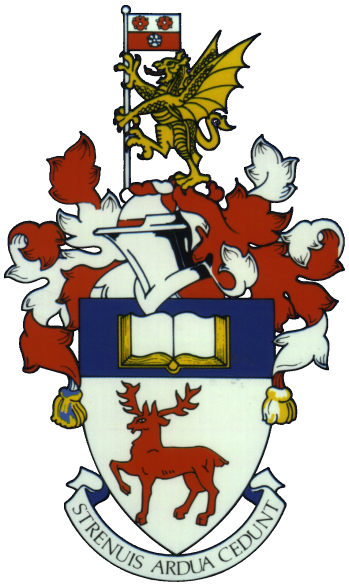


UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine
Human Genetics & Genomic Medicine

Next-Generation Sequencing Analyses in Human Diseases and Population Genomics

Mohammad Reza Jabal Ameli Forooshani

ORCID ID: 0000-0002-7762-0529

A thesis submitted for the degree of Doctor of Philosophy

October 2018

Abstract

Application of next-generation sequencing (NGS) in clinical diagnosis has enabled the efficient analysis of diverse genetic disorders. Rapid growth in the number of human genomes sequenced, underpinning a developing understanding of the disease-gene relationship. The high-throughput nature of NGS technology necessitates the need for a robust analytical framework for efficient and accurate genetic diagnosis. The higher degree of genetic variation identified in NGS applications often confounds the molecular diagnosis, and requires an enhanced strategy for the identification of causal variants. This thesis explores diverse applications of whole-exome and whole-genome sequencing at both the individual and population level for delineation of the human disease genome.

Design, implementation and benchmarking of efficient pipelines for analysing whole-exome and whole-genome sequencing data is first explored. Next, the diagnostic utility of the pipelines is examined in a range of rare disorders. This includes whole exome analysis of patients with hereditary nephrolithiasis, whole-exome and whole-genome analysis of a patient with severe skeletal dysplasia and targeted gene panel sequencing in a cohort of patients with syndromic cleft lip/palate (CLP). Through analyses of these cases, advantages and limitations of NGS analysis for establishing the molecular diagnosis in rare disorders are demonstrated. A novel method for ranking variants in the presence of phenotypic and genetic heterogeneity is introduced, and its diagnostic utility explored across syndromic CLP patients. While the application of variant-level attributes such as pathogenicity and conservation scores greatly facilitate molecular diagnosis, full resolution of the genetic architecture underlying disease genome depends on identification of factors that dictate the spatial distribution of pathogenic mutations across the genome. The non-random distribution of variants across the genome is the outcome of the complex interplay between selection, recombination and mutation which is reflected in genome-wide linkage disequilibrium (LD) patterns.

The final section of the thesis explores the possibility of delineating human disease genome from fine-scale LD maps in Sub-Saharan African populations (SSA). Extended population history in the SSA populations enables an unprecedented resolution for characterisation of LD patterns at sub-genic levels. LD maps constructed according to the Malécot-Morton model from the whole-genome sequence data of 295 individuals from major SSA populations correlates closely with the proposed models of Bantu expansion across Africa. Furthermore, the relationship between gene-ontology groups, gene essentiality and gene-age with the extent of LD is investigated, and a model for identification of the association between the LD extent and gene-group assignment is proposed.

Overall, this thesis demonstrates many applications of NGS technology and highlights the common limitations involved in the analysis and interpretation of variants revealed from high throughput NGS analysis.

Contents

List of Figures

List of Tables

Declaration of authorship

I, *Mohammad Reza Jabal-Ameli Forooshani*, declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

Title of thesis: **Next-Generation Sequencing Analyses in Human Diseases and Population Genomics**

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:

December 8, 2018

List of publications

1. **MR Jabalameli**, J Martinez, LA Uruena, IK Temple, RJ Pengelly, S Ennis, I Briceno, A Collins
   Diagnostic outcomes of exome gene panel sequencing in patients with unusual syndromic cleft lip/palate phenotypes. (2018)
   BioRxiv, doi: https://doi.org/10.1101/465179

2. A Karimi-Moghadam, S Charsouei, B Bell, **MR Jabalameli**
   Parkinson disease from Mendelian forms to genetic susceptibility: new molecular insights into the neurodegeneration process.(2018)
   Cellular and Molecular Neurobiology, 10.1007/s10571-018-0587-4.

3. RJ Pengelly, A Vergara Lope, D Alyusfi, **MR Jabalameli**, A Collins
   Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation(2017)
   Briefing in bioinformatics, 2017, 1-7.

4. RJ Pengelly, S Greville-Heygate, S Schmidt, EG Seaby, **MR Jabalameli**, SG Mehta, MJ Parker, D Goudie, C Fagotto-Kaufmann, C Mercer, A Debant, S Ennis, D Baralle
   Mutations specific to the Rac-GEF domain of TRIO cause intellectual disability and microcephaly (2016)
   Journal of medical genetics, jmedgenet-2016-103942.

5. **MR Jabalameli**, I Briceno, J Martinez, RJ Pengelly, S Ennis, A Collins
   Aarskog-Scott syndrome: phenotypic and genetic heterogeneity (2016)
   AIMS Genetics 3 (1), 49-59.

List of publications in preparation:

1. **MR Jabalameli**, RJ Pengelly, S Ennis, A Collins
   Fine-scale linkage disequilibrium daps in sub-Saharan African populations: functional clustering analysis and relationship to human disease genome

2. **MR Jabalameli**, D Hunt, D Bunyan, IK Temple, M Collins, S Ennis
   Distal arthrogryposis type 5D with novel clinical features and compound heterozygous mutations in ECEL1

Ethical approval

Analysis of patients detailed in Chapter 3 of this thesis has been approved by the East of England- Cambridge Central Research Ethics Committee (REC reference: 16/EE/0293). This study has also been approved by the Health Research Authority (PAT0295; IRAS project ID: 206718). This analysis was funded by Southampton Hospital Charity (Charity registration number: 1051543; Fund number: 0182).

In relation to the sample analysed in Chapter 4, the use of WES/WGS data in this patients in order to try and advance a molecular genetic diagnosis in a research setting is permitted within the governance framework of the NHS. This was undertaken with consent from the family for the interrogation of their DNA in a research setting with the assurance that any results obtained would be then verified by an NHS accredited laboratory, as took place for this family.

Analysis of patients detailed in Chapter 5 has been approved by the Research Ethics Committee at the Universidad de La Sabana, Bogota, Colombia (Comité de Ética en Investigación- Acta number: 29, 25 May 2012).

Permission to access, processing and analysis of African genome data used in Chapter 6 is certified under the Wellcome Trust Sanger Institute (WTSI) data access agreements (Request ID: 6187 and DAE-RJP-050116-APCDR).

Acknowledgements

Firstly, I would like to thank the British Council for providing me with the prestigious Chevening award to pursue my graduate studies in the United Kingdom. Without any doubt, it has been a fundamental step in my career, and I am certainly proud to be a member of this select community of scholars from all around the world.

I would like to thank my supervisors Professor Sarah Ennis, Professor Andrew Collins and Dr William Tapper for many hours of advice and guidance over the past three years. I was very fortunate to be involved in teaching during my PhD studies, and I would like to thank Sarah for providing me with this brilliant opportunity. My interest in population genetics emerged during many hours of exciting conversations with Andy for which I am sincerely grateful. Of course, without technically astute advice from Will many aspects of my research would have been impossible.

I would like to thank patients and their families who have taken part in the research detailed herein. In particular, much of this work is indebted to open access data from multiple consortia, including the African Genome Variation Project, 1000 Genomes and Exome Aggregation Consortium. I would like to extend thanks to all participants and organisers of these consortia.

I would like to thank Dr Valerie Walker for her continued inspiration and moral support over the past three years. I would also like to thank Professor Karen Temple and Dr Faisal Rezwan who have provided valuable insight and advice during my PhD. Special thanks to Nikki Graham for facilitating access to archival DNA storage and Dr Elena Vataga and David Baker at the IRIDIS high-performance computing facility for their invaluable help and support.

Many thanks to my lab mates who made this journey more enjoyable. In particular, I would like to thank Dr Marcin Knut for helping me with an introductory tutorial on command line tools for using the IRIDIS computer cluster when I started my PhD.

To my examiners, Prof. Diana Baralle and Prof. Nikolas Maniatis, thank you for taking the time to consider my thesis.

Finally, I would like to thank my family for the enormous amount of support they have given me. Mom and Dad, I would not have achieved this without your encouragement, inspiration and support.

Abbreviations

AAS	Aarskog-Scott syndrome
ACB	African Caribbeans in Barbados (the 1000 Genomes sub-population)
ACMG	American College of Medical Genetics and Genomics
AF	Allele frequency
AFR	African super-population in the 1000 Genomes Project
AGVP	African Genome Variation Project
AMCN	Arthrogryposis Multiplex Congenita with synostosis
AMR	Admixed American super-population in the 1000 Genomes Project
Arg	Arginine amino acid
ARVD1	Arrhythmogenic right ventricular dysplasia 1
AS	Autism spectrum
ASD	Atrial septal defect
Asn	Asparagine amino acid
Asp	Aspartate amino acid
ASW	Americans of African Ancestry in SW USA (the 1000 Genomes sub-population)
BAM	Binary Alignment Map
BEB	Bengali from Bangladesh (the 1000 Genomes sub-population)
BND	Breakend structural variation
BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion score
CCD	Central core disease
CDC42	Cell Division Cycle 42
CDLS1	Cornelia de Lange syndrome-1
CDX	Chinese Dai in Xishuangbanna (the 1000 Genomes sub-population)
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry (the 1000 Genomes sub-population)
CGH	Array Comparative Genomic Hybridisation
CHB	Han Chinese in Bejing (the 1000 Genomes sub-population)
Chr.	Chromosome
CHS	Southern Han Chinese (the 1000 Genomes sub-population)
CL	Cleft lip
CLM	Colombians from Medellin- Colombia (the 1000 Genomes sub-population)
CLP	cleft lip and/or cleft palate
CM	Complex-Mendelian gene group
CMA	Chromosomal microarray
CMT2V	Charcot-Marie-Tooth disease type 2V
CNM	Complex non-Mendelian gene group
CNV	Copy number variation
CP	Cleft palate
CPO	Cleft palate only
Cys	Cysteine amino acid
DA5D	Arthrogryposis 5D
dbSNP	Database of single nucleotide polymorphisms

dbVar	Database of Genomic Structural Variation
DD/ID	Developmental delay and/or intellectual disability
DDG2P	Development Disorder Genotype- Phenotype Database
DECIPHER	DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
Del	Deletion
DGVa	database of Genomic Variants archive
DOC	Depth of coverage
DP	Depth of coverage across all samples
EAS	East Asian super-population in the 1000 Genomes Project
EBOV	Ebola virus
END	Essential non-disease gene group
ER	Endoplasmic Reticulum
ESN	Esan in Nigeria (the 1000 Genomes sub-population)
ESP	Exome Sequencing Project (NHLBI)
EUR	European super-population in the 1000 Genomes Project
EVS	Exome Variant Server
ExAC	The Exome Aggregation Consortium
FATHMM	Functional Analysis through Hidden Markov Models
FCTCS	Cutaneous telangiectasia and cancer syndrome
FESD	Focal epilepsy with speech disorder
FIN	Finnish in Finland (the 1000 Genomes sub-population)
FLHS	Floating-Harbor syndrome
FN	False negative
FP	False positive
FS	Fisher's Strand bias
GBR	British in England and Scotland (the 1000 Genomes sub-population)
gDNA	genomic deoxyribonucleic acid
GEF	Guanine nucleotide exchange factor
GERP	Genomic Evolutionary Rate Profiling
GIAB	Genome in a Bottle project
GIH	Gujarati Indian from Houston- Texas (the 1000 Genomes sub-population)
Gln	Glutamine amino acid
Glu	Glutamate amino acid
GM-MDT	Genomic medicine multidisciplinary team
GRC	Genome Reference Consortium
GRIN2A	Glutamate Ionotropic Receptor NMDA Type Subunit 2A
GWAS	Genome-wide association studies
GWD	Gambian in Western Divisions in the Gambia (the 1000 Genomes sub-population)
Het	Heterozygous mutation
HGMD	Human Gene Mutation Database
His	Histidine amino acid
Hom	Homozygous mutation
HRI	Hill–Robertson interference
HSF	Human splice finder
HTS	High-throughput sequencing
HWE	Hardy-Weinberg equiliberium
IBD	Identity by descent

IBS	Iberian Population in Spain (the 1000 Genomes sub-population)
IGV	Integrative Genomics Viewer
INDEL	Insertion/Deletion mutation
Inv	Inversion
ITU	Indian Telugu from the UK (the 1000 Genomes sub-population)
JPT	Japanese in Tokyo- Japan (the 1000 Genomes sub-population)
KABUK1	Kabuki syndrome 1
KABUK2	X-linked dominant Kabuki syndrome 2
KHV	Kinh in Ho Chi Minh City- Vietnam (the 1000 Genomes sub-population)
KS	Klinefelter syndrome (47- XXY)
LD	Linkage disequilibrium
LDS5	Loeys-Dietz syndrome 5
LOVD	Leiden Open Variation Database
LPRD1	LEOPARD syndrome 1
LSDB	Locus Specific Mutation Databases
LWK	Luhya in Webuye- Kenya (the 1000 Genomes sub-population)
Lys	Lysine amino acid
M-CAP	Mendelian Clinically Applicable Pathogenicity score
MAF	Minor allele frequency
MCA	Multiple congenital anomalies
Met	Methionine amino acid
MNC	Mendelian non-complex gene group
MQ	Mapping Quality Scores
MRD4	Autosomal dominant mental retardation 4
MSL	Mende in Sierra Leone (the 1000 Genomes sub-population)
MXL	Mexican Ancestry from Los Angeles USA (the 1000 Genomes sub-population)
NDNE	Non-disease non-essential gene group
NGS	Next-generation sequencing
NHS	National health service
nm	Nanometre ($1e^{-9}$ Metre)
NMD	Nonsense mediated decay
NS1	Noonan syndrome 1
OFD1	Orofaciodigital syndrome 1
OMIM	Online Mendelian Inheritance in Man (OMIM Database)
ONT	Oxford Nanopore Technology
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDA	Patent ductus arteriosus
PE	Paired-end sequencing
PEL	Peruvians from Lima- Peru (the 1000 Genomes sub-population)
PJL	Punjabi from Lahore- Pakistan (the 1000 Genomes sub-population)
PUJ	Pelviureteric junction
PUR	Puerto Ricans from Puerto Rico (the 1000 Genomes sub-population)
QC	Quality Check
QD	Quality by Depth
RSTS1	Rubinstein-Taybi syndrome 1

SAM	Sequence Alignment Map
SAS	South Asian super-population in the 1000 Genomes Project
SC	Shawl scrotum
SCKL1	Seckel syndrome 1
SD	Skeletal dysplasia
SED	Soton Exome Database
Ser	Serine amino acid
SMRT	Single Molecule Real-Time sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single nucleotide variant
SR	Split-read
SSA	Sub-Saharan Africa
STU	Sri Lankan Tamil from the UK (the 1000 Genomes sub-population)
SV	Structural variation
TCGA	The Cancer Genome Atlas
Thr	Threonine amino acid
TP	True positive
Trp	Tryptophan amino acid
TSI	Toscani in Italy (the 1000 Genomes sub-population)
TSO	TruSight One sequencing capture-kit (Illumina- USA)
Tyr	Tyrosine amino acid
VCF	Variant call format
VEP	Variant Effect Predictor
VQSR	Variant Quality Score Recalibrator
VSD	Ventricular septal defect
VUS	Variant of uncertain significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WISH	Wessex Investigational Sciences Hub laboratory
WTCHG	Wellcome Trust Centre for Human Genetics
YRI	Yoruba in Ibadan- Nigeria (the 1000 Genomes sub-population)
ZMW	Zero-mode waveguide

Nomenclature

ϵ	Exponential decay of association with physical distance
ρ	Probability of association between SNPs acording to Malécot-Morton model
θ	Recombination fraction
d	Distance between each pair of SNPs in kilobases
D_E	Total number of risk alleles among affected individuals
D_N	Total number of alternative alleles among cases
D_{AB}	Coefficient of linkage disequilibrium
F_{ST}	Distance between each pair of SNPs in kilobases
H_E	Total number of risk alleles among matched controls
H_N	Total number of alternative alleles among matched controls
L	Residual association between a pair of SNPs at large distance
M	Probability of association at zero distance in Malécot model
N_e	Population effective size
NR	Number of non-recombinant individuals
OD	Optical density
P_AP_B	Probability of alleles A & B segregating independently
P_{AB}	Probability of alleles A & B occurring on the same haplotype
Q	Phred quality score
R	Number of recombinant individuals
Ti	Transition mutations
Tv	Transversion mutations
TB	Terabytes (2^{40} bytes)

Indeed to my parents,
for their endless love, support and encouragement.

"The beauty of science
is to make things simple"
Anonymous

Chapter 1

Introduction

In this chapter, I provide a general foundation for better understanding of the results that follow in the next chapters. Since human disease genome is the focus of this thesis, an overview of the mutational spectrum in the human genome is provided in the first section and approaches for disease gene identification is discussed in the second part. Finally, this chapter concludes with a thorough discussion of the limitations and challenges involved in Next-Generation Sequencing (NGS).

1.1 Variation in The Human Genome

For over 5,000 monogenic disorders that are documented to date, the underlying genetic factor for only about 50% of these conditions is described, and the genetic aetiology for the majority of these disorders remains yet to be identified [1]. Furthermore, an increasing number of complex conditions like intellectual disability and a range of psychological conditions that were previously thought to have a multifactorial aetiology are now thought to be a collection of diverse rare monogenic disorders [2,3].

Completion of the 1,000 Genomes project in October 2015 led to the discovery of over 88 million genetic variants including 84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertion/deletions (INDELs) and about 60 thousands structural variants across 2,504 individuals from 26 different populations [4]. Variation in the human genome is the consequence of historical mutational events that enabled survival, evolution and diversification of human race throughout the natural selection. Adaptive advantage conferred by this variation resulted in a 4.1 to 5 million site differences in a typical human genome when compared to the reference sequence [4]. The great majority of variation in our DNA appears to be without consequence since only a small percentage of our genome is functionally important [5,6,7]. Numerous changes outside exonic boundaries have no obvious consequence in the functionality of proteins and therefore are well tolerated. Restricting analyses to these functionally active regions substantially reduces the total number of non-reference sites to about 6×10^3 variants out of which 149-182 sites are identified to be protein truncating. About 10-12 thousand sites are presumed to have sequence-altering implications and 459-565K variants are discovered to overlap known regulatory regions [4]. The total number of non-reference sites greatly varies in different populations. Individuals of African descent harbour the greatest number of variant sites ($\sim$5M sites) per genome while this figure is depressed in recently bottlenecked populations (Figure 1.1).

It is noteworthy that not all of these variant sites are related to human disorders and in fact only a fraction of these variants have pathological implications. Genetic redundancy also confers some degree of protection against inactivating mutations. A classic example of this genetic redundancy is displayed by ribosomal RNA genes, where inactivating mu-

tations in one copy of an rRNA gene do not necessarily lead to rRNA impairment as many other functional copies compensate for the aberrant copy [8].

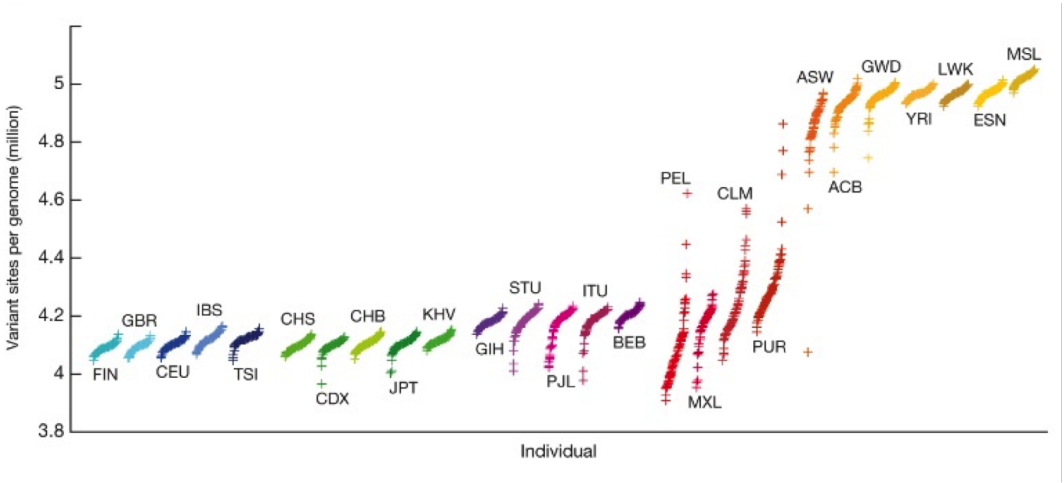


Figure 1.1: Total number of polymorphic variant sites catalogued across 26 populations from Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and America (AMR). Increased number of variant sites in populations with African origin confirms the out-of-Africa model of human evolution. *FIN:* Finnish in Finland; *GBR:* British in England and Scotland; *CEU:* Utah Residents (CEPH) with Northern and Western European Ancestry; *IBS:* Iberian Population in Spain; *TSI:* Toscani in Italy; *CHS:* Southern Han Chinese; *CDX:* Chinese Dai in Xishuangbanna, China; *CHB:* Han Chinese in Bejing, China; *JPT:* Japanese in Tokyo, Japan; *KHV:* Kinh in Ho Chi Minh City, Vietnam; *GIH:* Gujarati Indian from Houston, Texas; *STU:* Sri Lankan Tamil from the UK; *PJL:* Punjabi from Lahore, Pakistan; *ITU:* Indian Telugu from the UK; *BEB:* Bengali from Bangladesh; *PEL:* Peruvians from Lima, Peru; *MXL:* Mexican Ancestry from Los Angeles USA; *CLM:* Colombians from Medellin, Colombia; *PUR:* Puerto Ricans from Puerto Rico; *ASW:* Americans of African Ancestry in SW USA; *ACB:* African Caribbeans in Barbados; *GWD:* Gambian in Western Divisions in the Gambia; *YRI:* Yoruba in Ibadan, Nigeria; *LWK:* Luhya in Webuye, Kenya; *ESN:* Esan in Nigeria; *MSL:* Mende in Sierra Leone. Figure adapted from the 1000 Genomes Project Consortium [4]

1.2 Genetic Basis of Diseases

Mutational events at the DNA level generally lead to genetic disorders through one of the three major mechanisms:

1. Changes in the protein-coding sequence result in the total loss of function or alternatively gain of function. This class of mutations result in the gene products that are totally dysfunctional or may have acquired altered (or perhaps new) function. Mutations of this category necessarily result in changes in the conformation of protein structure.

2. Mutational events that result in changes in the genetic dosage also impair the functionality of cells and result in genetic disorders. These type of mutations generally result in unstable mRNA or aberrant gene copy number. Occasionally these mutations adversely affect *cis*-acting regulatory sequences that control gene expression. The resulting abnormal excess or deficiency of gene product results in pathological conditions.

3. Less frequently, mutations in regulatory elements of DNA indirectly impair the functionality of the downstream genes that are under the tight regulation of these elements and thereby lead to pathological conditions.

The mutation rate is not uniform across the genome and variant sites tend to cluster around certain DNA sequences that are prone to mutational events. Such regions are described as mutation hotspots where their inherent instability or chemical predisposition to nucleotide substitution results in greater frequency of mutational events [9]. Repetitive and low complexity DNA are among regions in which mutations involve large-scale deletions and duplications which potentially impact multiple genes [10]. In contrast, point mutations ubiquitously occur across the genome and based on their position they can result in a highly pathogenic or an entirely benign variant. In the next section, I will explore different categories of point mutations that underlie human disorders.

1.2.1 Pathogenic Point Mutations

Point mutations are the most frequent form of variation in the human genome. This category of mutations involve substitution, insertion or deletion of a single nucleotide in the DNA sequence. Single nucleotide substitutions are typically classified into two categories (Figure 1.2):

1. **Transitions (Ti)** are single nucleotide changes that substitute a nucleotide base with its similar kind (purine to purine or pyrimidine to pyrimidine). Transition mutations occur more frequently across the human genome.

2. **Transversions (Tv)**, in contrast, substitute a purine with a pyrimidine and vice versa and occurs with lower frequency across the human genome.

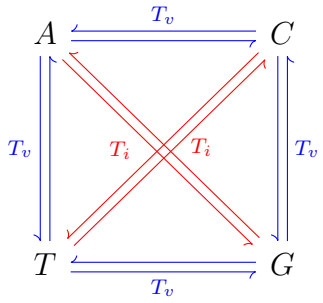


Figure 1.2: DNA substitution mutations; transitions (T_i) versus transversions (T_v). Although possible T_v are twice as many as T_i, T_i mutations in the human genome occur at a higher rate resulting in a genome-wide T_i/T_v ratio of $\sim$2.0 [11].

The transition to transversion ratio (T_i/T_v) is a measure of sequence change and is often used as a principal metric in evolutionary substitution models [12]. If the pairwise substitution rate for all possible nucleotide pair conversions was equal, then the T_i/T_v ratio was expected to be around 0.5 as there are 2 times as many possible transversions as transitions [11](Figure 1.2). In fact, the T_i/T_v ratio is not constant across the genomes of different species. For example, while there is a bias in favour of transitions over transversions in the genome of *Homo sapiens* and *Drosophila melanogaster*, no significant difference between transition and transversion rates are identified in the grasshopper genome [13]. In *Homo sapiens* mitochondrial genome the $T_i - T_v$ bias is even larger in favour of transitions and the T_i/T_v is estimated between 21 and 38 [14]. Given the genome-wide bias in the T_i/T_v ratio in humans, the ratio is frequently used as a quality check parameter in the high-throughput sequencing applications [15]. During the evolution of mammalian and avian orders, their genomes underwent a rapid GC increase across the coding parts that consequently resulted in the higher frequency of T_i conversions across the coding loci. It

is generally assumed that the T_i/T_v in human exonic region is ∼3.0 and outside the exome region is ∼2.0 [16].

When a single nucleotide substitution occurs in the coding sequence, it can lead to replacement of one codon in the mRNA with another codon. Considering the consequence of this replacement on the amino acid sequence of protein, single nucleotide variants (SNVs) are subcategorised to synonymous (silent), non-synonymous (missense) and nonsense mutations (Figure 1.3). As a result of substantial redundancy in the genetic code, synonymous (also known as silent) mutations are the most frequent type of SNVs. In fact, all amino acids (except Methionine (Met) & Tryptophan (Trp)) are encoded by multiple codons and substitutions at the single nucleotide level typically lead to substitution of different codons that effectively code for the same amino acid. Despite the fact that synonymous mutations are typically harmless, a minority of these silent mutations alter RNA splicing motifs and thereby result in pathologic conditions [17,18].

Non-synonymous substitutions (missense mutations) predominantly replace one amino acid with another amino acid and depending on the chemical properties of the amino acids that have been substituted, they represent a range of functional impact at the protein level. Substitution of polar amino acids (Arg, Lys, His, Asp, Glu, Asn, Gln, Ser, Thr, Tyr and Cys) with nonpolar amino acids effectively impairs the conformation and functionality of the protein and often have dramatic effects on the phenotype [19,20]. The chemical properties of substituted amino acids are used in predictive models to quantify deleteriousness of a mutation.

The relative frequency of synonymous and non-synonymous mutations vary according to their position of occurrence in the codon. For nuclear DNA, where 20 amino acids are coded by 61 codons, about $\frac{2}{3}$ ($\sim 70\%$) of substitutions at the third base position are usually silent, while by contrast, almost all changes at the second base pair are non-synonymous and about 96% of substitutions at the first base are identified as non-synonymous [21].

Nonsense mutations are the third category of single nucleotide substitutions in which replacement of a single nucleotide transforms an amino acid-specifying codon to a premature stop-codon. This type of mutation results in unstable RNA or truncated proteins. The pathogenic impacts of these mutations are tightly controlled by numerous cellular RNA surveillance mechanisms including *nonsense mediated decay (NMD)* [22].

Single nucleotide insertions and deletions are also an important group of pathogenic mutations that can disrupt the reading frame in exons and thereby indirectly lead to premature termination codons. These mutations are under the tight control of NMD mechanism, but occasionally if they skip NMD, result in truncated proteins that are either not functional or interfere with the function of the wild type proteins. These mutations are frequently associated with dominant negative conditions [23].

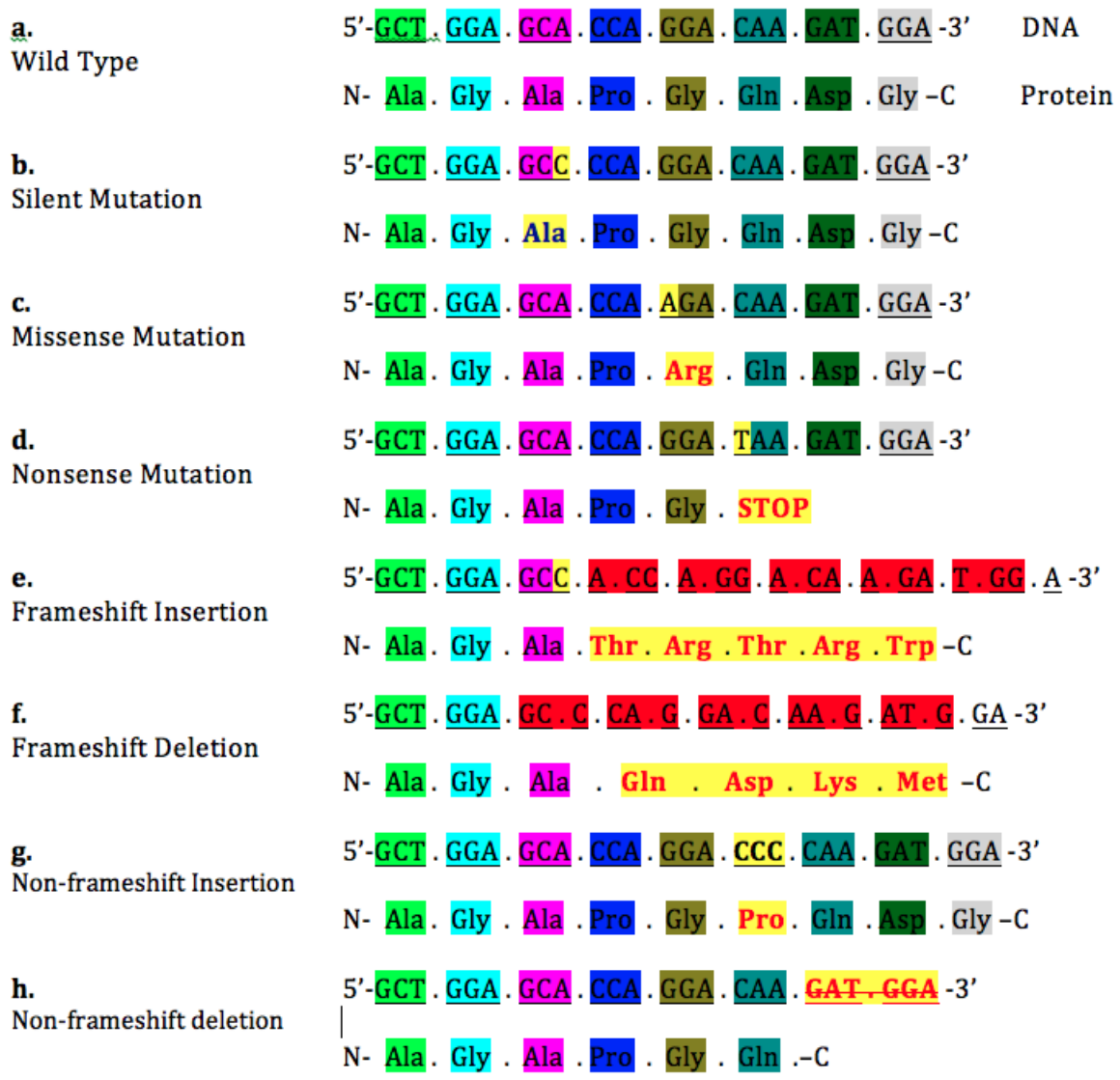


Figure 1.3: Overview of different types of mutations in the coding region of a gene. a. The open reading frame for eight amino acids in the wild type peptide is illustrated. The triplet nucleotides separated by dots represent the natural reading frame of the gene and colour of each triplet identifies the corresponding amino acid; **b.** The substitution of A to C at the third codon (highlighted in yellow) result in a silent mutation. Silent or synonymous mutations do not change the amino acid sequence, but they might affect mRNA stability and thus alter the protein expression [24]; **c.** The (G>A) at the 5th codon result in a missense mutation that lead to the substitution of Arg to Gly. This type of mutation alter the protein function by impairing the wild type peptide sequence; **d.** The C to T transition at the first base of codon 6 introduces a stop codon that result in a truncated peptide; **e.** Insertion of a cytosine nucleotide at the last base of the third codon impairs the reading frame and result in the change of amino acid sequence. Dots outside the highlighted triplets represent the new reading frame and dots inside the highlighted triplets identify boundaries of codons prior frameshift; **f.** Deletion of Adenine nucleotide at the third codon result in a frameshift deletion. Impaired reading frame result in the altered peptide sequence. In frameshift insertion/deletion (INDEL) translation continues until a stop codon is reached; **g. & h.** When the number of inserted or deleted nucleotides are multiples of 3, the reading frame does not change and depending on the nature of in-frame change (*i.e.* insertion or deletion), new amino acids might be introduced to or lost from the peptide.

1.2.2 Pathogenic Splicing mutations

Point mutations in the *cis*-acting elements of DNA are relatively common. These mutations impair RNA splicing and result in genetic disorders. Essential *cis*-acting regulatory elements that control RNA splicing are located adjacent to exon boundaries. Any single nucleotide changes in the highly conserved GT (at the 5' end) or AG (at the 3' end) have marked effect on RNA splicing and leads to exon skipping or intron retention. In addition to the highly conserved splicing acceptor and donor sites, single nucleotide changes in other *cis*-acting regulatory elements such as splicing enhancers and splicing silencers are also harmful. For example, mutations that impair exonic splicing enhancers (ESEs) result in skipping of the mutant exon by the splicing machinery and disrupt the gene product [17]. Since these mutations result in aberrant splicing they are not readily identified by DNA sequencing and usually expression level data is required to confirm their pathologic implication [25]. Occasionally, single nucleotide substitutions activate cryptic splice sites and result in new splice donor or splice acceptor sites. This class of mutational event impairs normal splicing and results in exon skipping or intron retention. These mutations impair the natural reading frame and result in truncated transcripts with premature termination codons that frequently undergo RNA degradation [26].

