0.05	21	◇					
ABCA3	G3052A	G1018S	0.001	0.45	0.064	0	1	56	2.165	4.65	12.09	0.062	22		◇				
ABCC6	C1963A	Q655K	.	1	0.004	0.061	1	53	3.499	4.97	4.84	0.017	23		◇				
MCPH1	A775C	K259Q	.	0.43	0.506	0.164	1	53	-1.592	-1.78	18.37	0.053	24	◇					
KMT2A	G10327A	A3443T	.	0.23	0.001	0.016	1	58	3.393	1.73	8.28	0.04	25	◇					
TRPS1	C2000T	S667L	.	0.01	0.024	0.418	1	145	1.658	4.9	7.06	0.044	26	◇					
PALB2	G265C	D89H	.	0.04	0.063	0.084	1	81	0.699	0.034	7.76	0.066	27			◇			
COL6A2	G316A	E106K	0.002	0.3	0.437	0.088	0.99	56	1.335	4.34	12.22	0.04	28	◇					
SZT2	G9611A	R3204Q	0.003	0.57	0.001	0.002	1	43	2.405	3.49	7.06	0.014	29	◇					
GJB6	A476G	N159S	0.002	0.76	0.038	0.007	0.91	46	1.71	3.95	6.08	0.015	30			◇			
MCPH1	T1273A	Y425N	.	0.75	0.001	0.096	1	143	0.073	-1.86	5.58	0.006	31	◇					

AF - AF in 1000 Genomes Project; PP-2 - PolyPhen-2; MT - MutationTaster; GS - Grantham score. Rank is based on sum of ranks for variants with PhyloP, GERP++, CADD and Logit scores and range from (predicted) most to least deleterious. ◇ indicates a heterozygous variant.

The variant identified in this patient corresponds to the same frameshift mutation identified as *de novo* in family 'I' by Bernier *et al.*[213] and Petit *et al.*[214] in their 'case 13'. The identification of the same mutation in three independent studies suggests that this may be one of the 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.*[215] 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.

8.3.1.2 Family SCLP2

The female proband presented with bilateral CLP together with a catalogue of other syndromic features (Figure 8.1), including abnormal nail pigmentation and cutaneous syndactyly. The family pedigree 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[17,255,256], whereas in this family affected males are known to have survived for 8–15 days. Facial clefting is a feature of the family phenotype, although a case of IP associated with bilateral CLP was described as 'unique'[255]. Familial IP is a rare condition arising approximately in 1 of 50,000 newborns[255], 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.*[257] 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) 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 dbSNP and have not previously been linked to clinical phenotypes. The second ranked variant 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[258]. 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 missense

mutation in exon 2 of the *IKBKG* gene on chromosome X. Smahi *et al.*[259] showed that cells of IP patients lack NF-κB function due to mutations in the *IKBKG* gene (NF-κB essential modulator).

Aradhya *et al.*[255] 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 E57K mutation found here is a substitution also reported by Aradhya *et al.* as only one of the two (of 29) small mutations that changed the amino acid identity. They 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.*[260] 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 Aradhya *et al.*[255], Conte *et al.* 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 p.E57K mutation we have identified here is described as presenting a 'milder' IP phenotype[255,260], although Aradhya *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.

8.3.1.3 Family SCLP3

The family (Figure 8.1) 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 shows micrognathia and, as a result, needed ventilator support in the ICU at birth and was treated with oral surgery (mandibuloplasty). Pierre Robin syndrome (PRS) 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. PRS 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[261]. It represents a causally heterogeneous series of events (micrognathia causing glossoptosis preventing palatal shelves to fuse) and is often referred to as the PRS. Tan *et al.*[262] 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[262].

Exome analysis (Table 8.1) 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 Van de Woude syndrome (VWS) and 80% of the causal mutations are found in exons 3, 4, 7 and 9, whereas mutations underlying popliteal pteringeum syndrome are more frequent in exon 4[247]. Wu-Chou *et al.*[263] found exon 5 mutations in 2 of 13 VWS cases. However, the SCLP3 family exhibits variable PRS features and lacks lip pits that are characteristic of VWS. Nikopensius *et al.*[264] 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. Vieira[265] describes positive associations of clefting with hypodontia with *IRF6*, although the role of this gene in PRS has not been previously described. Sanger sequencing confirmed carrier status for unaffected (transmitting) relatives II14 and III1.

8.3.2 Non-syndromic CLP

Table 8.3 lists 28 novel missense variants, each of which segregates within an individual family and is classed as deleterious by at least one predictive score, with Table 8.2 showing likely protein truncating and indel variants in the families. Table entries are ordered using combined ranks from most to least deleterious by predictive score. Four of the genes listed (*WNT7A*, *MSX1*, *CLPTM1* and *EVC2*, ranked 9, 10, 11 and 23 respectively) have previously been identified as containing variants implicated in NSCLP phenotypes. Family NSCLP1 has the 9th ranked variant in the *WNT7A* gene. Members of the *WNT* gene family have previously been associated with NSCLP phenotypes[266]. Specifically, a number of WNT signalling pathway genes including *WNT3A*, *WNT5A*,

WNT9B, and *WNT11* have been established as candidates[266] and mouse expression studies have shown roles for *WNT* genes in mid-facial formation and lip and palate development[267]. Chiquet *et al.*[266] tested 38 SNPs in seven *WNT* family genes within a large NSCLP cohort. Nominally significant associations within *WNT7A* were found but the strongest association were in *WNT3A*, *WNT5A* and *WNT11*.

Table 8.2: Novel protein truncating and indel variants in non-syndromic CLP cases

				NSCLP:						
Gene	**Nucleotide**	**Protein**	**ΔMaxEnt**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
DLG1	923_925del	308_309del	.				◇			
FRAS1	G7354T	E2452X	.						◇	
WDR11	2660_2662del	887_888del	.			◇				
IGF1R	3940_3941insCGTCCTCCC	L1314delinsPSSL	.	◇						
FBLN1	485-5C>-		22.14	◇						

The 10th ranked variant, found in family NSCLP4, is in the *MSX1* gene, and considered damaging by SIFT, PolyPhen-2 and MutationTaster, and has high GERP++ and CADD scores. Variants in this gene have been strongly implicated in NSCLP in several studies. Jezewski *et al.*[268] found mutations in 2% of cases and indicated that this has genetic counselling implications where autosomal dominant inheritance patterns are found. Exon 2 of *MSX1*, in which the p.P260T is located, has been found to be highly conserved with significantly fewer sequence variants compared with exon 1[268]. Functional validation of *MSX1* as a candidate is established through a cleft palate and foreshortened maxilla phenotype in knockout mice[269]. A number of association studies have also indicated involvement of *MSX1* in NSCLP.

In a study of 94 patients and 93 controls from Operation Smile, Colombia, four *MSX1* microsatellite alleles were analysed and a positive disease association was observed with CA polymorphisms in the gene[270]. An autosomal dominant *MSX1* mutation in a family with clefting and tooth agenesis indicates a familial pattern of segregating *MSX1* mutations. Jezewski *et al.*[268] sequenced the *MSX1* gene in 917 individuals with NSCLP and found potentially aetiological variation in 16 individuals including coding and non-coding variants. Diverse evidence establishes that *MSX1* promotes growth and inhibits differentiation. Mutations in *MSX1* can cause primary or secondary facial clefting within mouse models[269].

Table 8.3: Novel deleterious non-synonymous variants in non-syndromic CLP cases

													NSCLP:				
Gene	**Nucleotide**	**Protein**	**SIFT**	**PP-2**	**MT**	**GS**	**PhyloP**	**GERP**	**CADD**	**Logit**	**Rank**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
WDR35	C2161T	R721C	0	0.92	1	180	9.81	5.04	27.7	0.13	1	◇						
PTHLH	G71A	G24E	0	1	0.99	98	5.75	5.13	32	0.39	2		◇					
GPC6	T599A	F200Y	0	0.98	0.95	22	7.65	5.48	31	0.06	3	◇						
INPPL1	G349A	V117I	0	0.95	0.04	29	8.18	3.9	22.8	0.11	4	◇						
MYH3	G3869A	R1290H	0	0.1	0.94	29	4.95	4.84	21.3	0.13	5				◇			
AHDC1	C1996G	R666G	0	1	0.06	125	8.73	5.08	22.8	0.04	6		◇					
ABCA12	C254T	T85I	0.99	0.73	0	89	4.18	5.3	15.26	0.1	7			◇				
DEAF1	C1532G	A511G	0	0.59	1	60	9.01	3.03	17.71	0.08	8			◇				
WNT7A	G1019A	S340N	0	0.94	0.99	46	6.07	4.11	23.6	0.06	9	◇						
MSX1	C778A	P260T	0	0.61	0.99	38	5.96	4.76	27.6	0.04	10				◇			
CLPTM1	A1058G	N353S	0.04	0.6	0.99	46	6.6	3.01	17.19	0.09	11	◇						
IGF1R	C4030G	Q1344E	0	0.01	0.99	29	4.78	5.24	13.05	0.04	12	◇						
CFDP1	A535T	T179S	0	0.02	0.99	58	2.66	5.54	15.68	0.04	13	◇						
NBAS	G784A	G262S	0.01	0.09	0.86	56	4.26	4.15	13.81	0.07	14	◇						
COL17A1	T3434C	I1145T	0	0.15	0.31	89	5.46	4.39	12.18	0.06	15					◇		
CDON	A860G	N287S	0	0.34	0.64	46	3.1	5.01	15.32	0.04	16							◇
SNAP29	A427G	N143D	0.02	0.34	0.17	23	8.77	3.7	11.41	0.04	17		◇					
NOTCH2	G1465T	V489L	0	0.08	0.34	32	0.87	5.38	12.51	0.05	18				◇			
MASP1	G2087A	G696E	0.05	0.09	0.37	98	1.65	3.75	14.53	0.06	19						◇	
FREM2	A2512G	T838A	0	0	1	58	2.49	4.44	7.38	0.07	20				◇			
SPRY4	C856T	R286C	0	0.88	0.97	180	2.44	4.7	13.49	0.04	21			◇				
ZBTB24	A367G	K123E	0	0.05	0.32	56	1.52	4.16	14.67	0.03	22				◇			
EVC2	G2536A	E846K	0.1	0.67	0.27	56	1.14	2.85	16.13	0.03	23		◇					
SCN2A	T2204C	M735T	0.04	0	0.06	81	0.47	2.35	2.95	0.04	24			◇				
RYR1	G5459T	R1820L	0.04	0.01	0.71	102	0.93	1.71	8.87	0.03	25					◇		
WT1	C137T	A46V	0.02	0	0	64	0.33	0.81	12.21	0.02	26				◇			
INPPL1	T3563G	L1188R	0.1	.	0.01	102	0.44	1.47	10.2	0.01	27	◇						
COL6A2	G2470A	V824M	0	.	1	21	.	3.62	.	.	-					◇		

PP-2 - PolyPhen-2; MT - MutationTaster; GS - Grantham score. Rank is based on sum of ranks for variants with PhyloP, GERP++, CADD and Logit scores and range from (predicted) most to least deleterious. ◇ indicates a heterozygous variant.

The 11th ranked variant (from family NSCLP1) is in the *CLPTM1* gene (Cleft lip-and palate-associated transmembrane protein-1) which is situated at 19q13.3. A balanced translocation is this region was found in a multi-case CLP family[271] and this region is implicated in NSCLP by linkage at linkage and transmission disequilibrium test association studies[272]. However a *de novo* deletion of 0.8 Mb in this region associated with CLP, but not encompassing *CLPTM1*, has been reported[273]. As Kohli & Kohli[274] indicate the role of *CLPTM1* or other genes in this locus is uncertain and there is a need for further studies to elucidate the precise role of this region in NSCLP.

The 23rd ranked variant is in the *EVC2* gene (family NSCLP2) and belongs to the same two megabase chromosomal region as *MSX1* (4p16). Ingersoll *et al.*[275] found linkage and association signals in genes in this region by examining CLP cases and trios from a number of populations. They found suggestive evidence for linkage and association amongst cleft palate trios to *EVC2*. Mutations in *EVC2* can lead to Ellis-Van Creveld syndrome or Weyers acrofacial dysostosis[276]. The former is autosomal recessive and not usually associated with oral clefts but cases with 'partial hare-lip', and tooth anomalies have been reported[275].

8.4 Discussion

Linkage, candidate gene association and GWAS have been applied to investigate numerous multifactorial diseases, including NSCLP. As a result of these studies more than 11 genes and gene regions are now known or likely to have a role in NSCLP[46,247]. However, there is increasing evidence that NSCLP is a heterogeneous condition comprising a substantial multifactorial component but also a much smaller proportion of cases showing more Mendelian patterns of inheritance. The Gajdos *et al.*[277] segregation analysis indicated that the complex familial patterns observed in NSCLP is best explained as a mixture of monogenic cases, probably dominantly inherited, combined with others which have a multifactorial aetiology. The conclusions favour analyses of multiple-case pedigrees to reduce heterogeneity and help identify Mendelian sub-forms.

We have investigated 10 families, prioritising those with an extensive family history. In all three syndromic families, a monogenic cause of the disease was identified, consistent with their previous clinical diagnosis. Furthermore, in the seven non-syndromic families, we have identified novel, deleterious variants, in genes previously associated with CLP. It is possible that some of these variants have contributed to the high-penetrance non-syndromic CLP in these families; it is however impossible to confirm the precise role of these variants without functional evidence. In the case of the *MSX1*:p.P260T

variant however, the extensive evidence regarding the high penetrance pathogenicity of this gene makes it highly likely that this is aetiological in NSCLP4.

Chapter 9

Diagnostic Utility of Targeted Gene Panels in Kidney Disease

9.1 Background

9.1.1 Gene panels

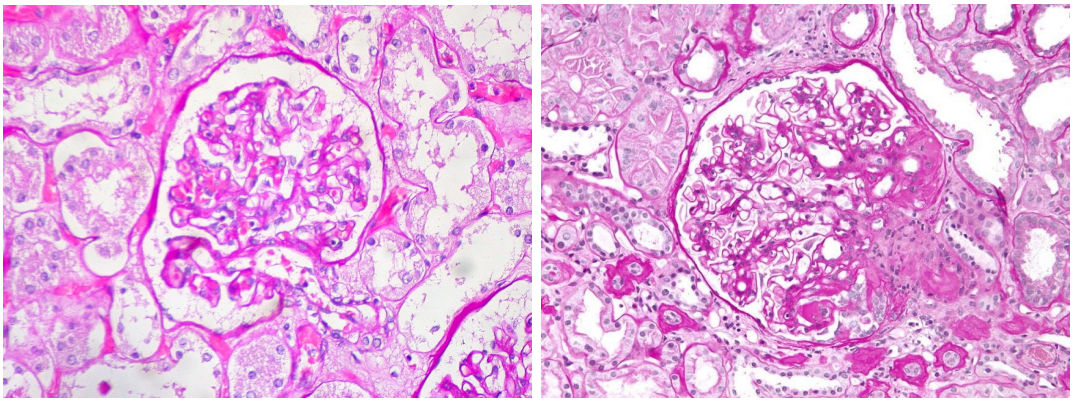
Targeted gene panels are an option for clinical diagnostics, providing a middle ground between traditional Sanger-based single gene sequencing and WES/WGS approaches. These more focused gene panels allow for the reduction in the required sequencing, facilitating the use of lower throughput sequencers, such as the Illumina MiSeq, as well as reducing the data analysis burden. However, sequencing only the *a priori* candidate genes for a patient limits options for the extension of the interrogation if required for the patient, possibly requiring further sequencing[278].

There is an increased availability of NGS based gene panels for clinical use. According to the UK Genetic Testing Network (UKGTN), there are 25 NGS gene panels currently approved for NHS testing (sequencing an average of 26 genes), with a further 60 panels recommended for approval as of April 2015[279]. Each panel is required to undergo a rigorous validation process prior to UKGTN approval. Due to this, alternatives such as clinical exomes, including ~5,000 clinically relevant genes, provide an umbrella panel, which may streamline the laboratory and validation workflow.

In this chapter I will describe the use of a custom NGS gene panel for focal segmental glomerulosclerosis (FSGS). This panel was designed based upon extensive curation of the literature and medical genetic databases and evaluated for its clinical utility when applied to heterogeneous, representative patient cohorts.

9.1.2 Focal segmental glomerulosclerosis

FSGS is a clinico-pathological diagnosis, which encompasses a spectrum of underlying disorders. Clinically, FSGS (as well as the similar diagnosis of steroid resistant nephrotic syndrome (SRNS)) presents with protein in the urine, and often progressive renal impairment, with characteristic eponymous findings on biopsy (Figure 9.1). The aetiology of FSGS often remains unknown, and these cases are classed as primary. There are several forms of secondary FSGS due to various causes, such as obesity, diabetes mellitus and hypertension with similar biopsy findings. A number of genes are also implicated in cases of hereditary FSGS. FSGS accounts for approximately 40% of nephrotic syndrome in adults in the U.S.[280].



(a) Healthy **(b)** FSGS

Figure 9.1: Comparison of healthy and sclerotic glomeruli. The glomeruli are the point of filtration of the kidneys. In FSGS, the glomeruli are scarred, reducing filtration capacity. Tissue is stained with periodic acid-Schiff stain, staining the basement membrane. The sclerosis is apparent in the bottom right setion of the glomerulus shown in b. Images used under Creative Commons Attribution licences, provided by Ed Uthman and Wikipedia user 'Nephron' for images a and b respectively.

Given the non-specific findings in FSGS, the diagnosis is sometimes given in error to other renal diseases presenting in a similar way. For instance, Gibson *et al.*[281] applied exome sequencing in a family diagnosed with FSGS, identifying a collagen variant (in *COL4A5*), changing the diagnosis to Alport disease (AD). AD has addition features such as hearing loss, as well as differing prognostic implications[281]. Following this case, it was decided to investigate the distribution of pathogenic variants in a large cohort of adults with a clinical diagnosis of FSGS.

9.2 Methods

83 patients with primary FSGS or SRNS registered with the Wessex Kidney Centre (catchment population two million) were recruited, after prioritizing those within the

registry with a family history (FHx) of renal disease. Clinical data were recorded from patient interviews and clinical records.

A gene panel containing 39 genes was designed on the Illumina TruSeq Custom Amplicon (TSCA) platform using the vendor DesignStudio software for an amplicon length of 250. Gene coverage was optimized by the manual adjustment of thresholds in problematic regions, as well as division of the panel into two kits to avoid unfavourable amplicon–amplicon interactions. The genes for inclusion on the panel (Table 9.1) were chosen based on a comprehensive literature review and information from the Human Gene Mutation Database (HGMD) Professional 2013.1 and 2013.3[229]. 22 additional SNPs were targeted to cover all 24 SNPs in the exome sample tracking panel[184] as two SNPs in *COL4A4* and *NPHS2* were already targeted. The final designed panel comprised two TSCA kits of 1,093 and 381 amplicons covering 137.2 and 45.9 kb respectively.

Sample processing was performed similarly to as described in Chapter 7 except where noted. gDNA was captured using the TSCA kits independently and pooled prior to sequencing on two lanes of the Illumina MiSeq, with a read length of 150 bp paired end. Per-base coverage of genes was calculated using BEDTools[195] and collated using custom scripts. All variants deemed potentially pathogenic with a read depth of < 80 were validated by Sanger sequencing. Putative splice variants within 10 bp of the intron–exon boundary were evaluated using *MaxEntScan*[211]; variants with a differential score of $|\geq 3|$ were deemed to be likely to disrupt splicing.

Recommendations by the American College of Medical Genetics (ACMG)[282] were followed to allocate variants into the categories 'definitely pathogenic', 'probably pathogenic', and 'possibly pathogenic'. Definitely pathogenic variants were listed in HGMD, consistent with the phenotype, and individually assessed to establish the strength of evidence for pathogenicity in the literature. Probably pathogenic variants included novel splice site, frame shift and nonsense variants, and variants listed as disease-causing in HGMD with insufficient or conflicting evidence in the literature to determine their definite pathogencity. Possibly pathogenic variants consisted of nonsynonymous variants with an AF < 0.05. Variant zygosity had to match its known pattern of inheritance, and be present in all affected relatives in the panel to be considered disease-causing.

Clinico-pathological parameters were compared between patients with pathogenic collagen variants and the remaining cohort. Statistical significance was determined by the χ^2, Fisher's exact test, or Mann-Whitney U test as appropriate, using SPSS v21 (IBM, Armonk, NY).

9.3 Results

82 patients with FSGS and one with SRNS were recruited into the study, with a median age at presentation of 37 (range 0–84); 61% male; all but 12 patients presented in adulthood. 75% of the cohort had progressed to end-stage renal disease, requiring renal replacement therapies (RRT) such as dialysis and transplantation. The cohort included nine individuals from within the region belonging to four families. To the best of our knowledge, the remaining individuals were unrelated, resulting in 76 independent families. All but two patients were Caucasian (one Black African and one Asian). The diagnosis of FSGS was based on eponymous biopsy findings in combination with proteinuria; except in five patients, where no biopsies were taken, and two patients with minimal change disease on biopsy. Their diagnosis of FSGS was supported by biopsies in similarly affected relatives and/or the clinical picture. A diagnosis of 'familial FSGS', requiring the diagnosis of FSGS in at least one relative, was established in 12 individuals from eight families.

98.9% of the coding region was targeted successfully across 39 genes (Table 9.1). Following sequencing and alignment, $> 94\%$ of the targeted region was covered to a depth of at least 10X, with a mean depth of > 300X (Table 9.1).

Table 9.1: Genes included in panel design and proportion successfully targeted

					% covered to median depth			
Gene	**Chr**	**Exons**	**Size (bp)**	**% Targeted**	**20X**	**30X**	**50X**	**100X**
ACSL4	X	15	2364	99.6	93.1	93.1	78.7	68.8
ACTN4	19	21	2736	99.5	91.5	86.4	72.2	48.3
ALG1	16	13	1395	91.2	98.2	97.8	95.5	94.7
APOE	19	3	954	99.9	96.4	96.4	87.9	77.8
APOL1	22	7	1289	99.8	67.0	66.7	66.7	48.5
ARHGAP24	4	12	2478	99.2	99.5	98.2	91.4	71.8
ARHGDIA	17	6	683	100.0	100.0	100.0	100.0	95.9
CD2AP	6	18	1920	99.2	92.4	92.4	92.2	58.2
CFH	1	23	3710	94.1	96.2	95.8	88.4	68.1
COL4A3	2	52	5013	99.1	98.5	95.6	93.9	90.0
COL4A4	2	47	5073	99.1	98.4	97.2	82.5	56.2
COL4A5	X	53	5383	99.2	83.4	74.1	64.6	34.4
COQ2	4	7	1266	99.5	100.0	99.2	90.8	68.2
COQ6	14	13	1495	99.5	99.7	94.9	90.0	80.3
INF2	14	23	3817	99.7	87.4	87.4	86.8	78.3
ITGB4	17	40	5628	99.6	97.0	97.0	95.0	92.0
LAMA5	20	80	11088	99.8	87.6	82.4	70.9	44.4
LAMB2	3	32	5397	99.9	99.0	99.0	95.8	81.6
LMNA	1	18	2452	91.4	94.6	94.6	94.6	94.6
LMX1B	9	10	1453	99.8	71.9	71.9	66.7	53.5
MYH9	22	40	5883	99.5	95.9	93.6	85.3	64.6
MYO1E	15	28	3327	99.2	98.2	95.9	90.9	66.2
NEIL1	15	11	1865	87.1	76.5	73.1	73.1	64.1
NPHP4	1	31	4505	99.4	92.0	86.9	77.9	53.9
NPHS1	19	29	3726	99.7	98.8	98.8	98.2	92.5
NPHS2	1	8	1152	99.5	84.5	80.4	80.4	73.9
NXF5	X	14	1098	99.5	83.9	83.5	75.2	64.6
PDSS2	6	8	1200	99.4	97.5	97.5	94.0	80.2
PLCE1	10	33	7300	99.6	97.6	97.6	97.1	86.0
PMM2	16	8	741	99.2	83.7	81.2	74.8	65.0
PODXL	7	9	1677	94.0	98.6	98.6	98.6	87.9
PTPRO	12	27	3655	99.3	100.0	100.0	93.1	79.6
SCARB2	4	12	1437	99.2	100.0	100.0	98.5	84.4
SMARCAL1	2	16	2865	99.5	99.6	99.2	94.8	86.2
SYNPO	5	5	7560	100.0	94.6	94.6	94.3	84.6
TRPC6	11	13	2796	99.6	97.2	95.6	95.6	86.1
WT1	11	12	1648	99.5	71.1	68.1	55.9	49.8
ZEB1	10	11	3445	99.8	98.3	96.7	93.1	87.8
ZMPSTE24	1	10	1428	99.4	100.0	100.0	100.0	85.6

562 on-target variants across the 39 genes were identified in the 83 patients. After filtering for functional effects and AF, 266 variants remained (Figure 9.2). 17 definitely pathogenic variants, five probably pathogenic variants, and 242 possibly pathogenic

variants were identified. The participant-centric Table 9.2 shows only those definitely or probably pathogenic variants which occurred in the zygosity reported to be disease-causing.

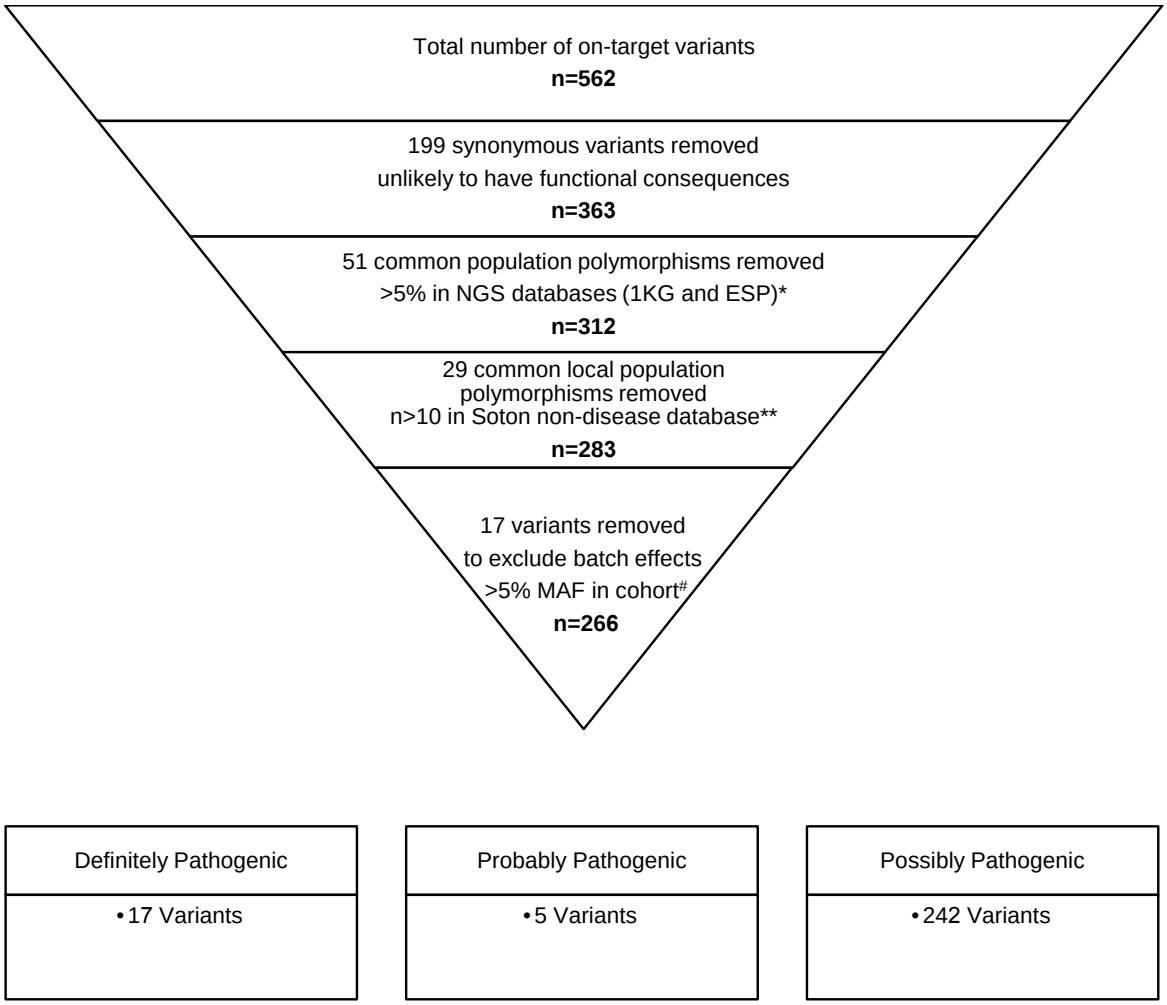


Figure 9.2: Variant attrition throughout filtering in FSGS cohort. *The 1000 Genomes Project (1KG) and Exome Server Project (ESP) are large genomic databases including populations with European ancestry. Variants with AF of $> 5\%$ in these databases were excluded from further study as they are likely to be common population polymorphisms with no significant functional consequences. **Variants found to be common in our own genetic non-disease database of unrelated whole-exome sequenced individuals ($n = 292$). #Variants with AF in the cohort of $> 5\%$ were excluded as they were likely to represent common local population polymorphisms or batch effects (artefacts).

Definitely pathogenic variants were found in 14 patients from 12 of the 78 families (Table 9.2). In order to establish diagnostic rates for a pure adult FSGS/SRNS cohort, we excluded a family with previously suspected AD, and patients with nail patella syndrome and congenital nephrotic syndrome from the statistics, leaving 75 families. Thus we achieved molecular diagnoses in 12% of families and 13% of the case series. Definitely and probably pathogenic variants combined were identified in 16 patients from 15 families, giving a diagnostic rate of 20%. Diagnostic rates for patients with and without FHx, and according to the age of disease onset are shown in Table 9.3.

Table 9.2: Aetiological variants identified in FSGS cohort

Participant	Path	Chr	Gene	Variant(s)	Inheritance	Sex	FHx[c]	Clinical Diagnosis	Age RRT	Age enrolled
F1 mother	def	X	*COL4A5*	p.G1170S[b]	X-linked	F	2.5	FSGS[d]	39	64
F1 son	def	X	*COL4A5*	p.G1170S[b]	X-linked	M	2.5	FSGS[d]	35	45
F2 brother[a]	def	2	*COL4A3*	p.[G818R;L1474P]	Recessive	M	1	likely AD[d]	-	33
F2 sister[a]	def	2	*COL4A3*	p.[G818R;L1474P]	Recessive	F	1	likely hereditary nephritis[d]	-	42
I1	def	X	*COL4A5*	p.G325R[b]	X-linked	M	3.75	FSGS[d]	36	49
I2[a]	def	9	*LMX1B*	p.W266C	Dominant	F	2.5	FSGS with NPS[d]	46	47
I3[a]	def	1	*NPHS2*	p.[R138Q;R138Q]	Recessive	F	2	FSGS[d]	6	48
I4	def	2	*COL4A4*	p.S969X	Recessive	F	2	FSGS[d]	-	43
I5	def	X	*COL4A5*	p.G1170S	X-linked	F	1.5	FSGS	64	66
I6	def	14	*INF2*	p.R218Q	Dominant	F	1	FSGS[d]	-	36
I7	def	2	*COL4A3*	p.L1474P	Recessive	M	0	FSGS[d]	57	66
I8	def	1	*NPHP4*	p.[R1192W[b];R848W]	Recessive	M	0	FSGS[d]	-	28
I9	def	6	*CD2AP*	p.K301M[b]	Dominant	F	0	FSGS[d]	30	34
I10	def	22	*MYH9*	p.M1651T	Dominant	M	0	FSGS[d]	60	66
I11	prob	20	*LAMA5*	p.G3685R[b]	Dominant	M	0.5	FSGS[d]	-	28
I12	prob	20	*LAMA5*	p.G3685R	Dominant	M	0	FSGS[d]	59	66
I13	prob	19	*ACTN4*	p.V801M	Dominant	F	0	FSGS[d]	30	47
I14	prob	11	*WT1*	c.1432+1G>C	Dominant	F	0	SRNS	16	44
I15	prob	14	*INF2*	c.1735+2T>G	Dominant	M	0	FSGS[d]	40	52
I16	prob	X	*NXF5*	c.860+2T>C[b]	X-linked	F	0	FSGS[d]	67	67

[a]Excluded from statistics due to prior clinical diagnosis.

[b]Confirmed by Sanger sequencing.

[c]FHx score is defined as $2\sum\Phi$ for all relatives with a similar clinical diagnosis

[d]Biopsy proven.

Table 9.3: Diagnostic rates in sub-cohorts

	Definitely pathogenic	**Probably & definitely pathogenic**
Total (unrelated)	$\frac{9}{75}$ (12%)	$\frac{15}{75}$ (20%)
Total	$\frac{10}{79}$ (13%)	$\frac{16}{79}$ (20%)
With FHx	$\frac{5}{23}$ (12%)	$\frac{6}{23}$ (26%)
Without FHx	$\frac{5}{52}$ (10%)	$\frac{10}{52}$ (19%)
Adult onset	$\frac{6}{69}$ (13%)	$\frac{14}{69}$ (20%)
Infantile onset	$\frac{1}{1}$ (100%)	$\frac{1}{1}$ (100%)
Childhood onset	$\frac{1}{3}$ (33%)	$\frac{1}{3}$ (33%)
Adolescent onset	$\frac{0}{7}$ (0%)	$\frac{1}{7}$ (14%)

9.3.1 Patients with collagen variants

Eight participants from six families with disease-causing *COL4A* variants were identified (Table 9.3), including one family with previously suspected AD (F2). Excluding this family, collagen variants represented 56% of all definitely pathogenic variants in the pure FSGS/SRNS cohort. The discovered *COL4* variants confirmed the diagnosis of AD in two patients (F2 brother and sister), changed it to AD in four (F1 mother and son, I1 and I5), and TBMN in two patients (I4 and I8). All participants had heavy proteinuria, resulting in nephrotic syndrome in I1.There was no documented microscopic haematuria in F1 mother, I1 (male) and I5 (female). Only one participant presented with hearing loss (F2 brother). In two others, hearing loss developed post-transplantation and was attributed to external factors.

Ophthalmic tests are only documented in I1 and were normal. Light microscopy showed FSGS in all biopsied participants, with the exception of F2 sister (normal light microscopy). As shown in table 6, F1 mother's first EM was normal, the second revealed glomerular basement membrane (GBM) lamellation possibly compatible with AD, but she had no associated clinical features or FHx at the time. The EM in F1 son was not diagnostic, but showed widespread podocyte foot process fusion, lamellation and splitting of the GBM. Clinical testing for AD was arranged at the time, but not completed. The two participants with single *COL4A3/4* variants (I4 and I8) had microscopic haematuria (with a FHx in I4), nephrotic range proteinuria, and FSGS on light microscopy. EM in I4 was normal.

9.3.2 Patients with non-collagen aetiological variants

Pathogenic variants in *CD2AP* and *INF2* causing autosomal dominant disease were found in patients I6, and I9; both had biopsy-proven FSGS. I2's *LMX1B* variant causes FSGS and congenital nail patella syndrome, matching her diagnosis. I3's homozygous

podocin (*NPHS2*) variant p.R138Q was responsible for FSGS in infanthood and the death of two siblings from SRNS (parents unaffected). Compound heterozygosity for two *NPHP4* variants in I8 with sporadic biopsy-proven FSGS and nephrotic range proteinuria changed his diagnosis to nephronophtisis. The *MYH9* variant p.M1651T in I10 with sporadic FSGS and intermittent macroscopic haematuria at presentation causes May Hegglin anomaly, characterized by platelet anomalies with the possible development of renal failure. No platelet abnormalities were noted in I10, but the clinical picture was confused by recurrent bleeding on anticoagulants, requiring multiple blood transfusions.

9.3.3 Patients with probably pathogenic variants

Probably pathogenic splice-site variants in *INF2*, *NXF5* and *WT1* were discovered in three individuals (I14, I15 & I16), all with biopsy-proven FSGS. We took the conservative approach of classifying the nonsynonymous variants in *LAMA5* in I11 and I12, and in *ACTN4* in I13 as probably pathogenic, instead of pathogenic despite their listing in HGMD, due to either conflicting or insufficient evidence in the literature regarding their pathogenicity.

9.3.4 Variants in families

No reported pathogenic variants were identified in two of the four families on the panel. F4 sisters 1 and 2 both have the novel *LAMB2* variant p.D181N, which would be expected to be recessive. No common variants were found for F3 father, daughter 1 and daughter 2. Apart from two pathogenic *COL4A3* variants, F2 brother and sister also share the *ACTN4* variant p.R310Q and the *APOL1* variant p.S324G.

9.3.5 Clinical characteristics associated with pathogenic variants

Clinical and histological features were analyzed for patients with pathogenic collagen variants, compared to the remaining patients (Table 9.4). Differences in gender, proteinuria, age at RRT, RRT requirement, renal transplantation, and biopsy findings were not significant. *COL4* variant patients were more likely to have FHx, haematuria, GBM abnormalities, and younger age at presentation.

Table 9.4: Clinical feature comparison between patients with identified variants

	COL4A	**Non-*COL4A***		**No variant identified**	**Significance level**
		Definitely pathogenic	**Probably pathogenic**		
	$n = 8$	$n = 6$	$n = 4$	$n = 57$	
Male Gender	4/8	4/6	1/4	39/57	$p = 0.448$
Age presentation	5–56, median 23	23–53, median 27	15–54, median 30	2–81, median 41	$p = 0.029$
Age at RRT	35–64, median 39	30–60, median 59	16–67, median 35	10–85, median 52.5	$p = 0.744$
Protein:creatinine ratio	300–900, median 500	53–1,352, median 329	212–2,000, median 434.5	13–2,740, median 730.5	$p = 0.867$
Nephrotic syndrome	1/8	2/6	1/4	29/57	$p = 0.071$
Haematuria	5/8	2/6	0/4	9/57	$p = 0.009$
Hearing deficit age < 40	2/8	1/6	0/4	1/57	$p = 0.054$
Biopsy shows FSGS	6/7	6/6	3/3	54/54	$p = 0.1$
GBM abnormalities	3/4	1/2	0/2	5/26	$p = 0.041$
ESRD	5/8	3/6	4/4	43/57	$p = 0.433$
Transplant	4/8	2/6	4/4	30/57	$p = 1$
Transplant recurrence	0/4	0/2	0/4	5/30	$p = 1$
FHx	7/8	2/6	0/4	17/57	$p = 0.001$

9.4 Discussion

We designed a customized NGS panel for the investigation of FSGS/SRNS, which is the first comprehensive FSGS gene panel in an adult cohort. Targeted panels have been recognized as a promising approach in the investigation of FSGS[283]. We have shown that this technique works well with excellent coverage of the targeted genes after manual optimization, and a high diagnostic rate. We resolved 12–20% of all FSGS/SRNS cases. This represents 22–26% of families with family history and 10–19% of those without. We solved 13–20% of adult onset cases, which is higher than the previously reported 8–14%, and explained by our use of NGS allowing the testing of a comprehensive gene panel.

Of strong clinical relevance is the frequency with which *COL4* mutations were found to underlie FSGS. Pathogenic *COL4* mutations were discovered in five of nine families (56%) with a definitely pathogenic gene mutation, and 7% of families in the cohort, representing the highest prevalence of any mutation. They were found in 38% of families with familial FSGS (3/8), and 3% of sporadic FSGS/SRNS (2/67). AD or TBMN had only been suspected in one family (F2) not included in the statistics.

Our prevalence of *COL4* mutations in familial FSGS is higher than the recently reported 10–12.5%[284], which is likely explained by our inclusion of *COL4A5*, where over half of our *COL4* mutations occurred. This is more consistent with the mutation distribution in AD[285]. Our cohort included more patients with sporadic than familial FSGS, thus also giving an estimated prevalence of *COL4* mutations in sporadic FSGS.

The simultaneous sequencing of 36 podocyte genes allowed us to rule out potential modifier mutations in *NPHS1/2*. The *ACTN4* variant p.R310Q in F2 brother and sister is thought to predispose to FSGS[286,287] and may have acted as a modifier. Theoretically others could have been missed by less than complete coverage and the unknown effect of novel variants.

We established six new diagnoses of AD in our FSGS cohort. AD can be difficult to diagnose due to variations in diagnostic features, both clinical and histological[288,289]. Diagnostic criteria for AD have been published recently[290]. They rely on the presence of (familial) haematuria with or without renal impairment, in combination with either characteristic EM biopsy changes, or specified *COL4* mutations—which established the diagnosis in all patients in our cohort. FSGS has been found to be the most common misdiagnosis in female patients with X-linked AD[289]. In these cases it can be difficult to distinguish biopsy changes of AD mimicking FSGS from a development of FSGS.

There is evidence that FSGS occurring in our patients with AD is more likely to represent FSGS phenocopy than merely the late development of FSGS. The first biopsy per family was taken 0–3 years after presentation when the excretory renal function was still (near) normal. GBM thickness was largely within normal limits in all four, with normal GBM appearance on F1 mother's first biopsy and FSGS on light microscopy.

Relevantly, the clinical features of the patients with *COL4* mutations were atypical for AD. Three patients (one male, two female) with *COL4A5* mutations had no documented haematuria. The absence of haematuria is reported in 5% of females, but 0% of males with X-linked AD[288]. It is possible that haematuria was present intermittently in male I1, but missed at clinical sampling, with his clinical records being incomplete following transfer from another unit. Hearing loss developed late in two of three males, and was not present in any female. This is compatible with 90% of males and 10% of females developing hearing loss before the age of 40[288]. The severe phenotype in F1 mother and female I5 with progression to ESRD occurs in 15% of female carriers[288], and is likely due to skewed X-inactivation[291].

Several pathogenic mutations were encountered in other patients in *CD2AP*, *INF2*, *LAMA5* and *ACTN4* reported to cause autosomal dominant FSGS of adult onset. We are only the second group to describe mutations in *LAMA5* in FSGS[292], with mouse models demonstrating *LAMA5*'s role in the formation and maintenance of the glomerular filtration barrier[293]. The mutation p.G6358R found in two of our patients was also identified by a previous targeted NGS panel[292]. Further work is needed to confirm the exact role of *LAMA5* in FSGS.

The discovery of two *NPHP4* mutations changed the diagnosis of I6 from FSGS to nephronophthisis, also known to be associated with biopsy findings of FSGS[294]. *MYH9* mutations causing May Hegglin anomaly can present with features similar to AD, as was the case in I8[295].

Assigning variants to the category "probably pathogenic" is fraught with difficulty[282]. The probably pathogenic splice-site mutations we identified in *INF2*, *NXF5* and *WT1* have a high ΔMaxEnt score, which indicates a high probability of disrupting canonical splicing. One variant in female I12 occurred in the same position as a published splicing change in WT1[296], known to cause isolated SRNS in females, and Frasier syndrome in males[297]. Without extensive functional studies, we cannot prove if all altered splice products are definitely pathogenic. The absence of a family history in some patients with presumed dominant pathogenic mutations may be explained by *de novo* mutations, incomplete penetrance, or a false negative family history.

We have performed confirmatory Sanger sequencing for all pathogenic variants with read depths below 80, confirming the accuracy of our NGS genotype calls, since targeted NGS has been shown to be equally reliable as conventional sequencing for read depths above 30[298]. Before any variants are reported back to participants, fresh blood samples will be sent to an approved National Health Service (NHS) laboratory for confirmatory sequencing to meet current diagnostic standards.

By comparing patient and biopsy characteristics between participants with and without pathogenic or collagen mutations, we demonstrated that these features poorly predict the underlying pathology. As could be expected, the presence of a positive FHx and younger age at presentation make a genetic aetiology more likely. When haematuria, hearing and GBM abnormalities are present, this can suggest an underlying *COL4* mutation. Our case histories confirm that EM can raise suspicion and guide genetic testing, and should become routine practice in all cases of FSGS[299]. A normal EM, however, should not give false reassurance as the typical changes have been found in only approximately 60% of AD[300].

The discovery of gene defects can have significant benefits for patients and their relatives, including genetic counseling, screening, avoiding unnecessary immunosuppression, and slowing the progression of renal disease through early treatment, with known benefits of early renin-angiotensin system blockade in AD[301,302]. The risk of graft loss in renal transplantation can be predicted as very low, due to disease recurrence for patients with podocyte mutations[303] or anti-GBM disease in AD[304].

Our targeted NGS panel produces fast, affordable and reliable results, verified by confirmatory sequencing. The cost was approximately €250 EUR per participant for all 39 genes, compared to conventional sequencing costs of €1,000 for a single collagen gene. Limitations of the technique are the intentionally absent coverage of intronic regions (also missed by conventional sequencing) unless these are actively included in the panel design, and being restricted to genes previously associated with the disease. Furthermore, the interpretation of novel variants remains challenging. We have chosen a conservative approach likely to under-diagnose pathogenic variants, rather than risk over-diagnosis. The identified 242 possibly pathogenic variants are likely to contain further disease-causing mutations and non-functional polymorphisms. This distinction cannot be clarified without extensive functional studies. The above reasons combined can explain why we did not identify pathogenic mutations in all of the investigated cases with family history.

In summary, NGS, as a targeted panel or whole exome sequencing, is an ideal approach for the genetic testing of FSGS with multiple possible underlying aetiologies.

We have demonstrated that not only *COL4A3/4*, but also *COL4A5* mutations should be considered in patients with FSGS, especially in the presence of a positive FHx, even if clinical and biopsy features are atypical.

Part III

Mapping of Linkage Disequilibrium

Chapter 10

Characterisation of LD Maps Generated from Whole-genome Sequencing Data

10.1 Background

Detailed analysis of the linkage disequilibrium (LD) structure of human populations has been vital for the successful mapping of many human disease genes, understanding mechanisms underlying genetic recombination and elucidating patterns of selection and population structure[32]. The development of array-based genotyping (ABG) panels of single nucleotide polymorphisms (SNPs) enabled genome-wide association studies (GWAS) to localise numerous genetic variants with roles in human disease. Recognition that the genome contains 'blocks' of low haplotype diversity[305] facilitated the selection of 'tagging' SNPs to enable cost-effective genotyping using panels of 500,000 to one million SNPs[306]. Extensive SNP genotyping enabled the International HapMap Project to characterise the LD structure of diverse human populations[32]. The first LD maps of human chromosomes showed a haplotype block structure punctuated by 'steps' aligning with recombination hotspots[122,307]. The strong alignment of linkage and LD maps confirms historical recombination as the major determinant of LD structure[106,122,308].

Array-based LD maps of human chromosomes contain regions with negligible apparent LD between adjacent markers, seemingly reflecting high regional recombination, which are not well defined in the maps. Service *et al.*[308] assessed the impact of increasing marker density in a number of these regions using ABG data and found that some, though not all, regions were resolved with increasing marker density. For chromosome 22, 53% of these regions were resolved using 27,060 *vs.* 9,658 SNPs. Differences between populations were apparent, with LD maps from isolated populations (therefore having more extensive LD) containing substantially fewer such regions. Tapper *et al.*[106]

constructed genome-wide LD maps using ~500,000 SNP genotypes from 60 HapMap samples with European ethnicity, identifying 3,144 poorly resolved regions genome-wide and estimated that ~40,000 markers per Morgan would be needed to fully characterise LD structure. Assuming the autosomal linkage map length is ~33 Morgans[309] this suggests that 1.3 million SNPs genome-wide would be sufficient to resolve these regions in this population. However, this assumes uniform marker spacing and LD intensity, whilst in reality much higher local marker density may be required for some of these regions. A particular difficulty exists for populations which have reduced LD due to extended population history, such as those from Sub-Saharan Africa, for which considerably higher marker coverage is required for complete coverage.

Given that whole-genome next generation sequencing (WGS) provides maximal genotype density, we consider the advantages of WGS-derived SNP genotypes for the characterisation of LD structure in different populations. We construct LD maps according to the Malécot-Morton model, using the program *LDMAP*[106,310]. This model is defined as:

$$\hat{\rho} = (1 - L)Me^{-\epsilon d} + L \tag{10.1}$$

where $\hat{\rho}$ is the association between SNPs, the asymptote L is the 'background' association between unlinked markers which is increased in small sample sizes and with residual population structure, M reflects association at zero distance, with values of 1 consistent with monophyletic origin and < 1 with polyphyletic inheritance, ϵ is the rate of LD decline, and d is the physical distance in kb between SNPs[122].

LDMAP constructs maps in linkage disequilibrium units (LDU, equal to ϵd) such that one LDU corresponds to the (highly variable) physical distance over which LD declines to background levels. LDU plotted against the chromosome location forms step-like patterns with intense breakdown in LD, canonically due to recombination hotspots, and plateaus for broader regions of low haplotype diversity (blocks). Overall LDU map lengths are proportional to time since an effective population bottleneck[308,311]. Hence, populations with shorter LDU maps have been founded more recently, experienced a more recent selective sweep, or have a smaller effective population size (such as some population isolates) compared to those with longer maps (such as Sub-Saharan African populations).

The close correspondence between LD patterns and the linkage map reflects the dominant role of recombination in LD structure. In contrast to linkage maps, which are derived from family data and describe recombination over recent generations, LD maps are constructed from population data and reflect the historical impacts of recombination, mutation, selection and population history. Our findings show that WGS based LD

maps provide greatly increased resolution of LD structure in both populations and indicate some genome regions in ABG-derived maps are incompletely covered. The findings have implications for interpretation in genome-wide association studies (GWAS) and support the use of WGS for association mapping and for establishing LD structure for studies of mechanisms underlying recombination and for identifying genomic regions subject to selection.

10.2 Methods

Publicly available 1000 Genomes Project[12] data derived from the Complete Genomics high depth whole-genome sequencing platform was used for WGS map generation[139]. WGS data for two population cohorts were used, namely the Utah Residents (CEPH) with Northern and Western European ancestry (CEU; 96 individuals), and Yoruba in Ibadan, Nigeria (YRI; 80 individuals). For comparison, array-derived HapMap Phase 3 release 3 data were also used[35]. ABG cohorts used were CEU (112 individuals), and YRI (147 individuals) samples. All individuals utilised for map generation were founders, and physical positions were defined according to GRCh37 (hg19) coordinates.

We consider here the region Chr22:20,000,000–51,304,566. The centromeric heterochromatin was excluded as these regions show very low density of polymorphic makers and complete LD, as well as a tendency for erroneous genotyping due to the repetitive nature of the sequences. Genotype data were filtered prior to map generation using *PLINK*[182] or *VCFtools*[164] to remove non-biallelic SNPs, SNPs with MAF within the dataset < 0.05, SNPs with Hardy-Weinberg equilibrium deviation p-value < 0.001[312] and SNPs with $> 5\%$ missing data. All statistical analyses were performed using *R*[313].

LD map generation was performed using the *LDMAP* program, with default parameters[106,310]. For sample size reproducibility investigations, random subsets of the full cohort were generated and LD maps generated from the resulting dataset for three regions (Chr22:20,000,000–25,000,000, Chr22:30,000,000–35,000,000 and Chr22:45,000,000–47,000,000; 12 Mb total size) with 20 pseudoreplicates generated for each region. We restricted these analyses to 12 Mb of the chromosome due to the computational intensity of LD map generation. Following subsampling, filtering and LD map generation with a range of sample sizes, a negative exponential cumulative model was fitted to the marker density data for each population and extrapolated to estimate sample sizes required for effective map saturation. We defined map saturation as the sample size at which an additional 10 individuals provides less than 1% increase in marker density.

We investigated regions of intense LD decline, which are canonically the product of high levels of historical recombination. Recombination hotspots are known to span just 1-2 kb[113,114]. For comparison of LDU maps we defined a hotspot as a region of maximum size 5 kb in which there was at least a one LDU change between two encompassed SNPs, as observed in previous studies[314]. Hotspots were deemed concordant between datasets if there was any physical overlap; these liberal definitions were required due to the differing marker composition and density of datasets.

10.3 Results

To investigate the impact of using WGS data for defining patterns of LD, we utilised publicly available WGS genotype data for chromosome 22 within the 1000 Genomes Project (henceforth referred to as the WGS dataset), and array-based genotype data from the International HapMap Project Phase 3 (henceforth the ABG dataset)[12,35]. Due to its small size, chromosome 22 exhibits the highest recombination intensity in the genome[106] whereby LD declines sharply with distance and the LD maps are thus particularly sensitive for demonstrating the impact of the increased marker density in WGS data. We analysed LD maps constructed from CEU (Utah Residents (CEPH) with Northern and Western European ancestry) and YRI (Yoruba in Ibadan, Nigeria) populations. These are representative of populations which have developed since the effective 'out of Africa' bottleneck (CEU) and Sub-Saharan Africans (YRI). SNP markers within these datasets were filtered as described in Methods; final marker counts for each are given in Table 10.1. A detailed breakdown of marker attrition through filtering is presented in Table 10.2.

Table 10.1: Number of individuals, component marker counts and LD map length using ABG and WGS data

		Individuals	Markers	Map Length (LDU)
ABG	**CEU**	112	15359	850.07
	YRI	147	16083	993.80
WGS	**CEU**	96	66704 (4.34)	1021.07 (1.20)
	YRI	80	91320 (5.68)	1569.46 (1.56)

Fold change *vs.* ABG data in parentheses.

Table 10.2: Marker counts throughout filtering for all datasets

	ABG				WGS			
	CEU		YRI		CEU		YRI	
	Count	FC[d]	Count	FC	Count	FC	Count	FC
Raw count	17938	-	18906	-	214399	-	279848	-
MAF[a]	15420	0.86	16142	0.85	74946	0.35	106910	0.38
HWE[b]	17923	1.00	18887	1.00	211048	0.98	275780	0.99
Missingness[c]	17872	1.00	18906	1.00	198911	0.93	258517	0.92
Final count	15359	0.86	16083	0.85	66704	0.31	91320	0.33

[a]Markers with minor allele frequency < 0.05 within the cohort excluded.

[b]Markers with a Hardy-Weinberg equilibrium deviation p-value < 0.001 within the cohort.

[c]Markers with $> 5\%$ data missing excluded.

[d]Fold change in comparison to the raw count for each filtering criterion in isolation.

10.3.1 LD map topography

LD maps produced using the ABG and WGS CEU datasets appear topographically highly similar when plotted, though with differing overall map lengths (Figure 10.1). Regions of concordant strong LD are apparent, seen as low gradient regions in the plot, as well as regions of weak LD, appearing as a steep gradient. In addition, both maps appear to have similar contours to the linkage map produced from European samples, with broad areas reflecting strong and weak LD/recombination[315]. It is noteworthy that there is an increased overall map length for the CEU WGS map compared to the ABG map (1.2 fold, Table 10.1). The change in map length is concurrent with much greater increases in marker density (4.3 fold) from ABG to WGS datasets.