As discussed in the earlier sections, the mutation frequency varies across the genome and GC-rich segments of the genome are generally more susceptible to single nucleotide substitutions. The genome-wide germline mutation rate is estimated around 1.18×10^{-8} ($\sigma = \pm 0.15 \times 10^{-8}$) per nucleotide per generation [27], and the genome of each individual collectively harbours 340- 400 putative loss of function variants, but not all of these mutations are pathogenic as they are largely frequent mutations in non-essential genes like blood group genes or Human Leukocyte Antigen (HLA) gene clusters [28].

1.2.3 Structural Variation

Structural variants (SVs) are abundant large scale changes within the genome that vary in size. These mutational events typically involve copy-number variation (often greater than 50bp) and are implicated in both Mendelian and complex disorders [29]. These large scale mutational events include:

1. *Deletions* which are considered the most abundant SVs in the genome ($\sim$ 79%) and involve removal of DNA segments (Figure 1.4a). Deletions result in the loss of genetic material and impact gene dosage [30].

2. *Insertions* involve addition of new sequence into the genome. Depending on the origin of added sequence, these SVs are classified into novel insertions (figure 1.4.b) or mobile element insertions (Figure 1.4.c). The mobile element insertions account for the second most frequent SVs across the genome ($\sim$ 20%) and novel sequence insertions account for only a mere 1% of SVs across the genome.

3. *Duplications* that result from propagation of native sequences of DNA within the genome. Depending on the arrangement and orientation of duplicated segments these type of SVs are classified as tandem (Figure 1.4.d) or interspersed duplications (Figure 1.4.e). Tandem duplications collectively account for less than 1% of SVs across the genome.

4. *Inversions* are described as chromosomal rearrangements in which a segment of DNA undergoes breakage at two loci and rearranged within itself (Figure 1.4f). During inversion no chromosome material is lost and thus this type of SVs are classified as balanced rearrangements.

5. *Translocations* involve exchange of chromosomal segments between nonhomologous chromosomes (Figure 1.4.g). Depending on whether exchange of chromosomal segments involve acquisition or loss of chromosome material, these SVs are further classified into balanced or unbalanced translations.

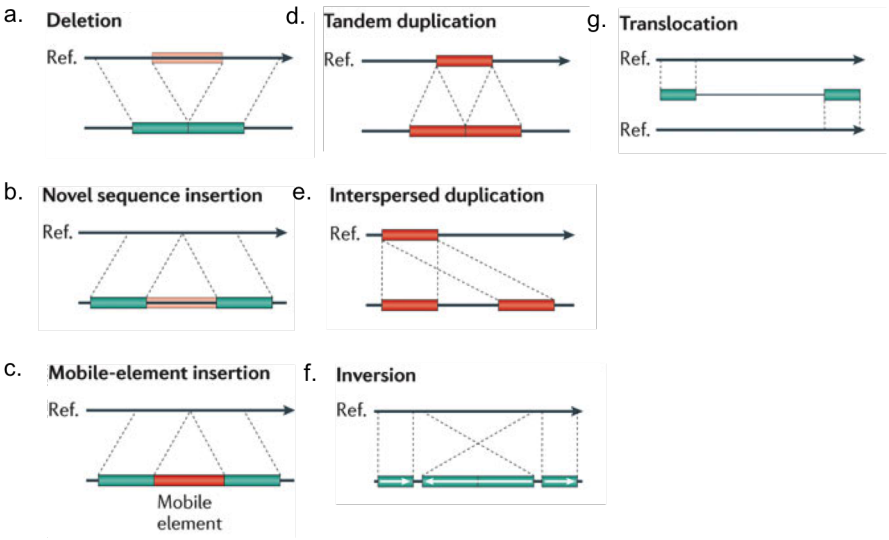


Figure 1.4: Schematic representation of different classes of structural variants (SVs); (Figure adapted from *Alkan et al.*[31]).

Structural variants collectively account for 0.5-1% of inherited genetic differences between individuals[32]. Recurrent *de novo* SVs are known to have high locus-specific formation rates and occur in $\frac{1}{7{,}000}$ live births[33]. Detection of genomic structural variants from next-generation sequencing data is explored in Chapter 4.

1.3 Approaches to Identifying Disease Genes

The idea that human disorders are influenced by hereditary factors was conceptualised in the mid to late 19th century. The Roux-Weismann theory of development during the 1880s proposed chromosomes as the principal determinants of heredity[34] and led to subsequent introduction of the chromosomal theory of Mendelism by Walter Sutton and Theodor Boveri in the early 20th century. A series of landmark studies in the mid-20th century led to molecular dissection of genetic disorders such as albinism and brachydactyly[35]. The search for the molecular component of heredity during this time established the central dogma and led to the molecular disease paradigm[36,37]. A brief overview of conventional methods that have been used in medical genetic research is provided in the following sections.

1.3.1 Linkage Mapping

Strategies for discovering disease genes where little is known about the gene product has relied on the use of linkage. In linkage analysis co-inheritance of marker alleles and disease phenotype is used to localise the causal genomic region. Markers that are close to each other on the same chromosome are more likely to be inherited together on a same stretch of DNA. This is because the chance of recombinational events that separate markers are reduced when marker alleles are positioned in close proximity to each other. In a fully

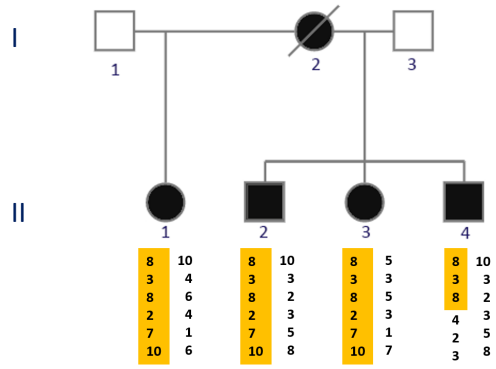


Figure 1.5: Linkage analyses in a pedigree with Darier-White disease. This dominantly-inherited skin condition is mapped to the long arm of chromosome 12. The haplotype within 12q segregating with the disease is shaded in yellow. The shaded 8-3-8 haplotype in individual II-4 denotes the genomic region on chromosome 12 that consistently segregates with the disease among all affected individuals. Figure reproduced from the linkage analyses in British pedigrees ascertained for Darier-White disease[38]

penetrant Mendelian disorder, linkage strategies can be used to identify a chromosomal section that contains mutated genes (Figure 1.5).

The statistical basis of linkage was initially developed in the 1930s for identifying genetic linkage in autosomal dominant disorders[39,40]. Nearly two decades later, Newton Morton introduced parametric logarithm of odds (LOD) score which laid the foundation for the fine mapping of human traits[41]. The principle of parametric linkage analysis revolves around the statistic that can fundamentally be expressed as:

$$lod = Z = \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{(0.5)^{(NR+R)}} \tag{1.1}$$

In this equation θ denotes the recombinant fraction under the inheritance model assumption (Recessive, Dominant or X-linked). NR denotes the number of non-recombinant individuals and R denotes the number of recombinant members. Markers that are not physically linked together have 50% recombination due to independent assortment and thus 0.5 is used in the denominator to calculate likelihood of recombinant offspring under no linkage.

Parametric linkage analyses are a successful approach for mapping genes in monogenic disorders. Parametric linkage analyses is essentially accomplished under the assumption of a specific model that gives details about parameters like the mode of inheritance, disease gene frequency and penetrance of the disease. Parametric linkage analyses, however, have limited power for complex disorders where the mode of inheritance or genotypic penetrance is less clearly defined. Model-free methods of linkage analysis (non-parametric) was proposed for analysing complex disorders[42]. In non-parametric linkage studies patterns of allele sharing among affected relatives are compared against chance expectations and a likelihood-based model is applied to test for linkage[43].

As with other methodologies, linkage analyses have limitations. In the absence of selective mating in studies of human disease, establishing informative pedigrees with multiple fully penetrant individuals is almost impossible and therefore generation of linkage maps for human traits is difficult. In cases where the disease has detrimental effect on fitness, establishing extended informative pedigrees with the trait of interest is practically impossible.

The discovery of restriction fragment length polymorphisms (RFLPs) in 1980 revolutionised human gene mapping and enabled the first human linkage maps to be produced without the need for recombinant phenotype[44]. The principles of genotyping and linkage mapping were successfully applied by genetic laboratories around the world and led to mapping of cystic fibrosis[45] and discovery of the Huntington's disease gene[46]. The principles of linkage mapping have also applied in studying complex disorders, but apart from some fortuitous discoveries like mapping of insulin gene (*INS*) locus in type 1 diabetes[47] and discovery of ApoE locus for Alzheimer's disease[48], these efforts have been largely futile in detecting genes underlying complex disorders. The greater degree of heterogeneity and low effect size in complex disorders requires a different approach for disease gene discovery that is discussed in the next section.

1.3.2 Association studies

The advent of the TaqMan platform in 1993 enabled efficient genotyping of polymorphic markers and small insertion/deletions (INDEL) and paved the way for denser maps of human polymorphisms to be produced at lower costs[49]. This opportunity enabled geneticists to genotype a greater number of loci in larger cohorts at significantly lower costs. This consequently resulted in a dramatic increase in adaption of genome-wide linkage studies. There was however inherent limitations with application of linkage analysis for studying complex disorders. Reduced effect size and low penetrance in complex disorders result in cross-family linkage analyses lacking statistical power.

The comprehensive comparison of linkage and association studies conducted by Risch and Merikangas in 1996[50] proved greater statistical power for association studies in almost all circumstances especially when the frequency of disease allele is low, and effect size is small (*i.e.* scenarios in which odd ratio (OR) is less than 2). This led to the idea of genome-wide association studies (GWAS) that hinged on the delivery of high-density polymorphism map by human genome project. In the same year, the first automated capillary sequencing machine (AB310) released by Applied Biosystems Inc. (ABI) that followed by the release of upgraded 96-capillary ABI 3700 series[51]. This technology was particularly important at the time as it enabled the first draft of human genome project to be released in 2001[52,53]. During the same period efforts in cataloguing human genome variation resulted in the establishment of dbSNP database[54]. Since the establishment of the dbSNP in 1999, the number of indexed variants for the human has sharply increased to 325 million in dbSNP build 150 (Figure 1.6).

The principle of association studies revolves around statistical testing of co-occurrence of alleles and phenotypes in individuals within population. Association studies essentially have a case-control design and disease risk for tested variants is typically measured by odds ratio (OR). In association studies, the frequency of genotypes and/or alleles in a large cohort of unrelated individuals with the trait of interest (*cases*) are compared against the frequency of genotypes/alleles in ethnically matched *controls*. The odds of being affected when possessing a specific variant is expressed as:

$$OR = \frac{D_E/D_N}{H_E/H_N} \tag{1.2}$$

In this equation D_E denotes total number of risk alleles among affected individuals; D_N denotes total number of alternative alleles among cases; H_E denotes total number of risk alleles among matched controls and H_N denotes total number of alternative alleles among controls.

Significant association between a genetic marker and the trait is not always due to genetic factors. Population substructure and history can play a substantial role in false

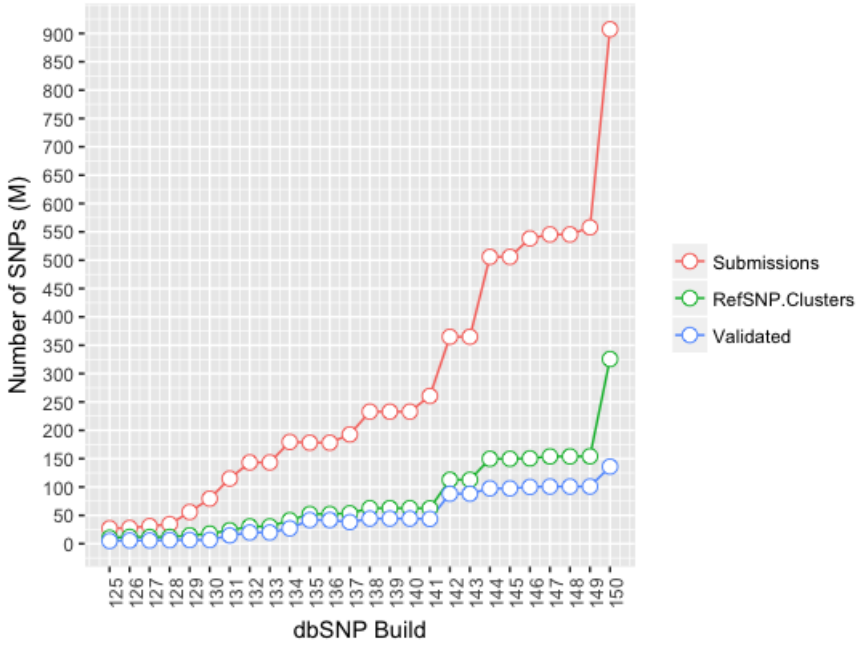


Figure 1.6: Growth of dbSNP for *Homo Sapiens* between Sep. 2005 to Feb. 2017. The dbSNP builds are identified across the x-axis and the number of SNPs are expressed in millions across the y-axis. The red line represents the total number of submitted SNPs (#ss). The green line represents the total number of unique, non-redundant reference SNP (#rs) and the blue line illustrates the total number of validated SNPs. Data extracted from dbSNP Summary statistics (`www.ncbi.nlm.nih.gov/projects/SNP/snp_summary`)

positive association signals. For instance variation in admixture proportions in admixed populations such as Latin American Mestizo populations result in false positive associations [55,56] or variations in allele frequencies in African populations can result in spurious association signals [57]. Inevitably, it is mandatory that cases and controls have the same population ancestry and origin to avoid stratification bias in association studies [58].

1.3.3 LD and genome-wide association studies

Human genome variants implicated in pathogenic conditions are either directly involved in the disease or tightly linked to a disease-susceptibility allele. In complex conditions where the causal variant (A) is linked to a susceptibility allele (B), the frequency of the haplotype containing both variants (F_{AB}) will be higher than the frequency of both alleles estimated under the independence assumption($F_A F_B$) among the affected individuals. This non-random association of two or more loci in the human genome is so called linkage disequilibrium (LD) and is quantified by the *coefficient of linkage disequilibrium* according to formula below:

$$D_{AB} = P_{AB} - P_A P_B \tag{1.3}$$

Where P_{AB} is the probability of both alleles occurring on the same haplotype and $P_A P_B$ is the probability of alleles segregating independently [59].

LD became very significant in the development of association mapping haplotype structures and paved the way for genome-wide association studies (GWAS) [60,61]. Linkage disequilibrium (LD) between two or more loci occurs when combination of alleles are positively selected due to the evolutionary advantage they confer, or simply when they are positioned at proximate loci where recombination is reduced [62]. Discovery of LD between SNPs enabled GWAS studies to be conducted with a manageable number of tag-SNPs

(Figure 1.7).

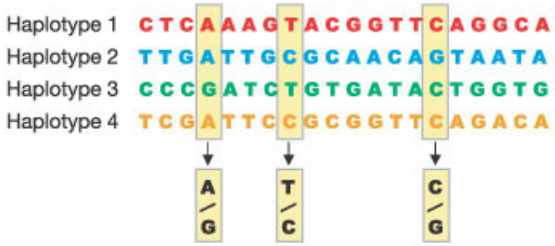


Figure 1.7: Haplotypes and tag-SNPs as the underlying principles of GWAS. LD at neighbouring SNPs result in a specific combination of alleles that are always inherited together. These blocks are known as haplotypes. Tag-SNPs are surrogate genetic markers that represent genomic haplotypes where adjacent SNPs are in LD with each other. Here four haplotypes are tagged only by the three bi-allelic markers (shaded in yellow). Genotyping of these three tag-SNPs is sufficient for identification of genetic variation and association to traits without the need for genotyping all SNPs; (Figure taken from the international HapMap consortium [63]).

As discussed earlier, human genetic variation differs across populations and LD patterns vary in different populations. To discover and catalogue genetic variation across different populations with ancestral origin in Africa, Asia and Europe, the HapMap project was established in 2002 [64]. Introduction of high-density genotyping arrays by Affymetrics and Illumina in combination with the data generated in the HapMap project enabled GWAS to be successfully applied in the context of complex disorders [65].

The first successful GWAS report was published in 2005 by Klein *et al.* and involved identification of two associated SNPs in the context of age-related macular degeneration [66]. Despite small sample size (*96 Cases vs. 50 controls*), a great deal of this fortuitous discovery was indebted to the large effect size rendered by the identified variants.

Genome wide association studies greatly contributed to our understanding of complex disorders, but their inherent limitations in detecting association for low frequency variants restricted their use [67]. The tag-SNP approach that facilitate GWAS is ill-suited for populations with different LD to HapMap population. Furthermore, variants with low frequency are not well tagged by common SNPs and therefore GWAS studies are deficient in detecting association for such variants. As proposed by Pritchard, low frequency variants play an important role in susceptibility to complex disorders and GWAS limitations in detecting such variants cannot be overlooked [68].

1.3.4 Next-generation sequencing

Alongside the development of microarray chips, the advent of 454 pyrosequencing method in 2004 enabled massively parallel sequencing and led to the birth of next-generation sequencing [69]. A year later, Illumina introduced the Solexa platform that was based on reversible termination sequencing method and ABI commercialised its SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform in 2007. The arrival of next-generation sequencing (NGS) technology enormously increased sequencing capacity and enabled sequencing of larger number of samples in a considerably shorter period of time [70].

Another technological progress that enabled cost effective sequencing of a subset of genome introduced by NimbleGene in 2007. Their sequence capture technology enabled enrichment of pre-specified part of the genome in microarray applications. This enabled enormous cost reduction through concentrating genotyping effort on the region of interest [71]. This advancement led to the birth of whole exome sequencing (WES), in which

only the coding region of the genome (exome) is selected for sequencing. The advent of enrichment technology and WES revolutionised gene hunting for Mendelian disorders [72].

These technological advancements collectively resulted in a new collaborative effort for fine mapping of human genome variation. As a result the 1,000 genomes project founded in 2007 with the aim of generating a near-complete map of genetic variation in a thousand human individuals [73]. While the initial release of the project comprised only 16 million variant sites (including SNPs, indels and structural variants) across 180 HapMap samples [28], the final release of the project consisted of over 88 million variants (84.7 million SNPs, 3.6 million indels and 60,000 structural variants) across 2,504 individuals from 26 different populations [4].

The introduction of whole exome sequencing (WES) in combination with the progress made in cataloguing low frequency variants by the 1,000 genome project resulted in a paradigm shift in disease gene identification for Mendelian disorders. Following the discovery of the causative mutation in Miller syndrome by exome sequencing [74], WES has become the dominant approach in identification of disease gene for Mendelian disorders.

The unifying characteristic of NGS platforms is their unprecedented ability to sequence millions of DNA fragments from different samples simultaneously. The NGS technology relies upon alignment or *de novo assembly* of numerous short overlapping reads generated from fragmented genomic DNA (gDNA) [75]. The application of NGS in disease gene discovery has two particular advantages over the traditional methods such as linkage and association studies. The sequence reads from NGS platforms provide direct information about variants in the target region and therefore contrary to linkage studies, no prior assumption about the genomic position of any pathogenic variant is required. Furthermore, clinical NGS can be successfully applied for GWAS studies to tackle the problem of power reduction arising from recombination (reduced linkage disequilibrium) between the tag SNPs and pathologic variant [76,77,78].

The limitations of GWAS and linkage meta-analysis for detecting rare variants with intermediate to low penetrance highlights the limitations of traditional disease gene discovery strategies. In fact, a disease with locus heterogeneity and/or reduced penetrance cannot be successfully identified by traditional strategies. Furthermore, rare disorders segregate in only a small number of cases or families and the availability of a limited number of cases usually restrict the power of disease gene identification by conventional approaches. Development of massively parallel sequencing in combination with targeted gene capture enabled cost-effective identification of rare variants in Mendelian disorders.

Disease gene discovery efforts in complex disorders have considerably benefited from the advancement in NGS technology. The power of NGS in detecting low-frequency variants primarily came to light in 2009 when the protective role of *IFIH1* low-frequency variants identified via direct sequencing of exons and splice site of 10 candidate genes [79]. Another landmark study into the genetics of inflammatory bowel disease (IBD) in 2011 that utilized NGS in studying 56 candidate genes led to the identification of novel risk factor in *NOD2* and two protective variants in *IL23R* [80]. Furthermore, the success of WGS in identifying hotspots for *de novo* mutations in complex psychiatric disorders such as autism [81] has proven the clinical utility of NGS in complex conditions. These primary successes of NGS in identifying low-frequency variants for complex disorders combined with its achievements in the identification of aetiological variants underlying Mendelian disorders [74] resulted in whole genome/exome sequencing to become the mainstream strategy for disease gene discovery. Whole genome sequencing (WGS) also enables high-resolution LD maps to be produced [82] that in turn will facilitate the design and implementation of imputation-based strategies for GWAS studies of low-frequency variants in complex disorders [83].

An overview of conventional strategies in disease gene discovery and respective study

design is provided in Figure 1.8.

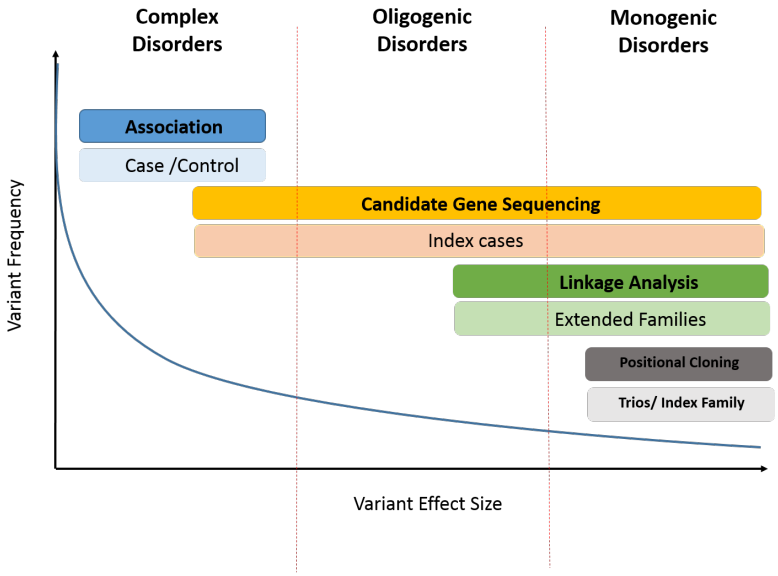


Figure 1.8: **Overview of conventional approaches in disease gene identification.** Common variants that underlie complex disorders are originated from mutations in the distant past that have subsequently reached moderate frequency in the population. Based on the common disease-common variant hypothesis these variants generally have low impact and in combination with other common variants contribute to complex disorders. Association studies detect such loci and have case/control design [84]. At the other end of the spectrum are fairly recent mutations with a very strong effect that underlie aetiology of Mendelian disorders. These variants are subject to strong selective forces that will limit the frequency of these mutations in the general population [85]. *De novo* mutations with large effect size can be readily identified in NGS experiments with trios/index family design [3]. Identification of variants with intermediate effect size is carried out by studying candidate genes in index cases.

1.4 Applications of NGS in clinical diagnoses and population genomics

Numerous aspects of NGS technology including cost-effectiveness, speed, high resolution and accuracy of data generated by massively parallel sequencing has made NGS the primary method of choice in genomic research. Falling costs of NGS resulted in a sharp increase in the number of studies utilising NGS capabilities in disease gene discovery (Figure 1.9).

NGS technology is increasingly used in the clinical setting not only for identification of pathological variants but also for *de novo* genome assembly of pathologic organisms for which a reference genome sequence is not available. In the recent Ebola virus (EBOV) outbreak in West Africa, NGS technology was successfully applied to produce assembly of full-length EBOV genomes and enabled rapid public health response to the crisis [86]. Furthermore, the application of NGS technology at population level in the African Genome Variation project enabled dense genotyping of 18 ethno-linguistic groups from sub-Saharan Africa [87].

In recent years, several implementations of NGS technology have been developed by leading companies such as Illumina (`https://www.illumina.com`), Pacific Bioscience (`http://www.pacificbiosciences.com`) and Oxford Nanopore (`http://www.nanoporetech.com`). While multiplexing sequencing reaction is the unifying feature of

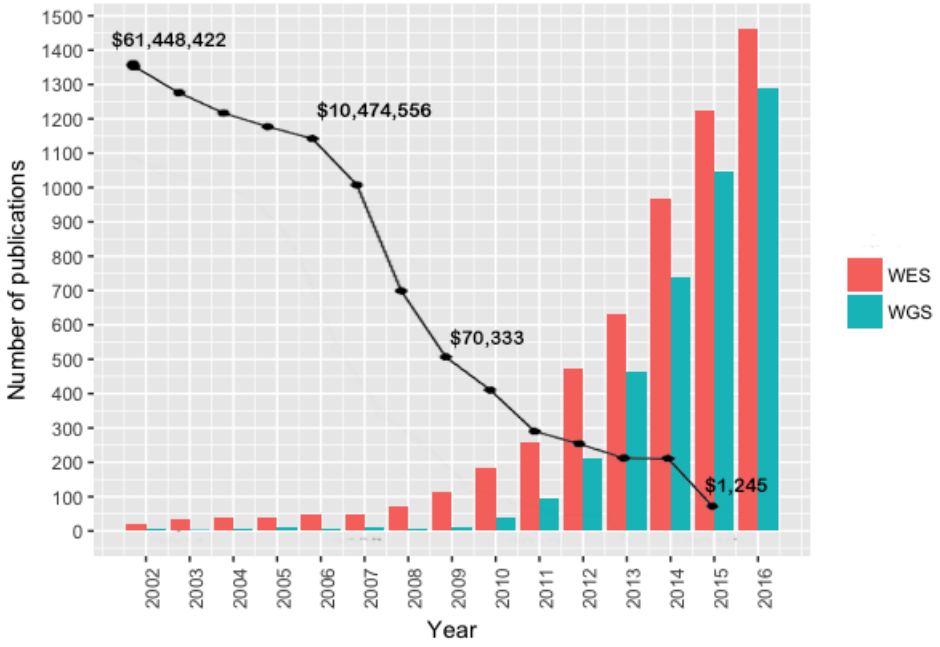


Figure 1.9: Number of studies utilising massively parallel sequencing technology from year 2002 to 2016. Falling cost of NGS enabled more genetic labs around the world to use NGS technology in disease gene discovery. Prices shown (black line) are the cost of sequencing per genome through 2002 to 2015 in US dollars. Data extracted from PubMed (`https://www.ncbi.nlm.nih.gov/pubmed`) and the National Human Genome Research Institute (`https://www.genome.gov/sequencingcosts`).

all these platforms, their throughput and read lengths vary greatly. Details of popular NGS platforms are summarised in Table 1.1.

The application of NGS in Mendelian disorders over the past decade has resulted in a substantial increase in our understanding of molecular basis of Mendelian disorders. The fact that NGS enables simultaneous analysis of multiple regions in the genome has made it a cost-effective alternative option for molecular diagnosis in the clinic. In situations where single gene testing fails or when genetic heterogeneity underlies the condition, WES/WGS can be successfully applied for establishing a diagnosis. The challenge, however, remains as to whether to use targeted gene panels or whole-exome/genome sequencing. Although WES/WGS assess a larger portion of the genome, targeted gene panels offer a cost-effective solution for analysing a particular set of genes with respect to a specific phenotype. In WES where only the protein-coding region of the genome (exons) are studied, the overall depth of sequencing data drops in regions with high GC content, sequence homology or repetitive sequences. In these scenarios, targeted gene panels offer a better solution for identifying aetiological variants[88]. Furthermore, the deeper coverage achieved in targeted sequencing provides greater confidence for variant detection across known phenotype-specific genes. The clinical utility of targeted approach for molecular diagnosis is discussed in Chapter 5.

The diagnostic utility of NGS is particularly pronounced when applied for the detection of rare variants underlying Mendelian disorders. Whole exome sequencing offers an ideal approach for identifying novel loci in Mendelian disorders. Due to the increased power of WES in identifying causal variants even in a small cohort it is readily applied in the research setting to identify causal genes in undiagnosed genetic conditions with varying mode of inheritance[89]. In studies carried out by Yang *et al.*, the application of WES for establishing the molecular diagnosis in patients with suspected genetic conditions achieved a 25% molecular diagnostic rate[90,91]. Application and analysis of WES data in establishing molecular diagnosis for nephrolithiasis patients is discussed in Chapter 3.

Rapid advances in sequencing technology and computation capabilities has made WGS

more accessible to the clinic. WGS offers an unprecedented opportunity to decipher all types of genetic variation in all parts of the genome. Genomics England exemplifies one of the largest collaborative research established to harness the potentials of WGS in the clinic. WGS not only provides a thorough picture of non-coding regions of the genome, but also outperforms WES in covering coding segments of the genome [92]. In a recent study, diagnostic yield of WGS for establishing molecular diagnosis in 217 patients for whom previous screening failed to identify the pathogenic variants is identified to be 34%. This figure increases to 57% in trios [93]. Furthermore, WGS allows identification of all forms of structural variation covered in the section 1.2.3. WGS is changing the foreground of healthcare and shifting medical practice toward personalised medicine. The application of WGS in identifying structural variation is explored in Chapter 4.

1.4.1 Limitations of Next Generation Sequencing

Although NGS technology has been widely adopted and both clinical and research laboratories around the world benefited from it's high throughput, NGS still has major limitations. Although, the error rate associated with the base calling varies depending on the NGS platform, it usually ranges from $\sim 0.1\%$ to 15% which is higher than traditional Sanger sequencing method [75]. This higher error rate is specifically problematic when it leads to false negative or false positive discoveries [94]. Algorithmic solutions such as GATK Base Quality Score Recalibration (BQSR) have been proposed for correcting the systemic errors made by sequencers [95], but base calling error rates are still identified to be correlated with minor allele frequencies of the variant site. In that sense the error rate of 4-6% for rare SNVs implies the necessity of corroborating NGS finding by Sanger sequencing [96]. Recently, a more sophisticated machine learning algorithm has been suggested to tackle this problem [97].

A second limitation of NGS platforms lies in their relatively short read length that makes them ill-suited for variant calling in repetitive and low complexity regions of the genome. The current best practice NGS analysis relies heavily on mapping-based variant calling where short-reads are uniquely mapped to the reference genome. In low complexity regions where tandem repeats stretch to hundreds of base pairs, short reads could not be uniquely mapped to the reference genome resulting in erroneous alignment and spurious variant calls [98,99]. A classical example of this problem in exome sequencing is highlighted by Fajardo *et al.* where misalignment in low complexity regions and paralogous genes or pseudogenes result in false positive signals [100].

The read length in the majority of conventional NGS platforms ranges from 35bp to 250bp (Table 1.1). Platforms with long-read sequencing capability have been developed [101], but the higher read lengths comes at the expense of higher prices which has been the limiting factor for their widespread adoption in the clinic and research laboratories [75].

Depth of coverage across the genome in NGS methods is not uniform and GC- and AT-rich segments of DNA are usually undercovered [102]. In addition due to short read lengths, the phase of variants cannot be resolved directly and imputation techniques must be applied [103].

Finally, the relatively short length of reads in NGS technology also impose algorithmic challenges for *de novo* assembly of genomes. The majority of assembly algorithms use the de Bruijn graph and Eulerian path approaches [104] which is inefficient in assembling segmental duplications and larger common repeats such as Alu repeats [105]. Efforts into *de novo* assembly of genome have led to improved algorithms [106], however full resolution of genome assembly still relies on larger read lengths.

	Illumina		Pacific Bioscience	Oxford Nanopore
Platform	HiSeq 2500 v3	NextSeq 500/550	MiSeq v2 High output	PacBio RSII	MK 1 MinION
Sequencing principle	Reversible terminator sequencing by synthesis			Real-time single molecule DNA sequencing	Nanopore exonuclease sequencing
Detection method	Fluorescence/Optical	Fluorescence/Optical	Fluorescence/Optical	Fluorescence/Optical	Electrical Conductivity
Read Length	36 (SE) 50 (PE) 100 (PE)	75 (SE) 75 (PE) 150 (PE)	25 (PE) 36 (SE) 150 (PE) 250 (PE)	~ 20 Kb	up tp 200Kb
Number of Reads	1.5 billion (SE) 3 billion (PE)	400 million (SE) 800 million (PE)	12-15 million (SE) 24-30 million (PE)	~ 55 K	>100K
Runtime	2 - 11 days	11- 29 hours	4- 39 hours	4 hours	up to 48 hours
Throughput	47-52 Gb (SE_36) 135- 150 Gb (PE_50) 270-300 Gb (PE_100)	25-30 Gb (SE_75) 50-60 Gb (PE_75) 100-120 Gb (PE_150)	750-850 Mb (PE_25) 540-610 Mb (SE_36) 4.5-5.1 Gb (PE_150) 7.5-8.5 Gb (PE_250)	500 Mb - 1 Gb	up to 1.5 Gb
Accuracy	>99% (<0.1% error in substitution)	>99% (<0.1% error in substitution)	>99% (<0.1% error in substitution)	84 - 85% (13% single-pass error + <1% error in indel)	~ 88% (~ 12% error in indel)
Instrument Cost	$ 740,000	$ 250,000	$ 125,000	$ 695,000	$ 1,000
Advantage	Very high throughput Cost effective	Wide range of applications Medium range price	Cost effective Short runtime	Short runtime Very long reads	Extremely long reads Low cost
Disadvantage	Long runtime Expensive instrument	Not suitable for gene expression profiling	Suitable for microbial applications Short read length Not Suitable for WGS	Suitable for de novo assembly Highest error rate Expensive instrument	Portable device High error rate

Table 1.1: Summary of commonly used NGS platforms. SE, single-end sequencing; PE, paired-end sequencing; indel, insertions and deletions. Data extracted from the Illumina fact sheet (`https://www.illumina.com/systems/sequencing-platforms`) and Goodwin *et al.*[75] & Lee *et al.*[107]

1.5 Third-Generation Sequencing

The limitations of NGS (especially short read lengths) necessitated a new generation of long-read sequencing platforms to be developed. In the third generation sequencing unlike NGS where sequencing is carried out by amplification and synthesis of millions of short read sequences, the long-read sequencing is achieved by base calling nucleotide sequence at the single molecule level [108].

The field of long-read sequencing is under active development and to date two major platforms including the PacBio Single Molecule Real-Time sequencing (SMRT) [109] and Oxford Nanopore Technology (ONT) [110] platforms have been commercialised. The core concept in the both technologies revolves around single molecule sequencing without the need for amplification and library preparation (in contrast to NGS).