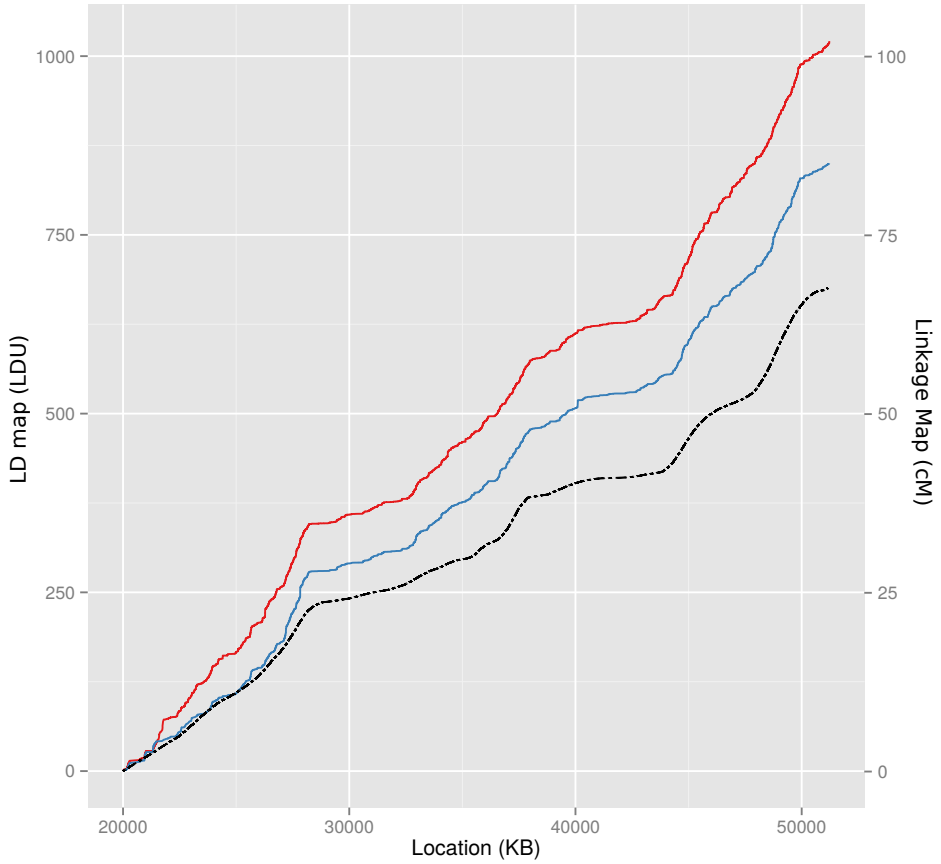


Figure 10.1: Comparison of WGS (red) and ABG (blue) CEU LD maps (left ordinate axis scale) and linkage map (black; right ordinate axis scale) for chromosome 22. Linkage map shown is from the June 2012 release of the Rutgers Map v3, interpolated using the Kosambi function (available at `http://compgen.rutgers.edu/download_maps.shtml`)[315].

LD maps for the two WGS populations also show close alignment in LD structure with broad shared regions of stronger and weaker LD. When the LDU maps are represented as a rate (LDU/kb) in 100 kb windows (Figure 10.2) the positions of the peaks, where LD declines rapidly, align closely between the two populations, as do regions with strong LD (low LDU/kb) . The much longer LDU map for the YRI population reflects population history with increased time to erode LD through recombination, mutation and other processes[311]. There is a particularly marked increase in length for the YRI map of 1.6 fold from ABG to WGS data sets (Table 10.1).

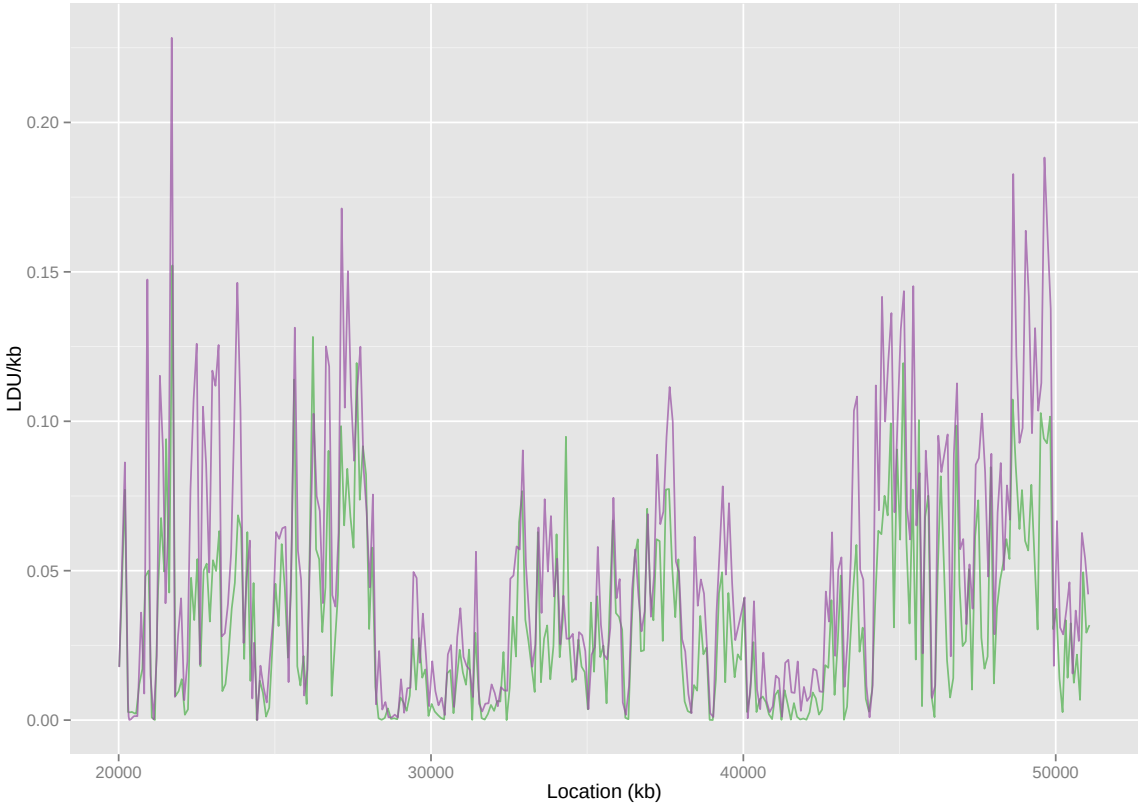


Figure 10.2: Comparison of regional rates of LD breakdown for CEU (green) and YRI (purple) populations using the WGS dataset for chromosome 22 for 100 kb windows. A very strong correlation between the LDU/kb for the two populations can be seen ($\rho = 0.91$, $p < 2.2 \times 10^{-16}$).

10.3.2 Marker density and frequency

The WGS data provides up to a 5.7 fold increase in number of markers compared to ABG data (Table 10.1 & 10.2). This increase in marker density allows greatly improved resolution of the LD maps in many regions. Although whole-chromosome LD map contours of ABG and WGS derived maps look very similar, noteworthy differences exist at higher resolution. Figure 10.3 shows an expanded view of a 250 kb region of the YRI population maps. The map of this region generated from the lower density ABG data failed to resolve 13 hotspots which are discernible in the WGS-based map. Many such narrow regions of high recombination can be far more accurately located using WGS maps.

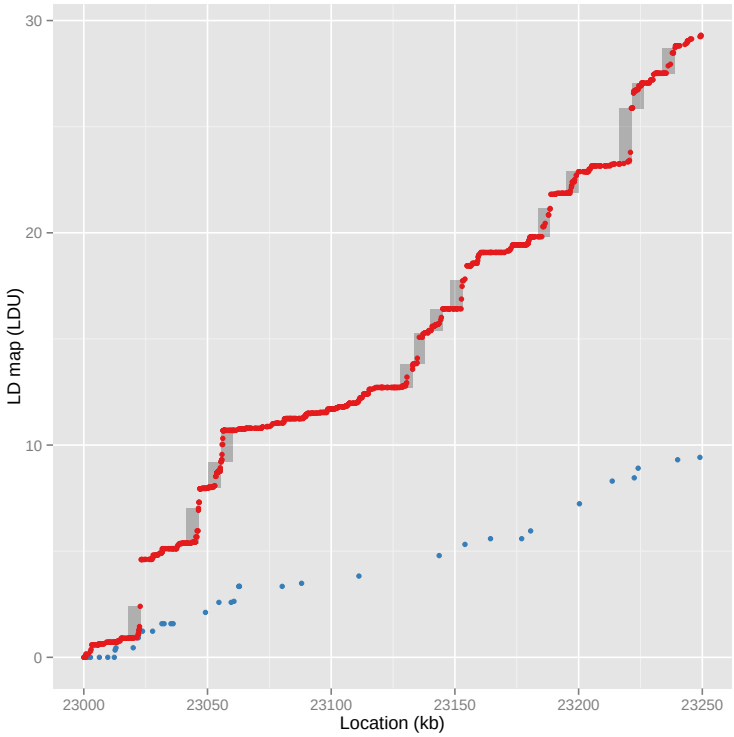


Figure 10.3: Fine detail comparison of WGS (red) and ABG (blue) LD maps for a 250 kb region of YRI chromosome 22. All markers are plotted individually; hotspots are highlighted in grey. Whilst 13 hotspots are identified within the WGS map for this region, the ABG map shows no hotspots.

As well as increased marker density in the WGS data, there is also a shift in the minor allele frequency (MAF) spectrum of the component markers (Figure 10.4). The WGS dataset shows a significant reduction in the median MAF compared to the ABG data ($p < 2.2 \times 10^{-16}$ for each population), with a far greater magnitude change in the YRI population compared to the CEU population (with a 35 and 18% reduction in median MAF respectively). These data illustrate that: 1) markers at the lower frequency end of the range are particularly underrepresented in the arrays used to genotype the HapMap samples; and 2) this underrepresentation is most pronounced for the YRI population.

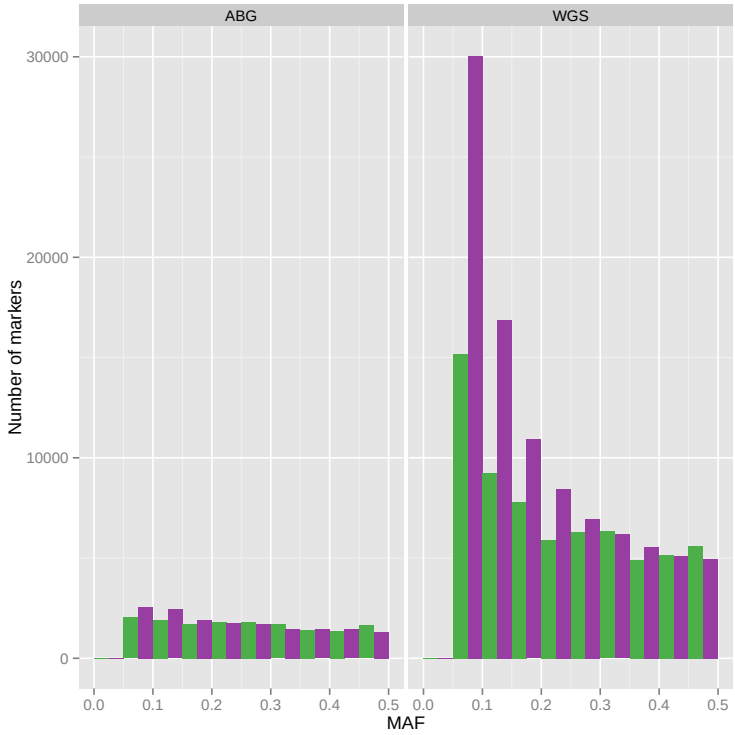
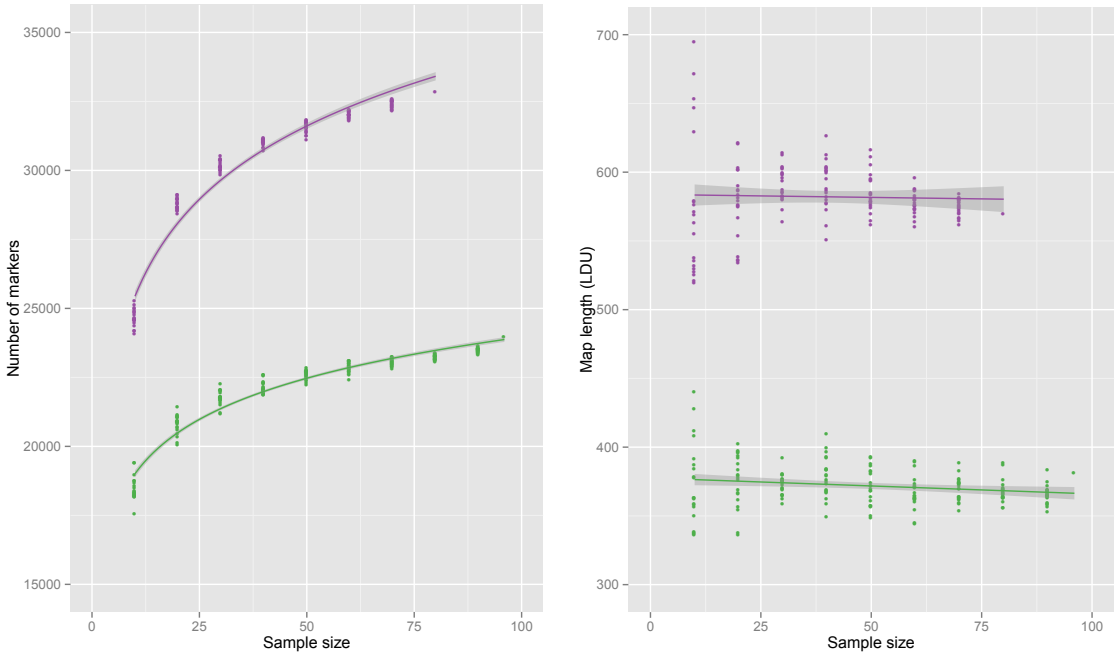


Figure 10.4: Histogram showing MAF distributions within ABG (left panel) and WGS (right panel) datasets for CEU (green) and YRI (purple) populations. A MAF bin width of 0.05 has been used. The median MAF for CEU is 0.25 and 0.21 for the ABG and WGS data respectively; the same metrics for the YRI are 0.23 and 0.15 respectively.

10.3.3 Effect of population size

We investigated the extent to which population sample size within the WGS datasets impacts the marker density available for map generation, as well as the length of the final LD maps. For 12 Mb of the chromosome we generated random subsets of the full datasets with varying sample size, and then performed marker filtering and map generation as described. With an increased sample size, a higher marker density is achieved for map generation, with diminishing returns with larger sample sizes (Figure 10.5a). From these data, we extrapolated the sample size for which the addition of 10 individuals increases marker density by $< 1\%$; this marker saturation is achieved with 90 and 110 individuals for the CEU and YRI populations respectively.



(a) Marker density **(b)** LDU length

Figure 10.5: Correlation between number of individuals sampled and number of markers (a) and LDU length (b) for a 12 Mb region, in the WGS data for CEU and YRI populations. For marker density, a negative cumulative exponential regression has been fitted ($r^2 > 0.94$, $p < 2.2 \times 10^{-16}$ in both populations. For LDU length, a linear regression has been fitted ($r^2 = 0.04$, $p = 0.0087$ for CEU, gradient is not significantly different from zero for YRI ($p = 0.69$). Shaded regions indicate 95% confidence intervals).

For maps from these data subsets, there is a weak, but significant, correlation between sample size and LDU length of the resultant CEU maps (Figure 10.5b); the YRI maps show no significant correlation. This indicates that overall map lengths are largely robust to variations in sample size. Due to the increased marker diversity of the YRI cohort compared to the CEU, a greater number of individuals need to be sampled for complete marker saturation. At smaller sample sizes however, the deviation of map lengths from average is much broader, reflecting increased sensitivity to heterogeneity within the dataset (Figure 10.6). Despite the increased map variability, the WGS map remains consistently longer than the corresponding ABG map. Even where maximal marker densities have been attained, larger sample sizes are likely to improve the population representativeness of the map.

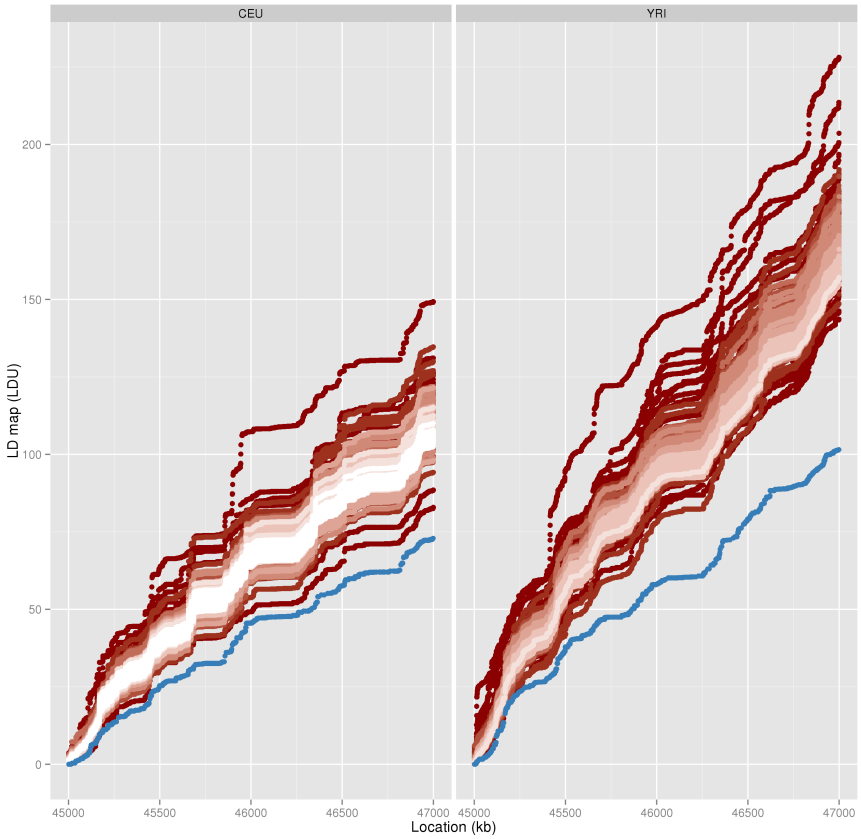


Figure 10.6: LD maps for a 2 Mb region constructed following random subsampling of WGS data for varying sample sizes (red to white with increasing sample size, range 10 to 90, with increments of 10 individuals) of both populations. For comparison, the ABG map is included (blue). Increasing variability in the WGS map can be seen in lower sample sizes, with the maps converging at larger sample sizes. Despite the increased variability at the smallest sample size of 10, the ABG map remains consistently shorter.

10.3.4 Fine map structure comparison between ABG and WGS

To compare LD structure between ABG and WGS maps we segmented the LD maps into non-overlapping 100 kb regions (Table 10.3). All LD maps show a very strong correlation with all other maps ($\rho > 0.87$), with stronger correlations within population.

Table 10.3: Spearman's rank correlations between LDU map lengths of 100 kb segments

CEU-ABG	**CEU-WGS**	**YRI-ABG**	**YRI-WGS**	**Linkage**	
1	0.92	0.88	0.87	0.56	**CEU-ABG**
	1	0.89	0.91	0.58	**CEU-WGS**
		1	0.94	0.60	**YRI-ABG**
			1	0.59	**YRI-WGS**
				1	**Linkage**

$p < 2.2 \times 10^{-16}$ for each correlation.

In all cases, the correlation with the linkage map is also strong ($\rho = 0.56 - 0.60$); this correlation is likely lower due to the lower resolution of the linkage map and components of the LD structure that are not due to recombination. We find a particularly strong correlation ($\rho = 0.94$, $p < 2.2 \times 10^{-16}$) in the lengths of these segments in LDUs between the two YRI data sources. The increase in LD map length for the WGS YRI map might be partly attributed to the greatly increased marker density, however there is only a relatively weak, though strongly significant, correlation between increase in marker density and increase in LDU length in these 100kb regions ($r^2 = 0.19$, $p < 2.2 \times 10^{-16}$; Figure 10.7). A total of 37.5% of 100 kb regions show negligible change in LDU length ($< |1|$) despite greatly increased marker density, suggesting a large proportion of the chromosome is approaching complete marker saturation in the ABG data. However, other regions show substantially increased LDU length (with many regions increased by over 5 LDU) with the higher marker density, suggesting they are poorly resolved in array-based maps.

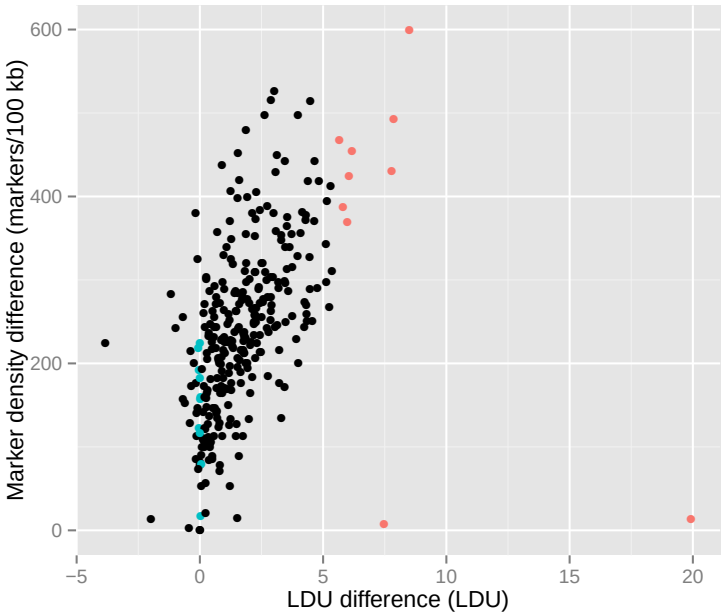


Figure 10.7: Scatter plot showing change in LDU *vs.* change in marker density for 100 kb regions between ABG and WGS map from YRI datasets. The 20 regions selected for further analysis as regions of largest magnitude change (red) and those with minimal length change (blue) are shown. Note that two of the selected regions span 23,000–23,200 kb, shown in Figure 10.3. A total of 312 regions were assessed in total.

The 100 kb regions in the YRI data which exhibit the largest and smallest magnitude LDU length change (10 of each; Figure 10.7) between ABG and WGS maps were further investigated. Regions with large LDU increase in the WGS data contain SNPs with a significantly higher MAF than regions with a small change ($p = 5.7 \times 10^{-7}$, median of 0.18 and 0.13 for the large and small magnitude change regions respectively), no significant difference between the MAF distributions of these regions was observed in

the ABG data ($p = 0.39$). This indicates that while there is particular enrichment of lower frequency markers using the WGS data, it is the inclusion of common variation absent from array panels which has the largest effect on the resulting LD map. The exclusion of highly LD informative common variation in array-based panels may reflect the ascertainment of tagging SNPs which is not optimised for all populations.

10.3.5 Hotspot identification

The LD landscape is known to comprise long regions of low haplotype diversity punctuated by very narrow regions of LD breakdown which align with recombination hotspots. WGS-based maps allow for more complete resolution of recombination hotspots compared to ABG-based maps (Figure 10.3). We therefore systematically evaluated hotspots identified in the four LDU maps. We defined hotspots as five kb regions containing SNPs which were separated by at least 1 LDU. In both populations, the WGS derived maps delimit a substantially increased number of hotspots (Table 10.4). The CEU maps show a 1.7 fold increase in resolved hotspots, compared to 2.8 fold increase in the YRI maps. This indicates that array-based genotyping only partially resolves the LD structure in both populations and resolution is particularly incomplete for the YRI population.

Table 10.4: Counts of hotspots in each dataset with corresponding hotspots identified in all other datasets

		ABG		**WGS**	
		CEU	**YRI**	**CEU**	**YRI**
ABG	**CEU**	**170**	86 (0.51)	137 (0.81)	119 (0.70)
	YRI	88 (0.50)	**176**	115 (0.65)	152 (0.86)
WGS	**CEU**	157 (0.53)	126 (0.43)	**296**	224 (0.76)
	YRI	149 (0.30)	187 (0.38)	244 (0.50)	**491**

Values shown indicate the number of hotspots in the dataset indicated with the row label with a corresponding hotspot(s) in the dataset indicated with the column label. Proportion of total hotspots recapitulated is shown in parentheses.

We also assessed concordance between hotspots identified in the datasets (Figure 10.8). The majority of hotspots identified in ABG data were also identified in the corresponding WGS maps (81 and 86% for CEU and YRI maps respectively). However, for YRI only 38% of hotspots identified in the WGS map were also represented in the corresponding ABG map. Furthermore, only 13% of identified hotspots showed concordance across the four datasets, with 29% of all hotspots only observed in the YRI WGS map. Of the 170 CEU hotspots identified in the ABG map the YRI ABG map identifies only 50% while, in contrast, the YRI WGS map detects 70%. This indicates that relatively poor resolution of the LD structure in the YRI array-based map suggests misleadingly low

concordance between hotspot locations across the two populations. Leveraging WGS data will therefore enable more effective characterisation of LD structure for YRI, and other populations with an extended population history, for disease gene mapping and the functional analysis of genomes.

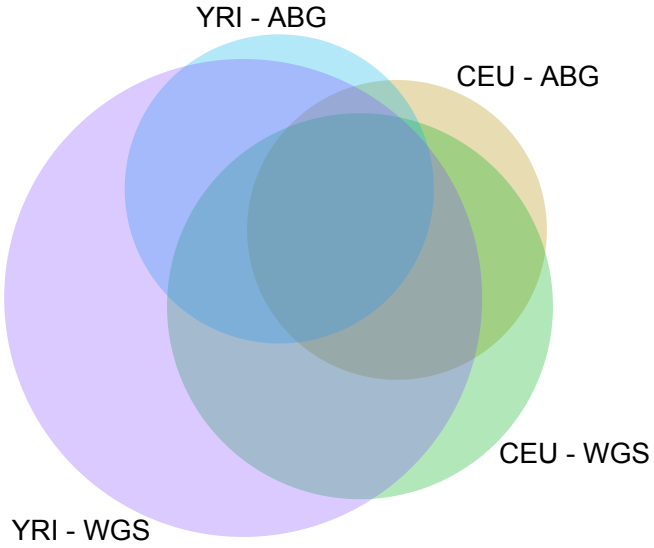


Figure 10.8: Euler diagram showing overlap between hotspots identified in each dataset. The area of all regions is proportional to the number of hotspots which are present in those sets; total area represents 629 independent hotspots across all datasets.

10.4 Discussion

We have shown that WGS-derived data enables superior resolution of LD structure in two populations with distinct histories. The increased marker density provides much improved delineation of regions of high and low recombination. Although some chromosome regions are well represented in array-based maps, population specific increases in map lengths of ~20-60% reflect improved WGS resolution of the LD structure in other regions. These seem likely to include regions highlighted as poorly characterised in earlier array-based maps[106,308]. Similarly, Lau *et al.*[316] observed a ~3% increase in map length when comparing maps generated from HapMap phases 1 and 2, with the associated increase in marker density.

We have shown that the YRI maps are improved by the greatest margin due to the inclusion of common variation excluded from the array-based genotyping panel. Array genotyping necessarily has a data acquisition bias; variants must be identified prior to array design, limiting the array capture to known variation which may be optimally informative for only the populations used for variant discovery. This ascertainment bias can cause issues in population genetic studies particularly where array data of a population not included in variation discovery is being investigated[317,318]. Recently developed arrays which include data from the three HapMap phases, along with variants

identified in the 1000 Genomes Project, achieve coverage of common variation of 92–93% for CEU but only 76% for YRI[319].

The evidence presented here indicates that the YRI LD structure is particularly poorly represented using array-based data, reflecting these unresolved biases in marker selection. While improvements in representativeness have been made, achieving good representation of all populations using ABG methodologies is intrinsically impracticable given technological and cost limitations on genotyping density. In contrast, using WGS there is negligible acquisition bias for variant discovery, though there can be bias where a population is highly divergent from the reference genome assembly; improvements in assembly and analytical tools should hopefully further reduce this bias in the near future[186]. Some regions are still however refractory to WGS analysis, such as repetitive regions, again, advances will continue to reduce these issues[320].

The total LD map length is relatively independent of number of samples. This indicates that although an increase in the number of homogenous individuals used in map generation improves accuracy, resolution and population representativeness, the underlying *LDMAP* algorithm provides robust maps with even small population samples as previously noted[310,311]. This may prove invaluable where the ascertainment of large data samples is impractical.

The high diversity of African populations, which reflects a much longer effective population bottleneck time, offers a rich resource for analysis of LD structure. Increased historical recombination makes sub-Saharan African populations ideal for GWAS studies, particularly for post-GWAS refinement, as well as for basic research into recombination biology and selection. Poor representation of African LD structure is considered likely to impact reproducibility of GWAS results. Marigorta & Navarro[321] investigated GWAS-derived disease variant reproducibility across 28 diseases. While most loci and SNPs discovered in Europeans have been extensively replicated in European and East Asian populations, replication in African populations is much less frequent. At least a proportion of these failed replications reflect heterogeneity in LD between causal variants and the tag SNPs used in GWAS panels so selection of alternative tags specific to the population used may improve reproducibility.

The incomplete resolution of LD structure in array-based LD maps which is evident even for the CEU population may have impacted the detection of disease variation in genome-wide association studies. With decreasing sequencing costs, WGS-based GWAS are becoming viable, with some successes reported[322]. These studies have the advantages of avoiding the marker ascertainment bias, and enable rare and common

variation to be interrogated contemporaneously. Such studies may improve GWAS reproducibility, as well as identification of additional disease variation underlying some of the 'missing heritability'[323].

LD maps have been used successfully in GWAS for refinement of candidate regions[111,324]. Sabatti *et al.*[324] defined regions of interest around nine newly identified disease genes underlying metabolic traits using a liberal four LDU window. Improvements in LD map resolution through the use of WGS data will substantially reduce the size of regions for targeted follow-up. To investigate the potential gains of using WGS-derived LD maps for fine mapping, we assessed the physical window size corresponding to four LDU for 172 GWAS association signals identified in European populations on chromosome 22[46]. We considered the physical distance between the two nearest markers up and downstream which are at least two LDU away from the GWAS signal SNP. For the CEU population map WGS-based four LDU windows were, on average, 17% smaller compared to the ABG map (262 *vs.* 316 kb respectively). Furthermore, if we presume these GWAS signals are reproducible in Sub-Saharan African populations, the average four LDU window is just 152 kb in the WGS YRI map, a further 42% reduction in candidate region size compared to the CEU WGS map.

Considerably greater resolution can be achieved in fine-mapping using a population with African ancestry by exploiting the weaker LD as has been recently demonstrated in African American populations[325]. African populations have been historically under-represented in population genetic studies but the African Genome Variation Project[326] is focussed on using whole-genome sequencing and other methods to refine the detection of disease variation in these populations. Construction of fully saturated whole genome LD maps from diverse African samples will undoubtedly improve efforts to map disease variants and help distinguish true population differences in genetic disease variation from those which have failed to replicate due to incomplete marker coverage in African samples.

We have herein discussed several improvements to LD mapping attained using WGS data. Firstly, WGS data allows complete resolution of LD structure, given the maximal marker density. Secondly, as there is no ascertainment bias in genotypes, the data are also far more representative of the population under study, particularly notable for Sub-Saharan African populations. Thirdly, data from a larger number of individuals is required to best interrogate LD patterns in diverse populations, particularly those with long population history. We have shown that array-based SNP panels incompletely represent the LD structure in both populations studied and this may have impacted the success of genome-wide association studies for detecting disease variation. Genome-wide

association studies using whole genome sequences may offer a route to capturing some of this additional variation.

Chapter 11

Evaluation of LD patterns between commercial chicken lines

11.1 Background

A detailed understanding of LD structure is essential for designing SNP genotyping arrays, successful association mapping of the genetic factors underlying traits of interest, establishing mechanisms underlying genetic recombination and elucidating patterns of selection and population structure. This is particularly true for commercial chicken (*Gallus gallus*) lines where LD analysis has the potential to establish the genetic mechanisms underlying selection and therefore contribute to further commercial development of lines.

The chicken genome comprises many chromosomes of varying properties, categorised primarily by size. The macrochromosomes (GGA1–5) span 50–200 Mb, intermediate chromosomes (GGA6–10) range from 20–40 Mb and 28 microchromosomes (GGA11–38) which average ~12 Mb[327,328]. The microchromosomes are characterised as having higher GC content, gene density and much higher recombination rates compared to macrochromosomes (~50–100 kb/cM versus ~300 kb/cM in macrochromosomes). The latter may reflect the requirement for a minimum of at least one chiasma for each chromosome per meiosis and a higher density of cohesin binding sites[329].

Previous studies of LD in the chicken have established that the micro chromosomes show reduced LD compared to macro chromosomes and these differences are almost completely explained by differences in the recombination rate[328]. Studies of egg laying chickens indicate higher levels of LD compared to broilers[330,331]. Despite relatively low levels of LD in broilers, Andreescu *et al.*[330] determined that there is significant overlap in LD for marker pairs across nine different commercial broiler lines.

Linkage disequilibrium maps constitute the LD analogue of the genetic linkage map and have been extensively utilised for human data[106,122]. LD maps are constructed from population data and reflect the historical impacts of recombination, mutation, selection and population history[109,311,314]. This approach to LD mapping has been previously successfully applied to other agricultural species, namely cattle[332]. Thus LDU maps of commercial chicken lines have the potential to provide new insights into patterns of recombination and selection.

Previous studies have begun to describe differences in recombination across *Gallus* genomes based on linkage and LD structure[328] and genome-wide LD maps have the potential to yield further insights. Here, we construct genome-wide LDU maps for three chicken lines types (broilers (BRO), white egg layers (WEL) and brown egg layers (BEL)) and contrast the LD structure across the three lines considering recombination hot spots, differences between chromosome types and motifs underlying major features of the maps.

Birds lack the zinc-finger protein PRDM9, required for recombination hotspot localisation in humans and other mammals[333]. Despite this, recent work by Singhal *et al.*[334] has shown that hotspots are highly concordant between wild populations of finch, due largely to the localisation of recombination to functional elements of the genome, namely CpG islands and transcription start sites (TSS).

Here, we construct genome-wide LDU maps for three chicken lines breeds (broilers (BRO), white egg layers (WEL) and brown egg layers (BEL)) and contrast the LD structure across the three lines considering recombination hot spots, differences between chromosome types and motifs underlying major features of the maps. High resolution mapping of LD in these commercial lines will facilitate array design to best capture the breed diversity with minimal data generation, which is of interest to commercial genome-led breeding operations.

11.2 Methods

Genotypic data used in this work are as reported in the validation populations of Kranis *et al.*[335], with all genomic coordinates based on the galGal4 reference assembly; 1,050,975 SNPs were genotyped passing initial QC in total. Only data from independent founders were included in these analyses. All pairwise samples were compared and wherever one individual of any pair showed $> 80\%$ genome-wide identity by similarity they were excluded. Individuals with $< 95\%$ genotyping completeness were also excluded. Filtering was performed using a modified version of *PLINK* v1.07[182] in

order to acomodate the additional chromosomes seen in the chicken. Multi-dimensional scaling (MDS), as implemented in *PLINK*, was undertaken using all autosomal markers in order to evaluate the population structure of the samples.

Once the three breed cohorts were defined, SNP marker filtering was undertaken independently for each population. Markers with $< 95\%$ genotyping completeness, MAF < 0.05 or Hardy-Weinberg equilibrium (HWE) deviation p-value of < 0.001 were removed to leave a dataset containing only common, high quality markers[312]. Within each breed the inbreeding coefficient (F) was calculated as detailed by Wright[336].

LD maps were generated for the assembled autosomes GGA1–28 on filtered data according to the Malécot-Morton model using *LDMAP*[122,123,310]. Where necessary, filtered genotype data were split into ~25,000 marker segments (with 200 marker overlap) to allow for parallelised processing. Overlapping map segments were then trimmed of the terminal 25 markers, and merged to form complete, contiguous whole-chromosome LD maps for the assembled autosomes. The order of markers in linkage maps from Elferink *et al.*[337] was revised in line with galGal4 from the native chicken assembly based upon SNP positions on this assembly within dbSNP 144. Following transition to the galGal4 marker order, a small number of markers in the linkage map were no longer sequential in the cumulative linkage map, and as such were manually removed.

To compare map structure between breeds, we focussed on the macrochromosomes GGA1–5, which were chosen to avoid confounding factors such as potentially incomplete reference assemblies, as well as varying recombination rates for the microchromosomes[327,329,337]. The Spearman's rank correlation of LDU lengths for all 40 kb regions between the three breeds was calculated (after Rubin *et al.*[338]).

For fine-scale interrogation of the LDU length of 5 kb regions, the concordance seen for the longest LDU regions, defined according to the top percentiles in order to allow for the differing extents of global LD for the breeds, was calculated. This analysis gives an indication of the extent to which narrow regions of intense LD breakdown are shared between pairs of samples. A large degree of concordance between long LDU segments would suggest there is a high proportion of shared recombination hotspots between the samples considered. As a final control, a randomised dataset was used for which an equal number of 5 kb regions were randomly selected independently for each dataset, and the concordance calculated; 100 pseudo-replicates were performed for each percentile cutoff.

For comparing LDU decline rates with genome features we focused on the BRO dataset due to the largest sample size. GC content was calculated directly from the

reference sequence for the 5 kb regions, CpG islands were defined according the UCSC genome browser, and Ensembl annotations were used to define TSS. *BEDTools* was used to calculate the distance between elements and regions[195].

11.3 Results

11.3.1 Input data

MDS of all samples shows distinct clustering of breeds, though with three distinct population clusters within each breed, corresponding to distinct commercial lines, and thus isolated populations (Figure 11.1)[335]. For LDU map construction population clusters were initially pooled within each breed. Counts of chickens and marker SNPs passing quality and frequency filtering are detailed in Table 11.1.

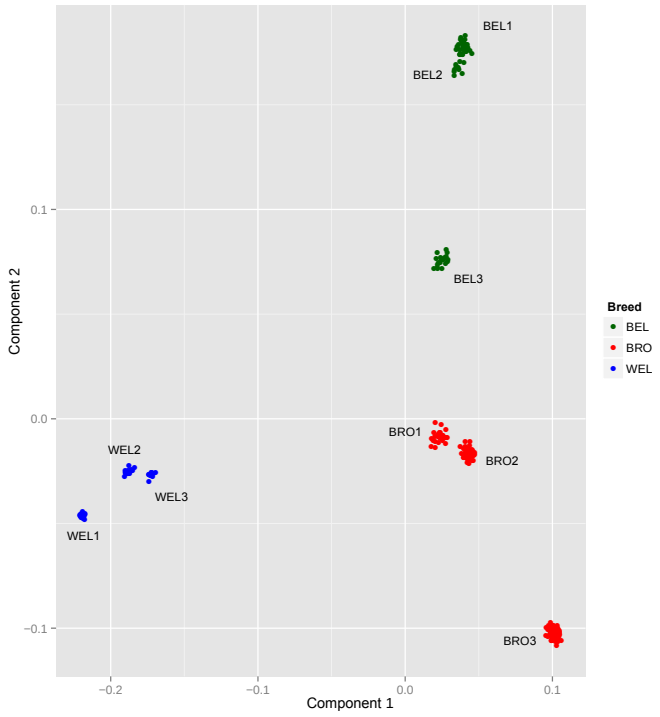


Figure 11.1: MDS for whole-genome genotype data for commercial chicken lines. Analysis includes 219 founder chickens. Chickens coarsely cluster within breeds, with three population clusters for each breed apparent, consistent with the three commercial lines genotyped for each breed. Population cluster designations are labelled on the plot.

Table 11.1: Number of individuals and component marker counts for analysed populations

		BRO	BRO3	BRO2[d]	BRO3a	BRO3b	BEL	WEL
Founders	**Males**	58	50	-	26	24	12	8
	Females	17	9	-	4	5	40	38
	Total	123[d]	59	48	30	29	52	46
SNPs	**Raw count**	966355	789359	790531	692467	778135	891200	691954
	MAF[a]	833639	638947	658548	631449	645713	796430	627294
	HWE[b]	760893	788284	789450	691625	777771	763931	420130
	Missingness[c]	966346	787732	788594	690038	776114	888903	690298
	Final count	630435	636535	655905	628382	643554	667605	354737

[a]Markers with minor allele frequency < 0.05 within the cohort excluded.
[b]Markers with a Hardy-Weinberg equilibrium deviation p-value < 0.001 within the cohort.
[c]Markers with $> 5\%$ data missing excluded.
[d]Sex data unavailable for BRO2 line.

11.3.2 Global map properties

LD maps were generated for all autosomal chromosomes for the three breeds (Figure 11.2). The physical map of the chromosome is represented on the x-axis, while the y-axis shows the LDU maps for each breed and and the linkage map in cM from Elferink *et al.*[337]. As found in human LDU and cM maps there is a large region showing little change in LD or cM, consistent with the location of the submetacentric centromere where recombination is suppressed and there is therefore intense linkage disequilibrium[106,339]. Summary length statistics for all autosomes are shown in Table 11.2.

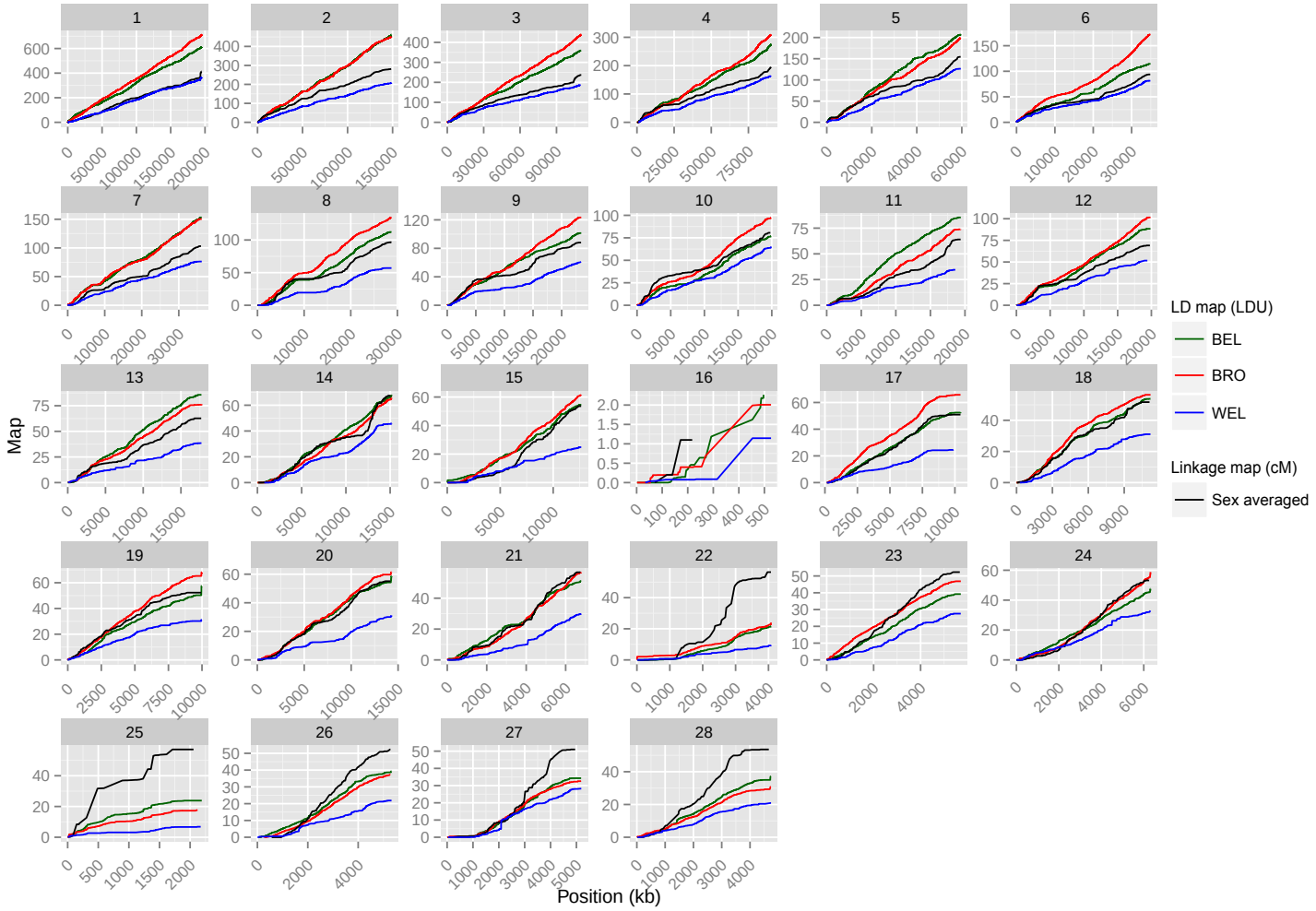


Figure 11.2: LD and linkage map plots for 28 autosomes of *G. gallus*. The broadly analogous structure of the linkage map and LD maps for the three populations can be seen. All maps contain a large plateau around 10,000 kb, corresponding to the centromere. Overall length of the LD maps is inversely related to the strength of LD within a breed. Broilers show the lowest LD overall reflecting relatively high haplotype diversity while white egg layers show strongest LD and lowest population haplotype diversity.

Table 11.2: Map lengths for autosomes of *G. gallus*

			BRO		BEL		WEL	
Chr	**Mb span**	**cM**	**LDU**	**LDU/cM**	**LDU**	**LDU/cM**	**LDU**	**LDU/cM**
1	195.2	413.5	713.7	1.7	612.4	1.5	363.8	0.9
2	148.8	281.3	452.1	1.6	462.8	1.6	207.8	0.7
3	110.4	236.9	439.2	1.9	359.4	1.5	186.3	0.8
4	90.2	195.2	309.8	1.6	277.1	1.4	164.1	0.8
5	59.5	154.4	198.6	1.3	207.1	1.3	126.3	0.8
6	34.9	93.8	171.9	1.8	114.7	1.2	81.4	0.9
7	36.2	103.1	150.3	1.5	153.1	1.5	76.3	0.7
8	28.7	96.6	134.5	1.4	112.2	1.2	57.0	0.6
9	23.4	88.1	123.4	1.4	101.4	1.2	61.0	0.7
10	19.9	80.6	97.4	1.2	77.0	1.0	65.3	0.8
11	19.3	64.0	73.9	1.2	85.4	1.3	34.5	0.5
12	19.9	69.1	101.4	1.5	88.4	1.3	51.6	0.7
13	17.7	62.7	76.3	1.2	85.7	1.4	38.5	0.6
14	15.1	67.4	65.9	1.0	68.1	1.0	46.2	0.7
15	12.6	53.6	61.6	1.1	54.6	1.0	24.8	0.5
16	0.5	59.1	2.0	0.0	2.3	0.0	1.1	0.0
17	10.3	50.9	65.7	1.3	52.6	1.0	24.5	0.5
18	11.2	51.7	56.4	1.1	53.9	1.0	31.1	0.6
19	10	52.3	67.9	1.3	57.6	1.1	31.3	0.6
20	14.2	55.1	61.5	1.1	58.6	1.1	31.1	0.6
21	6.8	56.9	56.8	1.0	51.1	0.9	29.8	0.5
22	4.1	56.4	21.9	0.4	21.7	0.4	9.5	0.2
23	5.7	52.3	46.6	0.9	39.2	0.7	27.6	0.5
24	6.2	53.2	57.6	1.1	47.6	0.9	33.0	0.6
25	2.1	57.1	17.8	0.3	23.8	0.4	6.8	0.1
26	4.9	52.3	37.4	0.7	39.5	0.8	21.9	0.4
27	5.2	51.0	32.6	0.6	34.3	0.7	28.2	0.6
28	4.7	53.6	30.4	0.6	37.1	0.7	20.9	0.4
Σ	**917.7**	**2762.2**	**3724.6**	**1.3**	**3378.7**	**1.2**	**1881.7**	**0.7**

LDU map lengths reflect haplotypic diversity within that population and can be compared with independent measures of population diversity such as F inbreeding coefficients[109,308,336]. The mean F inbreeding coefficients are 0.21, 0.26 and 0.51 for the BRO, BEL and WEL populations respectively, with the greater value for WEL indicating far more limited genetic diversity within the population. In comparison, the ratio of LDU/Mb is also variable between breeds (5.37, 5.13 and 2.85 LDU/Mb for BRO, BEL and WEL respectively across the autosomes). This ranking of the breeds by LDU length is consistent with the trend obtained from the F statistic, in line with expectations and previous literature[330,331,335]. This suggests that the array-based LD maps are appropriately estimating the population diversity.

In all breeds, the LDU length is also strongly correlated with the cM length of the linkage map ($\rho > 0.8$, $p < 2.5 \times 10^{-7}$ for all breeds); there is a breakdown in this correlation driven by particularly poor correlation for the smaller microchromosomes (Figure 11.3). There is also a general trend towards a lower LDU/cM ratio in the smaller chromosomes (Figure 11.3; Table 11.2). In humans, this ratio is consistently ~20 LDU/cM[106]. As expected, there is a very strong correlation between LDU and physical length of the chromosome ($\rho > 0.95$, $p < 2.5 \times 10^{-7}$ for all breeds); this is expected due to the d term in the Malécot-Morton model.

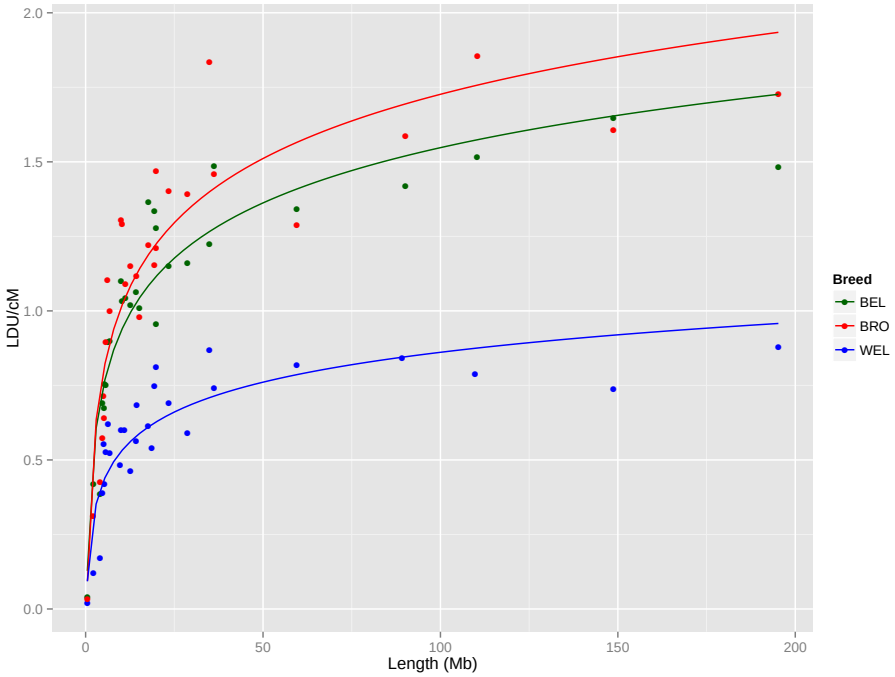


Figure 11.3: Relationship between physical chromosome length and LDU/cM ratio for all autosomes in the three breeds. There is a clear trend for the physically smaller chromosomes to exhibit lower LDU/cM ratios, with a negative exponential relationship. Lines indicate best fit for log_{10}(length) *vs.* LDU/cM ($r^2 > 0.75$, $p < 1.7 \times 10^{-9}$ for all breeds).

11.3.3 LD structure between breeds

Following map generation, we interrogated the fine map structure for the breeds, specifically to what extent patterns of LD were conserved between the breeds. There was a weak, though highly significant correlation between breeds for the LDU length of corresponding 40 kb regions ($\rho < 0.21$; $p < 2.2 \times 10^{-16}$ for all pairwise comparisons; Figure 11.4).

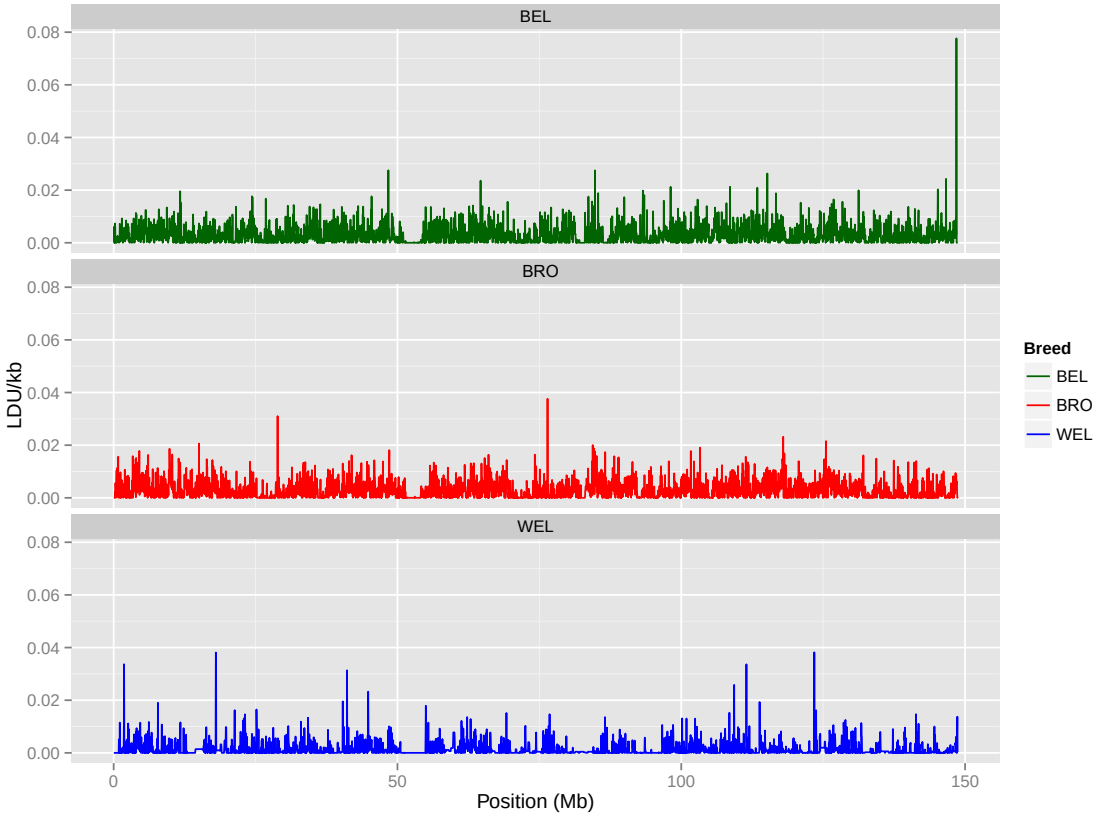


Figure 11.4: Comparison of LD breakdown intensity on GGA2 for the three breeds of *G. gallus*. LDU/40 kb is shown in sliding windows for the three breeds, a common region of high LD is seen following ~50,000 kb, corresponding to the centromere. There are minimal other trends apparent in the localisation of LD intensities between breeds.

Regions spanning a few kb in which there is strong breakdown of LD are known to align with recombination hotspots for which there is a high degree of concordance in, for example, human populations[114,115,128,340]. We investigated the extent to which narrow regions with LD breakdown are conserved across the three breeds. In humans, recombination hotspots span 1–2 kb[113], so we investigated whether 5 kb regions (in order to allow for the resolution of the genotyping array) with the longest LDU lengths within a breed were conserved between populations[106,109,308]. When comparing paired chicken breeds, there is a low concordance in the top LDU length percentile 5 kb windows, with ~5% concordance between breeds for the top 5 percentile (Figure 11.5). All breed pairwise comparisons show little concordance between LDU lengths although there remains greater concordance than expected in a randomised dataset.

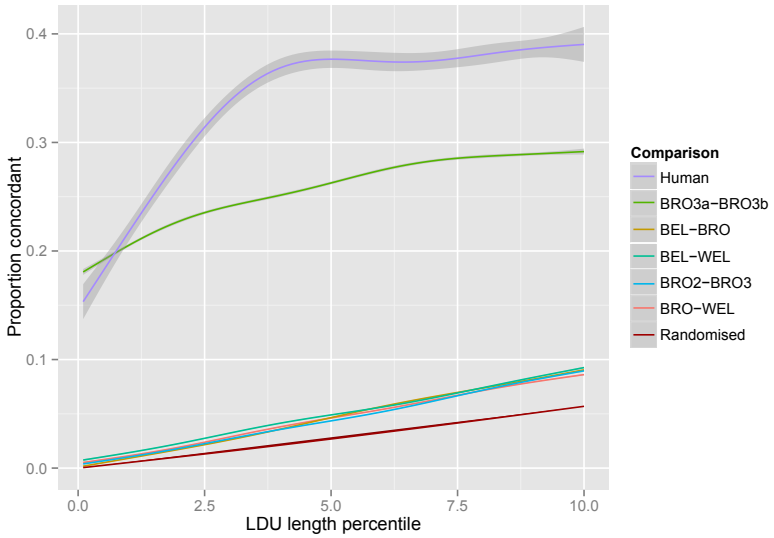


Figure 11.5: Pairwise concordance of regions of LD breakdown between populations. Shown is the proportion of regions in the top n percentile which intersect between the breeds. For pairwise comparisons between BRO/BEL/WEL ~5% of regions in the top 5 percentile are concordant. This proportion is also similar where two separate BRO lines (BRO2/BRO3) are compared. When the largest population is bisected (BRO3a/b), this proportion is ~27%, still far lower that the equivalent comparison between human populations[109]. All actual comparisons show a greater proportion of concordance than random, indicating some shared mechanism.