Single Molecule Real-Time (SMRT) sequencing is a synthesis based platform in which incorporation of differently labelled nucleotides into the DNA strand is optically monitored and translated into long-read sequences. The single polymerase reactions in the SMRT technology occurs in a tiny chambers known as Zero-mode waveguide (ZMW). These chambers have 100 nanometre (nm) depth and 70nm width and accommodate only one single strand of DNA. The release of a fluorescent label upon incorporation of a new nucleotide results in emission of a signal that is recorded in real-time. One SMRT-cells comprises ∼150,000 ZMWs and generates sequence reads of 10 kbp on average [109,111].

In contrast to SMRT technology where the sequencing reactions occurs inside aluminium nano chambers (ZMWs), in ONT, the sequencing occurs inside nanopores that are embedded in a phospholipid bilayer [112]. In this technology negatively charged single stranded DNA is forced to pass through the nanopore channel and the base sequence is detected in real-time from the change in the amplitude of ion current applied to the opposite sides of the bilayer [113].

Among the four platforms developed based on the nanopore technology (MinION, GridION$_{X5}$, PromethION, SmidgION), the MinION has been widely adopted by research laboratories due to its affordable price and portability. The MinION platform is a pocket-sized USB shape device comprised of 512 nanopore channels that generates long read sequences up to 150 Kbp [114]. The MinION device however has notoriously high error rate (∼25- 40%) that renders it unsuitable for clinical use. Recent advances in variant detection algorithms for the MinION nanopore sequencer promised some hope for its future applications in the human disease gene discovery [115,116].

The third generation sequencing has been successfully applied to microbiome analysis [117,118,119] and gene-fusion discovery [120], but it has not yet widely applied to the field of human genetics. In particular the high error rate and expensive price associated with the third generation sequencing have limited its application in the field of human genetics.

1.6 High Throughput Sequencing Challenges

The field of human genetics has enormously benefited from the advances in the high throughput sequencing technology. NGS technology in particular has enabled vast amounts of genomic data to be produced in a relatively short time frame at affordable cost. The great success of high throughput sequencing, however, has brought about several challenges including data storage bottleneck, computation requirements for *in silico* analysis, data interpretation, data protection and patients privacy.

The reducing cost of genome sequencing in the recent years (Figure 1.9) resulted in a huge amount of genomic data that requires intensive data storage and distributed computing resources for analysis. In the recent years, the cost of sequencing has dropped at a faster pace compared to the cost of data storage [121] (Figure 1.10). This has resulted in

a data storage bottleneck necessitating distributed storage and cloud computing solutions for tackling this challenge [122].

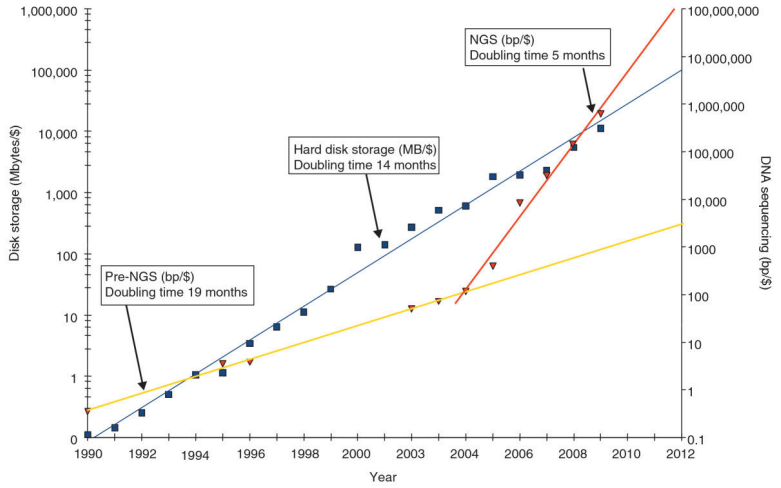


Figure 1.10: Relative historical trends in data storage capacity and DNA sequencing costs per US dollar. The blue squares represents the capacity of storage in megabytes per US dollar and the red triangles identify the number of base pairs that can be sequenced per US dollar (All numbers represented in logarithmic scale). The blue line shows exponential growth in the capacity of storage per US dollar. The historic trend in DNA sequencing capability per US dollar is represented by the yellow line and red line. The doubling time for sequencing capability until 2004 is slightly slower than the growth in data storage capacity (blue line vs. yellow line), however the advent of NGS technology in 2004 inflect the curve and in 2009 doubling time for sequencing capability exceeds that of storage capacity; (The figure adopted from the review paper by *Lincoln D Stein* [121])

In silico analysis of the large amount of data generated by NGS technology is computationally demanding and requires intensive processing capabilities such as supercomputers which are beyond the conventional computing resources available to the diagnostic labs. This has resulted in implementation and deployment of genomics pipelines on commercial cloud computing services such as Amazon EC2 [123].

Similar to conventional genetic test results, the biological and clinical interpretation of variants discovered by high throughput sequencing have considerable importance for the patients, their parents and their relatives. Unfortunately, the majority of variants identified by whole genome sequencing are either non-coding or coding variants of uncertain significance (VUSs) which require scalable functional studies to establish their biological impact [88].

Given the extent of genomic data generated by high throughput sequencing, incidental findings are common in clinical exome and genome sequencing. These incidental findings are unrelated to the primary indication for ordering sequencing but might have medical value for the patient health. The extent to which incidental findings should be reported to the patient has been a matter of considerable debate. In order to address this challenge, the American College of Medical Genetics and Genomics (ACMG) has published a guideline for reporting incidental findings [124].

Finally, there has been a growing concern over the privacy of the patients in clinical exome or genome sequencing. It has been suggested that a collective set of 75 statistically independent SNPs is adequate to uniquely identify the owner of the DNA sample [125]. It is estimated that individuals can be identified in a pooled mixture of genomic DNA even when they contribute as low as 1% to the DNA pool mixture [126]. In the era of big data, securing patients' identity imposes one of the greatest challenges in the field of high throughput sequencing. With the availability of public databases and distributed storage services, it is essential to ensure patients privacy.

1.7 Overview of Thesis

In this chapter an overview of genetic mutations underlying human disorders is provided. Different approaches to the human disease gene discovery is discussed and advantages and disadvantages of each approach is reviewed. The progress in high-throughput sequencing that enabled better understanding of rare Mendelian disorders is discussed and limitations of NGS technology outlined. Finally, the challenges that are faced in the era of high throughput sequencing and potential solutions are described.

In **Chapter 2**, the methods applied for analyses of samples are reviewed and the specificities of pipelines developed for *in silico* analysis of samples is discussed. Different aspects of the pipelines including alignment, variant calling and annotation is explained and strategies for identification of aetiological variants in WES and WGS applications are detailed. Finally, the performance of the pipelines used for variant analysis in this thesis is compared and strategies for increasing sensitivity and specificity of pipelines are discussed.

Chapter 3 describes the application of whole-exome sequencing in nephrolithiasis. This chapter explores how analytical pipelines discussed in Chapter 2 are applied in an oligogenic condition to discover the aetiological variants.

In **Chapter 4**, application of whole exome sequencing and whole genome sequencing for resolving a rare case of congenital anomaly is discussed. Furthermore, this chapter describes how WGS can be applied in the field of rare disorders to identify pathogenic structural variants.

In **Chapter 5**, the utility and application of gene-panels for genetic diagnosis in faciogenital dysplasia is reviewed. Phenotypic and genetic heterogeneity in faciogenital dysplasia is reviewed and limitations of targeted gene-panels for identification of causal variants in the context of disease discussed.

In **Chapter 6**, genome-wide pattern of linkage disequilibrium (LD) across the genome of sub-Saharan African populations is explored and the relationship between the gene-specific LD structure and numerous aspects of the gene including essentiality and tendency to contain disease causing variation is investigated.

Finally, the concluding remarks are detailed in **Chapter 7.**

Chapter 2

Methods and Analytical Pipeline

2.1 Introduction

Traditionally, studying rare disorders has been difficult, and lack of financial incentives concerning profit return had limited research on these disorders. According to the European Commission, disorders that have a frequency of less than 1 in 2000 are categorised under the umbrella term of rare disorders [127]. Rare disorders although rare individually, collectively have a prevalence of 6 to 8% in the general population and affect 27-36 million people around the world [128]. Studies show that ∼80% of rare disorders have an underlying genetic aetiology and comprise up to 20% of paediatric inpatient admissions [129]. It is estimated that rare disorders underlie 31% of death in paediatric medicine and 51% of neonates diagnosed with a rare disorder fail to thrive during the first year of life [130].

Studying rare disorders in the last few decades has enabled identification of novel genes and molecular pathways that were not known before and has enormously increased our insight into the aetiology of many rare conditions. As an example, studies into severe forms of familial insulin resistance by Barroso *et al.* enabled the identification of key genes that also underlie type 2 diabetes [131]. Despite these developments, rare disorders are generally under-diagnosed [93,132]. The latter notion is reflected in the 3,341 listed disorders in the OMIM database for which the molecular basis is unknown (as of October 2018). Given the significant burden of rare disorders on health systems (184% extra hospital charges compared to other diseases [133]) and long diagnostic odyssey for patients and their respective families, a paradigm shift in studying these disorders is required.

Recent progress in NGS technology enabled the successful discovery of genes related to rare genetic disorders [127]. Contrary to Sanger sequencing where only one locus is studied at a time, NGS enables mutational screening across the whole genome or whole exome. Three main NGS strategies applied in the context of rare disorders include whole genome sequencing (WGS), whole-exome sequencing (WES) and targeted gene panels. Each strategy has its own advantages and limitations, and the choice of each option depends on the nature of the disorder and the cost incurred. Some classical rare disorders such as ciliopathies with well studied morbid genome are good candidates for targeted gene panels [134], whereas rare disorders for which the genotype-phenotype correlation is less clear, WGS or WES approaches are more recommended [93].

2.1.1 Data Analysis: The NGS Bottleneck

As high-throughput sequencing (HTS) continues to improve and pave its way into routine clinical practice, the challenge of analysis and interpretation of data becomes immense. Some of the challenges in relation to the data analysis aspect of NGS is discussed in

Chapter 1 (Section 1.6), and here I briefly overview technical aspects of data handling in relation to clinical NGS data.

Currently, Illumina platforms that are based on short read sequencing dominate the sphere of clinical resequencing applications and the majority of data generated so far is based on short-read sequencing technology (Table 1.1). Despite the advantages offered by this technology, the sheer amount of data generated by short-read sequencing poses significant data storage and analysis challenge. For instance, a BAM (Binary Alignment MAP) file of a single human WGS at 30X coverage occupies ∼90GB of hard-disk space. This figure readily scales up to thousands of terabytes (TB) even in a modest sample-size project[135]. Two major factors that influence the size of a BAM file include the length of short-reads and the average number of reads per each base (known as coverage) across the target region. In general, the size of a BAM file is directly related to the size of the target region and the fold coverage across the target region. (Table 2.1). BAM files are usually converted to VCF (variant call format) files that only contain information about the sites that are different from the reference genome. Although these files are much smaller in size (∼30 Gb for WGS and up to ∼70 Mb for WES), the raw sequence data (FastQ files) and raw alignment data (SAM/BAM files) are essentially kept for future reference. Furthermore computing time and power spent on generating processed data (*i.e.* BAM and VCF files) from the raw sequence data (FastQ files) make them valuable to be retained on the hard-disk memory. All these easily scale up the memory requirements for NGS data analysis.

Platform	Target region	Coverage	No. of reads	Read length	BAM file size	Processed file storage (FQ+ BAM+ VCF)
Whole-genome sequencing (WGS)	6.6 Gb	34.1 X	727,228,800	147	91 GB	175 Gb
Whole- exome sequencing (WES)	71 Mb	62.7 X	40,070,889	75	5.6 GB	41 Gb
Targeted gene panels	12 Mb	66.7 X	47,123,868	151	4.9G	11Gb

Table 2.1: Comparison of file storage requirements for different strategies in clinical NGS applications.

The second major challenge in analysing clinical NGS data is the choice of best software and data analysis pipeline for reliable and reproducible interpretation of NGS data. The sensitivity of NGS pipelines extensively relies on the successful identification of noise and false positives (including systemic biases such as batch effect, strand-bias effect and coverage-bias effect among others) from true positives in a high dimensional data space. This computationally intensive task requires robust statistical and machine-learning models for optimal analysis of data generated through NGS applications. As of October 2018, over 10,300 tools for WGS and WES analysis are listed at OMICtools[136], for which a consensus over their clinical reliability is lacking. Given this large number of algorithms and tools available for NGS analysis, designing and implementing an optimal pipeline is challenging.

As discussed in Section 1.6, clinical interpretation of variants identified through NGS analysis, especially when it pertains to patient care, in the absence of functional analysis can be challenging. A typical WGS sample contains ∼3 million variants, and WES generates up to 30,000 variants per sample. All these variants are not pathogenic, and therefore a robust filtering strategy is required to discern causal mutations from benign variants. To guide interpretation of clinical NGS data several prediction algorithms and classifying strategies have been developed, but still, a consensus on the best practice is lacking[137]. This signifies the importance of a tailored pipeline specific for analysing Mendelian disorders.

Replication of results from clinical WGS/WES data relies upon access to the exact same pipeline, source codes and parameter settings (including software versions and

genome assembly release) used for initial analysis. In the absence of a standard pipeline or lack of consensus over best practice in clinical NGS analysis, replication and validation of NGS results appear extremely challenging[137]. All these hurdles necessitate design and implementation of a standard pipeline that enables reliable and reproducible interpretation of NGS data in the clinical setting.

2.1.2 Chapter overview

The aim of this chapter is to describe the *in-silico* pipeline and analytical methods used for analyses of NGS data across this thesis. Since the library preparation and sequencing was usually outsourced or carried out by academic collaborators, methods in relation to library preparation are only briefly reviewed.

In this chapter, the general workflow used for processing short-read sequencing data is introduced and a discussion about the different steps involved in analysing WES and WGS is followed. Since targeted gene panels are a specific form of WES in which only a subset of pre-selected genes is exome sequenced, the *in-silico* methods used for analysing this class of data is covered under WES analytical pipeline. Given that the majority of samples analysed in this thesis are of the WES type, the performance of three WES pipelines for variant discovery investigated and rationales for choosing the best pipeline are provided. In the final section, technical details of the WGS pipeline used for identification of structural variants is overviewed.

2.2 Methods

2.2.1 Samples and clinical phenotypes

The patients studied in my research were acquired through collaboration with local and international clinicians and researchers. The clinical phenotypes studied in my research include nephrolithiasis, skeletal dysplasia and syndromic CLP (with the provisional diagnosis of Aarskog-Scott syndrome) which are discussed in chapters 3, 4 and 5 respectively. For all samples, ethical approval of the respective organisation was in place and sample collection performed upon obtaining informed consent from the patients. The patients' data were processed according to the good clinical practice (GCP) guidelines set by the national institute for health research (NIHR). Table 2.2 provide an overview of cases and study designs covered in this thesis.

Chapter	Phentype	Study Design	Target region	Origin	Number of samples
III	Nephrolithaisis	Index cases	Whole Exome	UK	9
			KASP genotyping		48
IV	Skeletal dysplasia	Index case	Whole Exome	UK	1
			Whole Genome		
V	Aarskog-scott syndrome	Index cases	4813 clinically relevant genes	Colombia	13
VI	Healthy individuals (LD map study)	Population-wide	Whole Genome	Sub-Saharan Africa	320

Table 2.2: Summary of samples, clinical phenotypes and study designs discussed in this thesis.

2.2.2 DNA extraction and quality check

For samples from the UK, DNA extraction from blood or saliva was carried out locally at the Institute of Developmental Sciences (IDS, University of Southampton). For samples from the Colombian collaborators, extracted DNA was provided for use in NGS analyses. All samples underwent primary quality control checks to confirm the quantity and purity of extracted DNA. This step involved quantification of DNA by spectroscopic (Nanodrop spectrophotometer) and fluorometric methods (qubit fluorometer). In spectroscopic methods, the absorbance ratio of ultraviolet light at a wavelength of 260 nanometres (nm) is measured against absorbance at 280 nm, and purity of DNA extract is quantified by calculation of the optical density (OD) measure (Formulae 2.1).

$$OD = Log(A_{260}/A_{280}) \tag{2.1}$$

The A_{260} is also used to quantify the amount of extracted DNA. The widely accepted A_{260}/A_{280} threshold for pure DNA is $\sim$1.8 and samples with ratios below 1.8 are considered impure[138]. In fluorometric methods (using qubit fluorometer), DNA samples are tagged with fluorescent labels that selectively bind to double-stranded DNA (dsDNA) and the absorbance at the wavelength of 260nm is used as a measure of DNA concentration[139]. Minimal DNA amounts and concentrations required for NGS varies widely depending on the library preparation method and sequencing platform. Nevertheless, samples were discarded if they were not meeting minimal DNA concentration required for respective library preparation method.

2.2.3 Library preparation and sequencing

Library preparation is the fundamental step in the pre-sequencing workflow through which nucleic acid target (DNA in case of WGS/WES) is prepared to be used in the sequencing

machine. The size of target capture and library preparation method depends on the choice of capture kit and sequencing platform and slightly differs for different enrichment methods. Given that the majority of whole exome target capture in this thesis was carried out using Agilent SureSelect XT Human All Exon V5 and sequencing performed on Illumina paired-end platform, the target capturing and sequencing for this method is explained.

In WES applications, following DNA extraction 3μg of high-quality genomic DNA (gDNA) is acoustically sheared into small fragments using Covaris DNA Shearing System (Covaris, Woburn, MA, USA). Fragmented DNA is then size selected to obtain DNA fragments in 150-200bp range. Since 80% of human exons are typically shorter than 200bp [140], an average insert size of 250bp for paired-end sequencing is recommended to minimise the occurrence of overlapping reads on Illumina platforms [141].

Size-selected fragments are then used to generate Illumina paired-end DNA libraries. In brief, blunt end fragments are phosphorylated at the 5' end and adenylated at the 3' end to generate the 3'-dA overhang. Modified fragments are annealed to index paired-end adapters to generate adapter-modified fragments. Unligated adapters are subsequently removed and purified. The ligation products are subjected to multiple cycles of PCR using index primers to generate an indexed DNA library. Once amplification is complete, the library is purified from non-amplified fragments with magnetic beads (Figure 2.1).

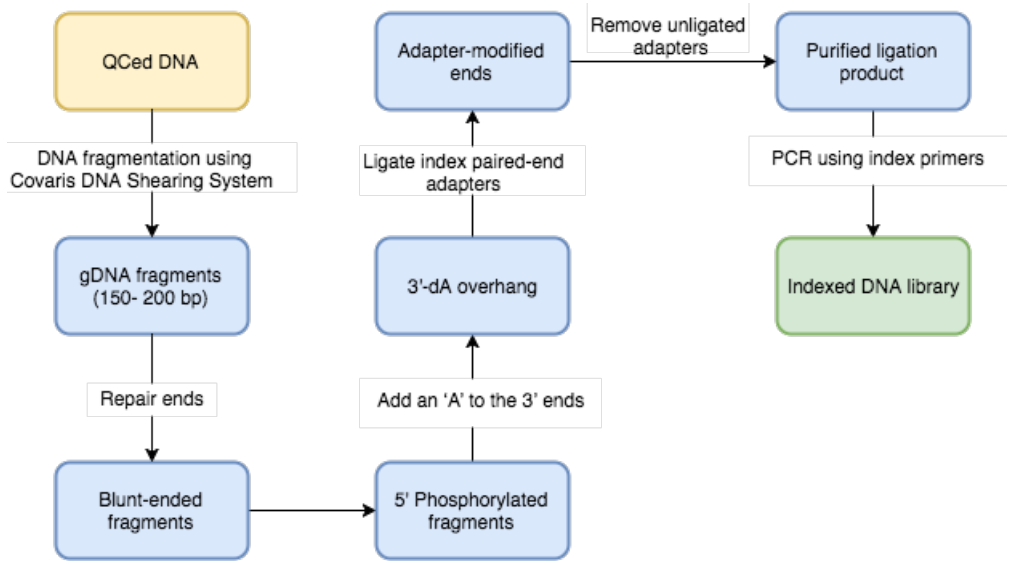


Figure 2.1: Sample preparation workflow for whole exome sequencing on Illumina paired-end sequencing platform. (Figure inspired by Illumina multiplexed paired-end sequencing manual [142])

Following the establishment of the *indexed library*, exonic regions of genes are recovered using Agilent SureSelect XT Human All Exon V5 capture kit. This is achieved by hybridising the exome-containing fragments to biotinylated library baits which are complementary to the exome. Non-covalent binding of streptavidin antibody-coated beads with biotinylated sequences enables capturing of exome-amplified fragments (Figure 2.2. The *enriched library* is subsequently sequenced on the Illumina paired-end sequencing platform (HiSeq 2500) according to the manufacturer's best practice guidelines (Figure 2.2). Paired-End (PE) reads with average size of $\sim$ 75 bp were used for *in-silico* analyses.

The workflow for library preparation in WGS is essentially the same as WES and follows the general steps described above. The exception in WGS applications is that the fragment sizes are generally larger and no capturing step is involved. Specific details for the whole genome sequencing (WGS) and TruSight One sequencing (TSO) panel are explained in the chapters 4 and chapter 5 respectively.

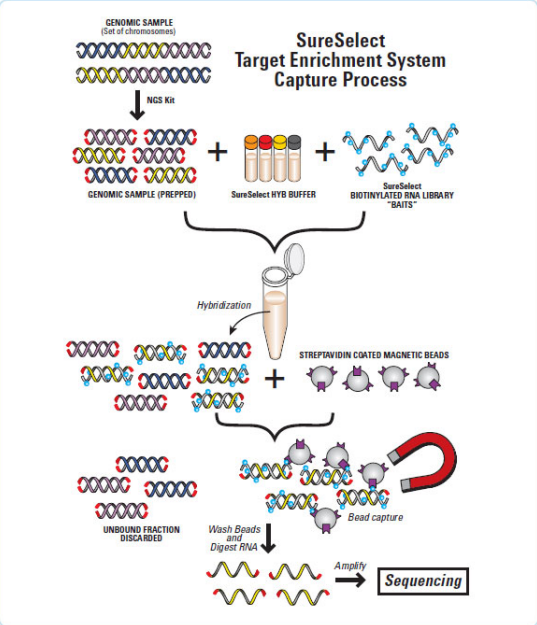


Figure 2.2: Schematic representation of Agilent SureSelect target enrichment technology used for pooling exome-amplified fragments in WES applications. (The figure obtained from `www.genomics.agilent.com/article.jsp?pageId=3083`)

2.2.4 Sample provenance control

In order to ensure data integrity and validity, concomitant to sequencing, all samples are genotyped at 24 SNPs using a KASP genotyping platform (LGC Genomics, Hoddeston, UK) to identify erroneous sample swaps on the same dispatch of DNA. The panel used for orthogonal genotyping of samples comprise 24 bi-allelic SNPs that are captured by all conventional exome capture kits[143]. The DNA fingerprint of each sample at these 24 loci is then compared to the genotypic profile of the respective sample obtained from short-read sequencing and discrepancies are interpreted as sample swap or contamination on the same dispatch DNA plate. Samples that fail this provenance check are flagged for exclusion or replacement.

2.3 *in-silico* analytical pipelines

The pipelines applied for WES/WGS analysis are generally comprised of three fundamental steps: 1) *Alignment*; 2) *Variant calling* and; 3) *Annotation*. The *in-silico* workflow integrates several bioinformatic software and algorithms and allows fine-tuning of parameters related to data analysis. Given the importance of reproducibility in clinical NGS data analysis, development and benchmarking of an automated pipeline is of paramount importance[144]. In this thesis, separate analytical pipelines were employed for analysing whole exome and whole genome data. Details of each pipeline is discussed in the following sections.

2.3.1 WES Analytical pipeline

For analysis of WES data that constituted the majority of samples in this thesis, the Soton Mendelian Pipeline (v.3.0 and v.4.0) in combination with a customised pipeline that integrates variant calls from different callers have been used. The Human Reference Genome build 37 (GRCh37) was used in all the three pipelines to call single nucleotide

variants (SNVs) and short (≤50bp) insertion and deletion variants (INDELs). Although a more recent release of the human reference genome assembly (GRCh38) is available since late 2013, still many software used in this thesis is incompatible with the GRCh38, and therefore the GRCh37 opted as a compromise. It worth mentioning that the possibility of coordinate conversion (liftover procedure) was also explored, but due to minor incompatibilities between the GRCh37 and GRCh38 annotations, to ensure reproducibility of results, all variants were defined according to the earlier version of the human genome assembly. The general workflow used in the three WES pipelines is essentially similar and comprised of same steps that are discussed below:

Raw sequence data QC & preprocessing

Raw sequence data are generated in FASTQ format. Raw sequence reads in FASTQ format is represented in plain text where corresponding quality scores for each base call is encoded as a single ASCII character. The quality score represents the probability of a wrong nucleotide call at each position and is expressed as a Phred (Q) score[145]. This score is logarithmically related to the base-calling error probabilities and expressed as formula 2.2 in which P denotes the probability of incorrect base call by the sequencing machine.

$$Q = -10 \times Log_{10}^{P} \tag{2.2}$$

Due to the nature of short-read sequencing in which adaptor sequences are ligated to DNA fragments during the library preparation process, 3'-adaptor sequences have to be removed prior downstream analysis. Cutadapt tool[146] is employed to remove undesired adaptor sequences from the raw sequence data. Following the exclusion of redundant sequences, the quality of raw sequences is inspected using FastQC software (Babraham Institute, UK)[147]. A per base quality score of 20 is used to exclude low-quality reads, and the over-represented k-mer flag is applied to identify positional bias within sequence reads. The K-mer module of the FastQC software screens for overrepresentation or depletion of predefined length K-mers (default= 7bp) within the library. To identify positional bias, the distribution of predefined length K-mers across all reads is investigated using a binomial test. Distribution of K-mers that significantly deviate from theoretically expected distribution under the binomial assumption is flagged for exclusion. This problem usually arises from random priming during library preparation or less frequently from unintended retention of adapter sequences in the trimmed reads. Using the comprehensive raw QC report from FastQC software, high quality of data is ensured before alignment step.

Sequence alignment

In the alignment step, clean short-reads are mapped to the human reference genome. This step is computationally intensive and fundamental to downstream analysis. Sequencing errors and low complexity regions of DNA usually result in mapping biases and therefore the selection of a good aligner that enables high-quality alignment, and minimal bias is of paramount importance for clinical validity of WES results.

Several mapping algorithms for alignment of short reads to the reference genome have been developed, but the majority of conventional short-read aligners apply hash tables or index-based algorithms for mapping short-reads to the reference genome[148]. The index-based aligners such as Novoalign (`http://www.novocraft.com`) are generally slower and use more computational memory, but they are more efficient in mapping reads with long gaps. Novoalign implements the Needleman-Wunsch algorithm[149] using the entire region of the sequences and uses affine gap penalties to find the optimal alignment for short reads. In comparison, heuristic-based aligners such as BWA[150] are generally faster and

well suited for short-read mapping[148]. BWA applies the Burrows-Wheeler Transform algorithm for alignment of short sequence reads (32-100 bp) against the large reference sequence[150].

Novoalign (v2.08.02) was the aligner of choice in Soton Mendelian v.3.0 pipeline and replaced with BWA (v.0.7.12) in the newer version of pipeline (Table 2.6 for more details).

The product of the alignment step is a SAM (Sequence Alignment Map) file that stores alignment information in the text format. Each short read in this file is represented by a single row containing 11 mandatory fields that summarise alignment details for the respective reads (Table 2.3). These 11 fields are identified by specific flags and used in post-alignment processing to remove duplicates, and low quality reads from the downstream analysis.

Field	Lable	Value Type	Description
1	**QNAME**	String	Query template name
2	**FLAG**	Integer	bitwise flag
3	**RNAME**	String	References sequence name
4	**POS**	Integer	1- based leftmost mapping position
5	**MAPQ**	Integer	MAPping Quality
6	**CIGAR**	String	CIGAR String
7	**RNEXT**	String	Ref. name of the mate/next read
8	**PNEXT**	Integer	Position of the mate/next read
9	**TLEN**	Integer	observed Template length
10	**SEQ**	String	segment SEQuence
11	**QUAL**	String	ASCII of Phred-scaled base QUALity+33

Table 2.3: Mandatory fields in SAM format (*Adopted from SAM/BAM specification available at* `https://samtools.github.io/hts-specs/SAMv1.pdf`.

To save hard disk memory and increase processing efficiency in downstream analysis SAM files are converted to their binary counterparts, known as BAM files, and post alignment processing is applied to the these binary sequence alignment files.

Post-alignment Processing

Several sources of error during library preparation, sequencing and alignment can give rise to erroneous calls and false positive/negative results further down the analysis pipeline. These may include, but are not limited to, *PCR artefacts*, *sequencing errors* and *mapping errors*. In order to minimise the impact of artefacts in variant calling, following alignment, BAM files are subjected to a three-step processing procedure that includes removal of duplicate reads, local re-alignment around indels and base quality score recalibration (BQSR). These steps ensure high quality of data for variant calling.

Duplicate removal: Many library preparation methods involve one or more PCR steps that potentially result in *PCR artefacts*. These PCR errors are presented as mismatches in the read alignment and when occur early during PCR amplification will eventually give rise to multiple reads with the mismatch that can falsely be interpreted as a variation. Moreover, PCR duplicates result in over-representation of the same pair of reads in the alignment that can skew coverage calculation and introduce bias for variant calling in heterozygous sites[151].

Reads that map to the precisely same coordinates in the alignment map represent duplicate reads. These duplicate reads arise from unequal amplification of DNA fragments during library preparation and, in order to eliminate the bias introduced by these PCR artefacts, alignment map (SAM/BAM) files are subjected to duplicate removal. Two major algorithms that are used for flagging and removal of duplicates include Picard

MarkDuplicates(`http://picard.sourceforge.net`) and SAMtools[152]. These programs identify the duplicate reads solely based on the 5' coordinates of overlapping reads and ignore the coordinates in the 3' end. Since base calling quality decreases toward the 3' end, alignment of bases at this end are generally less reliable and therefore will be ignored by the software for duplicate removal[153]. Furthermore, hard clipped reads, in which low-quality bases at the 3' end are trimmed, result in different mapping coordinates at the 3' end that can easily mislead the duplicate read identification.

It should be noted that duplicated reads do not necessarily share the same sequence content. Since PCR amplification might introduce some errors into the sequence information of the duplicate reads, the sequence content of the reads is not a good account for identifying duplication and therefore not considered during duplicate removal step.[154].

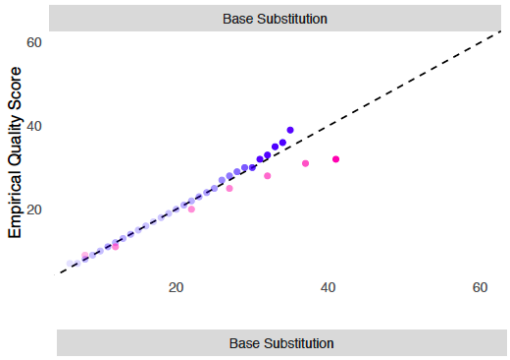
INDEL realignment Following duplicate removal, alignment of reads in regions containing INDELs must be improved. In conventional alignment algorithms, described earlier, the optimal mapping position for each read is determined independently. In contrast to sequence reads that simply contain SNVs, alignment of reads around INDELs is more challenging and requires more sophisticated methods such as gapped alignment. Since the aligner maps reads individually, the gap penalty among overlapping reads is not always the same, and local realignment around known INDELs is required to improve the accuracy of variant calls[98,155].

Major algorithms proposed for INDEL-based realignment either locally realign the group of gapped reads to the candidate haplotypes[156,157,95] or perform local *de novo* assembly of gapped reads around INDELs[158,159]. In the Soton Mendelian v3.0 pipeline in which Novoalign (v2.08.02) is used for alignment, gapped reads are realigned locally concurrent to global alignment using Smith–Waterman algorithm, whereas, in Soton Mendelian v4.0 and the customised pipelines, the INDEL realignment step is performed as part of variant calling procedure. In the latter pipelines, local realignment is postponed until variant calling step where GATK HaplotypeCaller is used for local *de novo* assembly of gapped reads.

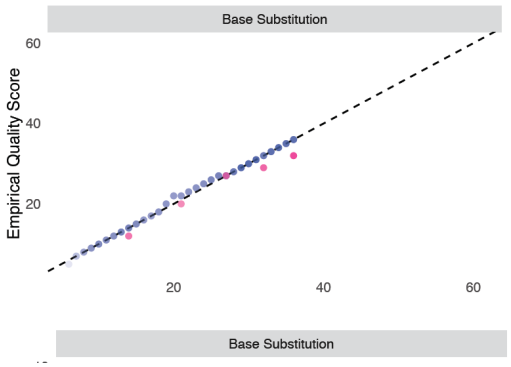
Local realignment of reads around INDELs have shown to improve the accuracy of INDEL calls substantially and it is recommended in the best practice guideline for variant discovery[160].

Base quality score recalibration (BQSR) Since the accuracy of variant calling relies extensively on base quality scores, the Phred scores generated by the sequencer need to be adjusted. In order to minimise systemic and platform-specific bias introduced by the sequencer in base quality scores assignment, sequencer-generated Phred quality scores (Q) are subjected to base quality score recalibration (BQSR). In this process reported quality scores from the sequencing machine in combination with valuable information from related covariates such as base position, neighbouring base content, sequencer cycle and the number of reads per lane is used to build a sophisticated model of empirical scores. Using a set of known variants to mask out real variation in the genome (usually dbSNP), reported scores from the sequencing machine are adjusted to empirical scores for each base in the read alignment (Figure 2.3).

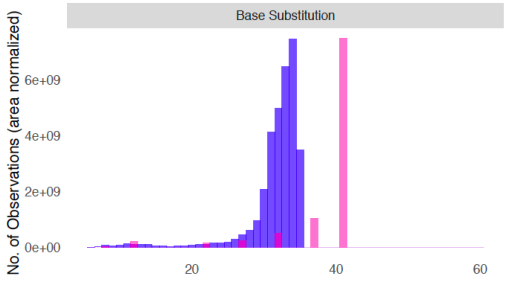
In the Soton Mendelian pipeline v3.0, the BQSR is carried out internally by Novoalign (v2.08.02) as a post-alignment processing option, whereas in the Soton Mendelian pipeline v4.0 and the customised pipeline, BQSR is implemented independently of alignment procedure via GATK Base Quality Score Recalibration (BQSR) module. The corrected base quality scores by the GATK suite are shown to significantly improve bias related to base quality assignment and increase the accuracy of variant calling[155].