11.3.4 Characteristics of regions of LD breakdown

Despite the low concordance seen in the inter-population comparisons for the chickens, the concordance seen is consistently almost 2-fold greater than that expected by chance. One potential reason for this is expected biases in recombination rate dependent upon sequence context[329,334]. One key determinant of recombination rate, GC content[329], was found to be significantly increased in 5 kb regions in the top 1 percentile of any breed when compared to regions of 0 LDU length in all breeds (42.0% and 39.2% respectively, $p < 2.2 \times 10^{-16}$). We further compared LDU/kb for the 5kb regions with distance to the nearest of CpG islands and TSS; this was found to be highly significant, though weak, negative correlation in our data for both TSS and CpG islands, consistent with findings in the finches (Figure 11.6)[334].

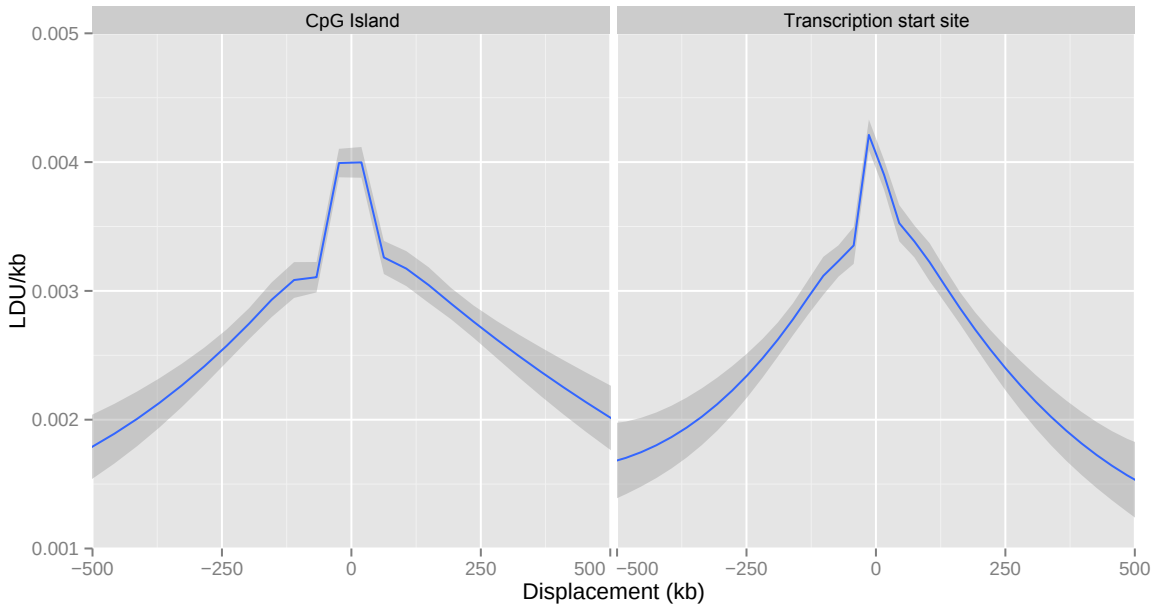


Figure 11.6: Association of LD breakdown with displacement from nearest functional element, namely CpG islands and TSS. There is a strong negative correlation between the distance from the functional elements and LDU/kb ($\rho = -0.12$ for CpG islands, $\rho = -0.10$ for TSS, $p < 2.2 \times 10^{-16}$ for both). Shown is the mean LDU/kb ratio for 5 kb bins, shaded area indicated 95% confidence interval.

In order to better characterise the relative contributions of TSS and CpG islands to recombination patterns we constructed a 2×4 contingency table for 5 kb regions exhibiting ≥ 0.003 LDU/kb against whether the regions are within 125 kb of a CpG island, TSS, both or neither (Table 11.3). These values were selected based upon the approximate points on inflection in Figure 11.6. There was a highly significant deviation from expected distributions under the null hypothesis ($p = 9.5 \times 10^{-224}$, χ^2 test). This shows that the increase in the number of regions with LDU/kb ≥ 0.003 where the nearest TSS and CpG island are both within 125 kb is greater than the sum of the increase where only one feature is within this range. This would indicate that it is an interplay of features which contribute to hotspot localisation, and that CpG islands have a greater effect than TSS.

Table 11.3: 2×4 contingency table of LDU/kb intensity and genomic features within 125 kb

	LDU/kb			
	≥ 0.003	**< 0.003**	**Odds ratio**	**Fold increase**
Neither	5255	12974	0.79	-
CpG only	1997	4208	0.88	1.12
TSS only	3746	8483	0.84	1.06
Both	33290	50836	1.08	1.37

11.4 Discussion

The analysis of LD maps for the three breeds indicates extensive LD genome-wide. Since one LDU represents the distance over which LD declines to background levels, the genome-wide Mb/LDU ratio gives an indication of the average physical extent of LD (termed the 'swept radius'). Figures for the three breeds are 246 kb for BRO, 272 kb for BEL and 488 kb for WEL (Table 11.2). In contrast, the corresponding figures for human populations are ~55 kb for human European populations and ~39 kb for African populations[316]. Although extensive LD is expected for chicken lines which have been subject to intense selection, profound differences in fine-scale LD structure between the three breeds are less expected.

Although some large scale genomic features such as centromeric regions which typically have extensive and intense LD are shared across breeds for some chromosomes (e.g. GGA2) there is relatively little concordance in LD structure genome-wide. The contours of the LD maps show many genome regions with widely divergent LD structure (Figure 11.4) and the overall correlation in LDU lengths of 40 kb windows is only $\rho = 0.21$. In contrast, the fine-scale LD structure of human populations is sufficiently concordant to support a 'cosmopolitan' LD map which recovers 91-95% of the information within population-specific maps[341].

The LDU/cM ratio of chromosome lengths is known to be virtually constant in human populations strongly suggesting that recombination is the primary determinant of LD structure[128]. However, for the three chicken breeds the linear relationship breaks down. Smaller chromosomes are shown to have a lower LDU/cM ratio. The breakdown in correlation between LDU and cM map lengths may be for several reasons. One possibility is that the linear relationship between LD and cM lengths breaks down under intense selection that has underpinned the three chicken breeds. If the breakdown in correlation is considered from a recombination standpoint it suggests that historical recombination intensity (based on LD maps) is lower than current recombination intensity (based on the linkage map) for the smaller chromosomes.

The possibility of complex interplay between historical and present day recombination intensity and selection is worthy of further study. Alternative possibilities recognise that the reference genome sequence is incomplete for several chromosomes, and particularly smaller chromosomes. Since the construction of linkage maps requires lower density markers and linkage extends much further than LD the construction of a complete linkage map of a chromosome is not highly sensitive to regions of missing or unreliable sequence[329]. In contrast LD is much shorter range and LD maps may be truncated in regions where SNP coverage is incomplete due to sequence gaps. Incomplete physical

maps for the smaller chromosomes may therefore contribute to truncated LDU maps as suggested by variable LDU/cM ratios[327]. This is however unlikely to be the sole explanation for the lower LDU/cM ratios in the smaller chromosomes due to the close negative exponential relationship between the physical chromosome size and LDU/cM ratio, indicative perhaps of an underlying biological mechanism as opposed to solely a technical artefact due to the incomplete assemblies.

Megens *et al.*[328] also found that recombination rates estimated from LD data were discordant with those obtained from the linkage map. Specifically they found that the recombination frequency for two microchromosomes (GGA26 and GGA27) estimated from LD was only 2.8 times greater than that of macrochromosomes (GGA1 and GGA2) when the expectation from the linkage map was 4.5 fold greater recombination on the microchromosomes[329]. This discrepancy was attributed to biases in fitting a model using effective population sizes computed in physical rather than genetic distance windows. The indication from this study that historical LD-based recombination rates appear discordant with the linkage map of different chromosomes is worthy of further investigation.

Our finding that the LD structure across the three breeds in highly discordant is in marked contrast to comparisons across human populations. Specifically narrow regions of LD breakdown which align with recombination hotspots in humans and are highly concordant across populations show little concordance across chicken breeds. Comparisons between major lines, and even between sub-populations with a major line (BRO2-BRO3, Figure 11.5) show alignment of such regions which is only slightly greater than 'random'. Concordance within a random split of a subpopulation (BRO3a-BRO3b) is much higher but even then does not approach the degee of alignment in the hotspot landscape human CEU and YRI populations. Although the different extent of LD genome-wide between the breeds has been known for some time[331] and the LD pattern between white and brown egg layers has been recognised as clearly different[342], this is the first study to recognise highly divergent fine-scale LD structure between breeds. This finding has implications for trait mapping since it suggests that to ensure coverage panels of tagging SNPs would be optimally selected only within breed and that, unlike in human analyses, a 'standard' linkage map may be less useful if it is not representative of the breed-specific recombination landscape.

Analyses of human LD maps have established that the recombination landscape can be recovered from LD structure[114,127,128]. From the derived recombination landscape the chromatin-modifying zinc-finger protein PRDM9 was shown to regulate recombination at 40% of human hotspots by binding to a degenerate 13 base pair motif[343]. Remarkably,

despite genomic similarity between humans and chimpanzees, there is virtually no sharing of recombination hotspot locations. Myers *et al.*[344] found that chimpanzee PRDM9 has a dramatically different predicted binding sequence. PRDM9 sequences are known to exhibit extremely rapid evolution which explains lack of hotspot conservation in other species which have PRDM9. However, chicken genomes, along with all other avian genomes tested (48 species) are known to lack PRDM9[333].

It may appear that our results herein are in direct conflict with the recent results of Singhal *et al.*[334], who showed that 73% of recombination hotspots were shared between the zebra finch (*Taeniopygia guttata*) and long-tailed finch (*Poephila acuticauda*). Like *G. gallus*, these finches lack PRMD9, a strong determinant of recombination localisation in humans and other mammals[115,334]. Due to this, Singhal *et al.*[334] posit that alternative binding motifs such as CpG islands and TSS play the analogous role of PRDM9 in humans. We see evidence of the clustering of low LD regions near these motifs, in agreement with them, however, it is possible that the overall LD landscape is so highly discordant between lines due to the short population histories and intense selection acting on these populations.

Chapter 12

Thesis Summary

Recent advances in genotype data acquisition technologies have facilitated huge leaps forward in genetic research. The decreasing costs, in addition to the increased throughput, quality and data density have all facilitated improved studies. As NGS technologies are increasingly validated for both quality and utility, they are making the transition into clinical genetics. Large national projects working to integrate NGS into clinical care, such as the Deciphering Developmental Delay and 100,000 Genomes Projects, are increasingly making the transition across the vague delineation between research and clinical practice.

A critical aspect in clinical application of any process is that samples to be analysed are appropriately tracked and managed so as to avoid sample mix-ups. Tools such as barcoded tubes and liquid handling robots minimise the risk. However, intrinsic markers of the sample are the least fallible approach. As such, I designed and evaluated the SNP panel reported in Chapter 6 to provide a suitable tool for tracking DNA samples directly, as well as tying them to the resultant WES data. This has been commercialised in partnership with LGC Genomics, who offer the panel as both a service and kit. The SNP panel has been utilised for several WES studies, and has been incorporated into custom targeted sequencing experiments by some groups, both internationally and locally.

Sensitivity is also of the utmost importance in clinical testing, arguably more so than the specificity of a test. The identification of the pathogenic variant(s) in a patient is clearly the primary goal of any clinical genetic investigations. As such, in Chapter 7 I detailed five exomes with pathogenic variation refractory to identification using standard analytical pipelines and filtering. The eventual identification of this variation, and the elucidation of issues associated with their identification will serve to inform future pipeline optimisation and analyses. These lessons will be applied within the Wessex Clinical Exome Pilot project, a joint venture between the University of

Southampton, University Hospital Southampton NHS Foundation Trust and Salisbury NHS Foundation Trust.

In Chapter 8 I discussed a number of families with cleft lip/palate phenotypes who were subjected to WES analysis. In these families a clear aetiology was identified in the families displaying syndromic presentations, with known pathogenic variants in *IRF6*, *IKBKG* and *SF3B4*. In patients with non-syndromic phenotypes however, no strong candidate pathogenic variants were identified. This is in accordance with expectation, with syndromic presentations usually being Mendelian in nature, and non-syndromic being complex. It is clear that studies with a far larger sample size than that analysed herein are required to truly evaluate the genetic contribution to non-syndromic cleft lip/palate. Where there are variants in families determined as possibly pathogenic candidates, such as the *MSX1*:p.P260T in Family NSCLP4, segregation analyses should be performed, providing that it is practicable to source further genetic material for the extended pedigree.

In Chapter 9, we utilised a targeted sequencing panel investigating FSGS, sequencing 39 genes in 83 patients. This TruSeq Custom Amplicon panel was shown to be an effective tool for the cost effective investigation of the genetic aetiology of FSGS. This was the first work to show that *COL4A5* underlies some cases of apparent FSGS, though other groups have recently reported the role of *COL4A3/4* in patients diagnosed with FSGS.

Overall, the work in clinical applications of NGS discussed in Part II serve to highlight the extensive utility of NGS technologies in clinical contexts. The routine evaluation of variants in order to assess pathogenicity is likely to require minimal input, though will be reliant on effective curation of databases such as HGMD. This will allow the refocussing of expertise onto the improvement of *in silico* analytical tools for improved variant detections. Furthermore, the ever increasing swathes of data will facilitate massive cohort studies, affording greater power for the detection of more subtle aetiological signals. This will allow for improved genotype–phenotype correlations, as well as the identification of associated variation contributing to common disease.

In Part III I described two bodies of work applying the Malécot-Morton model of LD to multiple populations. Firstly, in Chapter 10 I generated LD maps using both array-based and WGS data for CEU and YRI populations. It is clear from these analyses that WGS data provides significant information gains, particularly in diverse populations, where arrays do not fully capture the population variation. Following on from this work, I am generating LD maps for the whole genome using WGS data for

597 individuals from the Wellderly cohort. This work will aim to deliver the highest resolution mapping of LD to date, and interrogate this for features associated with recombination. Finally, the 100,000 Genome Project under Genomics England affords U.K. researchers unprecedented access to WGS data. Within the Population Genomics Genomics England Clinical Interpretation Partnership (GeCIP), we will have access to this data, allowing multiple population LD maps to be generated. With maps for multiple populations it will be possible to investigate differential selection between these populations.

Finally, in Chapter 11, I presented work done on the evaluation of LD patterns in commercial lines of *G. gallus*. In this, I identified highly discordant patterns of LD between the lines, even within breeds. This is at apparent odds with recent work that has shown that stable recombination hotspots in wild populations of finches lead to highly concordant patterns. Despite this, the association of recombination hotspots with sequence features seen in the wild still hold true for the commercial chickens. This discrepancy is likely due to the pressures imposed on the populations by commercial breeding, specifically population bottlenecks and selection pressures. This revelation is informative for commercial breeders, as it allows them to optimise their tools and analyses for genotype-led breeding programmes; as such, we have entered into a joint funding bid with a large international breeding company in order to further this work. Furthermore, through our collaboration with the Roslin Institute we have access to WGS data from a commercial breed. This will allow for refinement of the LD map, allowing maximal power to interrogate sequence features associated with recombination in the domestic chicken.

Overall, we have shown that the Malécot-Morton model continues to be a valuable tool for the mapping of LD using large-scale genomic data, and is further empowered by the increasing availability of WGS data on a population scale. These large datasets will facilitate the investigation of questions such as the regulation of recombination and population specificity of this regulation.

Appendices

References

1. Johannsen, W. The genotype conception of heredity. *Am Nat* **45**, 129–159 (1911) (cited on p. 2).
2. Mendel, G. Versuche über Pflanzen-Hybriden. *Verh Naturforsch Ver Brünn* **4**. (translated to English in *J R Hortic Soc* **26**, 1–32, 1901), 3–47 (1866) (cited on p. 2).
3. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. *J Exp Med* **79**, 137–158 (1944) (cited on p. 2).
4. Meselson, M. & Stahl, F. W. The replication of DNA in Escherichia coli. *Proc Natl Acad Sci U S A* **44**, 671–682 (1958). DOI: 10.1073/pnas.44.7.671 (cited on p. 2).
5. Watson, J. D. & Crick, F. H. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953) (cited on p. 2).
6. Alberts, B., Johnson, A., Lewis, J. & Raff, M. *Molecular Biology of the Cell: Reference edition* 5th Edition (Garland Science, Abingdon, UK, 2008) (cited on pp. 2, 86).
7. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970) (cited on p. 3).
8. Strachan, T. & Read, A. *Human Molecular Genetics* 4th Edition (Garland Science, Abingdon, UK, 2011) (cited on pp. 3, 4, 6, 7).
9. Genome Reference Consortium. *Human Genome Overview* URL: www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/ (2014) (cited on pp. 3, 11).
10. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007). DOI: 10.1038/nature05874 (cited on p. 4).
11. The ENCODE Project Consortium. An Integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). DOI: 10.1038/nature11247 (cited on pp. 4, 15).

12. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). DOI: 10.1038/nature11632 (cited on pp. 5, 16, 17, 42, 44, 49, 55, 58, 60, 61, 68, 90, 120, 121).

13. Lange, K. *Mathematical and statistical methods for genetic analysis* 70–84. DOI: 10.1007/978-1-4757-2739-5_5 (Springer, New York, 1997) (cited on p. 5).

14. Karigl, G. A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* **45**, 299–305 (1981). DOI: 10.1111/j.1469-1809.1981.tb00341.x (cited on p. 5).

15. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA Mutations in Human Disease. *Nat Rev Genet* **6**, 389–402 (2005). DOI: 10.1038/nrg1606 (cited on p. 7).

16. Pan, A., Chang, L., Nguyen, A. & James, A. W. A review of hedgehog signaling in cranial bone development. *Front Physiol* **4**, 61 (2013). DOI: 10.3389/fphys.2013.00061 (cited on p. 7).

17. John Hopkins University. *OMIM - Online Mendelian Inheritance in Man* URL: www.omim.org/ (2014) (cited on pp. 8, 54, 90, 96).

18. Morton, N. E. Sequential Tests for the Detection of Linkage. *Am J Hum Genet* **7**, 277–318 (1955) (cited on p. 9).

19. Cader, M. Z., Steckley, J. L., Dyment, D. A., McLachlan, R. S. & Ebers, G. C. A genome-wide screen and linkage mapping for a large pedigree with episodic ataxia. *Neurology* **65**, 156–158 (2005). DOI: 10.1212/01.wnl.0000167186.05465.7c (cited on p. 10).

20. Cui, Y., Li, G., Li, S. & Wu, R. English. in *Statistical Methods in Molecular Biology* (eds Bang, H., Zhou, X. K., Epps, H. L. & Mazumdar, M.) 219–242 (Humana Press, 2010). DOI: 10.1007/978-1-60761-580-4_6 (cited on p. 10).

21. MacDonald, M. E., Ambrose, C. M., Duyao, M. P. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993). DOI: 10.1016/0092-8674(93)90585-E (cited on p. 10).

22. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**, 228–237 (2003). DOI: 10.1038/ng1090 (cited on p. 10).

23. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). DOI: 10.1038/35057062 (cited on pp. 10, 11).

24. Collins, F. & Galas, D. A new five-year plan for the U.S. Human Genome Project. *Science* **262**, 43–6 (1993). DOI: 10.1126/science.8211127 (cited on pp. 10, 11).

25. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). DOI: 10.1038/nature03001 (cited on pp. 10, 11).

26. Tripp, S. & Grueber, M. *Economic Impact of the Human Genome Project* Booklet. [Booklet]. Battelle Memorial Institute, 2011 (cited on p. 11).

27. Oliver, M. F., Gumming, R. A., Geizerova, H. & Heady, J. A. Serum Cholesterol and ABO and Rhesus Blood-groups. *The Lancet* **294**, 605–607 (1969). DOI: 10.1016/S0140-6736(69)90322-5 (cited on p. 12).

28. Klitz, W., Aldrich, C., Fildes, N., Horning, S. & Begovich, A. Localization of predisposition to Hodgkin disease in the HLA class II region. *Am J Hum Genet* **54**, 497 (1994) (cited on p. 12).

29. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822 (2012). DOI: 10.1371/journal.pcbi.1002822 (cited on pp. 12, 15, 19).

30. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135–145 (2012). DOI: 10.1038/nrg3118 (cited on pp. 12, 16).

31. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37**, 413–417 (2005). DOI: 10.1038/ng1537 (cited on p. 12).

32. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003) (cited on pp. 13, 118).

33. International HapMap Consortium. A Haplotype Map of the Human Genome. *Nature* **437**, 1299–1320 (2005). DOI: 10.1038/nature04226 (cited on p. 13).

34. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). DOI: 10.1038/nature06258 (cited on p. 13).

35. International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010). DOI: 10.1038/nature09298 (cited on pp. 13, 17, 24, 42, 54, 120, 121).

36. Klein, R. J., Zeiss, C., Chew, E. Y. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385–389 (2005). DOI: 10.1126/science.1109557 (cited on pp. 13, 19, 24).

37. Terwilliger, J. D. & Hiekkalinna, T. An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur J Hum Genet* **14**, 426–437 (2006). DOI: 10.1038/sj.ejhg.5201583 (cited on p. 13).

38. Zhang, W., Collins, A. & Morton, N. E. Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum Genet* **115**, 157–64 (2004). DOI: 10.1007/s00439-004-1122-x (cited on p. 13).

39. Pereira, L., Pineda, M., Rowe, W. *et al.* The BRCA1 Ashkenazi founder mutations occur on common haplotypes and are not highly correlated with anonymous single nucleotide polymorphisms likely to be used in genome-wide case-control association studies. *BMC Genetics* **8**, 68 (2007). DOI: 10.1186/1471-2156-8-68 (cited on p. 13).

40. Zheng, H.-F., Ladouceur, M., Greenwood, C. M. & Richards, J. B. Effect of Genome-Wide Genotyping and Reference Panels on Rare Variants Imputation. *Journal of Genetics and Genomics* **39**, 545–550 (2012). DOI: 10.1016/j.jgg.2012.07.002 (cited on p. 13).

41. Grove, M. L., Yu, B., Cochran, B. J. *et al.* Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS One* **8**, e68095 (2013). DOI: 10.1371/journal.pone.0068095 (cited on p. 14).

42. Wagner, M. J. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics* **14**, 413–424 (2013). DOI: 10.2217/pgs.13.36 (cited on p. 14).

43. Huyghe, J. R., Jackson, A. U., Fogarty, M. P. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197–201 (2013). DOI: 10.1038/ng.2507 (cited on pp. 14, 15).

44. Martin, A. R., Tse, G., Bustamente, C. D. & Kenny, E. E. Imputation-based assessment of next generation rare exome variant arrays, 241–252 (2014). DOI: 10.1142/9789814583220_0024 (cited on p. 14).

45. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007). DOI: 10.1038/nature05911 (cited on p. 14).

46. Welter, D., MacArthur, J., Morales, J. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001–D1006 (2014). DOI: 10.1093/nar/gkt1229 (cited on pp. 14, 24, 101, 133).

47. Bennet, C. M., Baird, A. A., Miller, M. B. & Wolford, G. L. Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results* **1**, 1–5 (2010) (cited on pp. 14, 15).

48. *Concise encyclopaedia of bioinformatics and computational biology* 2nd (eds Hancock, J. H. & Zvelebil, M. J.) (John Wiley & Sons, Ltd, Chichester, UK, 2014) (cited on p. 15).

49. Cantor, R., Lange, K. & Sinsheimer, J. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet* **86**, 6–22 (2010). DOI: 10.1016/j.ajhg.2009.11.017 (cited on p. 15).

50. Manolio, T. A., Collins, F. S., Cox, N. J. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009). DOI: 10.1038/nature08494 (cited on p. 15).

51. Allen, H. L., Estrada, K., Lettre, G. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010). DOI: 10.1038/nature09410 (cited on p. 15).

52. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294–305 (2011) (cited on p. 15).

53. Yang, J., Benyamin, B., McEvoy, B. P. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010) (cited on p. 15).

54. Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* **21**, R1–9 (2012). DOI: 10.1093/hmg/dds387 (cited on pp. 15, 36).

55. Khoury, M. J., Janssens, A. C. J. W. & Ransohoff, D. F. How can polygenic inheritance be used in population screening for common diseases? *Genet Med* **15**, 437–443 (2013). DOI: 10.1038/gim.2012.182 (cited on pp. 15, 19).

56. National Human Genome Research Institute. *Sequencing costs* URL: www.genome.gov/sequencingcosts/ (2014) (cited on p. 16).

57. Cao, J., Schneeberger, K., Ossowski, S. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* **43**, 956–963 (2011). DOI: 10.1038/ng.911 (cited on p. 16).

58. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010). DOI: 10.1038/nature09534 (cited on pp. 16, 60).

59. Muddyman, D., Smee, C., Griffin, H., Kaye, J. & the UK10K Project. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med* **5**, 100 (2013). DOI: 10.1186/gm504 (cited on pp. 16, 55).

60. Wilkening, S., Tekkedil, M., Lin, G., Fritsch, E., Wei, W., Gagneur, J., Lazinski, D., Camilli, A. & Steinmetz, L. Genotyping 1000 yeast strains by next-generation sequencing. *BMC Genomics* **14**, 90 (2013). DOI: 10.1186/1471-2164-14-90 (cited on p. 16).

61. Pengelly, R. J. & Wheals, A. E. Rapid identification of Saccharomyces eubayanus and its hybrids. *FEMS Yeast Res* **13**, 156–61 (2013). DOI: 10.1111/1567-1364.12018 (cited on p. 16).

62. Scannell, D. R., Zill, O. A., Rokas, A., Payen, C., Dunham, M. J., Eisen, M. B., Rine, J., Johnston, M. & Hittinger, C. T. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 (Bethesda)* **1**, 11–25 (2011). DOI: 10.1534/g3.111.000273 (cited on p. 16).

63. MacArthur, D. G., Balasubramanian, S., Frankish, A. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–8 (2012). DOI: 10.1126/science.1215040 (cited on p. 16).

64. Ring, N., Meehan, T., Blake, A. *et al.* A mouse informatics platform for phenotypic and translational discovery. English. *Mamm Genome,* 1–9 (2015). DOI: 10.1007/s00335-015-9599-2 (cited on p. 16).

65. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–11 (2001). DOI: 10.1093/nar/29.1.308 (cited on pp. 17, 53, 68).

66. Krumm, N., Sudmant, P. H., Ko, A. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525–32 (2012). DOI: 10.1101/gr.138115.112 (cited on p. 17).

67. Need, A. C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., McDonald, M. T., Meisler, M. H. & Goldstein, D. B. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet* **49**, 353–361 (2012). DOI: 10.1136/jmedgenet-2012-100819 (cited on pp. 17, 19, 36).

68. Yang, Y., Muzny, D. M., Reid, J. G. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New Engl J Med* **369**, 1502–1511 (2013). DOI: 10.1056/NEJMoa1306555 (cited on pp. 17, 65, 88).

69. Yang, Y., Muzny, D. M., Xia, F. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014). DOI: 10.1001/jama.2014.14601 (cited on pp. 17, 65, 88).

70. Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. & Shendure, J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745–755 (2011). DOI: 10.1038/Nrg3031 (cited on pp. 17, 36, 37, 65).