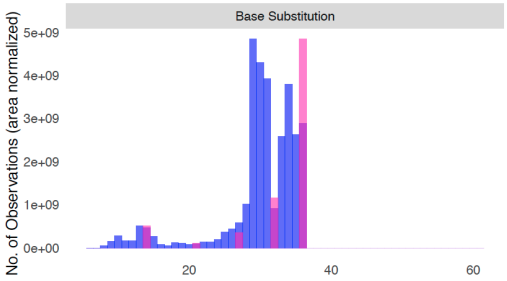
(a) Reported quality vs. empirical quality for WES data.



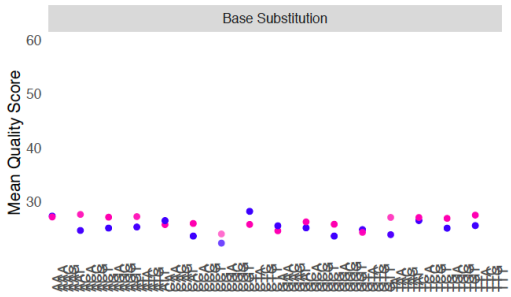
(b) Reported quality vs. empirical quality for TSO data.



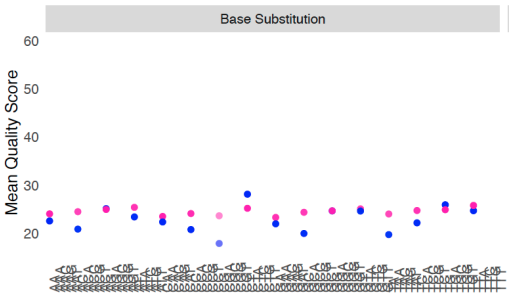
(c) Distribution of quality scores for WES data.



(d) Distribution of quality scores for TSO data.



(e) Mean quality scores for sequence context covariate (3 base suffix) for WES data.



(f) Mean quality scores for sequence context covariate (3 base suffix) for TSO data.

Figure 2.3: Comparison of pre and post recalibration quality scores for WES data (*left*) and clinical gene-panel (Illumina TruSight One panel) data (*right*). Scores before the recalibration process are identified in red and corrected scores after the BQSR are identified in blue; (a & b): representation of empirical versus reported quality scores for base substitution before and after the BQSR; (c & d): The distribution of quality scores for base substitution before and after the BQSR; (e & f): representation of mean quality scores before and after the BQSR for trinucleotide sequence context.

It is important to note that corrected base quality scores are generated under the assumption that any differences outside the dbSNP is a sequencing error. Given that over ∼25 million validated SNPs are indexed in dbSNP build 150 (Chapter 1), and only a few thousand variants are found in a typical human exome [161], this appears a statistically sound assumption. However, an obvious caveat should be taken into consideration in cases where only a small portion of the genome is sequenced. In such cases, the covariate model does not generalise well to the data and the BQSR may result in the exclusion of true positive variants [162,163,153].

Quality control checks on aligned reads

Inter-sample DNA exchange or contaminations are a source of error in NGS analysis. When DNA from two or more samples on the same dispatch DNA plate are mixed and sequenced together, sites at which samples are homozygous for alternative alleles are interpreted as heterozygous. This results in inflation of heterozygous ratios in the sample. In order to ensure reads from two samples are not mixed and verify heterozygous fractions match the expected ratios from the Hardy-Weinberg Equilibrium (HWE), BAM files are scrutinised in verifyBamID software [164] and the FREEMIX score is applied to investigate contamination level in the samples.

The FREEMIX score uses sequence only information to estimate the likelihood for contamination based on equation 2.3.1 [164].

$$\mathcal{L}(\alpha) = \prod_{i=1}^{M} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{R_i} \sum_{e_{ij}} \Big((1-\alpha) P(b_{ij}|g_i^1, e_{ij}) + \alpha P(b_{ij}|g_i^2, e_{ij}) \Big) P(e_{ij}) \right\} P(g_i^2) P(g_i^2) \tag{2.3}$$

In this equation, g_i denotes the true genotype for site i where b_{ij} is the base call for the j^{th} overlapping read. The e_{ij} denotes a latent indicator variable that takes either value 0 or 1 depending on whether b_{ij} is called correctly or not. In this equation, M denotes the total number of genotype sites in a single sample, and R_i identifies the total number of reads overlapping site i that has passed mapping and base quality thresholds [164]. The product of this equation is a single score that can be used as an estimate of contamination level in aligned reads. Samples with FREEMIX scores above 0.03% are considered contaminated and excluded from downstream analysis.

Variant Calling

Once BAM files are processed through post-alignment steps, they are subjected to variant calling. In this step, loci that are different from the reference genome are compiled and recorded in a text-based file known as variant call format (VCF). The VCF format is a standardised framework for storing all types of genetic variations including SNVs, INDELs and large structural variations (SVs) and used as the standard input file for downstream annotation software [165].

The structure of a VCF file consists of two main sections; the *header* that defines the fields (columns), and the *main body* that contains information about variant sites. The header consists of eight mandatory fields and an unlimited number of optional columns that describe the variant site (Table 2.4). In VCF format each variant site is represented by a single row, and the total number of rows in the body section corresponds to the total number of variants called (Figure 2.4).

Field	Lable	Value Type	Description
1	**CHROM**	Integer/ String	Identifier of the chromosom to which the variant is mapped.
2	**POS**	Integer	The 1-based position of the variation on the reference sequence.
3	**ID**	String	The identifier of the variant for example rsID for variants in dbSNP; Novel variants are coded as ".".
4	**REF**	Integer	Upercase reference base(s) at the site of variatio; In case of indels, this field must include base(s) before the event.
5	**ALT**	String	The list of alternative alleles at the variant site.
6	**QUAL**	Integer	A phred-score associated with the inference of alternative allele.
7	**FILTER**	String	A flag identifier indicating the variant site has PASSed all the filters.
8	**INFO**	String	An costomised list of semicolon seperated key-value pairs for describing the variant site. DP;VDB;RPB;AF1;AC1;DP4;MQ;FQ;PV4

Table 2.4: The eight mandatory fields in VCF format; The INFO-field arbitrary keys permitted for WES data include: DP: actual depth of coverage at the variant site; VDB: variant distance bias, used o check if the variant occurs at a random position along the aligned read and it is useful for identification of misalignment due to nearby SNVs; RPB: Mann-Whitney rank-sum test for tail distance bias; AF1: Maximum likelihood estimate of the site allele frequency under Hardy–Weinberg equilibrium (HWE) assumption; AC1: Maximum-likelihood estimate of alternative allele under no HWE assumption; DP4: Four comma separated numbers representing number of high quality reads covering or bridging reference forward, reference reverse, alternative forward and alternative reverse bases respectively; MQ: Root-mean-square mapping quality of reads covering variant site; FQ: Consensus quality expressed in phred-scaled probability; PV4: Four comma separated integers expressing p-values for strand bias, base quality bias, mapping quality bias and tail distance bias (*The information in the table is adopted from the Variant Call Format (VCF) Version 4.2 specification* [165]).

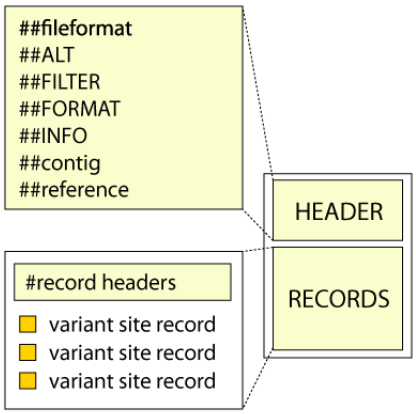


Figure 2.4: Basic structure of a variant call format (VCF) file (*Figure adopted from GATK forums;* `https://gatkforums.broadinstitute.org`).

In the Soton Mendelian pipeline v3.0, variant calling is carried out using SAMtools (v0.1.19) mpileup utility [152], whereas, in the Soton Mendelian v4.0 and the customised pipeline, the GATK (v3.6) HaplotypeCaller module [166] is used for variant calling. The SAMtools mpileup variant caller walks through the aligned region and calls genotypes for every single position based on the maximum likelihood estimation of variant site probabilities [152]. The GATK HaplotypeCaller on the other hand, not only calls SNVs and INDELs simultaneously but also performs local *de novo* assembly around INDEL regions and therefore renders upstream INDEL realignment unnecessary [162]. Furthermore, variant filtering in SAMtools is carried out through a set of predefined filters, while GATK HaplotypeCaller is able to learn from data and model per-read allele likelihoods for variant sites [153].

Quality control checks on variant callsets

In order to confirm the gender assignment and the relationship status of exome samples on the same dispatch DNA plate, variant callsets are subjected to further QC. Unexpected gender assignment and excess heterozygosity in variant callsets can be the result of cross-sample contamination. While the investigation of excess heterozygosity and contamination using VerifyBamID enables exclusion of contaminated samples before variant calling, it does not reveal the nature of cross-sample swaps and further checks after variant calling is required to identify samples with excess genotype similarity. Furthermore, the percentage of X chromosome heterozygosity from callsets can be utilised to confirm the gender status of exome samples. In the Caucasian population, which constitute the majority of clinical samples analysed in this thesis, X chromosome heterozygosity for female and male patients are estimated around ∼60-55% and ∼21-23% respectively. Apart from rare cases of Klinefelter syndrome (KS) in which male patients harbour an extra X chromosome (47, XXY), this method is able to reliably match reported genders from the lab with inferred gender status from variant calls. Similarly, excessive autosome heterozygosity (more than two standard deviation ($\mu \pm 2\sigma$) from samples mean) is interpreted as DNA contamination.

In this step, a similarity matrix based on the percentage of identical genotypes between samples is constructed, and unexpected similarities between a pair of exome callsets are interpreted as cross-sample swaps. The advantage of this method over the FREEMIX method for detection of contamination is its ability to identify precisely the samples that are swapped. Excess similarities between two or more callsets not expected from pedigree information (*i.e.* consanguinity) are interpreted as sample swaps. Application of this similarity matrix in callsets QC is presented in Chapter 3.

Annotation

Following variant calling, annotation attributes such as gene symbol, associated amino acid change or splicing impact, the frequency of the variant in the general population, pathogenicity and conservation metrics in relation to the variant site is determined through the annotation procedure. Since protein-altering variants such as non-synonymous SNVs and INDELs underlie 85% of Mendelian disorders[167,42], it is essential to discern synonymous (silent) variants from non-synonymous mutations early on in the analysis. Annotation facilitates the discovery of disease-causing mutations from background noise and common polymorphisms.

In the Soton Mendelian pipelines (v3.0 & v4.0), VCF files are subjected to gene-based and variant-based annotation using ANNOVAR software[168] (r.2013Aug23 & r.2015Dec respectively). In the gene-based annotation attributes such as genomic features (exonic/intronic), exonic function and exact amino acid change are assigned to related variant sites. The *gene-based* annotation is subsequently complemented with *variant-based* annotation in which information related to alternative allele frequency and variant functional prediction scores such as pathogenicity and conservation is integrated into the annotation. In both pipelines, the RefSeq transcripts are used for gene annotation, and all genomic coordinates are defined according to human reference assembly hg19 (GRCh37). Both pipelines utilise the dbSNP[54] build 142 to identify known variants. The minor allele frequencies (MAF) from three different populations including the 1000 Genome Project (global and continental), Exome Aggregation Consortium dataset (ExAC, non-TCGA samples)[5] and the NHLBI-ESP project (∼6500 exomes) are used for allele-frequency annotation. Pre-computed functional prediction scores including SIFT[169], PolyPhen2 HVAR[170], LRT[171], MutationTaster[172], FATHMM[173], CADD[174], GERP++[175], PhyloP[176] and SiPhy[177,178] are obtained from the dbNSFP v3.0[179]. The splicing impact of variants positioned up to three bases in the exonic region, and up to 8 bases

intronic of a splicing acceptor or splicing donor site in both pipelines are populated using the MaxEntscan software[180] and the human splice finder (HSF, v2.4.1)[181]. Finally, the frequency of observed genotypes in the Southampton Exome Database (SED) is determined and incorporated into the annotation record.

In the customised pipeline, variant sites are simultaneously annotated by the SnpEff[182] and the Variant Effect Predictor (VEP r.87)[183]. The general settings and filtering criteria used for annotation of variants in this pipeline are similar to Soton Mendelian pipelines with the exception that hard filtering is carried out prior annotation in order to reduce false positive calls in the callset. In this pipeline, VCF files are initially subjected to post variant calling filtering using GATK (v3.6) variant filtration module. The SNP clusters in which 3 or more SNPs occur within 15bp of each other are filtered out. This measure is somewhat arbitrary and open to debate, but statistical models recommend a collapsing window size of 15bp[184]. Furthermore, variant sites with normalised quality by depth less than 2 (QD< 2) and Fisher's Strand bias (FS> 30) are also excluded from the annotation analysis. The downstream annotation procedure and databases used for annotating variants are the same as Soton Mendelian pipelines.

Since VEP and SNPeff and ANNOVAR use slightly different terminologies for different classes of variants, sequence ontologies were unified across the three variant annotators so that results contain unique terms for each class of variants (Table 2.5).

	ANNOVAR (r.2013Aug23; r.2015Dec14)	**SNPeff** (v4.2)	**VEP** (r.87)	**UNIFIED annotation ontology**
1.	nonsynonymous_SNV	NON_SYNONYMOUS_CODING CODON_CHANGE NON_SYNONYMOUS_START CODNON_SYNONYMOUS_STOP START_LOST	missense_variant initiator_codon_variant stop_retained_variant incomplete_terminal_codon_variant	**nonsynonymous**
2.	synonymous_SNV	SYNONYMOUS_CODING SYNONYMOUS_START SYNONYMOUS_STOP	synonymous_variant	**synonymous**
3.	frameshift_deletion frameshift_insertion	FRAME_SHIFT	frameshift_variant	**frameshift**
4.	splicing	SPLICE_SITE_DONOR SPLICE_SITE_ACCEPTOR	splice_donor_variant splice_acceptor_variant	**splicing**
5.	stopgain	STOP_GAINED	stop_gained	**stopGain**
6.	stoploss	STOP_LOST	stop_lost	**stopLoss**
7.	nonframeshift_deletion	CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION	inframe_deletion	**inFrame_Del**
8.	nonframeshift_insertion	CODON_INSERTION CODON_CHANGE_PLUS_CODON_INSERTION	inframe_insertion	**inFrame_Ins**

Table 2.5: Different classes of variants annotated in WES pipeline; Different annotation software use different terminologies for identification of various classes of variants. In order to facilitate handling and manipulation of annotation output from different pipelines an arbitrary unified term was devised and applied to annotation outputs.

Upon annotation, experimental evidence in relation to the pathogenesis of variants is added to annotation output by querying curated mutation repositories including the Human Gene Mutation Database (HGMD v.2016.2)[185], Leiden Open Variation Database (LOVD v2.0)[71] and ClinVar[186].

The decision as to what annotation software and transcript set (NCBI RefSeq vs Ensemble or UCSC) should be used have a significant impact on the number and types of variants identified in WES analysis[187]. Variants that map to overlapping genes in the human genome might impact multiple transcripts in the two genes and possibly have a different outcome on each transcript. For instance, an in-frame insertion of a stop codon after the last triplet codon of an exon can be mutually interpreted as either a stop gain or

splice variant. While stop-gains have a punitive impact on the continuation of transcription in one transcript, it could simply have no effect on an alternatively spliced transcript; therefore a consensus on reporting these variants is required. The ANNOVAR algorithm always reports the most deleterious outcome based on an internal priority framework and also collapse start loss/gain variants into the single nonsynonymous category. While this approach resolves the ambiguity in variant impact, it might lead to oversimplification of a high impact start loss/gain mutations. This is specifically problematic in WES analysis of neoplastic tissue samples where gain or loss of function mutations in proto-oncogenes and tumour suppressors play an important role in the disease aetiology, and higher weight should be given to these variants[188]. Furthermore, subjective determination of worst outcome should be interpreted in the context adjacent variants. As an instance in a likely situation where a base insertion introduces a stop codon into the sequence, the variant could be classified as both frameshift insertion or stop-gain, but the decision as to whether it should be interpreted as frameshift or stop-gain is completely subjective and depends on the annotation tool. In this instance, if the frameshift is being offset by another preceding frameshift then the possible impact of the variant is more ambiguous and requires due diligence for interpretation.

Annotation tools as they stand today, all have their own specific advantages and limitations, and interpretation of results in the clinical settings requires careful examination of functional impact rendered by the variants. In the context of rare disorders, where rare coding variants are of particular interest, ANNOVAR provides a reasonably robust framework for classification of variants in WES applications[189].

2.3.2 Benchmarking of WES pipelines

In order to evaluate the performance of the three pipelines developed for analysing WES data, a batch of 40 exomes were processed through each pipeline and classification outputs from each pipeline compared against each other. The general workflow in all three pipelines is similar and includes *alignment*, *variant calling* and *annotation* steps as discussed earlier, but subtle differences in tools and processing methodology result in variability in output from the three pipelines. A general overview of tools used in each pipelined for processing WES data is provided in Table 2.6.

Step	**Soton Mendelian V.3.0**	**Soton Mendelian V4.0**	**Customised pipeline**
Alignment	Novoalign (v2.08.02)	BWA (v0.7.r12)	BWA (v0.7.r12)
Sort aligned reads	Samtools (v0.1.19)	Picard (v1.97)	Picard (v2.10.4)
Mark duplicate reads	Picard (v1.97)	Picard (v1.97)	Picard (v2.10.4)
Index aligned reads	Samtools (v0.1.19)	Picard (v1.97)	Picard (v2.10.4)
Base quality score recalibration (BQSR)	-	GATK (v3.6)	GATK (v3.6)
Generate recalibration plots	-	-	GATK (v3.6)
Calling variants (SNVs and INDELs)	Samtools (v0.1.19)	GATK (v3.6)-*HaplotypeCaller*	GATK (v3.6)-*HaplotypeCaller* Samtools (v1.3.2)
Variant quality score recalibration (VQSR)	-	-	GATK (v3.6)
Annotation	Annovar (r.2013Aug23)	Annovar (r.2015Dec14)	VEP (r.87) SnpEff (v4.2)

Table 2.6: Characteristics of pipelines used for analysing WES data.

Each pipeline outputs between 25,000 to 30,000 variants per exome on average. Samples that were processed through the customised pipeline tend to have a larger number

of SNVs compared to Soton Mendelian pipelines (Figure 2.5a). This can be attributed to the use of merged callset from the GATK haplotype-caller and Samtools in this pipeline. Conversely, the total number of INDEL variants identified by the customised pipeline is significantly lower than the Soton Mendelian pipelines (Figure 2.5b) which can be attributed to the use of variant quality score recalibration (VQSR) in this pipeline (Table 2.6). The GATK VQSR module[166] applies machine learning models to learn from the profile of highly-validated variants in the callset and build a Gaussian mixture model based on contextual annotation properties of these training datasets (DP, FS, QD, MQ among others). Finally, the model is used to summarise the annotation properties of novel variants in the dataset to a single VQSLOD score that can be used for robust filtering of false-positive variants. Given that the VQSR model works based on the probability estimation of annotation profiles in the callset, it is not applicable to small-scale exome analysis and sample sizes larger than 30 exomes is required[190].

On the other hand, both the Soton Mendelian v4.0 and the customised pipelines represent a superior performance for calling common INDELs (with MAF greater than 1%) (Figure 2.5f), that can be attributed to the robust power of GATK local realignment for minimising spurious calls around INDEL regions. This may reflect that both pipelines have better sensitivity for identifying true positive INDELs.

In relation to different versions of the Soton Mendelian pipeline, a lower number of stop gain/loss and splicing variants in the Soton Mendelian v4.0 could be attributed to the improved specificity (limited number of false positives) achieved through the base quality score recalibration (BQSR) step (Figure 2.5c). This observation is consistent with lower number of INDELs identified in the coding region by the Soton Mendelian v4 (Figure 2.5b). Conversely, the new version of pipeline presented a better performance for identification of common SNVs and INDELs in the callset (Figure 2.5e & 2.5f). The enhanced performance of the pipeline for calling frameshift variants is also reflected in the higher ratio of non-frameshift to frameshift variants (Figure 2.5d). These figures arguably emphasize the role BQSR in limiting alignment errors for variant discovery.

For all three pipelines, the T_s/T_v ratio was around the expected range $\sim$3 (Figure 2.5g) and the Het/Hom ratio was consistently identified around $\sim$1.5 (Figure 2.5h).

Since our conclusion about sensitivity and the precision of pipelines drawn from the above analysis was speculative, to robustly benchmark the performance of the pipelines against a highly validated callset, the high-coverage ($\sim$116x) WES data of individual *NA12878* obtained from the Genome in a Bottle (GIAB) project[191]. The exome target for this sample has been recovered using the SureSelect v2.0 capture kit (Agilent, Santa Clara, CA, USA) and sequenced in a HiSeq2000 sequencer (Illumina, San Diego, CA, USA). For the purpose of performance comparison between pipelines, true positive (TP) variants were defined as variant calls that are consistent with validated genotypes from the gold standard truth set. Also, false positives (FP) were defined as variant calls that are inconsistent with the truth set and the difference between the total number of variants in the truth set and TPs were defined as false negatives (FP). Comparison of precision and recall (sensitivity) metrics across the pipelines revealed that the Soton Mendelian v4.0 has slightly better sensitivity for identification of SNVs (Table 2.7). However, the customised pipeline revealed to perform better for identification of INDELs. The marginally reduced sensitivity of the customised pipeline for identification of SNVs could be attributed to the use of SNP cluster filtering through which three or more variant that occur within 15bp of each other are discarded. The issue of incorrect calls due to misalignment is less common in high-coverage high-quality exome samples, but it is a common source of false positive discovery in low-quality samples. This filtering step was intentionally introduced into the customised pipeline to lower the possibility of false positive discovery among a few low-quality samples that are discussed in Chapter 5. Moreover, the enhanced performance

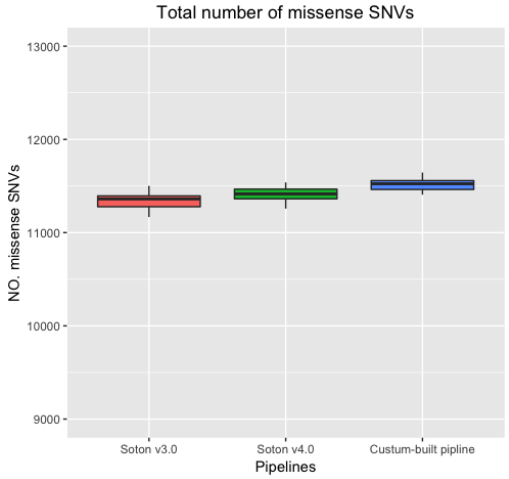
of the customised pipeline for INDEL discovery may be related to the use of the merged callset from both Samtools (v1.3.2) and GATK haplotypecaller (v3.6).

Pipelines	SNVs					INDEL				
	True Positive (TP)	False Positive (FP)	False Negative (FN)	Sensitivity	Precision	True Positive (TP)	False Positive (FP)	False Negative (FN)	Sensitivity	Precision
Soton Mendelian v3.0	24,779	498	1091	0.9578	0.9803	1,064	348	620	0.6318	0.7535
	Total number in truth set 25,870					Total number in truth set 1,684				
Soton Mendelian v4.0	24,847	314	1023	0.9605	0.9875	1,376	289	308	0.8171	0.8264
	Total number in truth set 25,870					Total number in truth set 1,684				
Custom-built pipeline	24,801	242	1069	0.9587	0.9903	1,403	222	281	0.8331	0.8634
	Total number in truth set 25,870					Total number in truth set 1,684				

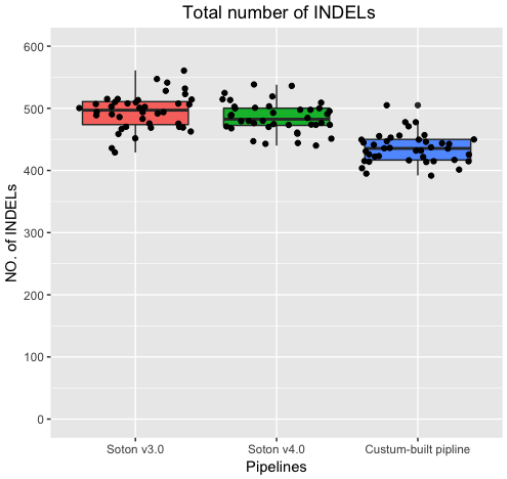
Table 2.7: Summary table for performance of WES pipelines benchmarked against the NA12878 gold standard truth set (The high-coverage whole-exome data of individual NA12878 obtained from the GIAB consortium accessable at: `http://genomeinabottle.org`)

The application of VQSR in the customised pipeline while providing better sensitivity and precision for variant detection (especially around INDEL regions) restricts the pipeline utility for small-scale analysis where the total number of sample is less than 30. Although I used the customised pipeline for exome-batch analysis throughout my research, for the benefit of consistency and reproducibility of analysis in individuals cases, results from the Soton Mendelian pipeline v4.0 is provided in the following chapters. The general workflow for the Soton Mendelian pipeline v4.0 is provided in Figure 2.6.

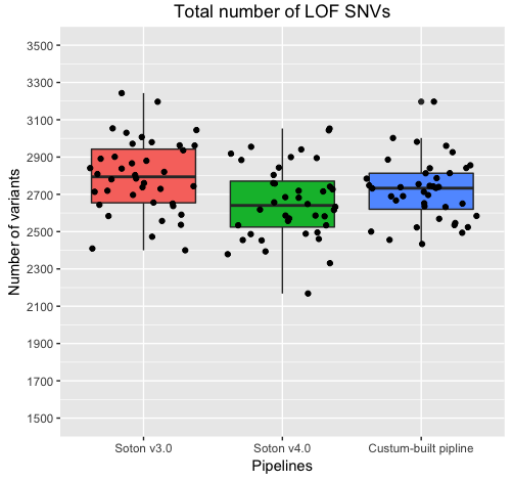
Throughout the following chapters, if the use of additional annotation software such as VEP or SnpEff revealed to be beneficial, the results from the Soton Mendelian pipeline v.4.0 has been complemented with the annotation outputs from the annotation module of the customised pipeline.



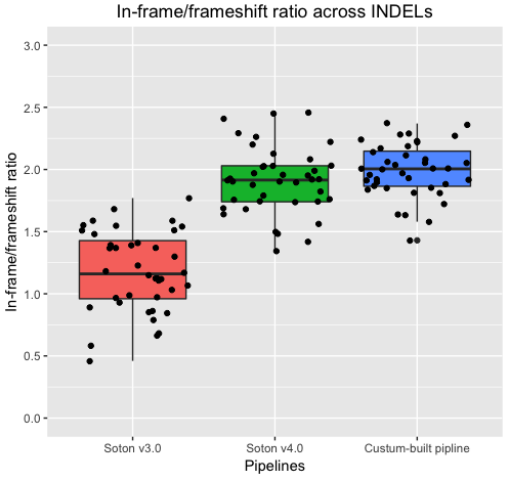
(a) Total number of missense SNVs; One-way ANNOVA *p-value*= 2.59e-16.



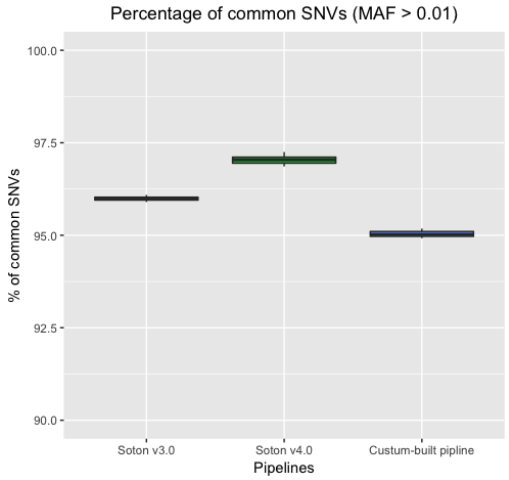
(b) Total number of INDELs per sample; One-way ANNOVA *p-value*= 2.20e-6.



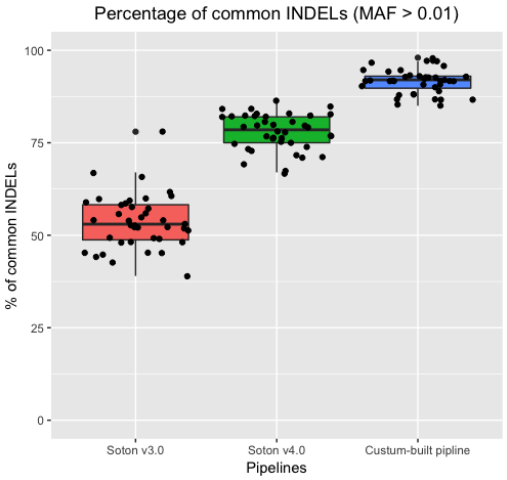
(c) Total number of SNVs with either stop gain/loss or splicing impact; One-way ANNOVA *p-value*= 0.005.



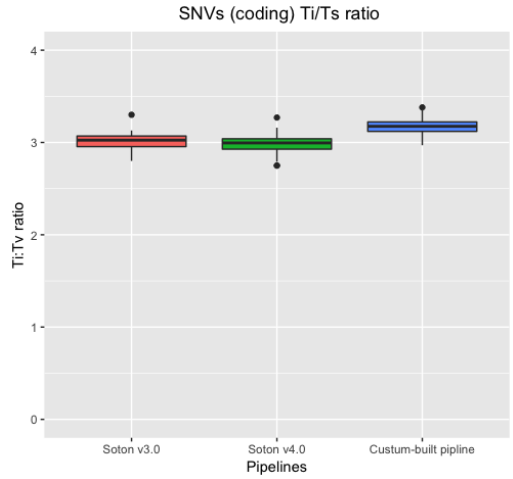
(d) Average ratio of non-frameshift to frameshift INDELs; One-way ANNOVA *p-value*= 6.29e-12.



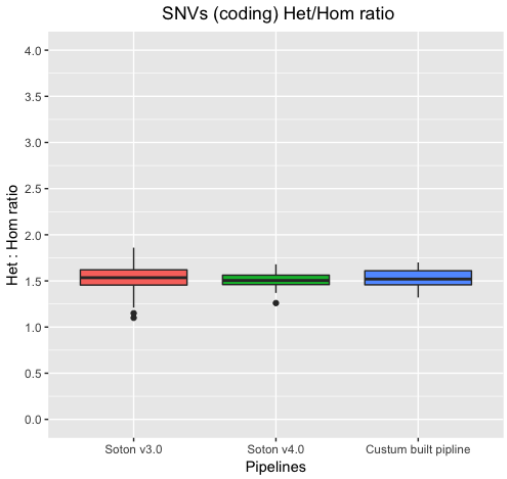
(e) The percentage of common SNVs per sample, One-way ANNOVA *p-value*= 8.27e-15.



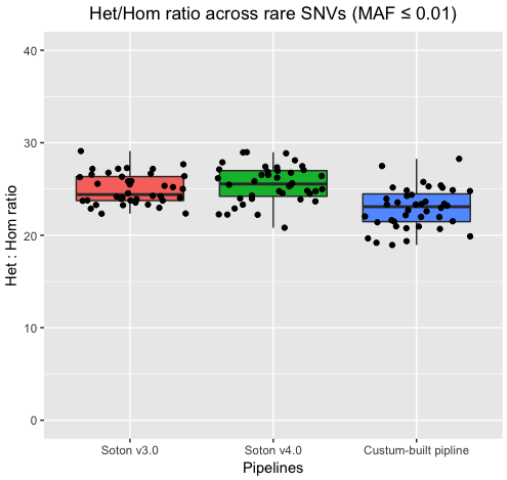
(f) The percentage of common INDELs per sample; One-way ANNOVA *p-value*= 3.42e-12.



(g) The T_i/T_v ratio across the coding SNVs; One-way ANNOVA *p-value*= 0.3818.



(h) The Het/Hom ratio across all coding SNVs; One-way ANNOVA *p-value*= 0.6754.



(i) The Het/Hom ratio across the rare variants (MAF $\leq$ 1%; One-way ANNOVA *p-value*= 3.589e-08)

Figure 2.5: Differences in the counts of coding single nucleotide variants (SNVs) and insertion-deletions (INDELs) between the Soton Mendelian Pipelines (v3.0 & v4.0) and the custom-built pipeline; (a) The larger number of missense variant called by the customised pipeline reflects the impact of merging call sets from different variant callers; (b) The lower number of INDELs called by the Soton v4.0 and the customised pipelines reflects the collective impact of BQSR and INDEL realignment in mitigating the rate of false positive calls for INDEL variants; (c) Reduced number of stop gain/loss and splicing variants called by the Soton v4.0 can be attributed to the enhanced annotation provided by the newer version of Annovar software. The slightly increased number of stop gain/loss and splicing variants called by the customised pipeline reflects the impact of the merged call set; (d) The enhanced performance of the Soton v4.0 and the customised pipelines for calling frameshift variants is reflected in the higher ratio of non-frameshift to frameshift INDELs ratio. This demonstrates the impact of INDEL realignment implemented by both pipelines to reduce false-positive frameshift calls; (e) All pipelines are able to call $\geq$95% of common SNPs indexed in dbSNP build 150. The lower detection rate for common variants in the customised pipeline is due to the stringent filtering criteria applied at the variant-calling step to remove SNP clusters in which three or more SNPs occur within 15 bp of each other; (f) The higher performance of the customised pipeline for identification of common INDELs reflects the collective impact of INDEL realignment along with merging call sets from different variant callers; (g) The T_i/T_v ratio across the coding SNVs for all pipeline is around $\sim$3; (h) The Hem/Hom ration across coding SNVs for all pipelines is around $\sim$1.5; (i) The lower Het/Hom ratio for rare coding variants in the customised pipeline reflects the utility of GATK hard-filtering step in reducing the number of false-positive heterozygous calls that arise from errors such as strand bias.