71. Taylor, J. C., Martin, H. C., Lise, S. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* **17**, 717–726 (2015). DOI: 10.1038/ng.3304 (cited on p. 17).

72. Belkadi, A., Bolze, A., Itan, Y. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**, 5473–5478 (2015). DOI: 10.1073/pnas.1418631112 (cited on p. 17).

73. Chan, I. S. & Ginsburg, G. S. Personalized medicine: progress and promise. *Annu Rev Genomics Hum Genet* **12**, 217–44 (2011). DOI: 10.1146/annurev-genom-082410-101446 (cited on p. 18).

74. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415–25 (2010). DOI: 10.1038/nrg2779 (cited on p. 19).

75. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628–40 (2011). DOI: 10.1038/nrg3046 (cited on p. 19).

76. Nelen, M. & Veltman, J. A. Genome and exome sequencing in the clinic: unbiased genomic approaches with a high diagnostic yield. *Pharmacogenomics* **13**, 511–4 (2012). DOI: 10.2217/pgs.12.23 (cited on pp. 19, 36).

77. Bollag, G., Hirth, P., Tsai, J. *et al.* Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* **467**, 596–9 (2010). DOI: 10.1038/nature09454 (cited on p. 19).

78. Austin, C., Pettit, S. N., Magnolo, S. K., Sanvoisin, J., Chen, W., Wood, S. P., Freeman, L. D., Pengelly, R. J. & Hughes, D. E. Fragment screening using capillary electrophoresis (CEfrag) for hit identification of heat shock protein 90 ATPase inhibitors. *J Biomol Screen* **17**, 868–76 (2012). DOI: 10.1177/1087057112445785 (cited on p. 19).

79. Scott, S. A. Clinical Pharmacogenomics: Opportunities and Challenges at Point of Care. *Clin Pharmacol Ther* **93**, 33–35 (2013). DOI: 10.1038/clpt.2012.196 (cited on p. 19).

80. Tornatore, K. M., Brazeau, D., Dole, K. *et al.* Sex differences in cyclosporine pharmacokinetics and ABCB1 gene expression in mononuclear blood cells in African American and Caucasian renal transplant recipients. *J Clin Pharmacol* **53**, 1039–1047 (2013). DOI: 10.1002/jcph.123 (cited on p. 19).

81. Coulthard, S. A., Rabello, C., Robson, J. *et al.* A comparison of molecular and enzyme-based assays for the detection of thiopurine methyltransferase mutations. *Br J Haematol* **110**, 599–604 (2000). DOI: 10.1046/j.1365-2141.2000.02218.x (cited on p. 19).

82. Hesselink, D. A., van Schaik, R. H., van der Heiden, I. P., van der Werf, M., Gregoor, P. J., Lindemans, J., Weimar, W. & van Gelder, T. Genetic polymorphisms of the CYP3A4, CYP3A5, and MDR-1 genes and pharmacokinetics of the calcineurin inhibitors cyclosporine and tacrolimus. *Clin Pharmacol Ther* **74**, 245–54 (2003). DOI: 10.1016/S0009-9236(03)00168-1 (cited on p. 19).

83. Johnston, J. J., Rubinstein, W. S., Facio, F. M., Ng, D., Singh, L. N., Teer, J. K., Mullikin, J. C. & Biesecker, L. G. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am J Hum Genet* **91**, 97–108 (2012). DOI: 10.1016/j.ajhg.2012.05.021 (cited on p. 19).

84. Stecklein, S. R. & Jensen, R. A. Identifying and exploiting defects in the Fanconi anemia/BRCA pathway in oncology. *Transl Res* **160**, 178–97 (2012). DOI: 10.1016/j.trsl.2012.01.022 (cited on p. 19).

85. Ravitsky, V. & Wilfond, B. S. Disclosing individual genetic results to research participants. *Am J Bioeth* **6**, 8–17 (2006). DOI: 10.1080/15265160600934772 (cited on pp. 19, 20).

86. Townsend, A., Adam, S., Birch, P. H., Lohn, Z., Rousseau, F. & Friedman, J. M. "I want to know what's in Pandora's Box": comparing stakeholder perspectives on incidental findings in clinical whole genomic sequencing. *Am J Med Genet A* **158A**, 2519–25 (2012). DOI: 10.1002/ajmg.a.35554 (cited on p. 20).

87. Lohn, Z., Adam, S., Birch, P., Townsend, A. & Friedman, J. Genetics professionals' perspectives on reporting incidental findings from clinical genome-wide sequencing. *Am J Med Genet A* (2013). DOI: 10.1002/ajmg.a.35794 (cited on p. 20).

88. Evans, J. P., Berg, J. S., Olshan, A. F., Magnuson, T. & Rimer, B. K. We screen newborns, don't we?: realizing the promise of public health genomics. *Genet Med* **15**, 332–334 (2013). DOI: 10.1038/gim.2013.11 (cited on pp. 20, 22).

89. Green, R. C., Berg, J. S., Grody, W. W. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**, 565–74 (2013). DOI: 10.1038/gim.2013.73 (cited on p. 20).

90. Allyse, M. & Michie, M. Not-so-incidental findings: the ACMG recommendations on the reporting of incidental findings in clinical whole genome and whole exome sequencing. *Trends Biotechnol* **31**, 439–41 (2013). DOI: 10.1016/j.tibtech.2013.04.006 (cited on p. 20).

91. Van El, C. G., Cornel, M. C., Borry, P. *et al.* Whole-genome sequencing in health care. *Eur J Hum Genet* **21**, 580–584 (2013). DOI: 10.1038/ejhg.2013.46 (cited on p. 20).

92. Schmidt, C. O., Hegenscheid, K., Erdmann, P. *et al.* Psychosocial consequences and severity of disclosed incidental findings from whole-body MRI in a general population study. *Eur Radiol* **23**, 1343–51 (2013). DOI: 10.1007/s00330-012-2723-8 (cited on p. 20).

93. Berg, J. S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C. P., Wilhelmsen, K. C. & Evans, J. P. An informatics approach to analyzing the incidentalome. *Genet Med* **15**, 36–44 (2013). DOI: 10.1038/gim.2012.112 (cited on p. 20).

94. West, K. M., Hopkins, S. E., Hopper, K. J., Mohatt, G. V. & Boyer, B. B. Found in translation: Decoding local understandings of genetics and heredity in a Yup'ik Eskimo community. *Public Underst Sci* **22**, 80–90 (2013). DOI: 10.1177/0963662510397224 (cited on p. 21).

95. 23andMe. *Exome 80x - 23andMe* URL: www.23andme.com/exome/ (2013) (cited on p. 21).

96. Hawkins, A. K. & Ho, A. Genetic counseling and the ethical issues around direct to consumer genetic testing. *J Genet Couns* **21**, 367–73 (2012). DOI: 10.1007/s10897-012-9488-8 (cited on p. 21).

97. Green, R. C. & Farahany, N. A. Regulation: The FDA is overcautious on consumer genomics. *Nature* **505**, 286–287 (2014). DOI: 10.1038/505286a (cited on p. 21).

98. Tobin, A. C. *Public perception of stem cell research and unproven stem cell therapies* MSc Dissertation. 2013 (cited on p. 21).

99. Chico, V. *Genomic negligence: an interest in autonomy as the basis for novel negligence claims generated by genetic technology* (Routledge, London, 2011) (cited on p. 21).

100. Gilbar, R. & Barnoy, S. Disclosure of genetic information to relatives in Israel: between privacy and familial responsibility. *New Genet Soc* **31**, 391–407 (2012). DOI: 10.1080/14636778.2012.687135 (cited on p. 21).

101. Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C. & Shendure, J. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013). DOI: 10.1038/nature12064 (cited on p. 21).

102. Landry, J. J., Pyl, P. T., Rausch, T. *et al.* The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3 (Bethesda)* **3**, 1213–24 (2013). DOI: 10.1534/g3.113.005777 (cited on p. 21).

103. Andrews, B. J. & Depellegrin, T. HeLa Sequencing and Genomic Privacy: The Next Chapter. *G3 (Bethesda)* **3**, vii (2013). DOI: 10.1534/g3.113.007427 (cited on p. 21).

104. Krejci, L., Altmannova, V., Spirek, M. & Zhao, X. Homologous recombination and its regulation. *Nucleic Acids Res* **40**, 5795–5818 (2012). DOI: 10.1093/nar/gks270 (cited on pp. 23, 25).

105. Morgan, T. H. *A Critique of the Theory of Evolution* (Princeton University Press, 1916) (cited on p. 23).

106. Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S. & Morton, N. E. A map of the human genome in linkage disequilibrium units. *Proc Natl Acad Sci U S A* **102**, 11835–9 (2005). DOI: 10.1073/pnas.0505262102 (cited on pp. 24, 25, 118–121, 131, 136, 139, 142, 143).

107. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299–309 (2002) (cited on p. 24).

108. Slatkin, M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 477–485 (2008). DOI: 10.1038/nrg2361 (cited on p. 24).

109. Pengelly, R. J., Tapper, W., Gibson, J., Knut, M., Tearle, R., Collins, A. & Ennis, S. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC Genomics* **16**, 666 (2015). DOI: 10.1186/s12864-015-1854-0 (cited on pp. 24, 25, 136, 141, 143, 144).

110. Collins, A. & Lau, W. CHROMSCAN: genome-wide association using a linkage disequilibrium map. *J Hum Genet* **53**, 121–126 (2008). DOI: 10.1007/s10038-007-0226-2 (cited on p. 25).

111. Elding, H., Lau, W., Swallow, D. & Maniatis, N. Refinement in Localization and Identification of Gene Regions Associated with Crohn Disease. *Am J Hum Genet* **92**, 107–113 (2013). DOI: http://dx.doi.org/10.1016/j.ajhg.2012.11.004 (cited on pp. 25, 133).

112. Ennis, S. Linkage Disequilibrium as a Tool for Detecting Signatures of Natural Selection. English. *Linkage Disequilibrium and Association Mapping. Methods in Molecular Biology* **376** (ed Collins, A. R.) 59–70 (2007). DOI: 10.1007/978-1-59745-389-9_5 (cited on p. 25).

113. Jeffreys, A. J., Holloway, J. K., Kauppi, L., May, C. A., Neumann, R., Slingsby, M. T. & Webb, A. J. Meiotic recombination hot spots and human DNA diversity. *Philos Trans R Soc Lond B Biol Sci* **359**, 141–52 (2004). DOI: 10.1098/rstb.2003.1372 (cited on pp. 25, 121, 143).

114. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* **310**, 321–324 (2005). DOI: 10.1126/science.1117196 (cited on pp. 25, 121, 143, 147).

115. Paigen, K. & Petkov, P. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* **11**, 221–233 (2010) (cited on pp. 25, 143, 148).

116. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005). DOI: 10.1093/bioinformatics/bth457 (cited on pp. 25, 26, 28).

117. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449–462 (2005). DOI: 10.1086/428594 (cited on p. 26).

118. Kumasaka, N., Nakamura, Y. & Kamatani, N. The textile plot: a new linkage disequilibrium display of multiple-single nucleotide polymorphism genotype data. *PLoS One* **5**, e10207 (2010). DOI: 10.1371/journal.pone.0010207 (cited on p. 26).

119. Mourad, R., Sinoquet, C., Dina, C. & Leray, P. Visualization of Pairwise and Multilocus Linkage Disequilibrium Structure Using Latent Forests. *PLoS ONE* **6**, e27320 (December 2011). DOI: 10.1371/journal.pone.0027320 (cited on p. 26).

120. Mueller, J. C. Linkage disequilibrium for different scales and applications. *Brief Bioinform* **5**, 355–364 (2004). DOI: 10.1093/bib/5.4.355 (cited on p. 27).

121. Malécot, G. *Les mathématiques de l'hérédité* (Masson, Paris, 1948) (cited on p. 29).

122. Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A* **99**, 2228–33 (2002). DOI: 10.1073/pnas.042680999 (cited on pp. 29, 118, 119, 136, 137).

123. Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y. & Collins, A. The optimal measure of allelic association. *Proc Natl Acad Sci U S A* **98**, 5217–21 (2001). DOI: 10.1073/pnas.091062198 (cited on pp. 29, 137).

124. Tapper, W. Linkage Disequilibrium Maps and Location Databases. *Methods in Molecular Biology* **376** (ed Collins, A. R.) 23–45 (2007). DOI: 10.1007/978-1-59745-389-9_3 (cited on p. 30).

125. Kingman, J. The coalescent. *Stochastic Processes and their Applications* **13**, 235–248 (1982). DOI: 10.1016/0304-4149(82)90011-4 (cited on p. 30).

126. Yang, T., Deng, H.-W. & Niu, T. Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. *BMC Bioinformatics* **15**, 3 (2014). DOI: 10.1186/1471-2105-15-3 (cited on pp. 30, 31).

127. Auton, A. & McVean, G. Recombination rate estimation in the presence of hotspots. *Genome Res* **17**, 1219–1227 (2007). DOI: 10.1101/gr.6386707 (cited on pp. 31, 147).

128. Tapper, W., Gibson, J., Morton, N. & Collins, A. A comparison of methods to detect recombination hotspots. *Hum Hered* **66**, 157–169 (2008). DOI: 10.1159/000126050 (cited on pp. 31, 143, 146, 147).

129. Babbage, C. *Passages from the Life of a Philosopher* (Longman, Green, Longman, Roberts & Green, London, 1864) (cited on p. 32).

130. Woods, C., Valente, E., Bond, J. & Roberts, E. A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J Med Genet* **41**, e101–e101 (2004) (cited on p. 34).

131. Frankel, A. Formalin fixation in the '-omics' era: a primer for the surgeon-scientist. *ANZ J Surg* **82**, 395–402 (2012) (cited on p. 34).

132. Paireder, S., Werner, B., Bailer, J., Werther, W., Schmid, E., Patzak, B. & Cichna-Markl, M. Comparison of protocols for DNA extraction from long-term preserved formalin fixed tissues. *Anal Biochem* **439**, 152–160 (2013). DOI: 10.1016/j.ab.2013.04.006 (cited on p. 34).

133. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**, 251364 (2012). DOI: 10.1155/2012/251364 (cited on pp. 34, 35).

134. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012). DOI: 10.1186/1471-2164-13-341 (cited on pp. 34, 35).

135. Research & Markets. *Next Generation Sequencing (NGS) Market - Global Forecast to 2017* 2013 (cited on p. 34).

136. Illumina, Inc. *Illumina's MiSeqDx^{TM} Receives FDA Premarket Clearance with Two Cystic Fibrosis Assays and Universal Kit for Open Use* URL: investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1878430 (2013) (cited on p. 34).

137. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008). DOI: 10.1038/nature07517 (cited on p. 35).

138. Martin, J. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671–82 (2011). DOI: 10.1038/nrg3068 (cited on p. 35).

139. Drmanac, R., Sparks, A. B., Callow, M. J. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78–81 (2010). DOI: 10.1126/science.1181498 (cited on pp. 35, 120).

140. Peters, B. A., Kermani, B. G., Sparks, A. B. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–5 (2012). DOI: 10.1038/nature11236 (cited on p. 35).

141. Eid, J., Fehr, A., Gray, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009). DOI: 10.1126/science.1162986 (cited on p. 36).

142. Schneider, G. F. & Dekker, C. DNA sequencing with nanopores. *Nat Biotechnol* **30**, 326–328 (2012). DOI: 10.1038/nbt.2181 (cited on p. 36).

143. Oxford Nanopore Technologies. *Oxford Nanopore introduces DNA 'strand sequencing' on the high-throughput GridION platform and presents MinION, a sequencer the size of a USB memory stick.* URL: www.nanoporetech.com/news/press-releases/view/39 (2012) (cited on p. 36).

144. Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J Med Genet* **48**, 580–589 (2011). DOI: 10.1136/jmedgenet-2011-100223 (cited on p. 36).

145. Ng, S. B., Turner, E. H., Robertson, P. D. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009). DOI: 10.1038/nature08250 (cited on p. 36).

146. Christodoulou, K., Wiskin, A. E., Gibson, J. *et al.* Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. *Gut* **62**, 977–84 (2013). DOI: 10.1136/gutjnl-2011-301833 (cited on pp. 36, 54, 66).

147. Hou, Y., Song, L., Zhu, P. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–85 (2012). DOI: 10.1016/j.cell.2012.02.028 (cited on p. 36).

148. Liu, P., Morrison, C., Wang, L. *et al.* Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* **33**, 1270–6 (2012). DOI: 10.1093/carcin/bgs148 (cited on p. 36).

149. Puffenberger, E. G., Jinks, R. N., Sougnez, C. *et al.* Genetic mapping and exome sequencing identify variants associated with five novel diseases. *PLoS One* **7**, e28936 (2012). DOI: 10.1371/journal.pone.0028936 (cited on p. 36).

150. Watkins, D., Schwartzentruber, J. A., Ganesh, J., Orange, J. S., Kaplan, B. S., Nunez, L. D., Majewski, J. & Rosenblatt, D. S. Novel inborn error of folate metabolism: identification by exome capture and sequencing of mutations in the MTHFD1 gene in a single proband. *J Med Genet* **48**, 590–2 (2011). DOI: 10.1136/jmedgenet-2011-100286 (cited on pp. 36, 76, 77).

151. Weston-Bell, N., Gibson, J., John, M. *et al.* Exome sequencing in tracking clonal evolution in multiple myeloma following therapy. *Leukemia* **27**, 1188–91 (2013). DOI: 10.1038/leu.2012.287 (cited on p. 36).

152. Yu, T. W., Chahrour, M. H., Coulter, M. E. *et al.* Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron* **77**, 259–273 (2013). DOI: 10.1016/j.neuron.2012.11.002 (cited on p. 36).

153. Sulonen, A.-M., Ellonen, P., Almusa, H. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94 (2011). DOI: 10.1186/gb-2011-12-9-r94 (cited on p. 36).

154. Kuhn, T. S. *The structure of scientific revolutions* 2nd Edition (University of Chicago Press, Chicago, 1970) (cited on p. 37).

155. Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J. & Snyder, M. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* **29**, 908–914 (2011). DOI: 10.1038/nbt.1975 (cited on p. 37).

156. Bodi, K., Perera, A. G., Adams, P. S. *et al.* Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech* **24**, 73–86 (2013). DOI: 10.7171/jbt.13-2402-002 (cited on p. 38).

157. Koboldt, D., Zhang, Q., Larson, D. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–76 (2012). DOI: 10.1101/gr.129684.111 (cited on pp. 38, 41, 45, 81).

158. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009). DOI: 10.1093/bioinformatics/btp324 (cited on p. 39).

159. Burrows, M. & Wheeler, D. J. *A block-sorting lossless data compression algorithm* tech. rep. (1994) (cited on p. 39).

160. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009). DOI: 10.1186/gb-2009-10-3-r25 (cited on p. 39).

161. Novocraft Technologies. *Novocraft.com* URL: www.novocraft.com/main/page.php?s=novoalign (2014) (cited on pp. 39, 66).

162. Ruffalo, M., LaFramboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790–2796 (2011). DOI: 10.1093/bioinformatics/btr477 (cited on p. 39).

163. Major, E., Rigó, K., Hague, T., Bérces, A. & Juhos, S. HLA typing from 1000 genomes whole genome and whole exome data. *PLOS One* **8**, e78410 (2013). DOI: 10.1371/journal.pone.0078410 (cited on p. 39).

164. Danecek, P., Auton, A., Abecasis, G. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011). DOI: 10.1093/bioinformatics/btr330 (cited on pp. 39, 49, 120).

165. Li, H., Handsaker, B., Wysoker, A. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009). DOI: 10.1093/bioinformatics/btp352 (cited on pp. 39, 41, 66, 81, 88).

166. McKenna, A., Hanna, M., Banks, E. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–303 (2010). DOI: 10.1101/gr.107524.110 (cited on pp. 39, 41, 87, 88).

167. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009). DOI: 10.1093/bioinformatics/btp394 (cited on pp. 40, 41, 74).

168. Hart, S., Sarangi, V., Moore, R., Baheti, S., Bhavsar, J., Couch, F. & Kocher, J. SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PLOS ONE* **16**, e83356 (2013). DOI: 10.1371/journal.pone.0083356 (cited on pp. 40, 41, 74).

169. Fromer, M., Moran, J. L., Chambert, K. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91**, 597–607 (2012). DOI: 10.1016/j.ajhg.2012.08.005 (cited on pp. 41, 74).

170. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**, S1 (2013). DOI: 10.1186/1471-2105-14-S11-S1 (cited on p. 41).

171. Kadalayil, L., Rafiq, S., Rose-Zerelli, M. J. J. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* (2014). DOI: 10.1093/bib/bbu027 (cited on p. 41).

172. Staaf, J., Lindgren, D., Vallon-Christersson, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol* **9**, R136 (2008) (cited on pp. 41, 42, 50, 67, 80, 81, 83, 84).

173. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004). DOI: 10.1093/biostatistics/kxh008 (cited on pp. 42, 81).

174. Pollard, K., Hubisz, M., Rosenbloom, K. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–21 (2010). DOI: 10.1101/gr.097857.109 (cited on pp. 42, 90).

175. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. & Batzoglou, S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010). DOI: 10.1371/journal.pcbi.1001025 (cited on pp. 42, 68, 90).

176. Rey, J., Deville, J. & Chabbert, M. Structural determinants stabilizing helical distortions related to proline. *J Struct Biol* **171**, 266–276 (2010). DOI: 10.1016/j.jsb.2010.05.002 (cited on pp. 43, 76).

177. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974). DOI: 10.1126/science.185.4154.862 (cited on pp. 43, 90).

178. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003). DOI: 10.1093/nar/gkg509 (cited on pp. 43, 68, 90).

179. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **76**, 7.20 (2013). DOI: 10.1002/0471142905.hg0720s76 (cited on pp. 43, 68, 90).

180. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred: II. Error probabilities. *Genome Res* **8**, 186–194 (1998) (cited on p. 44).

181. Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* **15**, 733–747 (2013). DOI: 10.1038/gim.2013.92 (cited on pp. 44, 63).

182. Purcell, S., Neale, B., Todd-Brown, K. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007). DOI: 10.1086/519795 (cited on pp. 45, 56, 120, 136).

183. Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M. & Kang, H. M. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839–48 (2012). DOI: 10.1016/j.ajhg.2012.09.004 (cited on pp. 45, 53, 69).

184. Pengelly, R. J., Gibson, J., Andreoletti, G., Collins, A., Mattocks, C. J. & Ennis, S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med* **5**, 89 (2013). DOI: 10.1186/gm492 (cited on pp. 45, 64, 66, 69, 105).

185. Schmeider, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS One* **6**, e17288 (2011). DOI: 10.1371/journal.pone.0017288 (cited on p. 46).

186. Church, D., Schneider, V., Steinberg, K. *et al.* Extending reference assembly models. *Genome Biol* **16**, 13 (2015) (cited on pp. 49, 132).

187. Westra, H.-J., Jansen, R. C., Fehrmann, R. S. N., te Meerman, G. J., van Heel, D., Wijmenga, C. & Franke, L. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–2111 (2011). DOI: 10.1093/bioinformatics/btr323 (cited on pp. 52, 53).

188. Lam, C. W. & Jacob, E. Implementing a laboratory automation system: experience of a large clinical laboratory. *J Lab Autom* **17**, 16–23 (2012). DOI: 10.1177/2211068211430186 (cited on p. 52).

189. Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R. & Kidd, K. K. SNPs for a universal individual identification panel. *Hum Genet* **127**, 315–24 (2010). DOI: 10.1007/s00439-009-0771-1 (cited on p. 52).

190. Zietkiewicz, E., Witt, M., Daca, P., Zebracka-Gala, J., Goniewicz, M., Jarzab, B. & Witt, M. Current genetic methodologies in the identification of disaster victims and in forensic analysis. *J Appl Genet* **53**, 41–60 (2012). DOI: 10.1007/s13353-011-0068-7 (cited on pp. 52, 53).

191. Freire-Aradas, A., Fondevila, M., Kriegel, A. K., Phillips, C., Gill, P., Prieto, L., Schneider, P. M., Carracedo, A. & Lareu, M. V. A new SNP assay for identification of highly degraded human DNA. *Forensic Sci Int Genet* **6**, 341–9 (2012). DOI: 10.1016/j.fsigen.2011.07.010 (cited on p. 52).

192. Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A. & Mittelman, D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**, e32 (2013). DOI: 10.1093/nar/gks981 (cited on p. 53).

193. Castro, F., Dirks, W. G., Fähnrich, S., Hotz-Wagenblatt, A., Pawlita, M. & Schmitt, M. High-throughput SNP-based authentication of human cell lines. *Int J Cancer* **132**, 308–314 (2013). DOI: 10.1002/ijc.27675 (cited on p. 53).

194. Xu, W., Gao, H., Seok, J., Wilhelmy, J., Mindrinos, M. N., Davis, R. W. & Xiao, W. Coding SNPs as intrinsic markers for sample tracking in large-scale transcriptome studies. *Biotechniques* **52**, 386–8 (2012). DOI: 10.2144/0000113879 (cited on p. 53).

195. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010). DOI: 10.1093/bioinformatics/btq033 (cited on pp. 53, 68, 105, 138).

196. Institute for Systems Biology. *Repeat Masker* URL: www.repeatmasker.org/ (cited on p. 54).

197. Kent, W. J. BLAT - The BLAST-Like Alignment Tool. *Genome Res* **12**, 656–664 (2002). DOI: 10.1101/gr.229202 (cited on p. 54).

198. Gibson, J., Tapper, W., Ennis, S. & Collins, A. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. *Hum Genet* **132**, 233–43 (2013). DOI: 10.1007/s00439-012-1243-6 (cited on pp. 54, 66).

199. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–9 (2011). DOI: 10.1093/bioinformatics/btq671 (cited on p. 55).

200. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–9 (2011). DOI: 10.1093/molbev/msr121 (cited on p. 55).

201. International Organization for Standardization. *Medical laboratories - Particular requirements for quality and competence.* ISO 15189 (2012) (cited on p. 63).

202. Wegman, E. J. On randomness, determinism and computability. *Journal of Statistical Planning and Inference* **20**, 279–294 (1988). DOI: 10.1016/0378-3758(88)90093-6 (cited on p. 63).

203. Hu, H., Liu, X., Jin, W., Ropers, H. H. & Wienker, T. F. Evaluating information content of SNPs for sample-tagging in re-sequencing projects. *Sci Rep* **5**, 10247 (2015). DOI: 10.1038/srep10247 (cited on p. 64).

204. LGC Genomics. *KASP exome sample tracking panel* URL: www.lgcgenomics.com/genotyping/kasp-genotyping-reagents/genotyping-panels (2014) (cited on p. 64).

205. Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. & Leunissen, J. A. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**, W71–W74 (2007). DOI: 10.1093/nar/gkm306 (cited on p. 66).

206. Pengelly, R. J., Upstill-Goddard, R., Arias, L. *et al.* Resolving clinical diagnoses for syndromic cleft lip and/or palate phenotypes using whole-exome sequencing. *Clin Genet* (2015). DOI: 10.1111/cge.12547 (cited on pp. 66, 89).

207. Broad Institute. *Picard* URL: picard.sourceforge.net/ (2009) (cited on p. 66).

208. Trager, E. H., Khanna, R., Marrs, A., Siden, L., Branham, K. E. H., Swaroop, A. & Richards, J. E. Madeline 2.0 PDE: a new program for local and web-based pedigree drawing. *Bioinformatics* **23**, 1854–1856 (2007). DOI: 10.1093/bioinformatics/btm242 (cited on p. 66).

209. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010). DOI: 10.1093/nar/gkq603 (cited on p. 68).

210. University of Washington. *NHLBI Exome Sequencing Project (ESP): Exome Variant Server* URL: evs.gs.washington.edu/EVS/ (2014) (cited on p. 68).

211. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377–394 (2004). DOI: 10.1089/1066527041410418 (cited on pp. 68, 105).

212. Babraham Bioinformatics. *FastQC - A Quality Control tool for High Throughput Sequence Data* URL: www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2012) (cited on p. 68).