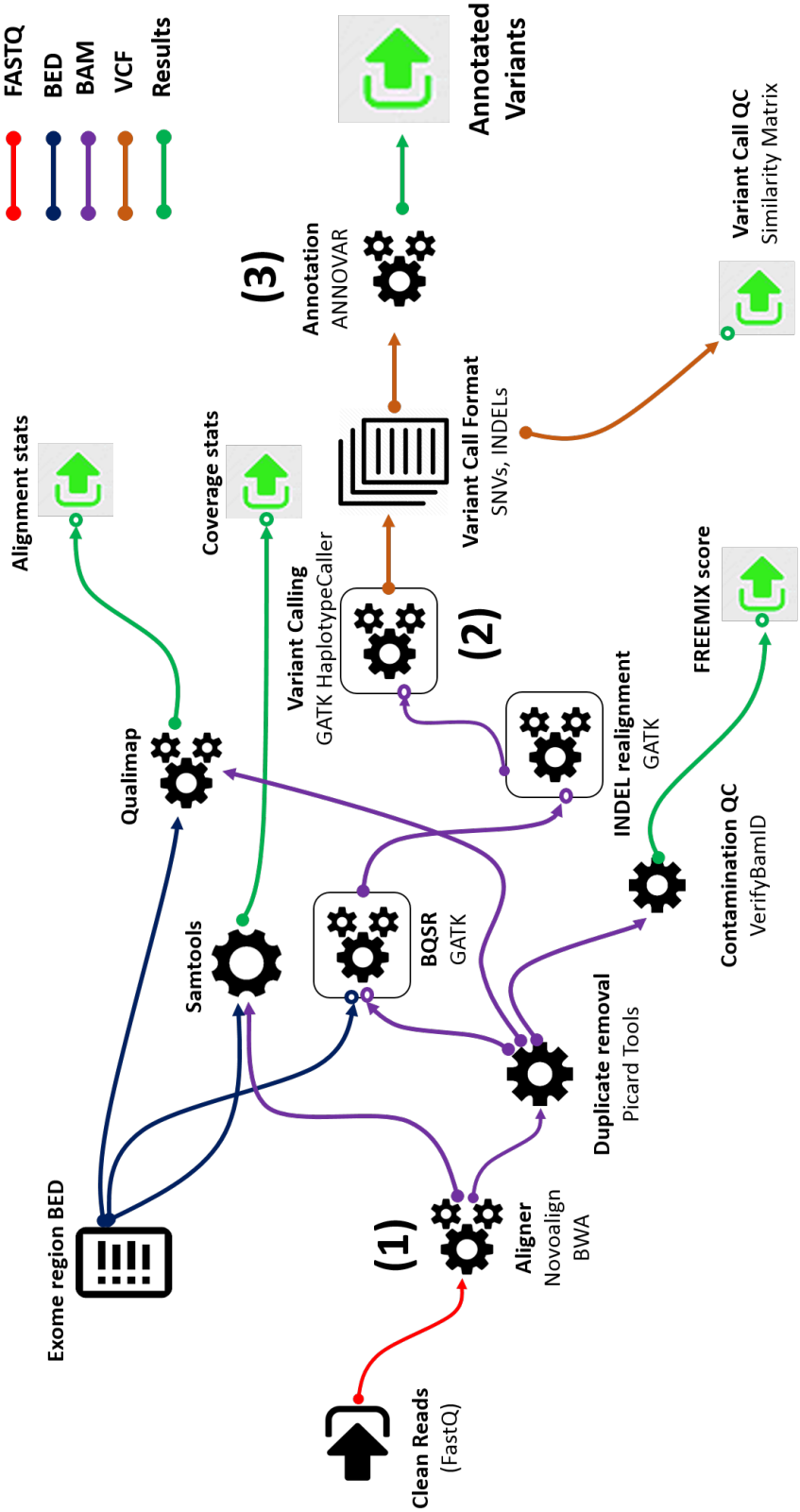


Figure 2.6: The general workflow for analysing whole-exome sequencing data; The pipeline applied for WES analysis is comprised of three major steps; 1) *Alignment*, 2) *Variant Calling* and 3) *Annotation*. In this schematic representation of the pipeline, the input and output to and from each module are colour coded to represent the format of the data.

2.3.3 WGS analytical pipeline

The general workflow used for analysing whole-genome data essentially follows the same three-step framework (*Alignment*, *Variant Calling* and *Annotation*) described for WES analysis. Due to the huge size of WGS data, raw FASTQ files are usually generated from multiple flowcells in the sequencer machine and therefore multiple FASTQ files for a single whole genome sample is common. In the pipeline developed for analysing WGS data (Figure 2.7), separate files were merged only after alignment and post alignment processing was carried out on the merged BAM file. This enables a better resolution for base quality recalibration and leaves no chance for duplicated reads to remain in the dataset.

Given longer read size in WGS applications (up to ∼150bp), alignment is carried out using BWA-MEM algorithm. This algorithm enables the identification of multiple non-overlapping alignment-hits in the genome that are potentially caused by structural variations (SVs). Following alignment, BAM files generated from independent FASTQ files are merged using BamTools (v2.3.0)[192] and subjected to Picard (v2.10.4) duplicate removal. Dedupe BAM files are then used for base quality score recalibration (BQSR) in GATK (v3.6) and passed to the GATK (v3.6) RealignerTargetCreator module to define intervals for local alignment around known INDELs. Subsequently, local realignment is performed in GATK (v3.6) IndelRealigner module. Processed BAM files are then used for calling SNVs and INDELs using GATK (v3.6) HaplotypeCaller. Given suitability of WGS data for studying structural variants (SVs), processed bam files are also passed to LUMPY[193] (v.0.2.11) and SVDetect[194] (v0.8b) for identification of large copy number variations (larger than 50bp in size).

Variant callsets from GATK (v3.6) HaplotypeCaller are directly used in Variant Effect Predictor (VEP r.87)[183] for annotating SNVs and INDELs. Since breakend (BND) variants (including deletion, inversion, duplication and translocations) cannot be annotated in VEP, variant calls from SVDetect are annotated in SnpEff (v4.3)[182]. Variant callsets from LUMPY (v.0.2.11) do not contain genotype information and require extra processing step to incorporate base details into the callset. Varian calls generated in LUMPY are therefore pass through a Bayesian SV genotyper developed by LUMPY team (available at `https://github.com/hall-lab/svtyper`). Processed VCF files containing SV calls are then annotated in SnpEff (v4.3) using the RefSeq transcripts (GRCh37/hg19).

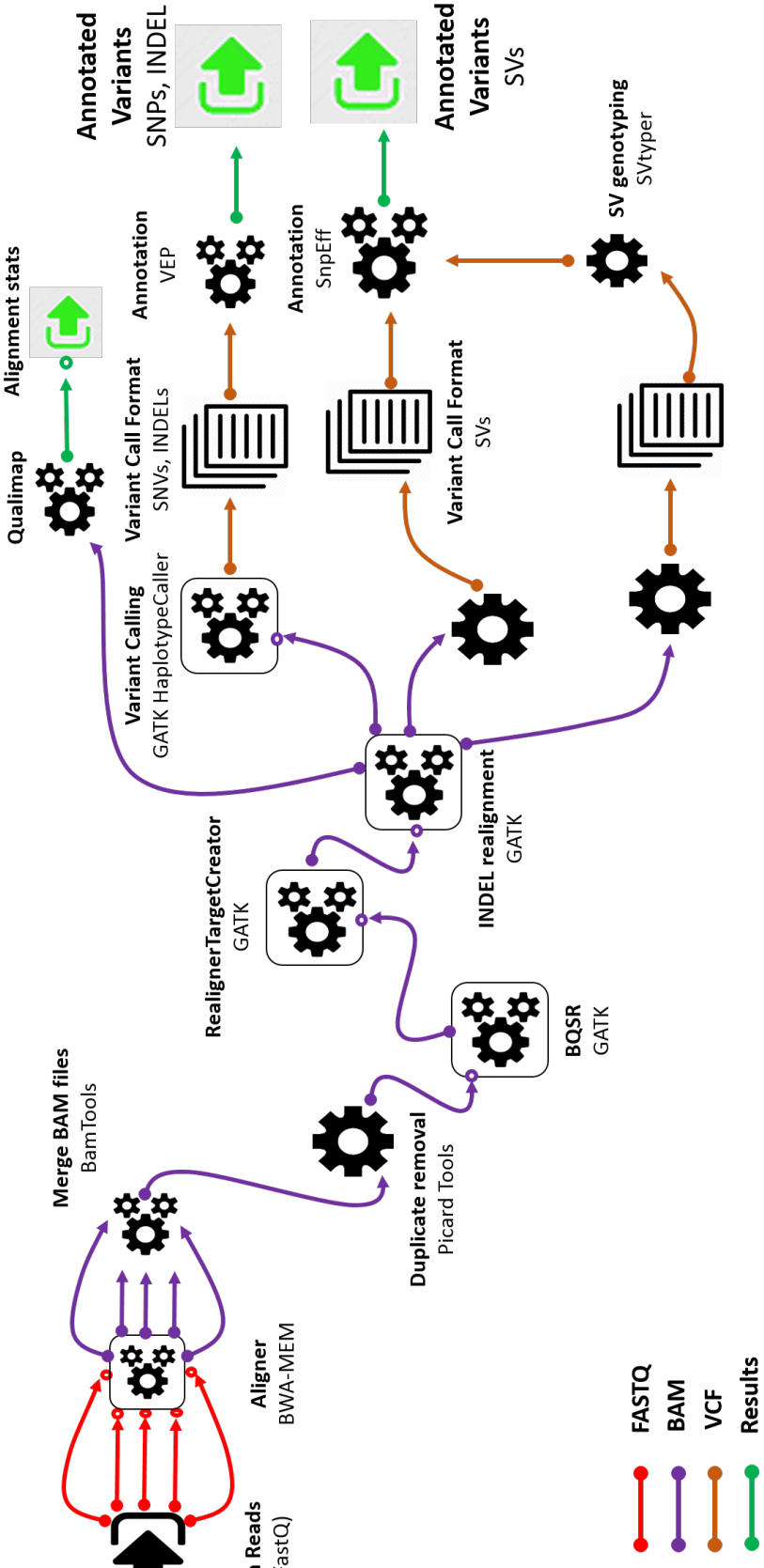


Figure 2.7: The general workflow for analysing whole-genome sequencing data; In this schematic representation of the WGS pipeline, the input and output to and from each module are colour coded to represent the format of the data

2.4 Overview of NGS approaches for SV discovery

Compared to conventional cytogenetic and molecular techniques, NGS applications have higher throughput for detection of SVs and do not require prior knowledge about the nature of chromosomal imbalance involved. The NGS methods developed for identification of SVs are generally categorised into *assembly*-based or *mapping*-based methods.

In *de novo* assembly methods, short reads are assembled together to form a complete jigsaw puzzle picture of the genome[195]. Assembly methods are a powerful approach for identifying all sorts of SVs, but they are computationally intensive and often underperform with short read data[105] (Figure 2.10, Assembly column).

The *mapping*-based methods are widely used in clinical genetics for identification of SVs[31,196]. In paired-end NGS applications, three strategies have been developed for detection of SVs.

Since in paired-end sequencing each fragment is sequenced from both ends, the distance and orientation of read-pairs (RP) can be used for identification of several classes of SVs. In this method, read pairs that map too far apart indicate a deletion and pairs that map more closely than expected from the insert size distribution identify insertions (Figure 2.8 A and B). In addition, orientation of read-pairs can be used to identify inverted regions (Figure 2.10, Read Pair column). Several algorithms for SV discovery have been developed based on this method, but RP method is ill-suited for identification of novel-sequence insertions with the length larger than the mean insert size[197,198,199,200]. In such cases, the alignment of read-pairs to the reference genome gives rise to clusters of reads that are not mapped in proper pairs. Unmapped reads are usually filtered out as part of the initial QC step, and therefore detection of this class of variants with the RP method is difficult (Figure 2.8 C).

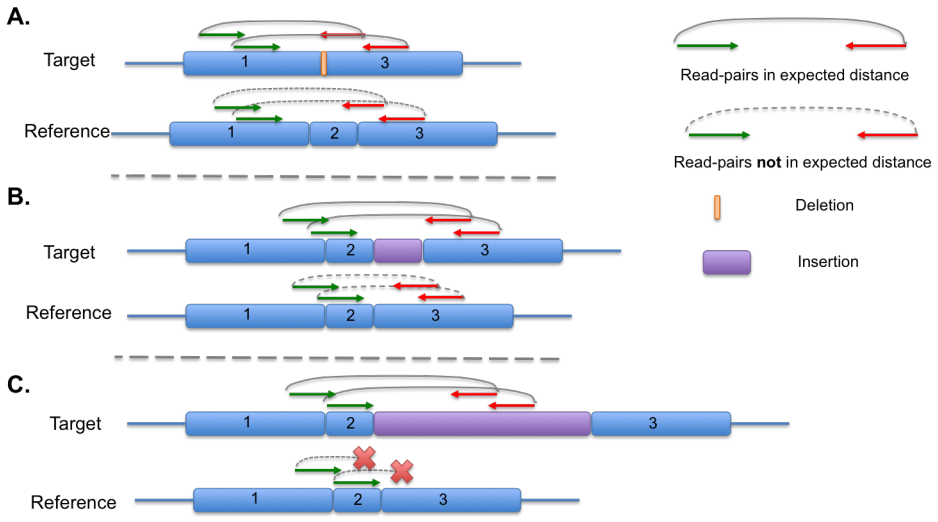


Figure 2.8: Schematic representation of mapped paired-end reads around INDELs; (A.) Read-pairs that span the deletion point map too far apart than expected in the reference genome; (B.) conversely read pairs that span an insertion map closer than expected in the reference genome; and (C.) insertions larger than the mean insert size cannot be readily detected since they produce hanging reads in which only a single read is mapped. Given that singly mapped and unmapped reads are usually discarded from downstream analysis detection of insertions larger than the mean insert size is difficult.

In high-quality NGS data, the distribution of read-depth across the genome/exome represent a Poisson or modified Poisson distribution which can be utilised to identify SVs that involve copy number changes[201]. Duplicated regions result in higher read-depth, and

deletions result in depletion of read-depth across the deleted region [201]. Due to the non-uniform depth of coverage especially in WES applications, algorithms developed based on this method rely on normalisation of read-counts across the target region which requires multiple samples to enable normalisation and true estimation of variability within the batch [202]. The read-depth method is unable to identify SVs that involve insertions or inversions (Figure 2.10, Read Count column).

Reads that partially map to two different regions of the genome are called split-reads (SRs) and can be used for identification of all types of SVs up to the base-pair resolution (Figure 2.10, Split Read column). In the SR method, the presence of an SV breakpoint is investigated using the alignment signature of split reads in the reference genome. Deletions introduce a gap between split fragments and insertions produce soft-clipped alignments between split fragments where bases do not align to the reference genome. The soft-clipped region between the two ends of reads is identified by stretches in the reference genome (Figure 2.9).

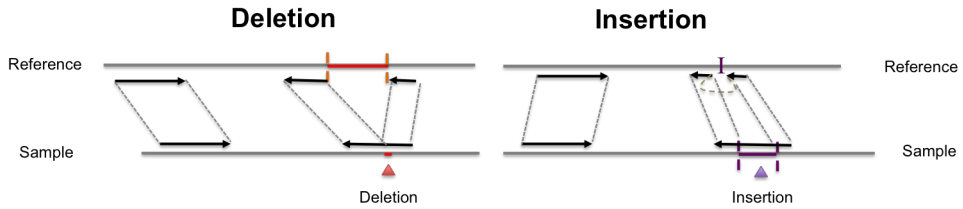


Figure 2.9: Schematic diagram for split-read signature around INDELs; Deletions in the patient's genome result in gaps between the split fragments in the reference genome and insertions cause unmapped sequences (soft-clipped alignment) between the two fragments of the split-reads.

Structural variants that involve translocations can also be identified by the SR method [203]. Tools developed based on the SR method use the one-end anchored reads strategy for identification of translocations. In this context, when a single read maps fully to the right position in the reference genome and the other pair splits to two partially mapped reads, the split reads are interpreted as a translocation [204].

All the *mapping*-based methods discussed above have their own advantages and limitations and therefore a tool that implements a combination of these algorithms can significantly boost the discovery rate in SV analysis. Recent algorithms such as SVDetect [194] and LUMPY [193] that combine multiple *mapping*-based strategies have been shown to achieve superior accuracy for detection of SVs in the human genome [193].

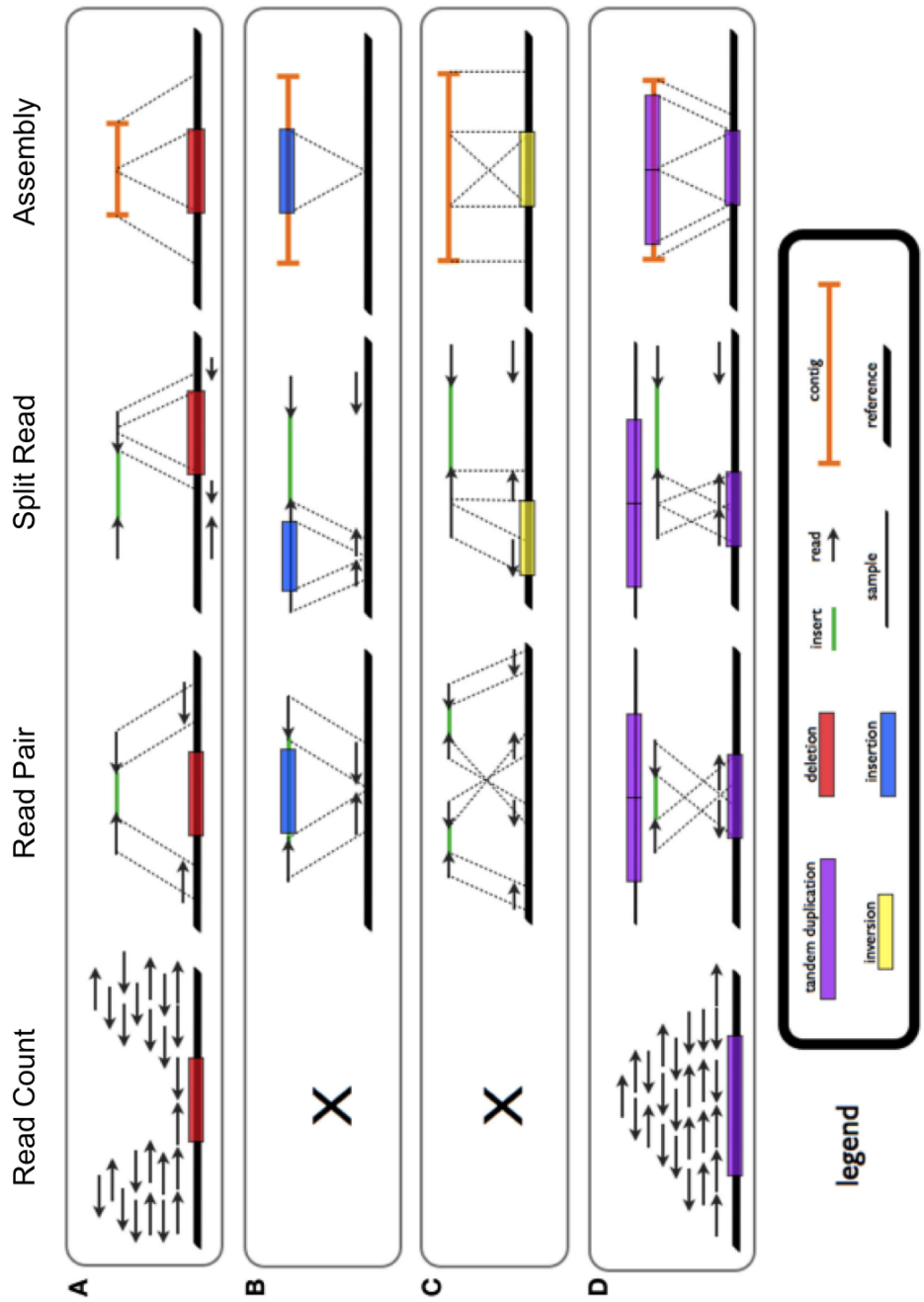


Figure 2.10: Schematic representation of structural variant (SV) discovery methods in NGS. (A), novel sequence insertion (B), inversion (C), and tandem duplication (D) in read count (RC), read-pair (RP), split-read (SR), and de novo assembly (AS) methods; (Figure adopted from Tattini *et al.*[205])

2.5 Discussion

In this chapter, I have described the tools and pipelines I used for analysing WES and WGS data throughout the thesis. Since the majority of data in this thesis is of WES type, the challenges in respect to sensitivity and specificity of WES pipelines has been discussed.

The issue of sensitivity and specificity in variant calling, stem from the underlying probabilistic statistical model implemented by different variant callers and therefore the choice of tools and pipelines is pivotal to successful identification of the causal variants. The mathematical and computational models used for alignment and variant calling are actively being developed, and therefore the best practice in relation to variant discovery is subject to change over time. Availability of high-confidence variant calls from the Genome in a bottle Consortium [191] enables benchmarking of pipelines and allows informed decisions about software and parameter options.

Since both the Soton Mendelian v4.0 pipeline and the customised pipeline perform quite comparably, to ensure consistency of methods for future follow ups, results from the Soton Mendelian v4.0 is presented in the following chapters. Where necessary, additional modifications to the main pipeline were applied to make relevant adjustments. For example, merging variant calls from different callers revealed to be useful for increasing the precision for INDEL discovery. Therefore, the relevant script from the customised pipeline has been utilised wherever necessary.

In conclusion, this chapter highlights the importance of design, implementation and benchmarking of pipelines for NGS analysis. Given the diversity of databases, algorithms and software available for variant analysis, once an efficient pipeline is developed all samples must be analysed using a single stable version of the pipeline so that reproducibility of results is ensured.

Chapter 3

Whole exome sequencing in Nephrolithiasis

3.1 Introduction

Nephrolithiasis, also known as renal stones, is a frequent condition affecting up to 15% of men and 6% women in industrialised countries [206]. Sporadic stones usually arise from environmental risk factors including diet, low fluid intake, heavy physical exercise and various medicines [206,207]. Stones can also be secondary to pre-existing conditions such as hyperparathyroidism, diabetes mellitus [208], hypertension [209] or gout [210]. In addition, anatomical anomalies including the horseshoe kidney can also predispose to kidney stones [211]. Inherited nephrolithiasis comprises a small but significant portion of the disease incidence worldwide [212]. Stones with genetic origin tend to segregate in families with either autosomal dominant, autosomal recessive or X-linked pattern of inheritance [213]. All types of stones result in a huge burden on the health care systems [214]. Identification of the genetic factors underlying inherited renal stones has implications for prognosis and early intervention. Furthermore, understanding the genetic basis of inherited nephrolithiasis helps in assessing risk to other family members.

3.1.1 Genetic basis of inherited nephrolithiasis

Inherited renal stones are often recurrent and bilateral and tend to segregate in families with a recent common ancestor [215]. Patients with inherited kidney stone disease usually present with additional metabolic imbalances such as Cystinuria (HP:0003131), Hypercalciuria (HP:0002150), Hypercitraturia (HP:0012406), Hyperoxaluria (HP:0003159), Hyperphosphaturia (HP:0003109) and Hyperuricosuria (HP:0003149). In families with history of recurrent stones, genetic defects in metabolic pathways involved with urinary promoters such as calcium, urate, cystine and sodium or urinary inhibitors (magnesium, citrate and nephrocalcin) have been found to underlie the condition [216]. Genetic defects underlying the most common inherited renal stones are broadly categorised into the four major classes that will be discussed in the following section:

1. Impairment of purine metabolism

Autosomal recessive forms of nephrolithiasis can results from deficiency of adenine phosphoribosyl transferase (APRT) and xanthine dehydrogenase (XDH) enzymes. Both enzymes are involved in purine metabolism and mutations in the genes coding for theses enzymes lead to accumulation of precursors immediately upstream of enzyme activity. The APRT enzyme is encoded by *APRT* gene on chromosome 16q4 and involved in biosynthesis

of adenosine monophosphate (AMP) from adenine and 5-phosphoribosyl-1-pyrophosphate (PRPP). A homozygous deficiency in the enzyme results in accumulation of adenine which in turn is metabolised to 2,8 dihydroxyadenine (DHA) by XDH enzyme. The XDH enzyme is a molybdenum-containing hydroxylases encoded by the *XDH* gene on chromosome 2p23 that catalyses oxidative metabolism of purines. Defective homozygous mutations of *XDH* lead to accumulation of xanthine. DHA and xanthine are both insoluble in water and promote formation of radiolucent stones. Pathogenic mutations in the *APRT* gene are estimated to occur in $\sim 1\%$ of healthy newborns[217].

2. Primary Hyperoxalurias

The primary hyperoxalurias (PH) are group of related renal disorders that lead to autosomal recessive renal stones. Three types of PH are described; PH1 (#OMIM: 259900) accounts for $\sim 80\%$ of cases and is caused by pathogenic mutations in the *AGXT* gene[218]. PH2 (#OMIM: 260000) and PH3 (#OMIM: 613616) are caused by homozygous mutations of the *GRHPR* and *HOGA1* genes respectively and account for the remainder 20% of the PH type nephrolithiasis[219]. In all PH cases pathogenic mutations lead to precipitation of calcium oxalate stones in the kidney. Strikingly, all types of PH have early age of onset and usually present during childhood[220].

3. Cystine transport impairment

Cysteine stones typically arise from mutations impairing the cysteine transport pathway[216]. Positively charged amino acids such as cysteine, lysine, ornithine and arginine are primarily reabsorbed across the apical membrane of the proximal renal tubule via amino acid transporters. The *SLC3A1* on chromosome 2p12 and *SLC7A9* on chromosome 19q13.11 encode components of the renal amino acid transporter system and pathogenic mutations in these two genes result in impairment of cysteine transport. In a recent study, *SLC7A9* mutations were estimated to account for 11% of adult and 21% of paediatric renal stones[221]. Impaired cysteine transport gives rise to cysteinuria that in turn leads to high urinary excretion of cysteine. This amino acid is highly insoluble and precipitates in the kidneys as the cysteine stones.

4. Renal tubular acidosis

The renal tubule cells play a very important role in homoeostasis of water and H^+ ions in the kidney. Disturbance to this regulation gives rise to renal tubular acidosis (RTA) where lowering of intracellular PH leads to increased reabsorption of citrate in the proximal section of the tubule. Urinary citrate is the most important inhibitor of calcium nephrolithiasis. Increased reabsorption of citrate in tubules results in precipitation of calcium phosphate stones in the kidneys. Mutations of basolateral anion exchanger AE1 encoded by the *SLC4A1* on chromosome 17 underlies the autosomal dominant RTA. The phenotypic spectrum of *SLC4A1* mutations ranges from the severe form of renal stones associated with growth delay to milder homeostatic imbalance that leads to stone formation[216].

3.1.2 Additional metabolic impairments in nephrolithiasis

In addition to the major molecular mechanisms explained above, at least 24 additional genes have been shown to underlie monogenic forms of nephrolithiasis[222,215].

Hypercalciuria is one of the most common metabolic abnormalities usually found in nephrolithiasis patients[223]. Affected individuals with hypercalciuria present with ex-

cess calcium excretion in the urine which leads to formation of calcium stones [224]. In nephrolithiasis patients, hypercalciuria occurs either in isolation or in combination with additional metabolic abnormalities (such as hyperphosphaturia) that collectively predispose to formation of renal calculi [225]. Twin studies suggest that heritability underlies 52% of hypercalciuria incidence [226] and 65% of patients with hypercalciuric nephrolithiasis are identified to have a positive family history [227]. Heritable forms of hypercalciuric nephrolithiasis commonly represent as a polygenic quantitative trait, however monogenic forms with autosomal dominant, autosomal recessive or X-linked pattern of inheritance are also described [228]. Table 3.1 provides a summary of monogenic nephrolithiasis that results from the impairment of calcium reabsorption mechanism. These genetic disorders are usually associated with calcium stones.

Disease	Gene	Chromosomal location	Mode of Inheritance	OMIM ID
Idiopathic hypercalciuria	ADCY10	1q24.2	AD	143870
	VDR	12q13.11	AD	
		9q33.2- q34.2	AD	
ADHH	CASR	3q13.33- q21.1	AD	601198
Bartter syndromes				
Type I	SLC12A1	15q21.1	AR	601678
Type II	KCNJ1	11q24	AR	241200
Type III	CLCNKB	1q36	AR	607364
Type IV(a)	BSND	1p32.3	AR	602522
Type V	CASR	3q21.1	AD	300971
Type VI	CLCN5	Xp11.23	XLR	
Dent's disease				
Type I	CLCN5	Xp11.23	XLR	300009
Type II	OCRL1	Xq26.1	XLR	300555
Lowe syndrome	OCRL1	Xq26.1	XLR	309000
HHRH	SLC34A3	9q34.3	AR	241530
NPHLOP1	SLC34A1	5q35.3	AD	612286
Hypomagnesemia				
Type III	CLDN16	3q28	AR	248250
Type V	CLDN19	1p34.2	AR	248190
Distal renal tubular acidosis	SLC4A1	17q21.31	AD	179800
Renal tubular acidosis with deafness	ATP6B1	2p13.3	AR	267300
Renal tubular acidosis	ATP6V0A4	7q34	unknown	602722

Table 3.1: Monogenic disorders associated with hypercalciuric nephrolithiasis;*ADHH*: autosomal dominant hypocalcemia with hypercalciuria, *HHRH*: Hypophosphatemic rickets with hypercalciuria, *NPHLOP1*: Nephrolithiasis, osteoporosis and hypophosphatemia, *AD*: Autosomal dominant, *AR*: Autosomal recessive, *XLR*: X-linked recessive. (The table modified from Stechman *et al.* [228])

In addition to hypercalciuria, metabolic impairments related to phosphate homeostasis have been implicated in kidney stones. Regulation of calcium and phosphate homeostasis is interdependent and impairments in regulation of one, will inevitably lead to perturbation of the other [229,230,231]. Regulation of phosphate homeostasis is largely mediated through the intestinal uptake or kidney reabsorption [232]. Both the intestinal and kidney absorption are controlled by a group of sodium-dependent phosphate co-transporters that belong to the solute carrier (SLC) gene family. Expression of these genes are under the tight control of a number of hormones and metabolic factors. Phosphaturic hormones such as parathyroid hormone (PTH), by decreasing renal phosphate reabsorption increases the phosphate leak into the urine. Secretion of PTH from parathyroid glands is controlled by a group of calcium sensing receptors (CaSR) on the extracellular surface of renal tubule cells. In the presence of high Ca^{2+} concentration in the renal tubular fluid, CaSRs are stimulated and secretion of PTH is suppressed through negative feedback regulation. When serum Ca^{2+} is low, CaSR is inactive and PTH is secreted which stimulates Ca^{2+} reabsorption in distal tubules through regulation of TRPV5. In addition secretion of PTH, stimulates the expression of *CYP27B1* in the kidneys which increases conversion of inactive vitamin D to active form (1,25 dihydroxy Vitamin D). This reaction occurs in the mitochondrial matrix and leads to increased Ca^{2+} and phosphate absorption from the small intestine and decreased reabsorption of phosphate from the renal filtrate that ultimately leads to

increased phosphate leak into the urine. PTH also inactivates sodium–hydrogen antiporter 3 (encoded by the *SLC9A3* gene) on the apical side of epithelial tubulus cells proximal to nephrons and lead to increased H^+ secretion and urine acidification [233,234]. Given the large amount of phosphate excreted daily into urine (0.9- $1gr/day$), maintaining the PH in a relatively acidic range is vital to the solubility of phosphate divalent ions ($H_2PO_4^-$) in the urine. Any impairment in either metabolic or homeostatic pathways that increases the urine PH, will lead to precipitation of excess phosphate and Ca^{2+} in the urine and formation of calcium phosphate crystals in the kidneys. As a matter of fact, the type of stones formed in the kidneys are fairly closely correlated with the urine PH. In general, patients with urine PH in acidic range (PH$\leq$ 6) tend to form pure Ca^{2+} oxalate or a mixture of calcium oxalate and uric acid stones, whereas Ca^{2+} phosphate stones are more frequent among patients with urine PH$>$ 6.2.

The amount of phosphate filtered into urine is generally expressed as $TmPO_4/GFR$ ratio and indicates the maximum rate of renal tubular phosphate reabsorption relative to the overall glomerular filtration rate. Low $TmPO_4/GFR$ ratio is indicative of large phosphate leak into urine and hence hyperphosphaturia. Studies into the genetic cause of hyperphosphaturia have shown that mutations of SLC gene family implicate in defective renal phosphate reabsorption and may explain the aetiology of low $TmPO_4/GFR$ ratio among stone formers [229]. Despite this, it is important to note that SLC gene mutations appear to occur too rarely to explain the widespread occurrence of hyperphosphaturia among stone formers, and it appears more likely that the low $TmPO_4/GFR$ ratio is the secondary outcome of a more subtle defect in renal tubular function [235].

Taken together, twin studies suggest that the contribution of genetic factors to the kidney stone phenotype is significant and accounts for up to 56% of disease incidence in the general population [236]. The heritability of monogenic stone phenotype across families with calcium excretion is estimated to be around $\sim$0.58 [237], however, twin studies indicated a larger heritability ($\geq$0.90) for urinary traits that predispose to kidney stone phenotype [238,239]. Early onset nephrolithiasis among children often has a recessive inheritance, whereas, dominant renal stones with incomplete penetrance is more frequent among adult patients [221]. Identifying the genetic components underlying hereditary forms of nephrolithiasis will help in reducing the cost burden associated with this disease in the general population; however, the complex biochemistry underlying heterogeneous forms of nephrolithiasis renders genetic diagnosis challenging.

3.2 Overview of the analysis

Given the significant role of heritability in nephrolithiasis, we sought to investigate the potential role of rare coding variants underlying familial nephrolithiasis across three extended pedigrees. As described earlier, aetiology of nephrolithiasis, especially among adult patients, ranges from purely environmental and dietary factors to highly penetrant genetic mutations. Considering segregation of the stone phenotype across multiple generations in the families studied here, we hypothesized that a genetic mutation may underlie patients' phenotype.

In order to investigate whether it might be possible to identify pathogenic coding variants that contribute to adult nephrolithiasis, we applied whole-exome sequencing (WES). In analysing the data, primarily, pathogenic mutations in genes involved in homeostasis pathways, were prioritised and consistency of segregation among affected patients in each individual families was investigated. Furthermore, the possibility that a rare ancestral haplotype might underlie the patients' condition was also investigated. Ultimately, in order to confirm the causal role of shortlisted variants in each pedigree, an extended number of individuals from each pedigree were genotyped across the selected regions. Variants

with substantial evidence implicating a role in these pedigrees were shared with colleagues with relevant expertise for functional investigation.