213. Bernier, F., Caluseriu, O., Ng, S. *et al.* Haploinsufficiency of SF3B4, a Component of the Pre-mRNA Spliceosomal Complex, Causes Nager Syndrome. *American J Hum Gen* **90**, 925–933 (2012). DOI: 10.1016/j.ajhg.2012.04.004 (cited on pp. 69, 70, 94, 96).

214. Petit, F., Escande, F., Jourdain, A. *et al.* Nager syndrome: confirmation of SF3B4 haploinsufficiency as the major cause. *Clin Genet* **86**, 246–251 (2014) (cited on pp. 69, 71, 94, 96).

215. Czeschik, J., Voigt, C., Alanay, Y. *et al.* Clinical and mutation data in 12 patients with the clinical diagnosis of Nager syndrome. *Hum Genet* **132**, 885–898 (2013). DOI: 10.1007/s00439-013-1295-2 (cited on pp. 70, 94, 96).

216. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013). DOI: 10.1093/bib/bbs017 (cited on pp. 70, 75).

217. Neuman, J. A., Isakov, O. & Shomron, N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform* **14**, 46–55 (2013). DOI: 10.1093/bib/bbs013 (cited on p. 71).

218. Parvaneh, N., Casanova, J. L., Notarangelo, L. D. & Conley, M. E. Primary immunodeficiencies: a rapidly evolving story. *J Allergy Clin Immunol* **131**, 314–23 (2013). DOI: 10.1016/j.jaci.2012.11.051 (cited on p. 71).

219. Chitasombat, M. N., Kofteridis, D. P., Jiang, Y., Tarrand, J., Lewis, R. E. & Kontoyiannis, D. P. Rare opportunistic (non-Candida, non-Cryptococcus) yeast bloodstream infections in patients with cancer. *J Infect* **64**, 68–75 (2012). DOI: 10.1016/j.jinf.2011.11.002 (cited on p. 72).

220. EMBL-EBI. *GO:0046655 folic acid metabolic process* URL: http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0046655#term=annotation (2014) (cited on p. 72).

221. Plagnol, V., Curtis, J., Epstein, M. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747–2754 (2012) (cited on p. 74).

222. Scotti, M., Stella, L., Shearer, E. & Stover, P. Modelling cellular compartmentation in one-carbon metabolism. *WIREs Syst Biol Med* **5**, 343–365 (2013). DOI: 10.1002/wsbm.1209 (cited on p. 76).

223. Allaire, M., Li, Y., MacKenzie, R. E. & Cygler, M. The 3-D structure of a folate-dependent dehydrogenase/cyclohydrolase bifunctional enzyme at 1.5 Å resolution. *Structure* **6**, 173–82 (1998) (cited on p. 77).

224. Tancredi, M., Sensi, E., Cipollini, G., Aretini, P., Lombardi, G., Di Cristofano, C., Presciuttini, S., Bevilacqua, G. & Caligo, M. Haplotype analysis of BRCA1 gene reveals a new gene rearrangement: characterization of a 19.9 KBP deletion. *Eur J Hum Genet* **12**, 775–777 (2004). DOI: 10.1038/sj.ejhg.5201223 (cited on p. 77).

225. Keller, M. D., Ganesh, J., Heltzer, M., Paessler, M., Bergqvist, A. G., Baluarte, H. J., Watkins, D., Rosenblatt, D. S. & Orange, J. S. Severe combined immunodeficiency resulting from mutations in MTHFD1. *Pediatrics* **131**, e629–34 (2013). DOI: 10.1542/peds.2012-0899 (cited on p. 77).

226. Watkins, D. & Rosenblatt, D. S. Inborn errors of cobalamin absorption and metabolism. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **157**, 33–44 (2011). DOI: 10.1002/ajmg.c.30288 (cited on p. 77).

227. Field, M. S., Shields, K. S., Abarinov, E. V. *et al.* Reduced MTHFD1 Activity in Male Mice Perturbs Folate- and Choline-Dependent One-Carbon Metabolism as Well as Transsulfuration. *Journal Nutr* **143**, 41–45 (2013). DOI: 10.3945/jn.112.169821 (cited on p. 78).

228. MacFarlane, A. J., Perry, C. A., Girnary, H. H., Gao, D., Allen, R. H., Stabler, S. P., Shane, B. & Stover, P. J. Mthfd1 Is an Essential Gene in Mice and Alters Biomarkers of Impaired One-carbon Metabolism. *J Biol Chem* **284**, 1533–1539 (2009). DOI: 10.1074/jbc.M808281200 (cited on p. 78).

229. Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D. & Cooper, D. N. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalised genomic medicine. *Hum Genet* **133**, 1–9 (2014). DOI: 10.1007/s00439-013-1358-4 (cited on pp. 79, 90, 105).

230. Hyakuna, N., Muramatsu, H., Higa, T., Chinen, Y., Wang, X. & Kojima, S. Germline Mutation of CBL Is Associated With Moyamoya Disease in a Child With Juvenile Myelomonocytic Leukemia and Noonan Syndrome-Like Disorder. *Pediatric Blood & Cancer* **62**, 542–544 (2015) (cited on p. 79).

231. Niemeyer, C. M., Kang, M. W., Shin, D. H. *et al.* Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat Genet* **42**, 794–800 (2010) (cited on p. 79).

232. Grand, F. H., Hidalgo-Curtis, C. E., Ernst, T. *et al.* Frequent CBL mutations associated with 11q acquired uniparental disomy in myeloproliferative neoplasms. *Blood* **113**, 6182–6192 (2009) (cited on p. 80).

233. Locatelli, F. & Niemeyer, C. M. How I treat juvenile myelomonocytic leukemia. *Blood* **125**, 1083–1090 (2015) (cited on p. 80).

234. Moy, R. L. Clinical presentation of actinic keratoses and squamous cell carcinoma. *J Am Acad Dermatol* **42**, S8–S10 (2000). DOI: 10.1067/mjd.2000.103343 (cited on p. 80).

235. Jayaraman, S. S., Rayhan, D. J., Hazany, S. & Kolodney, M. S. Mutational landscape of basal cell carcinomas by whole-exome sequencing. *J Invest Dermatol* **134**, 213–220 (2014) (cited on p. 81).

236. Rehman, I., Quinn, A., Healy, E. & Rees, J. High frequency of loss of heterozygosity in actinic keratoses, a usually benign disease. *The Lancet* **344**, 788–789 (1994) (cited on p. 81).

237. Foulkes, W. D., Black, D. M., Solomon, E., Trowsdale, J. & Stamp, G. W. Very frequent loss of heterozygosity throughout chromosome 17 in sporadic ovarian carcinoma. *Int J Cancer* **54**, 220–225 (1993) (cited on p. 84).

238. Kanehisa Laboratories. *KEGG PATHWAY Database* URL: http://www.genome.jp/kegg/pathway.html (2013) (cited on p. 86).

239. Keerthikumar, S., Raju, R., Kandasamy, K. *et al.* RAPID: Resource of Asian Primary Immunodeficiency Diseases. *Nucleic Acids Res* **37**, D863–7 (2009). DOI: 10.1093/nar/gkn682 (cited on p. 86).

240. Angulo, I., Vadas, O., Garçon, F. *et al.* Phosphoinositide 3-Kinase δ Gene Mutation Predisposes to Respiratory Infection and Airway Damage. *Science* **342**, 866–871 (2013). DOI: 10.1126/science.1243292 (cited on p. 86).

241. Lucas, C. L., Kuehn, H. S., Zhao, F. *et al.* Dominant-activating germline mutations in the gene encoding the PI(3)K catalytic subunit p110δ result in T cell senescence and human immunodeficiency. *Nat Immunol* **15**, 88–97 (2014). DOI: 10.1038/ni.2771 (cited on p. 86).

242. Zhang, J., Grubor, V., Love, C. L. *et al.* Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc Natl Acad Sci U S A* **110**, 1398–1403 (2013). DOI: 10.1073/pnas.1205299110 (cited on p. 86).

243. Crank, M., Grossman, J., Moir, S. *et al.* Mutations in PIK3CD Can Cause Hyper IgM Syndrome (HIGM) Associated with Increased Cancer Susceptibility. *J Clin Immunol* **34**, 272–276 (2014). DOI: 10.1007/s10875-014-0012-9 (cited on p. 86).

244. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., Wilkie, A. O., McVean, G., Lunter, G., Consortium, W. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912–918 (2014) (cited on p. 87).

245. Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. & Salit, M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246–251 (2014) (cited on p. 88).

246. Shkoukani, M., Chen, M. & Vong, A. Cleft Lip - A Comprehensive Review. *Front Pediatr* **1**, 53 (2013). DOI: 10.3389/fped.2013.00053 (cited on p. 89).

247. Collins, A., Arias, L., Pengelly, R., Martínes, 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* **1**, 10 (2013). DOI: 10.13172/2054-197X--1-987 (cited on pp. 89, 90, 98, 101).

248. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34**, E2393–E2402 (2013) (cited on p. 90).

249. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010) (cited on p. 90).

250. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561 (2009) (cited on p. 90).

251. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576 (2010) (cited on p. 90).

252. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014) (cited on p. 90).

253. Li, M.-X., Kwan, J. S., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q. & Sham, P. C. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* **9**, e1003143 (2013) (cited on p. 90).

254. Fuentes Fajardo, K. V., Adams, D., Mason, C. E., Sincan, M., Tifft, C., Toro, C., Boerkoel, C. F., Gahl, W. & Markello, T. Detecting false-positive signals in exome sequencing. *Hum Mutat* **33**, 609–613 (2012) (cited on p. 91).

255. Aradhya, S., 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* **10**, 2171–2179 (2001). DOI: 10.1093/hmg/10.19.2171 (cited on pp. 96, 97).

256. 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-$\varkappa$B activation. *Hum Mol Genet* **13**, 1763–1773 (2004). DOI: 10.1093/hmg/ddh192 (cited on p. 96).

257. Hadj-Rabia, S., Froidevaux, D., Bodak, N., Hamel-Teillac, D., Smahi, A., Touil, Y., Fraitag, S., de Prost, Y. & Bodemer, C. Clinical study of 40 cases of incontinentia pigmenti. *Arch Dermatol* **139**, 1163–1170 (2003) (cited on p. 96).

258. 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* **39**, 875–881 (2007) (cited on p. 96).

259. 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* **405**, 466–472 (2000) (cited on p. 97).

260. Conte, M. I., 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* **35**, 165–177 (2014) (cited on p. 97).

261. Evans, K. N., Sie, K. C., Hopper, R. A., Glass, R. P., Hing, A. V. & Cunningham, M. L. Robin sequence: from diagnosis to development of an effective management plan. *Pediatrics* **127**, 936–948 (2011) (cited on p. 98).

262. Tan, T. Y., Kilpatrick, N. & Farlie, P. G. *Developmental and genetic perspectives on Pierre Robin sequence* in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **163** (2013), 295–305 (cited on p. 98).

263. Wu-Chou, Y.-H., Lo, L.-J., Chen, K.-T. P., Chang, C.-S. F. & Chen, Y.-R. A combined targeted mutation analysis of IRF6 gene would be useful in the first screening of oral facial clefts. *BMC medical genetics* **14**, 37 (2013) (cited on p. 98).

264. Nikopensius, T., Jagomägi, T., Krjutškov, K. *et al.* 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 (2010) (cited on p. 98).

265. Vieira, A. Unraveling human cleft lip and palate research. *J Dent Res* **87**, 119–125 (2008) (cited on p. 98).

266. Chiquet, B. T., Blanton, S. H., Burt, A., Ma, D., Stal, S., Mulliken, J. B. & Hecht, J. T. Variation in WNT genes is associated with non-syndromic cleft lip with or without cleft palate. *Hum Mol Genet* **17**, 2212–2218 (2008) (cited on pp. 98, 99).

267. Lan, Y., Ryan, R. C., Zhang, Z., Bullard, S. A., Bush, J. O., Maltby, K. M., Lidral, A. C. & Jiang, R. Expression of Wnt9b and activation of canonical Wnt signaling during midfacial morphogenesis in mice. *Dev Dyn* **235**, 1448–1454 (2006) (cited on p. 99).

268. Jezewski, P., Vieira, A., Nishimura, C. *et al.* Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. *J Med Genet* **40**, 399–407 (2003) (cited on p. 99).

269. Satokata, I. & Maas, R. Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. *Nat Genet* **6**, 348–356 (1994) (cited on p. 99).

270. Otero, L., Gutiérrez, S., Cháves, M., Vargas, C. & Bérmudez, L. Association of MSX1 with nonsyndromic cleft lip and palate in a Colombian population. *The Cleft Palate-Craniofacial Journal* **44**, 653–656 (2007) (cited on p. 99).

271. Yoshiura, K.-i., Machida, J., Daack-Hirsch, S., Patil, S. R., Ashworth, L. K., Hecht, J. T. & Murray, J. C. Characterization of a novel gene disrupted by a balanced chromosomal translocation t(2;19)(q11.2;q13.3) in a family with cleft lip and palate. *Genomics* **54**, 231–240 (1998) (cited on p. 101).

272. Wyszynski, D. F., Maestri, N., McIntosh, I., Smith, E. A., Lewanda, A. F., Garcia-Delgado, C., Vinageras-Guarneros, E., Wulfsberg, E. & Beaty, T. H. Evidence for an association between markers on chromosome 19q and non-syndromic cleft lip with or without cleft palate in two groups of multiplex families. *Hum Genet* **99**, 22–26 (1996) (cited on p. 101).

273. Leal, T., Andrieux, J., Duban-Bedu, B., Bouquillon, S., Brevière, G.-M. & Delobel, B. Array-CGH detection of a de novo 0.8 Mb deletion in 19q13.32 associated with mental retardation, cardiac malformation, cleft lip and palate, hearing loss and multiple dysmorphic features. *Eur J Med Genet* **52**, 62–66 (2009) (cited on p. 101).

274. Kohli, S. S. & Kohli, V. S. A comprehensive review of the genetic basis of cleft lip and palate. *Journal of oral and maxillofacial pathology: JOMFP* **16**, 64 (2012) (cited on p. 101).

275. Ingersoll, R. G., Hetmanski, J., Park, J.-W. *et al.* Association between genes on chromosome 4p16 and non-syndromic oral clefts in four populations. *Eur J Hum Genet* **18**, 726–732 (2010) (cited on p. 101).

276. D'Asdia, M. C., Torrente, I., Consoli, F. *et al.* Novel and recurrent EVC and EVC2 mutations in Ellis-van Creveld syndrome and Weyers acrofacial dyostosis. *Eur J Med Genet* **56**, 80–87 (2013) (cited on p. 101).

277. Gajdos, V., Bahuau, M., Robert-Gnansia, E., Francannet, C., Cordier, S. & Bonaïti-Pellié, C. Genetics of nonsyndromic cleft lip with or without cleft palate: is there a Mendelian sub-entity? *Ann Genet* **47**, 29–39 (2004) (cited on p. 101).

278. Rehm, H. L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* **14**, 295–300 (2013). DOI: 10.1038/nrg3463 (cited on p. 103).

279. UK Genetic Testing Network. *NHS Directory of Genetic Disorders/Genes For Diagnostic Testing V12* URL: http://ukgtn.nhs.uk/fileadmin/uploads/ukgtn/Documents/Resources/Library/Reports_Guidelines/NHS_Directory_of_Genetic_Testing/UKGTN_Directory_of_Genetic_Testing_version_v12.pdf (2015) (cited on p. 103).

280. D'Agati, V. D., Kaskel, F. J. & Falk, R. J. Focal segmental glomerulosclerosis. *N Engl J Med* **365**, 2398–2411 (2011). DOI: 10.1056/NEJMra1106556 (cited on p. 104).

281. Gibson, J., Gilbert, R. D., Bunyan, D. J., Angus, E. M., Fowler, D. J. & Ennis, S. Exome analysis resolves differential diagnosis of familial kidney disease and uncovers a potential confounding variant. *Genetics Research* **95**, 165–173 (2013). DOI: 10.1017/S0016672313000220 (cited on p. 104).

282. Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., Lyon, E., Ward, B. E. & Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med* **10**, 294–300 (2008). DOI: 10.1097/GIM.0b013e31816b5cae (cited on pp. 105, 114).

283. Brown, E. J., Pollak, M. R. & Barua, M. Genetic testing for nephrotic syndrome and FSGS in the era of next-generation sequencing. *Kidney Int* **85**, 1030–1038 (2014) (cited on p. 113).

284. Malone, A. F., Phelan, P. J., Hall, G. *et al.* Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. *Kidney Int* **86**, 1253–1259 (2014) (cited on p. 113).

285. Martin, P., Heiskari, N., Zhou, J. *et al.* High mutation detection rate in the COL4A5 collagen gene in suspected Alport syndrome using PCR and direct DNA sequencing. *J Am Soc Nephrol* **9**, 2291–2301 (1998) (cited on p. 113).

286. Weins, A., Kenlan, P., Herbert, S., Le, T. C., Villegas, I., Kaplan, B. S., Appel, G. B. & Pollak, M. R. Mutational and biological analysis of α-actinin-4 in focal segmental glomerulosclerosis. *J Am Soc Nephrol* **16**, 3694–3701 (2005) (cited on p. 113).

287. Liu, Z., Blattner, S. M., Tu, Y., Tisherman, R., Wang, J. H., Rastaldi, M. P., Kretzler, M. & Wu, C. α-Actinin-4 and CLP36 protein deficiencies contribute to podocyte defects in multiple human glomerulopathies. *J Biol Chem* **286**, 30795–30805 (2011) (cited on p. 113).

288. Jais, J. P., Knebelmann, B., Giatras, I. *et al.* X-linked Alport syndrome: natural history and genotype-phenotype correlations in girls and women belonging to 195 families: a "European Community Alport Syndrome Concerted Action" study. *J Am Soc Nephrol* **14**, 2603–2610 (2003) (cited on pp. 113, 114).

289. Yao, X., Chen, X., Huang, G. *et al.* Challenge in pathologic diagnosis of Alport syndrome: evidence from correction of previous misdiagnosis. *Orphanet J Rare Dis* **7**, 100 (2012) (cited on p. 113).

290. Savige, J., Gregory, M., Gross, O., Kashtan, C., Ding, J. & Flinter, F. Expert guidelines for the management of Alport syndrome and thin basement membrane nephropathy. *J Am Soc Nephrol* **24**, 364–375 (2013) (cited on p. 113).

291. Guo, C., Van Damme, B., Vanrenterghem, Y., Devriendt, K., Cassiman, J.-J. & Marynen, P. Severe alport phenotype in a woman with two missense mutations in the same COL4A5 gene and preponderant inactivation of the X chromosome carrying the normal allele. *J Clin Invest* **95**, 1832 (1995) (cited on p. 114).

292. Chatterjee, R., Hoffman, M., Cliften, P., Seshan, S., Liapis, H. & Jain, S. Targeted exome sequencing integrated with clinicopathological information reveals novel and rare mutations in atypical, suspected and unknown cases of Alport syndrome or proteinuria. *PloS one* **8**, e76360 (2013) (cited on p. 114).

293. Goldberg, S., Adair-Kirk, T. L., Senior, R. M. & Miner, J. H. Maintenance of glomerular filtration barrier integrity requires laminin α5. *J Am Soc Nephrol* **21**, 579–586 (2010) (cited on p. 114).

294. Mistry, K., Ireland, J. H., Ng, R. C., Henderson, J. M. & Pollak, M. R. Novel mutations in NPHP4 in a consanguineous family with histological findings of focal segmental glomerulosclerosis. *Am J Kidney Dis* **50**, 855–864 (2007) (cited on p. 114).

295. Seri, M., Pecci, A., Di Bari, F. *et al.* MYH9-related disease: May-Hegglin anomaly, Sebastian syndrome, Fechtner syndrome, and Epstein syndrome are not distinct entities but represent a variable expression of a single illness. *Medicine (Baltimore)* **82**, 203–215 (2003) (cited on p. 114).

296. Miyoshi, Y., Santo, Y., Tachikawa, K., Namba, N., Hirai, H., Mushiake, S., Nakajima, S., Michigami, T. & Ozono, K. Lack of puberty despite elevated estradiol in a 46,XY phenotypic female with Frasier syndrome. *Endocr J* **53**, 371–376 (2006) (cited on p. 114).

297. Chernin, G., Vega-Warner, V., Schoeb, D. S., Heeringa, S. F., Ovunc, B., Saisawat, P., Cleper, R., Ozaltin, F., Hildebrandt, F. *et al.* Genotype/phenotype correlation in nephrotic syndrome caused by WT1 mutations. *Clinical Journal of the American Society of Nephrology* **5**, 1655–1662 (2010) (cited on p. 114).

298. Sikkema-Raddatz, B., Johansson, L. F., de Boer, E. N. *et al.* Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Hum Mutat* **34**, 1035–42 (2013). DOI: 10.1002/humu.22332 (cited on p. 115).

299. Adam, J., Connor, T. M., Wood, K., Lewis, D., Naik, R., Gale, D. P. & Sayer, J. A. Genetic testing can resolve diagnostic confusion in Alport syndrome. *Clinical kidney journal,* sft144 (2013) (cited on p. 115).

300. Mazzucco, G., Barsotti, P., Muda, A. O. *et al.* Ultrastructural and immunohistochemical findings in Alport's syndrome: a study of 108 patients from 97 Italian families with particular emphasis on COL4A5 gene mutation correlations. *J Am Soc Nephrol* **9**, 1023–1031 (1998) (cited on p. 115).

301. Gross, O., Licht, C., Anders, H. J. *et al.* Early angiotensin-converting enzyme inhibition in Alport syndrome delays renal failure and improves life expectancy. *Kidney Int* **81**, 494–501 (2012) (cited on p. 115).

302. Temme, J., Peters, F., Lange, K. *et al.* Incidence of renal failure and nephroprotection by RAAS inhibition in heterozygous carriers of X-chromosomal and autosomal recessive Alport mutations. *Kidney Int* **81**, 779–783 (2012) (cited on p. 115).

303. Vinai, M., Waber, P. & Seikaly, M. G. Recurrence of focal segmental glomerulosclerosis in renal allograft: An in-depth review. *Pediatr Transplant* **14**, 314–325 (2010) (cited on p. 115).

304. Cochat, P., Fargue, S., Mestrallet, G., Jungraithmayr, T., Koch-Nogueira, P., Ranchin, B. & Zimmerhackl, L. B. Disease recurrence in paediatric renal transplantation. *Pediatr Nephrol* **24**, 2097–2108 (2009) (cited on p. 115).

305. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–32 (2001). DOI: 10.1038/ng1001-229 (cited on p. 118).

306. Johnson, G. C., Esposito, L., Barratt, B. J. *et al.* Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233–7 (2001). DOI: 10.1038/ng1001-233 (cited on p. 118).

307. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**, 217–22 (2001). DOI: 10.1038/ng1001-217 (cited on p. 118).

308. Service, S., DeYoung, J., Karayiorgou, M. *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* **38**, 556–60 (2006). DOI: 10.1038/ng1770 (cited on pp. 118, 119, 131, 141, 143).

309. Lange, K. & Boehnke, M. How many polymorphic marker genes will it take to span the human genome? *Am J Hum Genet* **34**, 842–845 (1982) (cited on p. 119).

310. Kuo, T. Y., Lau, W. & Collins, A. R. LDMAP: the construction of high-resolution linkage disequilibrium maps of the human genome. *Methods Mol Biol* **376** (ed Collins, A. R.) 47–57 (2007). DOI: 10.1007/978-1-59745-389-9_4 (cited on pp. 119, 120, 132, 137).

311. Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., Bentley, D. R. & Morton, N. E. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci U S A* **101**, 18075–80 (2004). DOI: 10.1073/pnas.0408251102 (cited on pp. 119, 123, 132, 136).

312. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**, 887–93 (2005). DOI: 10.1086/429864 (cited on pp. 120, 137).

313. R Core Team. *R: A Language and Environment for Statistical Computing* URL: http://www.R-project.org/ (2014) (cited on p. 120).

314. Zhang, W., Collins, A., Maniatis, N., Tapper, W. & Morton, N. E. Properties of linkage disequilibrium (LD) maps. *Proc Natl Acad Sci U S A* **99**, 17004–17007 (2002). DOI: 10.1073/pnas.012672899 (cited on pp. 121, 136).

315. Matise, T. C., Chen, F., Chen, W. *et al.* A second-generation combined linkage physical map of the human genome. *Genome Res* **17**, 1783–6 (2007). DOI: 10.1101/gr.7156307 (cited on pp. 122, 123).

316. Lau, W., Kuo, T. Y., Tapper, W., Cox, S. & Collins, A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics* **23**, 517–9 (2007). DOI: 10.1093/bioinformatics/btl615 (cited on pp. 131, 146).

317. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27**, 2534–47 (2010). DOI: 10.1093/molbev/msq148 (cited on p. 131).

318. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**, 1496–502 (2005). DOI: 10.1101/gr.4107905 (cited on p. 131).

319. Charles, B. A., Shriner, D. & Rotimi, C. N. Accounting for linkage disequilibrium in association analysis of diverse populations. *Genet Epidemiol* **38**, 265–73 (2014). DOI: 10.1002/gepi.21788 (cited on p. 132).

320. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**. 10.1038/nrg3117, 36–46 (2012) (cited on p. 132).

321. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet* **9**, e1003566 (2013). DOI: 10.1371/journal.pgen.1003566 (cited on p. 132).

322. CHARGE Consortium. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899–901 (2013). DOI: 10.1038/ng.2671 (cited on p. 132).

323. Manolio, T. A., Chisholm, R. L., Ozenberger, B. *et al.* Implementing genomic medicine in the clinic: the future is here. *Genet Med* **15**, 258–267 (2013). DOI: 10.1038/gim.2012.157 (cited on p. 133).

324. Sabatti, C., Service, S. K., Hartikainen, A. L. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**, 35–46 (2009). DOI: 10.1038/ng.271 (cited on p. 133).

325. Gong, J., Schumacher, F., Lim, U. *et al.* Fine Mapping and Identification of BMI Loci in African Americans. *Am J Hum Genet* **93**, 661–71 (2013). DOI: 10.1016/j.ajhg.2013.08.012 (cited on p. 133).

326. Gurdasani, D., Carstensen, T., Tekola-Ayele, F. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015). DOI: 10.1038/nature13997 (cited on p. 133).

327. Hillier, L. W., Miller, W., Birney, E. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004) (cited on pp. 135, 137, 147).

328. Megens, H.-J., Crooijmans, R. P., Bastiaansen, J. W. *et al.* Comparison of linkage disequilibrium and haplotype diversity on macro-and microchromosomes in chicken. *BMC Genetics* **10**, 86 (2009) (cited on pp. 135, 136, 147).

329. Groenen, M. A., Wahlberg, P., Foglio, M. *et al.* A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res* **19**, 510–519 (2009) (cited on pp. 135, 137, 144, 146, 147).

330. Andreescu, C., Avendano, S., Brown, S. R., Hassen, A., Lamont, S. J. & Dekkers, J. C. Linkage disequilibrium in related breeding lines of chickens. *Genetics* **177**, 2161–2169 (2007) (cited on pp. 135, 141).

331. Heifetz, E. M., Fulton, J. E., O'Sullivan, N., Zhao, H., Dekkers, J. C. & Soller, M. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* **171**, 1173–1181 (2005) (cited on pp. 135, 141, 147).

332. Khatkar, M. S., Nicholas, F. W., Collins, A. R., Zenger, K. R., Cavanagh, J. A., Barris, W., Schnabel, R. D., Taylor, J. F. & Raadsma, H. W. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**, 187 (2008) (cited on p. 136).

333. Oliver, P. L., Goodstadt, L., Bayes, J. J. *et al.* Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet* **5**, e1000753–e1000753 (2009) (cited on pp. 136, 148).

334. Singhal, S., Leffler, E., Sannareddy, K. *et al.* Stable recombination hotspots in birds. *bioRxiv,* 023101 (2015) (cited on pp. 136, 144, 148).

335. Kranis, A., Gheyas, A. A., Boschiero, C. *et al.* Development of a high density 600K SNP genotyping array for chicken. *BMC genomics* **14**, 59 (2013) (cited on pp. 136, 138, 141).

336. Wright, S. Coefficients of inbreeding and relationship. *American Naturalist,* 330–338 (1922) (cited on pp. 137, 141).