3.3 Methods

3.3.1 Samples

Fifty patients from three extended pedigrees with recurrent stones were recruited for this study. These families were part of a larger cohort initially recruited in 1998/99 for linkage analysis at the University Hospital Southampton [240]. Primary linkage analysis by Damian Griffin did not yield a conclusive results and therefore they reselected for comprehensive genetic analysis using WES. All samples were of European descent residing in Hampshire, England. At the time of recruitment, recurrent kidney stones in all families were associated with urinary homoeostatic complications such as hyperphosphaturia and hypercalciuria. The characteristics of each pedigree is explained below:

Family A

At the time of recruitment seven members of this kindred across three generations (including four males and three females) were known to have developed stones (Figure 3.1). Four stone formers including III-1, III-20, IV-23 and IV-30 were alive at the time of recruitment and took part in the study. One characteristic feature of this family is that the stone-forming diathesis appears only in the proband's maternal side. Reviewing clinical history of the family revealed that individual IV-30 had multiple stones and in the majority of cases stone phenotype was presented between the age of 30 and 40 years. Biochemical inspection of plasma and 24h urine samples was indicative of prominent hyperphosphaturia across 6 family members including all the stone formers. In addition, two of the kidney stone formers (patients III-20 & IV-30) presented with hyperparathyroidism inferred from high serum PTH level. In the proband (individual III-1) mild hypercalciuria was associated with the stone phenotype. Details of biochemical test results for the 12 members of family A is provided in Supplementary Table 8.1. In order to minimise the impact of shared ancestry between the affected individuals and increase the specificity of variant discovery, DNA samples from individuals with furthest degree of relationship (III-1, IV-23 & IV-30) were selected for WES analysis. Patients who were alive at the time of recruitment and donated blood for research are identified by pink shaded circles in the Figure 3.1.

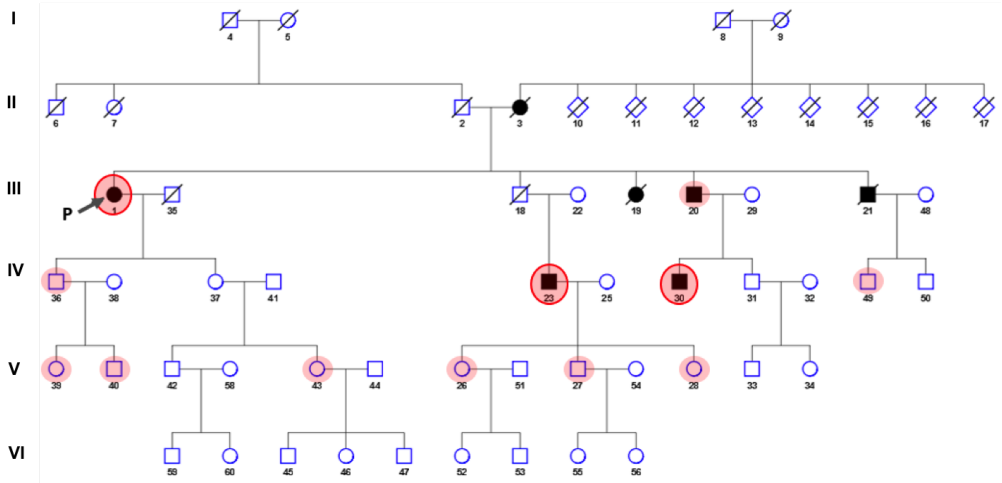


Figure 3.1: Pedigree showing an autosomal dominant pattern of inheritance for nephrolithiasis in family A. Individuals III-1, IV-23 and IV-30 underwent WES analysis. Shaded circles indicate individuals for whom DNA was available and underwent KASPar genotyping analysis. (Individuals III-20, IV-23,30, V-27,28 are presented with hypophosphatemia/phosphaturia; individuals III-20 & IV-30 are presented with hyperthyroidism (high serum PTH level), and hypercalciuria is present in individual III-1)

Family B

Nine members of this kindred (including 5 males and 4 females) were known to have developed stones across three generations. The stone-forming phenotype appears only in the proband's paternal side (Figure 3.2). At the time of recruitment only 6 of the stone formers were alive and 29 members of the family agreed to donate blood for research.

Hyperphosphaturia is the prominent abnormality in this family and appears to have an autosomal dominant pattern of inheritance. Five family members including the proband III-1, his father II-2, his aunt II-25 and two of his cousins (III-71 and III-75) presented with hyperphosphaturia (Supplementary Table 8.2 and Table 8.3). Borderline hypercalciuria is associated with kidney stone in patients III-1, III-32 and III-71.

While parathyroid hormone level was normal in all of the stone formers, five individuals including the 3 stone formers (III-1, III-2 and III-32) presented with hypervitaminosis D (Table 3.2). Reviewing patients' biochemical test results with the recruiting clinical biochemist (Dr Valerie Walker) revealed that the stone phenotype in this family is probably related to the renal phosphate leak which is also associated with the secondary absorptive hypercalciuria defect.

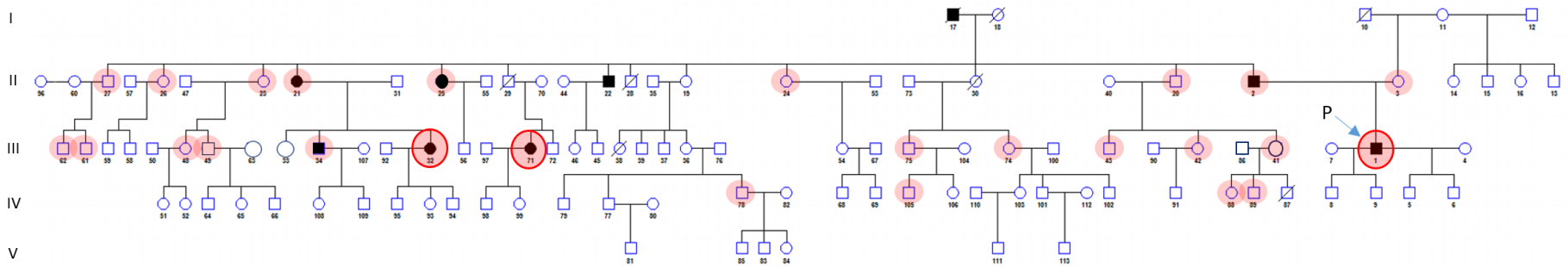


Figure 3.2: Pedigree showing an autosomal dominant pattern of inheritance for nephrolithiasis in family B. Individuals III-1, III-32 and III-71 underwent WES analysis. Shaded circles indicate individuals for whom DNA was available and underwent KASPar genotyping analysis.

	1	32	71	2	75	21	23	24	25	88	101	103
Hyperphosphaturia	o		o	o	o				o	o	o	o
Hypercalciuria	o	o	o					o				
Hypervitaminosis D	o	o		o			o	o				
Kidney Stone	o	o	o	o		o			o			

Table 3.2: Overview of major homeostatic abnormality among individuals of family B; Individuals 1, 32 and 71 were selected for exome sequencing.

Family C

At the time of recruitment, five members of this kindred including four males and one female were known to have presented with stones across the three generations. Reviewing patient's clinical history revealed that the proband (III-1), her son (IV-3) and her brother (III-8) all presented with multiple stone episodes. Inspection of biochemical test results (Supplementary Table 8.4) across individuals of this family revealed that hypercalciuria is the prominent trait in this family and in fact it is associated with the stone phenotype in two out of the three individuals selected for WES analysis (patients III-1 & III-3). In contrast to family A and B, hyperphosphaturia did not appear to be the prominent defect in this kindred, even though mild hyperphosphaturia was apparent in two family members including individuals III-1 and III-35. The inheritance of stone phenotype in this pedigree may be suggestive of X-linked recessive or dominant disease (Figure 3.3).

As for previous pedigrees, in order to account for IBD sharing and minimise the impact of shared haplotypes in variant discovery, individuals with farthest degree of relationship (*i.e.* III-8, III-34 and IV-3) were selected for WES analysis.

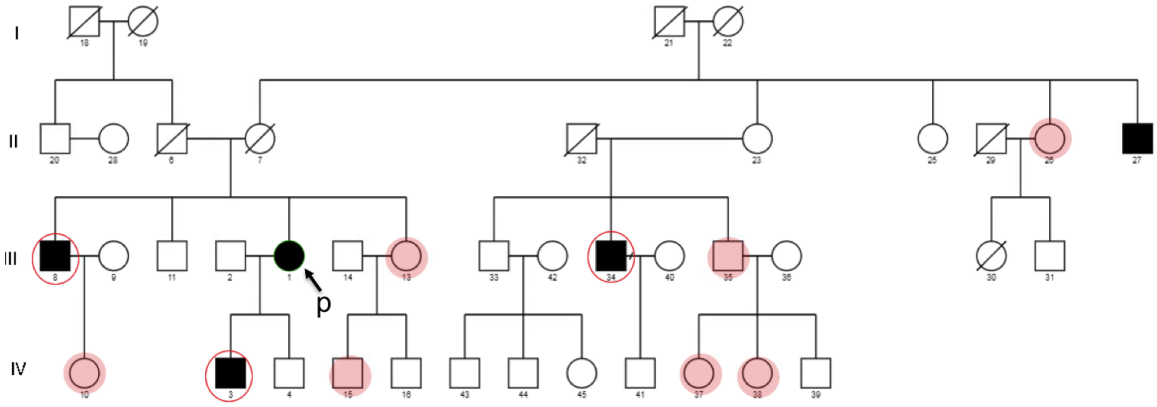


Figure 3.3: Pedigree suggestive of an X-linked pattern of inheritance for nephrolithiasis in family C. Individuals III-1, III-8 and IV-3 underwent WES analysis. Shaded circles indicate individuals for whom DNA was available and underwent KASPar genotyping analysis. (Individuals III-I, her sister III-13 and her first cousin III-35 are presented with idiopathic hypercalciuria. Individual III-8 and his first cousin III-34 also represent high 24-hour urinary Ca^{2+} level, but they are not hypercalciuric *per se*. Hypercalciuria is also evident across the individuals of last generation including family members IV-3,4,10 and 39. It is important to note that hypercalciuria trait across individuals III-13, IV-3 and IV-10 is also associated with hypervitaminosis D.

3.3.2 Sample processing and DNA quality control

Peripheral blood samples taken from 48 members across the three pedigrees at the time of recruitment, was used for genomic DNA (gDNA) extraction. DNA extraction was carried out in-house using the phenol-chloroform extraction (FC) method. Since samples where stored frozen for almost two decades they subjected to rigorous extraction QC in order to ensure samples met the minimum requirement for WES analysis.

3.3.3 Sequencing and *in-silico* data processing

Whole exome sequencing was carried out at the Wellcome Trust Centre for Human Genetics (WTCHG, Oxford) and raw sequence data were analysed in-house using the Soton Mendelian V4.0 pipeline with default settings as described in Chapter 2. In brief *in-silico* quality of raw sequence reads examined using FastQC software (v.0.11.5) and the sequence reads were aligned to the human reference genome GRCH37 (hg19) using the BWA aligner (v0.7.12). Following alignment, PCR duplicate reads were marked and discarded using Picard (v1.97). Base quality scores were recalibrated using GATK (v3.6) BaseRecaliberator and local realignment around INDELs was carried out using GATK IndelRealigner module. In order to verify that samples were not contaminated, BAM files were checked by VerifyBAMID (v.1.1.3) and FREEMIX contamination score was used to confirm the lack of contamination in samples.

In order to investigate possible sample swaps on the same dispatch of DNA, an aliquot of each sample's DNA was orthogonally genotyped at 24 loci and the concordance between the patients' genotypic profile yielded from exome sequencing and orthogonal genotyping was cross-checked. Following alignment, depth of coverage (DOC) analysis for family A and B was performed in BEDTools (v2.21) using Sureselect v5.0 (Agilent, Santa Clara, CA, USA) exome interval set totalling 51Mb. Exome enrichment in family C was carried out using the Sureselect v6.0 (covering 60Mb of the genome) and therefore respective BED coordinates were used for coverage analysis in this family.

Variant calling for each sample was contemporaneously carried out in Samtools v1.3.2 and GATK (v3.6) HaplotypeCaller. The resultant VCF files from the two variant callers from each sample were then merged and annotated using ANNOVAR (2015Dec14 release). Variant sites were defined according to the GRCh37 genome build and known polymorphisms were annotated according to dbSNP build 139. The 1,000 Genomes project phase 3 dataset along with NHLBI Exome Sequencing Project (ESP) and Exome Aggregation Consortium (ExAC) datasets were used for filter-based annotation. Furthermore, Soton Exome Database (SED) was also queried to incorporate allele frequencies (AF) from ~ 460 in-house WES analysed patients unaffected by nephrolithiasis. Human RefSeq transcript dataset (GRCh37.p10) was used for gene-based annotation and pathogenicity scores and conservation scores for variant sites were compiled using the PolyPhen-2, SIFT, GERP++ and Phylop algorithms. For variant sites within 10bp of intron-exon boundaries, ΔMaxEnt score was applied to inspect possible implication of base change on splicing.

3.3.4 Filtering and variant analysis

Annotated exome sequence data were cross-referenced against the panel of candidate genes (Supplementary Figure 8.1). The tier-1 panel consisted of 24 candidate genes from HGMD Professional v2016.3 using keywords "nephrolithiasis", "phosphaturia" and "hypercalciuria". The tier-2 panel consisted of 155 non-redundant genes extracted from extended search of OMIM database (December 2015 update) using the same keywords as above and the tier-3 panel was comprised of 188 additional genes with implication on homeostasis and

kidney function. The tier-3 gene set was prepared from an extended search of literatures by our clinical expert Dr Valerie Walker.

The filtering strategy for analysing variants involved exclusion of variants; (1) with read-depth less than 10; (2) located in genes identified as highly mutable (according to the Fuentes *et al.* list[100]) and (3) had a strand bias or low base-quality score (BQ<20) as described in Chapter 2.

Since nephrolithiasis is a relatively common condition in the general population, minor allele frequency of less than or equal to 2% (MAF≤0.02) was applied to retain rare and low frequency variants. This threshold is arbitrary, however it was decided based on the estimated prevalence of renal stone in the Hampshire region as suggested by the clinical colleague with relevant expertise in the field.

Following analysis of all variants in the list of 367 genes, aggressive filtering was applied to the remaining segregation consistent variants with MAF≤ 0.02 and variants were further excluded if they: (1) were either synonymous, non-frameshift insertion/deletion or non-coding RNA (ncRNA) splicing; or (2) had frequency of greater than 2% in the in-house SED database. The relevance of each variant to the homeostatic pathways was investigated using the DISEASE database[241] and variants in genes implicated in kidney related disorders were selected.

Variant discovery was finally complemented with genotype segregation analysis across additional members of each family (identified by filled circles in Figure3.1, Figure3.2 and Figure3.3). Genotyping procedure in individuals without WES data, carried out using KASPAR genotyping system[242].

3.4 Results

Principal component analysis (PCA) revealed that all samples are of European descent (Figure 3.4) and exhibited expected *identical by state (IBS)* sharing between pedigree members (Figure 3.5).

Figure 3.4: PCA projection of the nine nephrolithiasis samples (identified by squares) onto the 1000 genomes data; (**AFR:** African; **AMR:** Ad Mixed American, **EAS:** East Asian, **EUR:** European, **SAS:** South Asian)

	FM.A-1	FM.A-23	FM.A-30	FM.B-1	FM.B-32	FM.B-71	FM.C-3	FM.C-8	FM.C-34
FM.A-1	100	46.29	51.27	43.45	43.03	43.12	43.56	43.89	43.74
FM.A-23	46.83	100	50.86	44.47	44.02	44.06	43.91	44.23	44.13
FM.A-30	51.89	50.88	100	43.52	43.12	44.02	43.85	44.02	43.94
FM.B-1	43.82	44.34	43.37	100	47.36	48.44	44.61	44.54	44.31
FM.B-32	44.02	44.49	43.56	48.01	100	46.63	44.58	44.34	44.56
FM.B-71	43.75	44.19	44.12	48.73	46.28	100	44.48	44.18	44.58
FM.C-3	43.24	43.61	43.23	44.09	44.14	44.51	100	51.58	49.13
FM.C-8	43.75	43.84	43.13	44.51	44.23	44.57	51.46	100	47.61
FM.C-34	43.34	43.54	43.08	44.64	44.13	44.46	48.95	47.61	100

Figure 3.5: Similarity matrix analysis. The pairwise comparisons of genotypic similarity between samples is used to investigate potential deviations from expected IBS ascertained from pedigree information; In this IBS heatmap plot, unrelated individuals represent 43-44% genotypic similarity, whereas first cousins (including *IV-23* and *IV-30* in family A, *III-1*, *III-32* and *III-71* in family B and *III-8* & *III-34* in family C) represent 46-48% genotypic similarity and aunt-nephew relationship such as (*FM.A-1* - *FM.A-23* and *FM.A-1* - *FM.A-30* is identified by 50-51% genotypic similarity); All samples exhibited expected IBS between pedigree members.

For all samples, ≥80% of reads were correctly mapped to the reference genome and the mean coverage per sample across the target region was at least 48 reads (Table 3.3). The cumulative coverage plot demonstrated that greater than 85% of capture target was covered to a depth of at least 20X (Figure 3.6). The 20X threshold for capture coverage is rather arbitrary, but we assumed that this threshold ensures sufficient depth for reliable variant calling. The percentage coverage for the 367 tier gene panel is provided in the Supplementary Table 8.5.

Sample ID	Inferred gender	%X Het	% Het	Number of variants	% Reads mapped to target	Mean coverage	Verify BamID score
Family A-1	Female	59.5	61.67	24783	82.25	56.57	0.00129
Family A-23	Male	17.63	60.47	24791	81.77	62.03	0.00057
Family A-30	Male	14.63	60.74	25080	82.09	65.23	0.00032
Family B-1	Male	17.48	61.25	24719	82.58	67.44	0.0005
Family B-32	Female	65.34	61.66	24530	82.35	48.31	0.00043
Family B-71	Female	60.69	60.85	24865	82.42	54.29	0.00087
Family C-3	Male	20.49	64.52	36257	88.78	56.65	0.00031
Family C-8	Male	11.27	63.51	35725	90.21	44.40	0.00022
Family C-34	Male	16.66	64.31	36154	88.74	52.53	0.00001

Table 3.3: Alignment and coverage QC results; The greater number of variant sites across individuals of Family C is due to the use of enhanced version of capture kit (Agilent SureSelect V6.0).

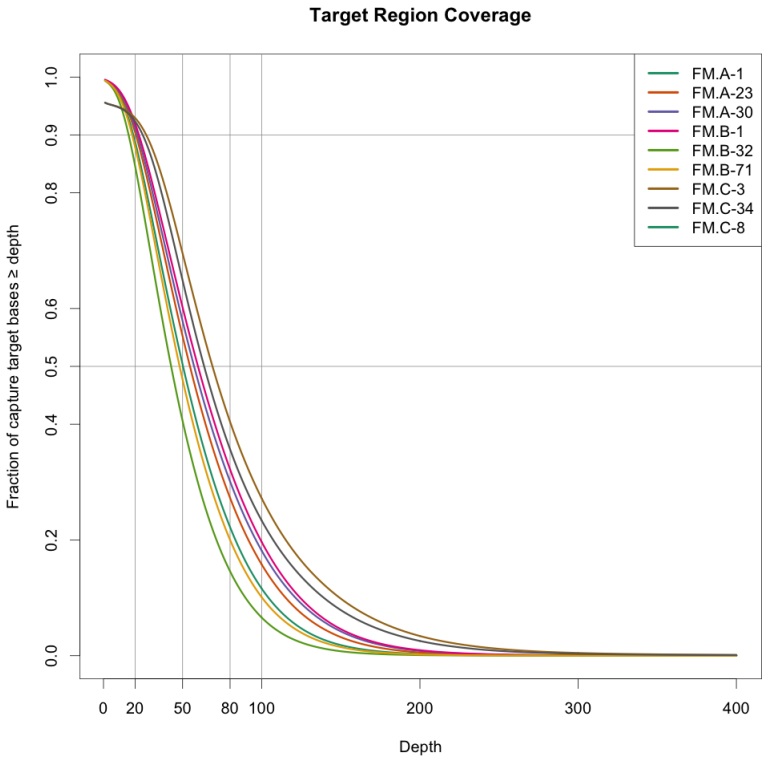


Figure 3.6: Cumulative depth of coverage (DOC) across capture target region.

3.4.1 Family A

As discussed in earlier sections, hyperphosphaturia appears to be the prominent abnormal biochemical trait in this family. At the time of recruitment the proband III-1 and her nephew IV-23 were both hyperphosphaturic. Since the proband's brother III-18 was not alive, his biochemical status could not be tested, however given the fact that his son IV-23 and his grandchildren V-26 and V-27 all manifested the trait, it appears likely that the proband's brother III-18 was also hyperphosphaturic. Furthermore, the proband's other brother III-20 presented with low renal threshold phosphate concentration (TmPo4/GFR) which is indicative of hyperphosphaturia (Supplementary Table 8.1). Considering that the proband's son IV-36 is also showing a borderline TmPo4/GFR value, it would be reasonable to assume that the renal hyperphosphaturia trait in this family demonstrates an autosomal dominant pattern of inheritance.

Gene panel results

Exome analysis and tiered filtering shortlisted a total of 72 variants (Supplementary Table 8.6) of which, five variants were shared among all the three affected individuals (Table 3.4). Among the variants shortlisted in this family, three variants including *NBPF3*:c.1563G>C, *ADRA2B*:c.891-892insGAAGAGGAG and *AKAP12*:c.4595-4596insGGA had a very high frequency in the SED database and therefore were excluded from further analysis. Variants with high frequency in the SED database represent regional common variants or more frequently they are the product of erroneous variant call across low complexity regions of DNA; such variants appear unlikely to underlie the hyperphosphaturia trait in this family. Furthermore, these variants have not been reported in highly curated databases such as the 1000 Genomes, EVS or ExAC, hence they probably denote a systematic bias in alignment or variant calling. The fact that these three variants fall in homopolymer tracts, gives further support to the notion that these variants are spurious and therefore deprioritised for further follow-up.

Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.A-1	p.A-23	p.A-30
NBPF3	ns	c.G1563C	p.Q521H	-	-	-	-	-	0.3641	hom (R.D.=8)	hom (R.D.=8)	hom (R.D.=8)
ADRA2B	nfi	c.891_892ins GAAGAGGAG	p.E298insEEEE	-	-	-	-	-	0.4652	het (R.D.=32)	het (R.D.=33)	het (R.D.=28)
AKAP12	nfi	c.4595_4596 insGGA	p.D1532insED	-	-	-	-	-	0.7641	hom (R.D.=109)	hom (R.D.=128)	hom (R.D.=146)
MAP3K5	ns	c.G3943A	p.D1315N	0.998946	4.37	0.0037	0.009302	0.006569	0.0098	het (R.D.=153)	het (R.D.=168)	het (R.D.=167)
SLC25A25	ns	c.G1047C	p.Q349H	0.999532	5.08	0.0018	0.002209	0.002637	0.0022	het (R.D.=48)	het (R.D.=57)	het (R.D.=52)

Table 3.4: Variants identified through tier filtering in family A; (ns: Non-synonymous; nfi: Non-frameshift insertion; PolyPhen2: Predictive score for estimating possible functional impact of amino acid change on the structure and function of human protein- PolyPhen2 scores ranges from 0.0 (tolerated) to 1.0 (deleterious); GERP^{++}: The Genomic Evolutionary Rate Profiling (GERP) score for quantifying variant-site conservation- GERP score ranges from -12.3 to +6.17 with scores $\geq$ 2 indicating high conservation rate; MAF: Minor allele frequency; EVS: The Exome Variant Server; ExAC: The Exome Aggregation Consortium; SED: The Soton Exome Database; R.D.: Read Depth; The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red).

Evaluation of pathogenicity and conservation scores across the shortlisted variants in the Table 3.4 revealed two highly damaging non-synonymous variants in the *MAP3K5* and *SLC25A25* genes.

The *MAP3K5*:c.3943G>A variant maps to the second exon of the gene and results in substitution of aspartic acid (Asp) with asparagine (Asn) in the amino acid sequence. Apoptosis signal-regulating kinase 1 also known as mitogen-Activated Protein Kinase Kinase Kinase 5 (MAP3K5) is a highly conserved kinase that plays an important role in

the MAPK signalling cascade. The gene consists of 30 exons that code for 11 essential kinase subdomains[243]. MAP3K5 acts upstream of p38 MAPK and JNK pathways and it has been identified to be active in pathological kidney conditions[244,245]. Studies suggest that activation of MAP3K5 induced p38 MAPK signalling cascade in the kidneys occurs in response to hyperosmotic stresses, and activation of p38 MAPK is deemed essential for protecting epithelial cells in glomeruli and tubules from osmotic stresses[246,247]. The non-synonymous *MAP3K5*:c.3943G>A maps to the second exon of the gene that codes for a domain of unknown function (Family: DUF4071 (PF13281)). Given the inferred damaging impact of this substitution (PolyPhen2^{HDIV} score= 0.99 and GERP^{++} score= 4.37), the variant was selected for follow up segregation analysis in all family members with available DNA.

The non-synonymous *SLC25A25*:c.1047G>C maps to the exon 8 of the gene and results in substitution of glutamine (Gln) to histidine (His). The *SLC25A25* gene consists 11 exons that code for a calcium-binding mitochondrial carrier known as solute carrier family 25 member 25[248]. The product of the gene functions as an ATP-Mg/Pi transporter in the inner membranes of mitochondria and regulates trafficking of ATP into or out of the mitochondrial matrix[249]. Microarray expression analysis of kidney stone in model animals identified 15% decrease in expression of *SLC25A25* during experimental course of induced nephrolithiasis in mouse[250]. Given the important role of *SLC25A25* in mediating phosphate balance in the cell cytoplasm and also considering the inferred damaging impact of this mutation (PolyPhen2^{HDIV} score= 0.99 and GERP^{++} score= 5.08) we proceeded to follow-up this variant as a plausible candidate for the hyperphosphaturia trait in this family.

Pan-genomic Result

Following analysis of the gene panel, an aggressive filtering was applied to the remaining variants with MAF$\leq$ 0.02 as per the procedure describe in Section 3.3.4. Pan-genomic analysis shortlisted a total of 19 variants that are shared by all the affected individuals (Table 3.5).

Chr.	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	ExAC	SED (n=460)	p.A-1	p.A-23	p.A-30
2	**PKP4**	ns	c.C2120T	p.A707V	0.978622	3.79	0.0009	0.00314	0.00138	-	HET (R.D.=72)	HET (R.D.=96)	HET (R.D.=99)
2	**PLA2R1**	ns	c.A1814C	p.H605P	0.997364	5.01	-	0.001744	0.001122	-	HET (R.D.=47)	HET (R.D.=34)	HET (R.D.=40)
3	**LOC100132146**	stg	c.C136T	p.R46X	-	-	0.01	-	-	0.0065	HET (R.D.=129)	HET (R.D.=177)	HET (R.D.=189)
5	**DOCK2†**	sp	c.1555+3G>A		MaxEnt= -0.82		-	-	-	-	HET (R.D.=42)	HET (R.D.=51)	HET (R.D.=59)
5	**HAVCR1**	ns	c.A1050G	p.Q339R	-	-0.02	0.0002	0.0009	0.001646	-	HET (R.D.=47)	HET (R.D.=40)	HET (R.D.=42)
6	**L3MBTL3†**	sp	c.1966+5C>G		MaxEnt= -2.81		-	0.000116	0.0000578	-	HET (R.D.=31)	HET (R.D.=40)	HET (R.D.=41)
9	**KIAA1045**	ns	c.G388A	p.D130N	0.999648	5.29	0.0023	0.001886	0.00157	0.0043	HET (R.D.=53)	HET (R.D.=45)	HET (R.D.=46)
9	**TLE1**	ns	c.G1468A	p.V490M	0.999241	5.61	-	-	-	-	HET (R.D.=64)	HET (R.D.=68)	HET (R.D.=70)
10	**MYO3A**	fsd	c.637_638del	p.213_213del	-	-	-	-	-	-	HET (R.D.=51)	HET (R.D.=98)	HET (R.D.=104)
10	**NEBL**	ns	c.C2654T	p.S885F	0.995291	4.63	0.0014	0.00407	0.002207	0.0011	HET (R.D.=58)	HET (R.D.=81)	HET (R.D.=83)
10	**CTNNA3**	ns	c.C2524T	p.R842W	0.965169	1.49	-	-	0.00006591	-	HET (R.D.=132)	HET (R.D.=106)	HET (R.D.=141)
12	**TAS2R30**	ns	c.T842G	p.L281W	-	-	-	-	0.002384	0.0011	HET (R.D.=176)	HET (R.D.=247)	HET (R.D.=218)
13	**MYCBP2**	ns	c.T8363C	p.L2788S	0.998524	5.14	.	0.001395	0.001146	0.0033	HET (R.D.=70)	HET (R.D.=75)	HET (R.D.=86)
15	**LOC283710**	fsd	c.67delC	p.P23fs	-	-	-	-	-	0.0087	HOM (R.D.=26)	HOM (R.D.=27)	HOM (R.D.=25)
15	**ACSBG1**	ns	c.G1975C	p.E659Q	0.983482	3.74	0.0018	0.005823	0.003377	0.0098	HET (R.D.=64)	HET (R.D.=60)	HET (R.D.=75)
16	**TXNDC11**	ns	c.C1299G	p.H433Q	0.025643	-0.281	0.0041	0.016628	0.008947	0.0109	HET (R.D.=61)	HET (R.D.=88)	HET (R.D.=89)
17	**MYH1**	ns	c.G1918A	p.G640S	0.999045	5.4	0.0009	0.00593	0.003272	0.0043	HOM (R.D.=108)	HOM (R.D.=100)	HET (R.D.=129)
19	**FAM71E2**	ns	c.A2333G	p.Q778R	-	-	-	-	-	0.0207	HET (R.D.=42)	HOM (R.D.=43)	HOM (R.D.=63)
19	**ODF3L2**	ns	c.C824T	p.T275M	0.028838	-7.13	0.0037	0.008048	0.006038	-	HET (R.D.=18)	HET (R.D.=19)	HET (R.D.=11)
20	**VPS16**	ns	c.T1750C	p.W584R	0.997977	5.29	0.0023	0.008372	0.006781	0.0011	HET (R.D.=49)	HET (R.D.=74)	HET (R.D.=71)

Table 3.5: Pan-Genomic variants with MAF $\leq$2% in the SED database identified across family A; ns: Non-synonymous SNV, sp: Splice-site variant, fsd: Frameshift deletion, stg: Stop-gain mutation, R.D.: Read depth, EVS: The Exome Variant Server, ExAC: The Exome Aggregation Consortium, SED: The Soton Exome Database; †: For splicing variants the MaxEnt scores are reported- Scores $>$ |3| are considered to disrupt splicing, The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red.

Across the shortlisted pan-genomic variants (Table 3.5), four candidates including '*PKP4*:c.2120C>T', '*PLA2R1*:c.1814A>C', '*NEBL*:c.2654C>T' and '*VPS16*:c.1750T>C' were selected for follow up segregation analysis. The remaining variants were excluded on the basis of low pathogenicity or lack of plausible biological relevance to kidney stones.

Segregation Analysis

A total of 7 variants in family A were considered for segregation analysis (Table 3.6). Based on consistency of segregation across available family members, variants were given higher priority if they were overrepresented among stone formers and they were absent in non-stone formers. According to this criterion, the *NEBL* variant ($NEBL : NM-006393 : exon26 : c.C2654T$) appeared to consistently segregate across all stone formers and is only present in one non-stone former with hyperphosphaturia (FM A-27).

NEBL encodes a nebulin like protein that is actively expressed in heart. Nebulin binds to actin and interacts with thin filaments and Z-line associated proteins in striated muscles. Nebulin mutations have been primarily reported in the context of dilated cardiomyopathy and endocardial fibroelastosis [251]. A recent report identified that nebulin expression increases in monocytes upon exposure to calcium oxalate [252].

Similarly, the *VPS16* variant ($VPS33B : NM - 018668 : exon18 : c.C1342T : p.P448S$) and *HAVCR1* variant ($HAVCR1 : NM-012206 : exon8 : c.A1050G : p.A350A$) consistently segregate among all stone formers and are present in only one asymptomatic carrier (individual FM A-26 for the *VPS16* variant and individual FM A-49 for the *HAVCR1* variant). *VPS16* encodes a vacuolar protein storing (VPS) complex that plays an important role in segregation of intracellular molecules into distinct organelles.

The *HAVCR1* gene encodes a cell surface receptor for T-cell immunoglobulin and mucin domain 1. Expression of this membrane receptor increases in response to kidney injury and it is suggested as a biomarker for kidney injury [253].

The *VPS33B* variant ($VPS33B : NM - 018668 : exon18 : c.C1342T : p.P448S$) does not consistently segregate with the stone phenotype in Family A, and it is unlikely to be causal in the context of stone phenotype.

The *MAP3K5* variant ($NM - 005923 : exon28 : c.G3943A$) consistently segregates across all stone-formers and three of non-stone formers including individual FM A-27 with hyperphosphaturia phenotype. Mutations of mitogen-activated protein kinase kinase kinase 5 (MAP3K5) are primarily shown to underlie range of malignancies including renal cancer.

The remaining variants in *PLA2R1*, *SLC25A25* and *PKP4* are present among majority of non-stone formers and thus are either unrelated to the stone phenotype or they represent a pathogenic variant with reduced penetrance which is a likely presumption in the context of nephrolithiasis in this family.

Variants	Family A											
	Stone formers				*	Non-stone former						
	FM A-01	**FM A-20**	**FM A-23**	**FM A-30**	**FM A-27**	**FM A-28**	**FM A-26**	**FM A-36**	**FM A-39**	**FM A-40**	**FM A-43**	**FM A-49**
PLA2R1:c.A1814C (rs151215519)	HET (G:T)	HET (G:T)	HET (G:T)	HET (G:T)	HET (G:T)	HET (G:T)	HOM (T:T)	HET (G:T)	HET (G:T)	HOM (T:T)	HET (G:T)	HET (G:T)
MAP3K5:c.G3943A (rs41288957)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)
SLC25A25:c.G1047C (rs140777921)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HET (G: C)	HOM (G:G)	HET (G: C)
NEBL:c.C2654T (rs143584663)	HET (A:G)	HET (A:G)	HET (A:G)	HET (A:G)	HET (A:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)
VPS16:c.T1750C (rs61729229)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HOM (T:T)	HOM (T:T)	HET (T:C)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)
PKP4:c.C2120T (rs140419507)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HET (T:C)	HOM (C:C)	HET (T:C)	HET (T:C)	HOM (C:C)	HET (T:C)	HET (T:C)
HAVCR1:c.A1050G (rs201441165)	HET (C:T)	HET (C:T)	HET (C:T)	HET (C:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HET (C:T)

Table 3.6: Segregation results for prioritised variants in family A. Heterozygous variants are identified by red colour and homozygous reference variants are highlighted in green; $*$: Individual *FM A-27* presented with hyperphosphaturia, but he has not developed kidney stones.

3.4.2 Family B

The most striking homeostatic imbalance among stone formers in this family is hyperphosphaturia (Supplementary Tables 8.2 and 8.3). Five individuals (II-1, II-2, II-25, III-71 and III-75) including three of the stone formers (II-1, II-2 and III-71) had very low TmPo4/GFR ratio indicating persistent hyperphosphaturia. Furthermore, four individuals (II-1, III-71, III-32 and II-25) had high 24-urinary calcium excretion; among whom individuals II-1, III-71, III-32 were stone formers (Supplementary Table 8.2). The three stone formers who exome analysed, presented with multiple biochemical abnormalities that predispose to nephrolithiasis. High calcium excretion was the consistent feature among the three patients. In relation to the stone phenotype in this family, considering that majority of the stone formers presented with hyperphosphaturia it is likely that autosomal dominant hypophosphatemic nephrolithiasis underlies the biochemical trait in this family.

Gene panel results

Tier filtering in family B shortlisted a total of 65 variants (Supplementary Table 8.7) of which five variants were shared among all affected individuals (Table 3.7).

Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.B-1	p.B-32	p.B-71
ADCY10	ns	c.C3599T	p.P1200L	0.158114,	-1.31	0.02	0.04	0.03	0.0315	het (R.D.=71)	het (R.D.=55)	het (R.D.=62)
ADRA2B	nfi	c.891_892ins GAAGAGGAG	p.E298del insEEEE	-	-	-	-	-	0.4652	het (R.D.=103)	het (R.D.=91)	het (R.D.=92)
SLC4A4	ns	c.A3220C	p.I1074L	-	-	0.16	0.186	0.26	0.1837	het (R.D.=20)	het (R.D.=13)	het (R.D.=13)
AKAP12	nfi	c.4595_4596 insGGA	p.D1532 insED	-	-	-	-	-	0.7691	hom (R.D.=167)	hom (R.D.=102)	hom (R.D.=125)
MAP3K14	u	UNKNOWN		-	-	-	-	-	0.9196	hom (R.D.=101)	hom (R.D.=64)	hom (R.D.=75)

Table 3.7: Rare variants identified across tier filtering for family B; (ns: Non-synonymous; nfi: Non-frameshift insertion; PolyPhen2: Predictive score for estimating possible functional impact of amino acid change on the structure and function of human protein- PolyPhen2 scores ranges from 0.0 (tolerated) to 1.0 (deleterious); GERP^{++}: The Genomic Evolutionary Rate Profiling (GERP) score for quantifying variant-site conservation- GERP score ranges from -12.3 to +6.17 with scores ≥ 2 indicating high conservation rate; MAF: Minor allele frequency; EVS: The Exome Variant Server; ExAC: The Exome Aggregation Consortium; SED: The Soton Exome Database; Frequencies in the SED database is expressed as number of alternative Homs, Hets, Total; R.D.: Read Depth; The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red.).

Across the variants identified through tier filtering all apart from the *ADCY10*:c.3599C>T had a high frequency in the SED database. As for the *SLC4A4*:c.3220A>C variant, however $GERP^{++}$ score indicating nucleotide change occurs within a conserved region ($GERP^{++}$=4.42), the low Polyphen2 score for this variant refutes its involvement in the disease. In addition, the high frequency of the variant in the SED database (MAF_{SED}= 0.18) exempts the variant as being causal in the context of kidney stones. Despite this, the variant was selected for follow up segregation analysis as a due diligence measure. For the remaining variants including ADRA2B:c.891_892insGAAGAGGAG, *AKAP12*:c.4595_4596insGGA and the *MAP3K14* variant with an unknown effect, their high frequency in the SED database and also lack of reported MAF in the curated databases rendered them less likely to underlie the stone trait in this family and therefore excluded from further follow-up.

The non-synonymous *ADCY10*:c.3599C>T has a low frequency in the general populations (MAF_{1000G}= 0.02, MAF_{EVS}= 0.04 & MAF_{ExAC}= 0.03). Low Polyphen2 and Gerp++ scores indicate that the variant is not deleterious and does not fall within a conserved region. Given that kidney stones are not lethal, low pathogenicity and conservation

scores for the *ADCY10* variant may not be inconsistent with causality. Therefore, this variant was selected for follow up segregation analysis.

Pan-genomic Result

Pan-Genomic analysis in family B identified 19 variants (Table 3.8). Inspection of pan-genomic result revealed that family B harbours a rare ancestral haplotype that stretches between Chr19:40Mb- 49Mb and segregates with the disease. It is notable that pan-genomic variants were shortlisted regardless of their chromosomal location and are listed solely on the basis of their inferred pathogenicity scores. Three damaging variants with high conservation scores (*PTPN13*:c.3563C>T, *PLD3*:c.694G>A and *PVRL2*:c.1093G>A) were shortlisted for further follow-up through segregation analysis.

Chr.	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.B-1	p.B-32	p.B-71
4	**PTPN13**	ns	c.C3563T	p.T1188I	0.999	5.74	0.0014	0.001574	0.001485	-	**HET** (R.D.=18)	**HET** (R.D.=16)	**HET** (R.D.=9)
4	**STOX2**	ns	c.A1094G	p.K365R	-	-	-	-	-	-	**HET** (R.D.=67)	**HET** (R.D.=40)	**HET** (R.D.=58)
8	**ASAP1†**	sp	c.1975+5G>A		MaxEnt= 3.12		0.01	0.03	0.02418	0.0098	**HET** (R.D.=48)	**HET** (R.D.=37)	**HET** (R.D.=34)
8	**TPD52**	ns	c.T25C	p.Y9H	0.172572	-1.03	0.01	0.004535	0.004602	0.0054	**HET** (R.D.=92)	**HET** (R.D.=52)	**HET** (R.D.=68)
10	**DCLRE1A**	stg	c.C412T	p.R138X	0.959824	3.11	0.0018	0.00314	0.002752	0.0054	**HET** (R.D.=142)	**HET** (R.D.=95)	**HET** (R.D.=123)
12	**ANKS1B**	ns	c.G1456A	p.A486T	.		0.01	0.015781	0.01214	0.0087	**HET** (R.D.=33)	**HET** (R.D.=35)	**HET** (R.D.=32)
14	**PACS2†**	sp	c.1985+9G>T		MaxEnt= NA		0.01	0.009797	0.01016	0.0098	**HET** (R.D.=133)	**HET** (R.D.=114)	**HET** (R.D.=102)
14	**C14orf166B**	ns	c.G328A	p.V110M	.		0.01	0.00593	0.004745	0.0076	**HET** (R.D.=23)	**HET** (R.D.=17)	**HET** (R.D.=18)
16	**DNAAF1†**	sp	c.1645-3C>T		MaxEnt= 0.07		-	-	-	-	**HET** (R.D.=88)	**HET** (R.D.=66)	**HET** (R.D.=68)
19	**PLD3**	ns	c.G694A	p.V232M	0.993514	4.23	0.0005	0.004884	0.003077	0.0033	**HET** (R.D.=64)	**HET** (R.D.=48)	**HET** (R.D.=54)
19	**HIPK4**	ns	c.C911A	p.T304N	0.088914	-8.8	.	0.008839	0.005435	0.0076	**HET** (R.D.=126)	**HET** (R.D.=80)	**HET** (R.D.=100)
19	**TMEM143**	ns	c.G427A	p.D143N	0.996452	3.86	0.01	0.01	0.00812	0.0109	**HET** (R.D.=77)	**HET** (R.D.=44)	**HET** (R.D.=63)
19	**DMPK**	ns	c.C1591G	p.Q531E	0.981423	2.62	-	-	-	-	**HET** (R.D.=51)	**HET** (R.D.=33)	**HET** (R.D.=38)
19	**BCAM**	ns	c.C824T	p.S275F	0.996651	2.77	0.0005	0.002093	0.001609	0.0011	**HET** (R.D.=73)	**HET** (R.D.=52)	**HET** (R.D.=73)
19	**NUMBL**	ns	c.T1470G	p.F490L	0.236576	1.59	.	0.004304	0.01043	0.0011	**HET** (R.D.=36)	**HET** (R.D.=20)	**HET** (R.D.=19)
19	**PVRL2**	ns	c.G1093A	p.V365M	0.992148	3.58	.	0.002942	0.01022	0.0011	**HET** (R.D.=35)	**HET** (R.D.=20)	**HET** (R.D.=24)
19	**EMR2**	ns	c.C2288T	p.T763M	0.010446	-8.28	0.01	0.010581	0.00669	0.0109	**HET** (R.D.=39)	**HET** (R.D.=19)	**HET** (R.D.=25)
19	**ARHGEF1**	sp†	c.842-7G>A		MaxEnt= 0.41		0.0018	0.004767	0.005181	0.0033	**HET** (R.D.=27)	**HET** (R.D.=29)	**HET** (R.D.=32)
X	**FAM120C**	ns	c.G2802A	p.M934I	0.888578	1.14	0.01	0.021998	0.01152	0.0141	**HOM** (R.D.=11)	**HET** (R.D.=14)	**HET** (R.D.=22)

Table 3.8: Pan-Genomic variants with MAF ≤2% in the SED database identified across family B; ns: Non-synonymous SNV, sp: Splice-site variant, stg: Stop-gain mutation, R.D.: Read depth, EVS: The Exome Variant Server, ExAC: The Exome Aggregation Consortium, SED: The Soton Exome Database; †: For splicing variants MaxEnt scores are reported- Scores > |3| are considered to disrupt splicing, The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red.

The *PTPN13* gene contains 48 exons and encodes a member of protein tyrosine phosphatase (PTP) family that is involved in a variety of cellular mechanisms including cell growth and oncogenic transformation [254]. No direct link between pathogenic mutations of *PTPN13* and homeostatic imbalance has been found to date, but pathogenic mutations of the gene have been reported in the context of renal cell carcinoma [255].

PLD3 on chromosome 19q13.2 contains 15 exons and encodes a member of the phospholipase D (PLD) family that catalyse hydrolysis of amyloid-beta precursor protein. The gene undergoes extensive splicing and is highly expressed in the central nervous system [256,257]. Pathogenic mutations of the *PLD3* have been primarily reported in the context of Alzheimer disease (AD19, #OMIM 615711). Similar to *PTPN13*, no direct link between the mutations of PLD and kidney stones have been reported yet.

PVRL2 on chromosome 19q13.32 encodes for a specific immunoglobulin-like adhesion molecules that act as adherence junctions in the cell membrane and facilitate calcium inde-

pendent cell-cell adhesion [258]. Similar to previous genes, no direct link between pathogenic mutations of *PVRL2* and kidney stone have been reported yet.

Given the lack of biological relevance between the shortlisted variants and nephrolithiasis, it is unlikely that these variants underlie the stone phenotype in this family and they could be regarded as incidental findings, however as a conservative measure they were considered for the follow up segregation analysis. Furthermore, both the *PLD3* and *PVRL2* are located on chromosome 19q13 and they might flag an ancestral haplotype that might underlie the stone phenotype in this family.

Segregation Analysis

Across the five variants considered for segregation analysis in family B, only the *ADCY10* variant ($NM-001167749:exon26:c.C3599T$) consistently segregated among all stone formers and the remaining variants did not satisfy the segregation criteria. Heterozygous mutations of *ADCY10* are reported in the context of monogenic nephrolithiasis [259] and therefore its possible role in disease pathology should not be ruled out solely based on its segregation among non-stone formers (Table 3.9).

Variants	Family B																								
	Stone Former						*	Non-stone Former																	
	FM B-01	FM B-02	FM B-21	FM B-32	FM B-71	FM B-25	FM B-27	FM B-03	FM B-20	FM B-23	FM B-24	FM B-26	FM B-34	FM B-41	FM B-42	FM B-43	FM B-48	FM B-49	FM B-61	FM B-62	FM B-74	FM B-75	FM B-78	FM B-88	FM B-89	FM B-101	FM B-103	FM B-105
ADCY10:c.C3599T (rs72697797)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HOM** (G:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HET** (A:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)
SLC4A4:c.A3220C (rs1062677)	**HET** (C:A)	**HET** (C:A)	**HOM** (A:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HOM** (A:A)	**HOM** (A:A)	**HET** (C:A)	**HET** (C:A)	**HOM** (A:A)	**HET** (C:A)	**HOM** (A:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HOM** (C:C)	**HET** (C:A)	**HET** (C:A)	**HOM** (C:C)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HET** (C:A)	**HOM** (A:A)
PVRL2:c.G1093A (rs201037066)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	???	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)
PTPN13:c.C3563T (rs186651045)	**HET** (C:T)	**HET** (C:T)	**HET** (C:T)	**HET** (C:T)	**HET** (C:T)	**HOM** (C:C)	**HET** (C:A)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HET** (C:T)	**HET** (C:T)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HET** (C:T)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HOM** (C:C)	**HET** (C:T)	**HOM** (C:C)
PLD3:c.G694A (rs145999145)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HET** (G:A)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HET** (G:A)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)	**HOM** (G:G)

Table 3.9: Segregation results for prioritised variants in family B. Heterozygous variants are identified by red colour and homozygous reference variants are highlighted in green; *: Individual *FM B-27* presented with hyperphosphaturia, but he has not developed kidney stones.

3.4.3 Family C

Idiopathic hypercalciuria appears to be the dominant trait predisposing to renal calculi in this family. All stone formers are either hypercalciuric (individuals III-1 and IV-3) or they have 24-hour urinary calcium levels very close to the maximum limit of normal range (individuals III-8 and III-34). In addition, in three family members including III-13, IV-3 and IV-10, hypercalciuria is associated with hypervitaminosis D which might be an indication for absortive hypercalciuria.

Gene panel result

The tiered filtering approach in this family shortlisted a total of 37 variants (Supplementary Table 8.8) of which four variants were shared among the three affected individuals (Table 3.10). Three variants including *SLC9A3*:c.2368T>C, *SLC26A2*:c.1721T>C and *GPSM1*:c.c.23T>C appeared to have a very high frequency in the SED and were therefore excluded from downstream analysis. Despite the low conservation and pathogenicity scores for the heterozygous *SLC20A1*:c.G1020C variant, it was selected for segregation follow up. The rationale behind selecting this variant was solely based on its significant role in absorbing phosphate from intestine. Given that mild hyperphosphaturia is associated with stone phenotype in individual III-1, we hypothesized that impaired functioning of the *SLC20A1* might be related to the stone phenotype in this family.

Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.C-3	P.C-8	p.C-34
SLC20A1	ns	c.G1020C	p.E340D	0.004	-1.31	3.69	.	.	0.012	het (R.D.= 20)	het (D.P.= 10)	het (D.P.= 27)
SLC9A3	ns	c.T2368C	p.C790R	0	3.84	-	0.87	0.8253	0.8144	hom (R.D.= 94)	het (D.P.= 50)	hom (R.D.= 115)
SLC26A2	ns	c.T1721C	p.I574T	0	4.37	.	0.99	0.9928	0.9931	hom (R.D.= 106)	hom (D.P.= 55)	hom (D.P.= 67)
GPSM1	ns	c.T23C	p.V8A	.	.	.	.	1	0.9986	hom (R.D.=15)	hom (D.P.= 19)	hom (D.P.= 16)

Table 3.10: Variants identified through tier filtering in family C; (ns: Non-synonymous; PolyPhen2: Predictive score for estimating possible functional impact of amino acid change on the structure and function of human protein- PolyPhen2 scores ranges from 0.0 (tolerated) to 1.0 (deleterious); GERP^{++}: The Genomic Evolutionary Rate Profiling (GERP) score for quantifying variant-site conservation- GERP score ranges from -12.3 to +6.17 with scores ≥ 2 indicating high conservation rate; MAF: Minor allele frequency; EVS: The Exome Variant Server; ExAC: The Exome Aggregation Consortium; SED: The Soton Exome Database; R.D.: Read Depth; The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red).

Pan-genomic Result

Since individuals of this family were exomed using a newer version of the capture kit (Agilent SureSelect V6.0) which effectively provides coverage across an extra 9Mb region (compared to V5.0), a great majority of pan-genomic variants identified in this family included exonic variants in non-coding genes or interstitial variants in either intronic or upstream domains. In order to not bluntly exclude important regulatory variants in upstream of exonic regions, the generic hard-filtering criteria according to the GATK best practice guideline was used to exclude unreliable calls. This step included exclusion of variants with Mapping Quality (MQ)$\leq$ 40, Quality by Depth (QD)$\leq$ 2 and Fisher's strand bias (FS)$\geq$ 60. Although these thresholds are non-conservative, this step ensures exclusion of many false positive calls and retains only reliable variants across the extended coverage region.

Pan-Genomic analysis in family C identified 20 variants with consistent non-reference zygosity across all individuals (Table 3.11). Potential implication of shortlisted variants

to kidney function and homeostatic mechanisms were thoroughly investigated in collaboration with our clinical expert. Ultimately, three variants including *DNAJA4*:c.161T>C, *ARSD*:c.1100A>G and *ARSD*:c.992G>A were selected for segregation follow up.

Chr.	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.C-3	p.C-8	p.C-34
1	NBPF10	ns	c.C536A	p.A179D	-	-1.67	-	-	9.98E-05	0.0002	**HET** (R.D.=206)	**HET** (R.D.=140)	**HET** (R.D.=205)
2	REG3A	ns	c.A149C	p.H50P	0.389	2.69	-	-	8.60E-03	0.0003	**HET** (R.D.=154)	**HET** (R.D.=111)	**HET** (R.D.=133)
2	ANKRD36B	ns	c.A2758G	p.K920E	0.361	1.25	-	-	2.14E-02	0.00748	**HET** (R.D.=75)	**HET** (R.D.=64)	**HET** (R.D.=84)
4	SLC9B1	stg	c.A1318T	p.K440X	-	3.47	-	-	-	-	**HET** (R.D.=44)	**HET** (R.D.=39)	**HET** (R.D.=33)
4	SLC9B1	stg	c.C1234T	p.R412X	-	2.61	-	-	-	-	**HET** (R.D.=56)	**HET** (R.D.=44)	**HET** (R.D.=44)
4	POU4F2	ns	c.C417A	p.D139E	0.899	5.77	2.00E-03	9.80E-03	9.70E-03	0.0132	**HET** (R.D.=78)	**HET** (R.D.=62)	**HET** (R.D.=67)
9	SPATA31A3	ns	c.G937C	p.E313Q	0.519	1.21	-	-	-	-	**HET** (R.D.=69)	**HOM** (R.D.=34)	**HET** (R.D.=28)
9	FOXD4L6	ns	c.C1247G	p.P416R	-	-	-	-	-	0.0014	**HET** (R.D.=9)	**HET** (R.D.=10)	**HET** (R.D.=13)
9	FOXD4L6	ns	c.G1097A	p.R366K	-	-	-	-	-	0.0011	**HET** (R.D.=48)	**HET** (R.D.=49)	**HET** (R.D.=56)
11	MUC6	ns	c.C5521A	p.P1841T	0.899	-3.93	-	-	5.40E-03	0.0011	**HET** (R.D.=634)	**HET** (R.D.=564)	**HET** (R.D.=698)
11	MUC6	ns	c.T5507A	p.L1836H	0.899	-3.34	-	-	5.00E-03	-	**HET** (R.D.=685)	**HET** (R.D.=613)	**HET** (R.D.=737)
11	MUC6	ns	c.C5272T	p.H1758Y	0.637	0.656	-	-	4.30E-03	-	**HET** (R.D.=805)	**HET** (R.D.=576)	**HET** (R.D.=709)
12	TAS2R31	ns	c.T869A	p.F290Y	0.358	2.41	-	-	4.90E-03	-	**HET** (R.D.=205)	**HET** (R.D.=119)	**HET** (R.D.=189)
15	DNAJA4	ns	c.T161C	p.V54A	0.899	5.63	7.59E-03	1.30E-02	1.05E-02	0.0024	**HET** (R.D.=43)	**HET** (R.D.=35)	**HET** (R.D.=52)
15	GOLGA6L10	ns	c.A638G	p.E213G	-	-	-	-	9.00E-04	0.0076	**HET** (R.D.=12)	**HET** (R.D.=11)	**HET** (R.D.=10)
17	MYO15A	ns	c.G2642A	p.R881Q	0.432	3.16	-	-	-	0.0002	**HET** (R.D.=48)	**HET** (R.D.=43)	**HET** (R.D.=40)
X	ARSD	stl	c.T1147C	p.X383Q	-	-	-	-	-	-	**HET** (R.D.=44)	**HET** (R.D.=26)	**HET** (R.D.=50)
X	ARSD	ns	c.A1100G	p.K367R	-	-	-	-	1.00E-03	-	**HET** (R.D.=72)	**HET** (R.D.=41)	**HET** (R.D.=61)
X	ARSD	stg	c.G992A	p.W331X	-	3.68	-	-	-	-	**HET** (R.D.=41)	**HET** (R.D.=27)	**HET** (R.D.=33)
X	ARSD	ns	c.G959A	p.G320D	0.899	3.68	-	-	7.09E-05	-	**HET** (R.D.=48)	**HET** (R.D.=32)	**HET** (R.D.=33)

Table 3.11: Pan-Genomic variants with MAF ≤2% in the SED database identified across family C; ns: Non-synonymous SNV, stg: Stop-gain mutation, R.D.: Read depth, stl: Stop-loss mutation, EVS: The Exome Variant Server, ExAC: The Exome Aggregation Consortium, SED: The Soton Exome Database; The zygosity status for each variant across the three individuals is provided in the last columns and homozygous variants are highlighted in red.

Since the stone diathesis predominantly affects males in this pedigree, we hypothesized that damaging mutations of X-chromosome might be significantly relevant in this family. Hence, across the four *ARSD* mutations identified through pan-genomic analysis, two mutations including the stop-gain mutation on the exon 6 (c.G992A) and the non-synonymous (c.A1100G) on the exon 7 were preferentially prioritised for segregation analysis. Since the non-synonymous (c.G959A) on the exon 6 was in-phase with the stop-gain (c.G992A) mutation, and also the stop-loss (c.T1147C) mutation on the exon 7 was in phase with the non-synonymous (c.A1100G) variant, we restricted the genotyping effort to only these two loci. Apart from high expression of the gene in kidneys and also the apparently X-linked transmission of the phenotype in this family, there was no indication as to functional relevance of *ARSD* mutations to renal stones. The *ARSD* gene maps to the distal part of chromosome X (Xp22.33) in the vicinity of pseudo-autosomal region 1 (PAR1) and encodes an essential member of sulfatase family which is active in bone and cartilage tissues [260].

DNAJA4 on chromosome 15q25.1 contains 8 exons and encodes a member of DnaJ family of chaperones that regulate protein folding and trafficking in the cytosol. Members of DnaJ family actively prevent aggregation of misfolded proteins in endoplasmic reticulum and exhibit relatively high-expression in kidneys [261]. Given comparably high pathogenicity and conservation score for this variant and also possible implication of heat-shock chaperones (HSP) in kidney stone formation, we sought to follow up this variant through segregation analysis.

It is noteworthy to mention that, possible implication of the *SLC9B1* stop-gain mu-

tations in the stone phenotype was carefully considered, but due to lack of functional relevance and also restricted expression of the gene, which is limited to testis, we decided to abandon further follow-up for the *SLC9B1* variants.

Segregation Analysis

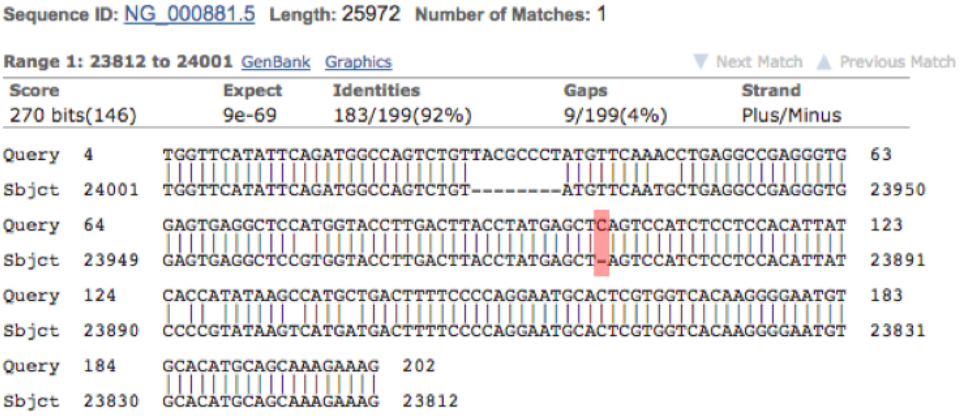
Across the four variants prioritised for segregation analysis, none appeared to be private to stone formers only (Table 3.12). Although the consistent heterozygous status for the *SLC20A1*:c.G1020C variant replicated through the KASPAR sequencing, presence of the heterozygous genotype in three non-stone formers including individuals C-10, C-13 and C-26 was inconsistent with the disease segregation. Despite both individuals C-10 and C-13 presenting with hypercalciuria, the status of 24h urinary calcium for individual C-26 could not be established due to lack of 24-hour urine sample availability (Supplementary Table 8.4).

Perhaps the most interesting insight, although not related to the stone phenotype, revealed from the genotyping assay was the status of *ARSD* variants across the individuals of this family. In contrast to WES result that was indicative of heterozygous variants at the *ARSD*:c.1100A>G and *ARSD*:c.992G>A, KSPAR genotyping assay was indicative of homozygous reference variants for the both loci across almost the entire pedigree (except for individual C-15). Since this was in sheer contrast with the WES result, and given that *ARSD* is not positioned in the PAR region, we sought to scrutinise the source of this discrepancy by looking into the read alignments across the entire *ARSD* region. Our analysis revealed that, the sequence homology between the *ARSD* gene and a related pseudogene on chromosome Y result in the false positive heterozygous calls across this region. The homologous sequence on the pseudoautosomal region of the Y chromosome share 91% sequence similarity with the *ARSD* gene. Due to the nature of enrichment process in which DNA fragments are pooled based on their unique sequences, inevitably *ARSD* related fragments are pooled with their pseudogene counterpart on the PAR region of chromosome Y. During the alignment, sequence differences across the remaining 9% of the gene result in erroneous heterozygous calls in males. For instance, the BLAST comparison of sequence composition 50bp upstream and downstream of the alternative C/T variant site at *ARSD*:c.992 highlights the source of false heterozygous call (Figure 3.7).

Variants	Family C									
	Stone formers			Non-stone former						
	FM C-03	**FM C-08**	**FM C-34**	**FM C-10**	**FM C-13**	**FM C-15**	**FM C-26**	**FM C-35**	**FM C-37**	**FM C-38**
SLC20A1:c.G1020C	HET (C:G)	HET (C:G)	HET (C:G)	HET (C:G)	HET (C:G)	HOM (G:G)	HET (C:G)	HOM (G:G)	HOM (G:G)	HOM (G:G)
ARSD:c.A1100G	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HOM (T:T)	HET (C:T)	HOM (T:T)	?	HOM (T:T)	HOM (T:T)
ARSD:c.G992A (rs111939179)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HET (T:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)	HOM (C:C)
DNAJA4:c.T161C (rs61752771)	HET (C:T)	HET (C:T)	HET (C:T)	HOM (T:T)	HET (C:T)	HET (C:T)	HET (C:T)	?	HOM (T:T)	HOM (T:T)

Table 3.12: Segregation results for prioritised variants in family C. Heterozygous variants are identified by red colour and homozygous reference variants are highlighted in green

(a) *Homo sapiens* arylsulfatase D (ARSD) on chromosome X



(b) *Homo sapiens* arylsulfatase D pseudogene 1 (ARSDP1) on chromosome Y

Figure 3.7: BLAST comparison of sequence composition 50bp upstream and downstream of the alternative [C/T] change at $ARSD : exon6 : c.G992A$; Un-aligned alternative base at the variant position is highlighted in red, paired nucleotide immediately left to the highlighted rectangles represent the wild type base (*i.e.* C for *ARSD* and T for *ARSDP1*)

The only exception to this consistent homozygous reference status for the *ARSD* variants across members of this family, was individual C-15 who revealed to be heterozygous at both loci (Table 3.12). Given that this individual is male and his mother's genotype identified to be homozygous reference at the both loci, his heterozygous status can be potentially attributed to genotyping error. Another possibility, although less likely, would be an XXY condition (Klinefelter syndrome (KS)) in this patient, which due to the lack of available data on his father's genotype (C-14) could not be reliably established. Nonetheless, the evidence in support of erroneous call for *ARSD* variants was compelling enough to reject possible implication of these two variants in causing familial phenotype.

Finally, inspection of genotypes for the *DNAJA4*:c.161T>C variant revealed that the heterozygous mutation is not unique to stone formers only (Table 3.12). Three individuals including C-13, C-15 and C-26 were identified to be asymptomatic carrier for the mutation. Inspection of biochemical test results across these individuals did not reveal a consistent biochemical abnormality and therefore this variant is most likely not related to the patients phenotype.

Taken together, segregation analysis in this family did not provide conclusive evidence for involvement of shortlisted variants in patients' phenotype.

3.5 Discussion

Here we applied WES to identify aetiological variants predisposing to nephrolithiasis in three unrelated kindreds. By targetting 4,763 exons across the 367 genes we sought to identify a shortlist of variants likely to underlie the renal stone phenotype specific to each pedigree. In order to distinguish benign variants from disease causing mutations across the genes shortlisted for segregation analysis, four criteria were employed: (1) Consistency of segregation among all affected individuals and lack of thereof among healthy relatives; (2) The prediction of functional impact of the variant on homeostatic mechanisms related to kidney function; (3) High level expression of the gene harbouring the variant in the kidney cortex or liver; (4) and previous reports of the gene harbouring the variant in nephrolithiasis.

Segregation results alone were not clearly indicative of an aetiological variant in any of the families studied. In particular presence of asymptomatic carriers in each pedigree complicates molecular diagnosis. In addition, limitations including the lack of available fresh DNA samples and also an up-to-date biochemical test results from the research participants further complicated molecular diagnosis. The DNA samples analysed here were extracted from the blood samples stored for about 20 years and as a result the capture coverage across some important genomic regions was not optimal (Supplementary Table 8.5). This inevitably might result in oversight of causal variants that might have otherwise been identified from fresh samples. A second limitation concerning patients' data was the lack of available biochemical test results for all members of the family that could perhaps hint at the broader implications of homeostatic imbalances beyond the stone phenotype alone. The late onset kidney stones in these families could well be the consequence of a more subtle health condition that due to lack of comprehensive information about the patients health went unnoticed. For example, hypertension and diabetes both predispose to kidney stones[262], but due to incomplete clinical information about all members of the pedigrees, possible implication of these condition on patients' stone phenotype could not be conclusively excluded.

Despite this, recurrent stone phenotype across multiple generations in these pedigrees led us to speculate that a highly penetrant genetic factor(s) may underlie the disease phenotype in these families. Given this primary conjecture, we adopted an index based WES method to analyse the genetics of nephrolithiasis. The gene variants identified here are implicated in a broad range of cellular processes which may be linked to stone predisposition.

3.5.1 Clinical correlates of variants in Family A

All variants shortlisted for follow-up in this family consistently segregated among the stone formers (Table 3.6). Four rare variants in *PLA2R1* ($NM-001007267 : exon11 : c.A1814C$), *MAP3K* ($NM-005923 : exon28 : c.G3943A$), *SLC25A25* ($NM-001006641 : exon8 : c.G1047C$) and *PKP4* ($NM-001304969.1 : exon13 : c.C2120T$) were also present among non-stone formers and therefore were unlikely to implicate the causal variant. The *PLA2R1* gene encodes both transmembrane and soluble form of the phospholipase A2 receptor (PAL2R) which regulates inhibition of the PAL2 signalling pathway by clearing phospholipase A2. Polymorphisms of this gene have been suggested to associate with idiopathic membranous nephropathies[263]. *PLA2R1* is highly expressed in the kidney cortex. Six of the 8 non-stone formers were identified as heterozygous for this variant and reported to remain disease free in the recent follow-up.

The *MAP3K5* gene code for the Mitogen-Activated Protein Kinase (MAPK) that involves in phosphorylation and activation of downstream kinases in signalling pathways. Given that the gene is not highly expressed in the kidney cortex or liver and also the fact

that three individuals among non-stone formers are heterozygous for this variant, it is very unlikely that the variant is causal in this pedigree.

The *SLC25A25* encodes calcium-binding mitochondrial carriers that are involved in the transport of metabolites, nucleotides and cofactors from the cytoplasm to matrix through mitochondrial inner membrane. These carriers function as magnesium (Mg^{2+}) facilitated ATP/Pi transporters and play a significant role in the regulation of the net ATP efflux to or from mitochondria. The *SLC25A25* is highly expressed in the liver and has a moderate expression in the kidney cortex. Seven non-stone formers were identified to be heterozygous at this site and therefore represented the highest enrichment among the non-stone formers.

The *PKP4* encodes an Armadillo-like protein that belongs to the p120(ctn)/plakophilin subfamily and is involved in gulating junctional plaque organization and cadherin function. Homozygous mutations of *PKP4* have been reported to underlie familial arrhythmogenic cardiomyopathies in both ventricles[264]. Six individuals among the non-stone formers were identified to be heterozygous for the *PKP4* ($NM-001304969.1:exon13:c.C2120T$) variant.

A thorough investigation of patients' clinical history in family A revealed that stone formers were also presented with a range of cardiovascular complications and were on medication that can potentially predispose to kidney stones for several years. As an instance, furosemide is a diuretic drug prescribed to treat fluid build up in the body due to numerous conditions including hypertension or congestive cardiac complications. Individual FM A-30 was on furesemide for years prior to developing the kidney stone and other stone formers in family A were also presented with high blood pressure and a range of cardiovascular complications. It is likely that the late onset kidney stone phenotype in this family is a consequence of drug side effects prescribed for the cardiovascular condition. This hypothesis is further supported by the fact that the *NEBL* variant consistently segregates among all stone formers and one asymptomatic carrier with hyperphosphaturia condition (FM A-27). *NEBL* mutations have been identified to underlie numerous cardiomyopathies[265] and therefore the impact of the identified variant on the cardiovascular phenotype in this family and thereby indirect implication for nephrolithiasis merits further investigation.