337. Elferink, M. G., van As, P., Veenendaal, T., Crooijmans, R. P. & Groenen, M. A. Regional differences in recombination hotspots between two chicken populations. *BMC genetics* **11**, 11 (2010) (cited on pp. 137, 139).

338. Rubin, C.-J., Zody, M. C., Eriksson, J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010) (cited on p. 137).

339. Krasikova, A., Deryusheva, S., Galkina, S., Kurganova, A., Evteev, A. & Gaginskaya, E. On the positions of centromeres in chicken lampbrush chromosomes. *Chromosome Res* **14**, 777–789 (2006) (cited on p. 139).

340. Jeffreys, A. J., Murray, J. & Neumann, R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell* **2**, 267–73 (1998) (cited on p. 143).

341. Gibson, J., Tapper, W., Zhang, W., Morton, N. & Collins, A. Cosmopolitan linkage disequilibrium maps. *Hum Genomics* **2**, 20–7 (2005) (cited on p. 146).

342. Qanbari, S., Hansen, M., Weigend, S., Preisinger, R. & Simianer, H. Linkage disequilibrium reveals different demographic history in egg laying chickens. *BMC genetics* **11**, 103 (2010) (cited on p. 147).

343. Ségurel, L., Leffler, E. M. & Przeworski, M. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol* **9**, e1001211 (2011) (cited on p. 147).

344. Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G. & Donnelly, P. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**, 876–879 (2010) (cited on p. 148).

Appendix A

Pertinent Code Custom-written for Analyses

A.1 Code Developed for Chapter 6

generate_fingerprint.pl

```
#!perl;
use warnings; use v5.8.8; use strict;

## Script to generate SNP fingerprint profiles from input of minor-allele frequency data for multiple SNPs to be used for population simulation of WES identification panel.
## Usage: "perl generate_fingerprint.pl [MAF file] [n fingerprints]".

open SOURCE, $ARGV[0] or die; ## Call source MAF file path as first argument on calling. File layout should be: "^rsID\tMAF\n"
chomp(my @data = <SOURCE>);
close SOURCE;

## Generate cut-offs for genotypes according to Hardy-Weinberg equilibrium.
## 1 = p^2 + 2pq + q^2
my (%het, %ref, @rsIDs);
foreach my $line (@data) {
  my @split = split /\t/, $line;
  $het { "$split[0]" } = (2 * ($split[1] * ( 1 - $split[1]))); ## 2pq
  $ref { "$split[0]" } = ((1 - $split[1]) * ( 1 - $split[1])); ## p^2
  push @rsIDs, $split[0];
}

my $n = $ARGV[1] or die "Enter desired number of fingerprints to be generated on calling\n"; ## User input for number of desired
```

```
    fingerprints to be generated.
my $iteration = 1;
while ($iteration <= $n) {
  my @fingerprint;
  foreach my $rsID (@rsIDs) {
    my $state;
    my $determinant = rand;
    if ($determinant <= $het{$rsID}) { ## Take heterozygotes.
      $state = 'TA'; ## T used for ref, A for alt arbitrarily, so can
          still utilise phylogenetic software for downstream analysis.
      push @fingerprint, $state;
      next;
    }
    if (($determinant > $het{$rsID}) and ($determinant <=
        ($het{$rsID} + $ref{$rsID}))) { ## Take reference
        homomozygotes.
      $state = 'TT';
      push @fingerprint, $state;
      next;
    } else {
      $state = 'AA'; ## Presume remainder to be homozygous for
          alternative.
      push @fingerprint, $state;
      next;
    }
  }
  my $finger_to_print = join "", @fingerprint;
  print "$finger_to_print\n";
  $iteration++;
}
```

Appendix B

Supplementary Data

B.1 Supplementary Data for Chapter 6

Table B.1: Candidate SNPs for inclusion in WES tracking panel

						Distance			Allele Frequencies				
Chr	**Position**	**rsID**	**Gene**	**Strand**	**Alleles**	**5' SNP**	**3' SNP**	**GC**	**KG**	**HM-CEU**	**HM-CHB**	**HM-JPT**	**HM-YRI**
1	**179520506**	**rs1410592**	***NPHS2***	**+**	**A/G**	**84**	**154**	**48.1**	**0.43**	**0.59**	**0.62**	**0.54**	**0.53**
1	228431095	rs1771455	*OBSCN*	-	C/T	61	140	53.5	0.37	0.73	0.57	0.60	0.66
1	209968684	rs2013162	*IRF6*	+	A/C	273	69	52.3	0.40	0.35	0.50	0.62	0.25
1	167849414	rs203849	*ADCY10*	-	C/T	87	244	47.9	0.48	0.47	0.30	0.43	0.64
1	209811886	rs2076356	*LAMB3*	+	G/T	57	168	47.5	0.48	0.65	0.30	0.29	0.35
1	**67861520**	**rs2229546**	***IL12RB2***	**+**	**A/C**	**620**	**677**	**54.7**	**0.45**	**0.64**	**0.36**	**0.44**	**0.58**
1	158582646	rs2251969	*SPTA1*	+	C/T	94	192	44.9	0.45	0.60	0.67	0.58	0.37
1	45973928	rs2275276	*MMACHC*	+	A/G	93	231	50.9	0.49	0.47	0.51	0.57	0.37
2	44502788	rs3738985	*SLC3A1*	-	G/T	227	854	54.1	0.33	0.77	0.37	0.45	0.77
2	75115108	rs10194657	*HK2*	+	A/G	269	271	47.7	0.45	0.60	0.43	0.58	0.43
2	49381585	rs1394205	*FSHR*	-	A/G	332	85	44.7	0.36	0.30	0.48	0.52	0.25
2	**169789016**	**rs497692**	***ABCB11***	**-**	**A/G**	**542**	**290**	**45.5**	**0.48**	**0.55**	**0.65**	**0.51**	**0.22**
2	170092395	rs2229267	*LRP2*	-	C/T	263	260	42.7	0.41	0.23	0.27	0.35	0.67
2	179454394	rs1560221	*TTN*	-	C/T	120	813	41.5	0.45	0.20	0.64	0.66	0.62
2	179455207	rs2163009	*TTN*	-	A/G	813	1940	42.5	0.45	0.20	0.64	0.66	0.62
2	215820013	rs10498027	*ABCA12*	+	A/G	160	547	40.1	0.33	0.40	0.21	0.22	0.26
2	**227896976**	**rs10203363**	***COL4A4***	**+**	**C/T**	**90**	**100**	**46.7**	**0.45**	**0.46**	**0.44**	**0.36**	**0.57**
2	219941063	rs897477	*NHEJ1*	+	A/G	117	164	45.9	0.35	0.75	0.44	0.47	0.67
3	4712413	rs2306875	*ITPR1*	+	A/G	179	297	47.7	0.40	0.65	0.26	0.40	0.69
3	148727133	rs4938	*GYG1*	+	A/G	166	303	40.9	0.31	0.30	0.30	0.37	0.33
3	45989044	rs2234358	*FYCO1*	+	G/T	1064	79	47.7	0.46	0.49	0.64	0.53	0.25
3	**4403767**	**rs2819561**	***SUMF1***	**-**	**C/T**	**149**	**50**	**51.7**	**0.33**	**0.56**	**0.73**	**0.73**	**0.72**
4	86915848	rs10003909	*ARHGAP24*	+	C/T	141	75	42.5	0.32	0.21	0.29	0.32	0.43
4	88534235	rs2736982	*DSPP*	+	A/G	95	528	44.3	0.33	0.64	0.64	0.63	0.71
4	**5749904**	**rs4688963**	***EVC***	**-**	**A/G**	**90**	**57**	**44.1**	**0.46**	**0.33**	**0.65**	**0.67**	**0.52**
4	86844835	rs6824722	*ARHGAP24*	+	A/G	353	113	41.7	0.35	0.24	0.23	0.42	0.47
5	13719022	rs30169	*DNAH5*	-	A/C	295	67	41.3	0.41	0.41	0.41	0.45	0.51
5	13829799	rs1348689	*DNAH5*	-	C/T	188	134	41.1	0.46	0.39	0.67	0.50	0.50
5	13845045	rs10041113	*DNAH5*	+	A/G	171	62	41.5	0.47	0.46	0.66	0.47	0.58

Selected SNPs in bold typeface. Distance - distance to SNP with AF $\geq$ 0.01; GC - % GC of flanking 250 bp; KG - 1000 Genomes Pilot average; HM - HapMap Phase 3

Table B.1: Candidate SNPs for inclusion in WES tracking panel continued

						Distance			Allele Frequencies				
Chr	**Position**	**rsID**	**Gene**	**Strand**	**Alleles**	**5' SNP**	**3' SNP**	**GC**	**KG**	**HM-CEU**	**HM-CHB**	**HM-JPT**	**HM-YRI**
5	40981689	rs1061429	*C7*	+	A/C	291	79	50.7	0.35	0.42	0.32	0.31	0.36
5	53751988	rs7823	*HSPB3*	+	C/T	87	430	45.7	0.42	0.65	0.40	0.35	0.77
5	55155402	rs1009639	*IL31RA*	+	C/T	407	188	41.7	0.39	0.71	0.43	0.49	0.63
5	**82834630**	**rs309557**	***VCAN***	**-**	**A/G**	**331**	**915**	**45.5**	**0.48**	**0.49**	**0.34**	**0.52**	**0.50**
5	135392426	rs4669	*TGFBI*	+	C/T	102	81	49.1	0.39	0.27	0.31	0.35	0.73
5	138456815	rs3088052	*SIL1*	-	A/G	148	101	54.3	0.43	0.44	0.27	0.20	0.73
5	171849471	rs17074773	*SH3PXD2B*	+	A/G	182	182	50.1	0.38	0.30	0.70	0.62	0.25
6	152464839	rs2256135	*SYNE1*	+	A/G	472	773	45.7	0.34	0.50	0.71	0.69	0.78
6	152466674	rs2747662	*SYNE1*	+	C/T	92	212	46.7	0.34	0.26	0.30	0.26	0.31
6	**146755140**	**rs2942**	***GRM1***	**+**	**A/G**	**176**	**184**	**54.3**	**0.48**	**0.54**	**0.49**	**0.55**	**0.47**
6	56471402	rs9382658	*DST*	+	A/G	299	127	40.1	0.37	0.40	0.23	0.26	0.50
6	152675854	rs9397102	*SYNE1*	+	A/G	523	912	42.9	0.40	0.64	0.58	0.64	0.53
7	34009946	rs10265207	*BMPER*	+	C/T	714	66	48.1	0.43	0.45	0.45	0.41	0.34
7	100804140	rs1048303	*AP1S1*	+	C/T	82	290	54.1	0.50	0.58	0.43	0.37	0.39
7	**48450157**	**rs17548783**	***ABCA13***	**+**	**C/T**	**60**	**97**	**52.5**	**0.48**	**0.46**	**0.72**	**0.53**	**0.48**
7	50742180	rs1800504	*GRB10*	-	A/G	62	365	54.7	0.49	0.53	0.40	0.31	0.65
7	55214348	rs2072454	*EGFR*	+	C/T	133	57	50.3	0.45	0.54	0.44	0.31	0.50
7	127250907	rs712700	*PAX4*	-	A/G	76	281	45.7	0.34	0.78	0.36	0.34	0.69
7	43846603	rs7738	*BLVRA*	+	A/G	659	1085	44.7	0.42	0.33	0.43	0.33	0.61
7	151254175	rs8961	*PRKAG2*	+	C/T	272	268	41.9	0.42	0.65	0.78	0.75	0.21
8	104337096	rs3808554	*FZD6*	-	C/T	308	271	40.1	0.46	0.48	0.57	0.63	0.25
8	**94935937**	**rs4735258**	***PDP1***	**+**	**C/T**	**2314**	**595**	**41.7**	**0.50**	**0.40**	**0.64**	**0.66**	**0.46**
9	**100190780**	**rs1381532**	***TDRD7***	**-**	**C/T**	**963**	**192**	**42.7**	**0.43**	**0.48**	**0.59**	**0.50**	**0.58**
9	136304497	rs3124768	*ADAMTS13*	-	C/T	265	231	54.9	0.47	0.46	0.78	0.79	0.38
9	80919756	rs3739474	*PSAT1*	+	G/T	153	199	44.5	0.50	0.67	0.26	0.30	0.50
9	104184022	rs4577	*ALDOB*	-	C/T	53	476	45.1	0.40	0.42	0.42	0.45	0.31
9	27202870	rs639225	*TEK*	-	C/T	143	132	41.1	0.44	0.51	0.41	0.41	0.38
9	77415284	rs7859201	*TRPM6*	+	A/C	284	465	40.9	0.45	0.34	0.37	0.42	0.77
9	97365642	rs9695	*FBP1*	-	C/T	100	78	50.3	0.50	0.49	0.48	0.40	0.55
10	85972043	rs10749482	*CDHR1*	+	A/G	259	555	50.9	0.40	0.28	0.32	0.23	0.80
10	**100219314**	**rs10883099**	***HPSE2***	**+**	**A/G**	**52**	**60**	**51.3**	**0.46**	**0.52**	**0.52**	**0.53**	**0.62**
10	78944590	rs1131824	*KCNMA1*	-	C/T	290	202	44.1	0.38	0.32	0.26	0.26	0.66

Selected SNPs in bold typeface. Distance - distance to SNP with AF $\geq$ 0.01; GC - % GC of flanking 250 bp; KG - 1000 Genomes Pilot average; HM - HapMap Phase 3

Table B.1: Candidate SNPs for inclusion in WES tracking panel continued

						Distance			Allele Frequencies				
Chr	**Position**	**rsID**	**Gene**	**Strand**	**Alleles**	**5' SNP**	**3' SNP**	**GC**	**KG**	**HM-CEU**	**HM-CHB**	**HM-JPT**	**HM-YRI**
10	95791763	rs17109674	*PLCE1*	+	A/G	150	702	42.1	0.40	0.31	0.49	0.48	0.54
10	117884950	rs2245020	*GFRA1*	+	A/G	235	310	53.7	0.49	0.57	0.61	0.45	0.22
10	104814162	rs2275271	*CNNM2*	+	C/T	1265	1106	43.5	0.43	0.38	0.53	0.45	0.30
10	113920465	rs2277207	*GPAM*	-	C/T	294	187	41.1	0.45	0.61	0.50	0.62	0.54
10	69926097	rs2673794	*MYPN*	-	A/G	164	222	48.5	0.50	0.68	0.24	0.42	0.29
10	73856984	rs3312	*ASCC1*	-	C/T	340	607	41.1	0.44	0.44	0.30	0.30	0.71
10	104596924	rs6163	*CYP17A1*	+	A/C	528	57	54.9	0.42	0.38	0.55	0.46	0.29
10	105819956	rs805701	*COL17A1*	+	A/G	600	479	49.1	0.41	0.65	0.72	0.70	0.21
11	6629665	rs1043388	*ILK*	+	C/T	135	73	47.5	0.31	0.31	0.34	0.37	0.42
11	**16133413**	**rs4617548**	***SOX6***	**+**	**A/G**	**190**	**1118**	**42.1**	**0.47**	**0.52**	**0.65**	**0.61**	**0.51**
11	30255185	rs6169	*FSHB*	+	C/T	502	638	43.5	0.40	0.38	0.67	0.69	0.79
12	8757481	rs2028373	*AICDA*	+	A/G	188	1062	41.9	0.48	0.63	0.62	0.42	0.32
12	52200742	rs60637	*SCN8A*	+	A/C	714	927	52.7	0.43	0.74	0.47	0.50	0.24
12	**993930**	**rs7300444**	***WNK1***	**+**	**C/T**	**246**	**84**	**44.9**	**0.44**	**0.46**	**0.55**	**0.48**	**0.28**
13	25466955	rs3742165	*CENPJ*	+	C/T	181	186	44.7	0.49	0.54	0.67	0.69	0.23
13	**39433606**	**rs9532292**	***FREM2***	**+**	**A/G**	**477**	**183**	**45.9**	**0.40**	**0.29**	**0.41**	**0.44**	**0.54**
14	76045858	rs2287016	*FLVCR2*	+	A/G	167	105	54.7	0.44	0.22	0.59	0.57	0.69
14	**50769717**	**rs2297995**	***L2HGDH***	**+**	**A/G**	**262**	**142**	**42.9**	**0.43**	**0.55**	**0.65**	**0.67**	**0.59**
14	74992800	rs699374	*LTBP2*	+	A/G	217	61	51.1	0.37	0.29	0.41	0.33	0.50
14	64637147	rs7161192	*SYNE2*	+	A/C	584	837	49.1	0.39	0.29	0.57	0.52	0.34
14	35871217	rs8904	*NFKBIA*	-	C/T	124	190	41.5	0.42	0.39	0.37	0.27	0.70
15	89401615	rs3825994	*ACAN*	-	A/C	236	199	53.3	0.42	0.77	0.44	0.52	0.29
15	**34528948**	**rs4577050**	***SLC12A6***	**+**	**A/G**	**360**	**329**	**44.9**	**0.39**	**0.68**	**0.75**	**0.63**	**0.32**
15	89402596	rs698621	*ACAN*	-	A/C	357	300	52.5	0.49	0.66	0.40	0.45	0.20
16	68729785	rs17715450	*CDH3*	+	A/C	259	581	53.9	0.50	0.60	0.48	0.61	0.28
16	**70303580**	**rs2070203**	***AARS***	**-**	**C/T**	**162**	**79**	**54.7**	**0.46**	**0.53**	**0.28**	**0.51**	**0.49**
16	68713823	rs2296408	*CDH3*	-	G/T	93	633	52.5	0.37	0.62	0.58	0.68	0.69
16	68713730	rs2296409	*CDH3*	-	C/T	157	93	51.5	0.37	0.62	0.58	0.68	0.69
16	70546234	rs3762171	*COG4*	-	C/T	555	162	47.3	0.44	0.34	0.70	0.49	0.28
17	**71197748**	**rs1037256**	***COG1***	**+**	**A/G**	**309**	**455**	**53.1**	**0.44**	**0.50**	**0.67**	**0.65**	**0.56**
17	71192663	rs1052706	*COG1*	+	A/G	195	210	51.7	0.49	0.50	0.67	0.65	0.27
17	71192873	rs11544800	*COG1*	+	A/G	210	82	53.1	0.49	0.48	0.69	0.59	0.24

Selected SNPs in bold typeface. Distance - distance to SNP with AF $\geq$ 0.01; GC - % GC of flanking 250 bp; KG - 1000 Genomes Pilot average; HM - HapMap Phase 3

Table B.1: Candidate SNPs for inclusion in WES tracking panel continued

						Distance			Allele Frequencies				
Chr	**Position**	**rsID**	**Gene**	**Strand**	**Alleles**	**5' SNP**	**3' SNP**	**GC**	**KG**	**HM-CEU**	**HM-CHB**	**HM-JPT**	**HM-YRI**
17	7192091	rs222842	*YBX2*	+	C/T	1030	107	52.1	0.50	0.66	0.36	0.31	0.27
17	10542471	rs2285475	*MYH3*	-	A/C	579	237	44.9	0.50	0.74	0.38	0.39	0.22
17	10536018	rs2285479	*MYH3*	-	C/T	257	166	49.1	0.49	0.74	0.40	0.39	0.24
17	42449789	rs5910	*ITGA2B*	-	C/T	273	277	51.9	0.40	0.38	0.48	0.34	0.46
18	**21413869**	**rs9962023**	***LAMA3***	**+**	**C/T**	**565**	**231**	**48.7**	**0.34**	**0.67**	**0.81**	**0.75**	**0.51**
18	12351342	rs11080572	*AFG3L2*	-	A/G	837	210	41.1	0.37	0.77	0.67	0.60	0.48
18	47455923	rs2298628	*MYO5B*	+	C/T	67	400	49.9	0.42	0.58	0.33	0.27	0.37
19	55494740	rs10412915	*NLRP2*	+	C/T	89	141	54.5	0.31	0.25	0.27	0.31	0.25
19	33353464	rs11084673	*SLC7A9*	+	A/G	107	390	53.9	0.30	0.36	0.27	0.27	0.27
19	**10267077**	**rs2228611**	***DNMT1***	**-**	**A/G**	**66**	**175**	**42.7**	**0.46**	**0.47**	**0.73**	**0.56**	**0.48**
19	38994910	rs2229144	*RYR1*	+	A/G	230	242	52.1	0.36	0.28	0.33	0.26	0.45
19	13445208	rs2248069	*CACNA1A*	-	A/G	64	161	47.9	0.37	0.70	0.69	0.69	0.44
19	12989560	rs2293682	*DNASE2*	+	A/G	1132	400	54.5	0.36	0.24	0.69	0.67	0.24
19	55441902	rs269950	*NLRP7*	+	C/T	92	54	47.7	0.43	0.48	0.72	0.68	0.52
19	16591464	rs9305079	*CALR3*	+	A/G	226	231	42.7	0.36	0.69	0.67	0.65	0.52
20	**6100088**	**rs10373**	***FERMT1***	**-**	**C/T**	**57**	**142**	**46.7**	**0.49**	**0.54**	**0.31**	**0.35**	**0.58**
20	19970705	rs2076584	*RIN2*	-	A/G	404	183	48.3	0.45	0.35	0.67	0.76	0.44
20	2413320	rs2076652	*TGM6*	-	A/G	194	58	54.1	0.48	0.33	0.46	0.53	0.70
20	52786219	rs2296241	*CYP24A1*	+	A/G	146	187	42.1	0.47	0.54	0.44	0.42	0.46
20	35865054	rs4608	*RPN2*	+	C/T	325	297	46.9	0.31	0.76	0.48	0.58	0.70
21	46908355	rs11702425	*COL18A1*	+	C/T	118	214	49.1	0.32	0.30	0.24	0.32	0.35
21	47773103	rs2249057	*PCNT*	+	A/C	248	74	51.9	0.31	0.39	0.47	0.40	0.23
21	**44323590**	**rs4148973**	***NDUFV3***	**+**	**G/T**	**129**	**130**	**49.9**	**0.48**	**0.65**	**0.33**	**0.38**	**0.73**
22	**21141300**	**rs4675**	***SERPIND1***	**+**	**C/T**	**510**	**92**	**51.3**	**0.45**	**0.46**	**0.62**	**0.51**	**0.57**
22	37469591	rs4820268	*TMPRSS6*	+	A/G	391	230	52.1	0.46	0.52	0.49	0.37	0.79

Selected SNPs in bold typeface. Distance - distance to SNP with AF $\geq$ 0.01; GC - % GC of flanking 250 bp; KG - 1000 Genomes Pilot average; HM - HapMap Phase 3

B.2 Supplementary Data for Chapter 7

Table B.2: Technical details for whole-exome datasets

					Reads									Variants		
							On Target (%)			Target covered (%)					Heterozygosity (%)	
Family	**ID**	**Kit**	**Sequencing**	**FastQC**[a]	**Total**	**Aligned**	**Baits**	**± 150**	**DOC**	**20 X**	**10 X**	**5 X**	**1 X**	**Called**	**Autosomal**	**ChrX**
A	Proband	V5	HiSeq 2000	GC	46,341,632	42,707,333	75.2	86.7	56.9	87.8	96.3	98.4	99.3	25,139	61.5	13.1
B	I1	V5	HiSeq 2000		55,689,034	53,089,120	77.6	85.5	70.6	86.3	95.8	98.7	99.8	23,488	59.8	6.5
B	II2	V5	HiSeq 2000		53,054,034	50,490,602	77.4	85.2	67.0	85.2	95.4	98.6	99.8	24,577	**63.6**	6.6
B	II3	V5	HiSeq 2000		45,684,120	42,871,207	81.2	89.4	59.7	82.3	94.3	98.3	99.8	23,886	**62.7**	6.8
C	Proband	V5	HiSeq 2000		49,451,613	46,401,925	76.2	88.4	58.5	86.3	95.0	98.0	99.6	24,955	61.3	15.1
D	Lesion[b]	V5	HiSeq 2000	GC	91,045,365	83,122,511	88.7	91.9	100.0	95.8	98.3	99.0	99.4	-	-	-
D	Germline	V5	HiSeq 2000	GC	45,115,828	40,501,403	87.7	91.3	48.4	84.3	95.5	98.2	99.3	25,663	62.9	62.3
E	I1	V5	HiSeq 2000		49,372,504	47,046,558	79.1	86.7	63.9	84.0	94.9	98.4	99.8	23,055	60.5	8.2
E	I2	V5	HiSeq 2000		53,246,902	50,168,377	78.9	86.5	68.0	85.6	95.5	98.5	99.7	23,797	59.9	64.0
E	II1	V5	HiSeq 2000		47,342,860	45,248,551	79.4	87.2	61.6	83.5	94.8	98.3	99.8	23,297	60.0	9.2

Quality values outside of range are highlighted in bold and discussed in main text. Kit - whole-exome capture kit utilised: V4/V5 - Agilent SureSelect Human All Exome V4/V5; FastQC - criteria of read quality failures: GC - per sequence GC content, Quality - Per base sequence quality; Aligned - reads mapping to reference with an alignment phred-scaled quality ≥ 20; ± 150 - bait regions padded by 150 bp; DOC - mean depth of coverage for baited regions of genome.

[a]Note that FastQC results are highly sensitive; a 'Fail' does not constitute an issue in the use of data in downstream analyses, provided that no other quality issues are apparent.

[b]Data not subjected to variant QC due to somatic nature of the tissue.

B.3 Supplementary Data for Chapter 8

Table B.3: Technical details for whole-exome datasets

	Reads					Target covered (%)				
Sample	**Total**	**Aligned**	**Unique**	**Mapped ±150**	**Mapped target**	**1 X**	**5 X**	**10 X**	**20 X**	**DOC**
SCLP1	46,341,632	45,995,191	45,230,995	86.5	82.9	99.3	98.4	96.3	87.8	56.9
SCLP2-1	102,525,932	101,596,662	100,154,970	89.5	85.3	99.3	99.0	98.6	97.4	127.9
SCLP2-2	92,357,604	91,426,256	90,117,254	89.1	84.6	99.4	99.1	98.6	96.9	113.7
SCLP2-3	102,409,746	101,335,363	99,904,666	89.8	85.5	99.4	99.1	98.7	97.4	127.3
SCLP3-1	88,367,648	87,481,021	86,218,963	88.6	83.3	99.3	99.0	98.5	96.9	108.0
SCLP3-2	66,687,586	66,633,760	62,443,807	79.9	73.0	98.5	96.8	93.6	83.3	56.8
NSCLP1-1	54,956,672	54,558,769	53,763,584	91.1	76.8	99.3	98.8	97.7	92.9	68.8
NSCLP1-2	74,355,880	73,326,222	72,878,819	95.2	84.4	99.9	99.3	98.1	93.7	94.6
NSCLP2-1	53,630,678	53,245,840	52,373,414	88.2	74.8	99.3	98.7	97.1	90.7	65.2
NSCLP2-2	46,509,220	46,179,091	45,401,788	87.9	75.3	99.2	98.4	96.4	88.3	57.3
NSCLP3-1	43,363,414	43,032,505	42,300,126	87.9	75.3	99.2	98.3	95.7	85.9	52.6
NSCLP4-1	41,622,776	41,312,749	40,618,226	88.2	75.5	99.2	98.2	95.5	85.0	50.9
NSCLP4-2	46,087,918	45,762,033	44,987,196	87.6	74.4	99.2	98.4	96.3	87.8	56.6
NSCLP5-1	49,254,976	48,875,805	48,031,749	84.5	72.0	99.3	98.5	96.4	88.2	59.2
NSCLP6-1	100,068,158	99,055,310	97,692,151	91.5	86.6	99.4	99.1	98.6	96.9	109.2
NSCLP6-2	99,974,938	98,831,948	97,384,564	87.2	81.8	99.4	99.1	98.8	97.4	119.1
NSCLP7-1	95,428,304	94,564,501	93,168,229	88.3	84.7	99.3	99.0	98.5	97.1	118.5
NSCLP7-2	97,299,918	96,408,611	95,022,993	89.8	85.7	99.3	99.0	98.6	97.3	123.1