Both the *VPS16* ($NM-080413:exon18:c.T1750C:p.W584R$) and *HAVCR1* ($NM012206:exon8:c.A1050G:p.A350A$) variants consistently segregate with stone phenotype among all the stone formers and are only present in one non-stone former. The VPS16 is involved in intracellular trafficking of molecules and plays an important role in segregation of vesicles into distinct organelles. Homozygous mutations of the *VPS16* have been reported in the context of adolescent-onset primary dystonia.

The *HAVCR1* gene encodes a cell surface glycoprotein belonging to the T cell transmembrane, immunoglobin and mucin gene family. The protein product of this gene known as KIM-1 is localised at a very high concentration on the apical membrane of proximal tubules and is found to be markedly upregulated following a proximal tubular injury from renal ischaemia and other damaging insults[266].

Both the *VPS16* and *HAVCR1* are highly expressed in the kidney cortex and liver. Given the significant role of these genes in molecular trafficking, especially across the apical membrane of proximal tubules, collective burden of these two mutations in predisposition to kidney stone merits further investigation.

During analysis of this family, it came to our attention that stone phenotype segregates with exactly same variant ($SLC25A25{:}NM-001006641:exon8:c.G1047C$) in a separate Italian family (Supplementary Figure 8.4). As described earlier *SLC25A25* regulates the flux of ATP in and out of mitochondria in response to cytosolic Ca^{2+} concentration. *In vitro* functional analysis have shown that inhibition of mitochondrial ATPase promotes

the influx of ATP-Mg^{2+} to mitochondria, and when cytosolic levels fall below the natural threshold (∼1mM), ATP-Mg^{2+} efflux from mitochondria to cytosol prevail. Under normal conditions, the net changes in ATP-Mg^{2+} transport reaches an equilibrium and will be maintained in a near-steady state[267]. The renal tubules have a heavy consumption of ATP and they are extremely vulnerable to mitochondrial dysfunction. Glomerular sodium-potassium ATPase energizes reabsorption of Na^+ and facilitate active transport of other ion molecules and uncharged solutes in kidneys. This requires a huge amount of energy, and in fact, daily ATP consumption of human kidneys is estimated to be around 2Kg/day[268]. The relationship between ATP depletion and predisposition to kidney stone is not clear, but it has been shown that deficient ATP supply in renal tubular cells leads to dramatic changes in cell permeability and cell adhesion molecules (CAMs) organisation[269,270]. In particular, ATP depletion results in mis-localisation of tubular Na^+-K^+ ATPase from basolateral membrane to cytosol or even apical surface of the cell. We speculated that loss of tubulous cells polarity in response to impaired ATP trafficking may impair the natural urine concentration and composition and promote the formation of renal stones.

Given the significant role of *SLC25A25* in ATP trafficking, we sought to investigate the functional impact of this mutation in collaboration with Dr Kunji's group at MRC mitochondrial biology unit, Cambridge University. To this end, the disease variant was expressed, solubilised and purified from *S. cerevisiae* mitochondria. Results from this experiment carried out by Fiona Fitzpatrick was indicative of a lower and less stable protein yield from the mutant strains (Supplementary Figures 8.2 & 8.3a). Preliminary results of the transport kinetics of the mutant SLC25A25 revealed that the protein is active, but the exact differences between the rates of wild-type and mutant protein could not be established (Supplementary Figure 8.3b). Taken together, considering that majority of non-stone formers were also identified to be heterozygous for this variant, it is possible that this mutation alone is not sufficient for the full penetrance of the disease and probably additional mutations play a role in predisposition to stone phenotype in this family.

3.5.2 Clinical correlates of variants in Family B

We identified a rare variant in *ADCY10* ($NM-001167749 : exon26 : c.C3599T$) that consistently segregates among all stone formers and also present in 12 out of 22 non-stone formers in Family B. One non-stone former with *ADCY10* variant (individual FM B-27) is also presented with hyperphosphaturia. The *ADCY10* gene encodes a soluble form of adenylyl cyclase that is regulated by bicarbonate and is insensitive to G-protein or forskolin regulation. Mutations of this gene have been identified in absorptive hypercalciuria[271]. Given the low pathogenicity of the identified *ADCY10* variant, the deleterious contribution of this variant as the monogenic cause of the stone phenotype in this family is unsubstantiated. Considering that stone diathesis in this family is associated with hypercalciuria and hypervitaminosis D, the possible role of this variant in pseudohypoparathyroidism merits specific attention. In this condition, parathyroid hormone (PTH) that is responsible for homeostasis of calcium, phosphorus and vitamin D is produced at the normal level, but the body is unable to respond to PTH. It is estimated that hyperparathyroidism accounts for about 3% of renal stone incidence[272]. Given the implications of *ADCY10* in pseudohypoparathyroidism, the role of the identified variant in predisposition to renal stones merits further investigation.

The remaining variants considered in segregation analysis do not show a consistent pattern of segregation across the stone formers and thus their possible implication in the stone phenotype could not be established.

3.5.3 Clinical correlates of variants in Family C

Although idiopathic Hypercalciuria (IH) is primarily a polygenic condition with significant environmental contribution [273], since transmission of the stone phenotype in this family appeared to follow a Mendelian X-linked pattern, we sought to investigate whether a single penetrant variant can be identified through WES analysis. As noted earlier none of the variants considered for segregation analysis in this family provides compelling evidence for being causal. Analysis of variants in this family, although did not reveal any insight regarding kidney stone, hinted a significant technical limitation of WES for analysing variants across the genes with homologous counterparts on the pseudo-autosomal regions.

3.6 Conclusion

Collectively, results identified here reflect the complexity of identifying causal variants in the context of kidney stones. Given the scope of homeostatic perturbation that can lead to the kidney stones, it would be more productive to focus on urinary phenotype rather than stone phenotype alone. This requires careful phenotyping of patients and related individual family members, a task that we were not able to accomplish due to lack of access to patients and their families.

Secondly, the majority of variants associated with metabolic perturbation have a low effect size and contribute minimally to the overall presentation of the phenotype [274]. Here we applied WES analysis to identify rare to low-frequency variants (MAF$\leq 2\%$) that predispose to supposedly monogenic nephrolithiasis. Given incomplete penetrance of renal stones among adult patients, modifier role of common variants in disease expression should not be neglected. Examples of reduced penetrance in renal stone are not unprecedented. In a WES analysis of 80 healthy individuals for identification of rare variants accounting for perturbed metabolite levels [275], one individual with a pathogenic mutation in the xanthine dehydrogenase gene (*XDH*:p.R1296W) had a normal hypoxanthine and urate levels reflecting the impact of modifier genes underlying reduced penetrance in this individual.

More importantly, as noted in the introduction section of this chapter, the kidney stones are the outcome of perturbed metabolism in diverse homeostatic pathways. Without concentrating on the metabolic profile of patients, identification of causal variants underlying kidney stone proves difficult if not impossible. Effectively there might be mutant individuals with perturbed metabolism in the homeostatic pathways that haven't undergone to develop a kidney stone yet. By focusing on the stone phenotype alone, we run the risk of overlooking the heritable factors underlying these metabolic imbalances that have phenotypic presentation (*i.e.* renal stone) among some family members. Conducting such analysis requires longitudinal measurements of urinary metabolites in these pedigrees, which due to lack of access to all patients across multiple generations, proved impossible. Should such data become available, the age of patients at the time of measurement should also be included as a covariate to account for the impact of variation in metabolite levels during life. Given the significance of common and low-frequency variants in the human metabolite profile [276,277,274], application of polygenic transmission disequilibrium test [278] for identification of causal variants as a future direction for this work merits special attention. Investigation of causal variants through this method requires a larger sample size that includes the WES data from not only patients but also their parents and unaffected siblings.

Some of the incidental findings revealed here might have clinical significance for the relatives of cases analysed. As an instance, the *PKP4*:c.2120C>T variant identified in Family A might be related to the cardiovascular phenotype in this family. Considering that a related gene (*PKP2*) is enlisted by the ACMG for incidental finding reporting in

clinical exome analyses, the potential implication of this variant for individuals of this family merits specific attention.

Taken together our WES analysis in adult nephrolithiasis patients revealed a broad range of cellular processes which may be linked to stone predisposition. These were not predictable from the familial clinical phenotypes.

Chapter 4

Diagnostic Application of NGS in a Patient with Congenital Skeletal anomaly and Dysplastic Features

4.1 Introduction

Children with congenital anomalies comprise a substantial portion of cases referred for genetic evaluation [279]. Rare disorders that manifest in childhood impose a significant burden on the patients, their families and the health system [280]. It is estimated that 50 to 80% of resources in paediatric inpatient services is used by patients with a pre-existing genetic condition [281].

It comes as no surprise that identification of the underlying genetic cause is key to the optimal management of these patients. Unfortunately, however, genetic diagnosis in rare disorders is not a straightforward task especially when the phenotypic presentation of the patient does not clearly implicate a known gene (or a set of known genes) or when the differential diagnosis suggests a range of rare congenital disorders.

Chromosomal microarray (CMA) that is proposed as the first-tier diagnostic test for patients with congenital malformations and developmental disorders, is only able to resolve the molecular diagnosis in 15-20% of cases and the genetic aetiology in 80-85% of cases with multiple congenital anomalies (MCA), developmental delay/intellectual disability (DD/ID) and autism spectrum (AS) disorders remain unexplained [282]. Whole exome sequencing and gene-panels have enabled physicians to complement CMA with targeted testing to achieve a greater diagnostic yield. Successful identification of the genetic cause in hypothesis-driven approaches like targeted gene-panels relies extensively on the accurate phenotyping and diagnosis and also efficient capture of all disease-associated loci. This approach often has a low diagnostic yield in the context of multiple congenital anomalies where distinctive diagnostic features and minimal locus heterogeneity is lacking [283,91,284].

The recent progress in cost and reliability of WES enabled widespread adoption of the technology in clinical genetics for investigation of coding variants that underlie rare disorders [285,286]. Limitations of WES in capturing smaller copy number variations (CNVs less than 100kb) and complex structural rearrangement has resulted in a modest diagnostic yield (∼25%) for congenital anomalies [91,284]. Given the impact of CNVs and structural variants in congenital malformations, more thorough testing is required.

Whole genome sequencing (WGS) enables identification of all types of genetic variation in a single test. Several studies demonstrated that WGS achieves a higher diagnostic yield in developmental disorders [287,288] and congenital anomalies [289,290]. Given the significance of SVs in rare congenital anomalies and low diagnostic yield of targeted approach in

resolving molecular diagnosis, WGS offers an optimal solution for resolving molecular diagnosis in rare conditions.

The present chapter discusses the application of WES and WGS sequencing in resolving the molecular diagnosis for a patient with MCA referred for NGS analysis from clinical collaborators.

4.1.1 Structural variants in rare disorders

Structural variations (SVs), reviewed in Section 1.2.3, are an important category of variants in the human genome and are identified to underlie developmental disorders. While SVs occurs less frequently compared to smaller-scale genetic variations, they usually have a greater functional impact due to their relatively large size (>50 bp).

These structural variants were traditionally studied through karyotyping methods such as chromosomal R-banding and G-banding which were arduous and had low resolution. The advent of fluorescence *in-situ* hybridization (FISH)[291,292] in the late 20th century resulted in a significant progress in resolution and throughput of SV discovery and paved the way for multilocus imaging of chromosomes. Conventional FISH probes developed for classical genetic disorders such as Smith-Magenis syndrome (SMS; #OMIM:182290) where deletion of chromosome 17p11.2 results in the disorder[293] or Charcot-Marie-Tooth disease, type 1A (CMT1A, #OMIM 118220) where duplication of chromosome 17p12 underlies the genetic condition[294], enabled rapid detection of carriers and resulted in marked progress in cytogenetic diagnosis. The principles of FISH technique relies on fluorescent labelling of metaphase chromosomes which requires cells to be alive and dividing. This limitation hampers the application of FISH on postmortem tissues or paraffin fixed samples from solid tumours[295]. Furthermore, identification SVs using FISH requires a pre-existing hypothesis as to what region of chromosome might be affected so that appropriate probes can be designed and applied. These major limitations necessitated a more high throughput method to be employed in clinical diagnosis[296,297].

The development of array comparative genomic hybridisation (CGH) enabled genome-wide screening of chromosomal imbalances and contributed significantly to the diagnosis of segmental aneuploidies[298,299]. Contrary to FISH, array CGH does not require prior knowledge about the chromosomal imbalance that is involved and can be applied to non-metaphasic chromosomes, but it is ill-suited for identification of balanced translocations and inversions. The application of array CGH, therefore, is limited in the diagnosis of SVs that do not involve genomic imbalances[300].

Structural variants that involve loss or gain of genomic material underlie an important set of genetic disorders[301]. To date, 42,040 SVs are catalogued in DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources[302,303]) to be involved in 112,767 phenotypic complications. Since benign SVs account for $\sim 1.2\%$ of variation among different populations[32], the database of Genomic Variants archive (DGVa) has been developed to catalogue non-pathogenic SVs in the human genome. To support the clinical interpretation of pathogenic SVs, the Database of Genomic Structural Variation (dbVar) has been developed to catalogue clinically relevant SVs[304]. It is estimated that unbalanced translocations comprise 99.8% of pathogenic SVs in dbVar[305]. Balanced translations are also frequently observed in pathological conditions[306,307,308]. The recent advances in genome sequencing and NGS technology enabled rapid and precise detection of SVs in the human genome.

4.1.2 The Genetics of Skeletal Dysplasia

Distributions in normal skeletal growth and development result in a heterogeneous category of congenital anomalies known as skeletal dysplasia (SD). Skeletal dysplasias usually arise as a result of pathogenic mutations in genes involved in bone and cartilage metabolism [309] and have a prevalence of 1 in every 5,000 births in the general population [310]. To date, 40 groups of genetic skeletal diseases including 456 unique skeletal anomaly have been identified for which the underlying genetic cause in only 316 conditions has been described [309].

The severity of SDs ranges from embryologically lethal to those with minimum morbidity and the clinical diagnosis of the specific disease entity is usually challenging. Despite the framework proposed by the International Skeletal Dysplasia Society to classify SDs based on clinical and radiological patterns [311,312], due to limited experience of individual clinicians with rare forms of SDs classification of the skeletal anomaly into a specific syndrome is difficult. Moreover, the extensive genetic heterogeneity underlying SDs renders the molecular diagnosis challenging. In that sense, mutations in different genes result in similar phenotypes or conversely the phenotypic heterogeneity arising from mutations of the same gene complicate diagnosis. In some SDs, characteristic skeletal manifestations are age-dependent, and they tend to disappear after maturation, or they are simply not present at the time of clinical evaluation. Finally, the phenotypic spectrum of the overwhelming majority of SD entities are not yet thoroughly established and disease variation (specifically in different ethnicities) complicate the clinical diagnosis [309,313].

Given the complexities involved in the clinical diagnosis of specific SD entities based on patients' skeletal manifestations, there has been a recent shift toward 'molecular ontology' for SD classification [313]. Considering that mutations in 226 genes have been already identified to underlie 316 SD entities, WES and WGS offers a cost-effective solution for the screening of a large spectrum of variants (such as SNVs, INDELs and large SVs) that may underlie the patient skeletal phenotype irrespective of clinical diagnosis [314].

4.1.3 Case History

The proband is a 18-year-old male patient from the Channel Islands. He presented with multiple skeletal dysmorphisms and a *de novo* balanced translocation at (8,10)(q22.1; q26.3) identified by karyotyping in the first month after birth in 1999. Array-CGH did not reveal any cryptic chromosomal imbalance. The patient is intellectually normal, and his height is within the normal range. The patient underwent several surgeries for talipes (HP:0001883) and bilateral radioulnar synostosis (HP:0002947). Investigation of the patient's family history showed two siblings with Bardet-Biedel syndrome (HZ *BBS1* mutation) at *BBS1*: c.1169T> G (p.Met390Arg). Inheritance appears to be either *de novo* dominant or recessive. The proband's parents are unaffected and reported as unrelated (Figure 4.1).

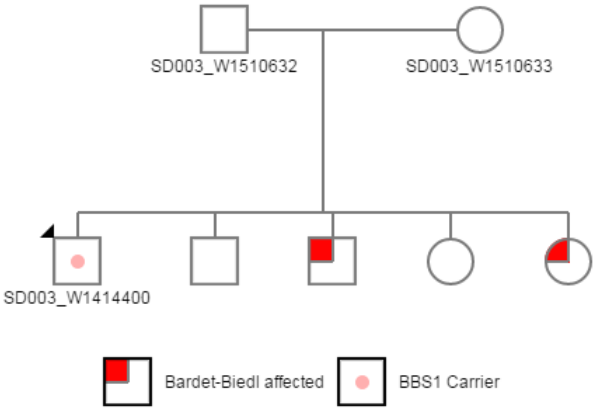


Figure 4.1: Pedigree showing inheritance of congenital arthrogryposis in Family SD003. Proband is identified by arrow.

Additional skeletal and non-skeletal features of the patient are detailed in Table 4.1.

Phenotype Category	Feature	HPO id.
Skeletal		
	Small skull vault but the patient is not microcephalic	-
	Extensive cervical fusion with restricted neck movement.	HP:0002949
	Lunotriquetral joint fusion (Joint contracture of hand)	HP:0009473
	Platyspondyly (flattened vertebral bodies throughout the axial skeleton)	HP:0000926
	Moderate kyphosis	HP:0002808
	Hypoplastic glenoid fossa (shoulder joint sockets are slightly small with unusual clavicles like handle bar)	HP:0006633
	Thin upper chest	-
	Short broad femoral neck (coxa vara)	HP:0002812
Non-skeletal		
	Pelviureteric junction (PUJ) obstruction	HP:0002812
	Cryptorchidism (Undescended testes)	HP:0000028
	Micrognathia	HP:0000347
	Marked webbing of the neck	HP:0000465
	Single palmar crease	HP:0010489

Table 4.1: Skeletal and non-skeletal features of patient SD003.

Figure 4.2: Radiograph depicting main skeletal features including: **(a)** Extensive cervical fusion (HP:0002949); **(b)** Lunotriquetral joint fusion (HP:0009473); **(c)** Thin upper chest; **(d)** Very narrow pelvis. **(e)** radioulnar synostosis (HP:0002947) & **(f)** Talipes (HP:0001883).(*Radiographs courtesy of Dr Amanda Collins, University Hospital Southampton*)

4.1.4 Overview of the Analysis

This chapter describes the application of WES followed by WGS sequencing for resolving molecular diagnosis in a patient with multiple congenital anomalies. Given the broad spectrum of pathogenic variants that underlie rare disorders (reviewed in Section 1.2), a two-stage study design was used for analysing the patient.

In the first stage, in order to investigate whether it is possible to identify a single penetrant SNV that underlies the patient's condition, gDNA was subjected to detailed, in-depth WES analysis and clinical correlate of pathogenic variants identified through WES analysis was explored in the context of the disease.

Since a multidisciplinary team meeting at the time gave the opinion that none of the putative variants identified through WES analysis was likely causal, WGS was employed in the second stage to ascertain potential role of structural variants (SVs) in the patient's phenotype. Using WGS, the precise breakpoints in relation to the reported *de novo* balanced translocation was identified, and the clinical relevance of additional SVs called in the patient's genome evaluated.

4.2 Methods

In order to identify the aetiology of present malformation, WES followed by WGS was carried out. *In vitro* sample processing involved isolation of gDNA from whole-blood by the salting-out method. Downstream sample processing and sequencing were carried out by an external service provider as described below:

4.2.1 Whole Exome Analyses

Exome sequencing was carried out at the Wellcome Trust Centre for Human Genetics (WTCHG). In brief, ultrasonication was applied to shear gDNA into smaller fragments and size-selection was performed to give a mean fragment size of ∼200 bp. Whole-exome enrichment was performed using the SureSelect Human All Exon V5 kit (Agilent, Santa Clara, CA, USA) and sequencing carried out on the HiSeq 2500 platform (Illumina, San Diego, CA, USA) using PE method (Read length ∼ 100bp) (see Section 2.2.3 for further details).

In silico data processing

Raw data analysis was performed using the WES pipeline with default parameters as described in Chapter 2. The quality of raw sequence data was inspected in FastQC v0.11.5, and reads were aligned to the reference genome GRCh37 (hg19) using Novoalign MPI v3.02.03. Aligned reads were subjected to standard quality check using Picard v2.8.3 in order to flag and remove duplicate reads. Variant sites were called using the SAMtools v1.3.2 mpileup command.

Coverage QC & *post-hoc* sample tracking

Following alignment, depth of coverage (DOC) analysis was performed in BEDTools v2.21. Mean DOC and percentage of coverage at 1, 5, 10 and 20X across the whole target region applied as QC-metric for inspecting the quality of coverage across exonic regions. Furthermore, the quality of the alignment procedure was ensured through inspection of on-target/off-target alignment proportions. To confirm gender, the X-chromosome and autosomal heterozygosity were calculated following genotype calling. An aliquot of DNA

dispatched for contemporaneous genotyping across a panel of 24 selected SNPs and consistency of genotype calls with the exome data ensured as a criterion for *post-hoc* tracking of the sample. VerifyBamID was utilized in order to check for contamination with exogenous DNA, and a FREEMIX score of less than 2% was considered acceptable.

Annotation of called variants

Called variants were annotated using ANNOVAR (2015Dec14 release). All variant coordinates were defined according to the GRCh37 genome build (hg19), and dbSNP build 142 was used for annotation of known polymorphisms. Filter-based annotation was carried out using alternative allele frequencies in the 1000 Genomes Project dataset, the NHLBI Exome Sequencing Project (ESP) and in-house allele frequencies (AF) among the ∼460 WES analysed patients in Soton Exome Database (SED). Gene-based annotation was carried out using RefSeq transcript database for protein-altering variants and prediction of deleteriousness for non-synonymous variants was inspected using PolyPhen-2 and SIFT scores. Moreover, GERP++ and Phylop scores were used to measure conservation of the variant site. The MaxEntScan score was computed for variant within 10bp of intron-exon boundaries and ΔMaxEnt score of ≥ 3 was considered as indicative of a variant with potential splicing impact.

Filtering and variant Analyses

A tiered filtering approach was adopted to prioritise variants as potentially causal for further analysis. A collective set of 67 genes previously reported to be involved in the context of present phenotype, prepared in two tiers as follows:

- **Tier 1:** A primary list of 55 candidate genes prepared by literature search nucleating from the Human Gene Mutation Database (HGMD professional v.2015.3) using the terms 'Arthrogryposis', 'KBG syndrome' and 'multiple pterygium syndrome' as the search keywords.

- **Tier 2:** A complementary list of 13 genes (not covered in Tier-1) was recovered from extensive search of OMIM database (December 2015 update).

Variants in these 68 candidate genes were subsequently subjected to deductive filtering using the following criteria:

1. Synonymous and non-frameshift indels were excluded. Synonymous variants within 10bp of intron–exon boundaries were intentionally retained for investigation of splicing.

2. Variants with MAF ≥ 0.01 in the 1000 Genomes and ESP database were excluded to focus on only rare variants.

3. Alternative allele genotype frequencies in the Southampton Exome Database (SED) was used to exclude variants with ≥ 10 frequency among the ∼250 in-house exome samples in regional patients without skeletal phenotype.

The 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.

4.2.2 Whole Genome Analyses

Whole genome sequencing was carried out by Novogene (Beijing Novogene Bioinformatics Technology Co., Ltd, China). *In vitro* sample processing involved library generation using DNA HT Sample preparation Kit (Illumina, San Diego, CA, USA) and DNA fragmentation by ultrasonication. Fragments with the mean size of ∼350bp were then 3′ poly-adenylated and ligated with the full-length adapter for Illumina sequencing. Upon cluster generation, libraries were sequenced on Hiseq X platform (Illumina, San Diego, CA, USA) using PE method (Read length ∼ 150bp).

In silico data processing

Raw sequence data was subjected to primary quality control in order to discard low quality reads with adapter contamination. Moreover, read pairs with more than 10% uncertainty in base calling and reads with low-quality bases greater than 50% were discarded. The quality of cleaned raw data was investigated in FASTQC and alignment to the reference genome (hg19, build 37) carried out using BWA MEM (v0.7.12) with default parameters. Following alignment, DOC analyses performed in Qualimap (v2.2.1).

Structural variant analyses & Annotation

In order to use WGS to detect SVs and identify breakpoints related to the balanced (8,10)(q22.1; q26.3) translocation, split-reads and discordant read pairs were investigated in LUMPY (v0.2.11) [193] and SVDetect (v0.8b) [194]. Evidence for potential translocations, inversions and large deletions/insertions were systematically investigated across the whole-genome using sliding-windows used to partition the genome. In order to increase sensitivity for identifying a balanced translocation, an optimal window size of 705 bp was calculated according to the formula prescribed by the software best practice guideline (μ denotes the mean insert size and σ denotes calculated standard deviation from the distribution of normally mapped reads.):

$$window - size = 2\mu + 2\sqrt{2\sigma} \tag{4.1}$$

Discordant paired-end reads that align to different chromosomes, with discordant strand orientation and at a distance deviating from the pre-defined insert length (∼ 324bp) provide strong evidence for translocation [315]. A Poisson model that uses: 1) supporting evidence from the local frequency of anomalously mapped read pairs; 2) size of the SV and; 3) coverage, was applied to calculate a confidence score for each variant. Significant SV predictions supported by at least 10 independent read-pairs were subsequently visually inspected in Integrative Genomics Viewer (IGV, v2.3).

Read pairs that mapped to different chromosomes at each end (chimeric reads) were considered to cover the translocation breakpoint and soft-clipped reads mapped to discontinuous sections of the reference genome were predicted to contain translocation point (Figure 4.3).

Called structural variants were processed with snpEff (v4.3) and Variant Effect Predictor (VEP, release 87) to provide annotation using RefSeq transcript database. Recurrent SVs were determined using DGVa gold standard variants [316] and excluded from downstream analysis. For structural variants that include coding DNA, HGMD (professional v.2015.3) database was queried for associated phenotypes. In addition, snpEff variant effect prediction was used to flag SVs with high impact. Shortlisted variants were cross-referenced with the DECIPHER (v9.15) [302] database to identify possible overlapping patients/phenotypes in the Developmental Disorder Genotype- Phenotype Database

(DDG2P), Locus Specific Mutation Databases (LSDB) and OMIM database. Only variants with pathogenicity description of "definitely pathogenic" were selected. Finally, the dbVar database was queried to confirm the pathogenic classification of the variant.

Ultimately, segregation of variants that were identified to be plausibly causal in the context of patient phenotype was followed up in the trio (parents and the patient) and confirmed by Sanger sequencing. To this end, PCR amplification of each exon was initially carried out using the Failsafe^{TM} PCR system (Epicentre, USA) with Failsafe^{TM} Buffer J according to the manufacturer's protocol. PCR products were then sequenced using the Big-Dye® Terminator Cycle Sequencing Kit (V.1.1) according to the standard protocol (Applied Biosystems, USA). Capillary sequencing was carried out using an ABI 3130x/ Genetic Analyzer (Applied Biosystems, USA) and sequence traces were investigated in Mutation Surveyor software (version 3.1, SoftGenetics, USA). (Validation of variants kindly carried out by Dr David Bunyan at the Wessex Regional Genetics Laboratory - Salisbury NHS Foundation Trust).

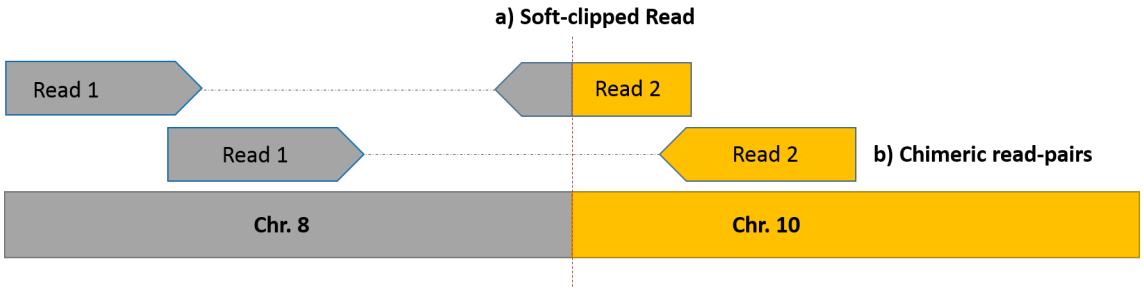


Figure 4.3: Breakpoint analysis using WGS data. **(a)** Soft-clipped read that maps to two different sequences identify the breakpoint. **(b)** Readpairs that map to different chromosomes (chimeric reads) cover the translocation breakpoint.

4.3 Results

WES and WGS were successfully applied to the DNA sample from the male patient with multiple congenital anomalies. NGS data of good quality was then used for variant analysis and SV calling.

4.3.1 WES Results

Quality scores for raw sequence data were concordantly above 30 (phred scale) and there was no evidence of contamination based on VerifyBamID results (Freemix score= 0.0019).

Exonic heterozygosity percentage (∼%61) was within the expected range for non-consanguineous mating. Sample relatedness was investigated using the similarity matrix as per the procedure described in Chapter 2 and no apparent relatedness to any other unrelated sample on same dispatch DNA plate was observed.

92.61% of the coding region was captured successfully across the 68 target genes. Following sequencing and alignment, ≥68.23% of the targeted region was covered to a depth of at least 20X (Table 4.2).

					% Covered to median depth				
Gene	**Location**	**# CE**	**CD size**	**% accessible target**	**1x**	**5x**	**10x**	**20x**	**Source**
ABCA3	16p13.3	30	49793	80.42	92.39	88.050	86.11	78.42	OMIM
ADCY6	12q13.12	21	14844	97.97	98.33	92.720	91.76	90.49	HGMD
ADGRG6	6q24.2	25	141186	100.00	60.13	58.290	56.74	54.10	HGMD
ANKRD11	16q24.3	11	48542	78.76	86.77	75.880	73.39	68.52	HGMD
B3GAT3	11q12.3	5	6247	100.00	97.58	83.800	75.69	66.30	HGMD
CHRNA1	2q31.1	10	16271	70.17	86.67	80.850	75.35	68.42	OMIM
CHRNB1	17p13.1	11	11596	95.94	85.89	75.900	70.51	64.91	HGMD
CHRND	2q37.1	12	9097	76.22	78.22	66.950	61.87	53.70	OMIM
CHRNG	2q37.1	12	5969	100.00	85.37	79.030	77.43	71.32	HGMD
CHST3	10q22.1	2	2629	95.22	81.77	43.060	37.74	31.13	HGMD
CNTNAP1	17q21.2	24	16081	95.27	100.00	99.710	97.70	93.71	HGMD
COL11A1	1p21.1	67	230160	100.00	86.19	81.900	79.46	72.44	HGMD
COL2A1	12q13.11	54	30915	100.00	96.06	93.020	91.31	85.88	HGMD
DNASE1L2	16p13.3	6	1602	100.00	100.00	81.080	76.27	69.25	OMIM
E4F1	16p13.3	12	11959	100.00	100.00	98.900	96.07	83.51	OMIM
ECEL1	2q37.1	17	6501	96.36	97.68	87.280	79.98	68.12	HGMD
ECI1	16p13.3	7	11588	100.00	100.00	89.380	83.68	64.15	OMIM
ERBB3	12q13.2	28	21,755	100.00	80.15	69.320	68.06	63.59	HGMD
ERCC5	13q33.1	15	29637	100.00	98.86	89.570	87.90	81.20	HGMD
ERCC6	10q11.23	20	74150	96.95	48.02	38.590	37.63	36.14	HGMD
FGF9	13q12.11	3	29523	95.63	39.63	22.610	20.57	18.56	OMIM
FGFR2	10q26.13	17	113,961	86.00	84.99	74.330	73.57	71.55	OMIM
FGFR3	4p16.3	15	13,328	100.00	91.23	77.060	72.51	62.19	HGMD
FLNA	Xq28	43	22,316	100.00	99.53	98.650	97.82	82.48	HGMD
FLNB	3p14.3	45	162198	100.00	92.32	85.240	84.36	81.19	OMIM
GBA	1q22	11	6,118	88.26	78.99	75.210	74.71	70.04	HGMD
GBE1	3p12.2	16	271,111	82.87	92.91	88.770	88.32	84.33	HGMD
GDF5	20q11.22	2	4,020	100.00	99.07	83.260	81.79	78.22	OMIM
GLE1	9q34.11	16	36,365	100.00	92.91	88.770	88.32	84.33	HGMD
IRF6	1q32.2	7	12,994	91.61	93.66	80.310	77.24	68.27	HGMD
LBR	1q42.12	13	20,773	78.91	72.00	55.120	53.42	45.31	HGMD
MAB21L2	4q31.3	1	1,080	52.44	68.02	51.020	47.29	44.28	HGMD
MTM1	Xq28	14	78,981	98.08	62.17	57.400	53.49	40.32	HGMD
MYBPC1	12q23.2	30	89,325	100.00	100.00	98.200	97.25	91.21	HGMD
MYH3	17p13.1	39	26,412	95.40	98.19	98.190	97.67	92.56	HGMD
MYH8	17p13.1	38	29,774	89.28	98.74	95.480	94.52	90.25	HGMD
NALCN	13q33.1	43	343,832	97.83	95.96	87.420	86.68	83.42	HGMD
NEB	2q23.3	148	247,397	82.79	90.96	87.890	87.35	85.77	HGMD
NOG	17q22	1	699	100.00	83.05	49.420	44.25	40.38	OMIM
PGM3	6q14.1	13	22,035	53.58	71.08	55.650	50.44	44.88	HGMD
PI4KA	22q11.21	55	150,679	95.49	95.17	89.760	85.32	77.73	HGMD
PIEZO2	18p11.21	52	477,064	99.37	79.45	73.180	71.97	68.55	HGMD
PIP5K1C	19p13.3	18	67,224	45.00	88.36	54.470	52.91	49.97	HGMD
POR	7q11.23	15	32,489	100.00	93.30	86.030	81.63	77.08	OMIM
PTH1R	3p21.31	14	20,097	91.94	97.83	95.080	91.62	76.81	HGMD
RAPSN	11p11.2	8	10,991	100.00	100.00	93.790	83.10	80.90	HGMD
RBBP8	18q11.2	18	89,389	87.04	96.16	92.550	91.66	89.87	HGMD
RIPK4	21q22.3	8	26,204	100.00	79.04	66.150	64.51	60.00	HGMD
RNPS1	16p13.3	7	10,680	96.53	94.98	86.900	78.16	56.48	OMIM
RYR1	19q13.2	106	153,591	99.82	99.81	98.690	94.97	84.58	HGMD
SLC35A3	1p21.2	8	51,940	91.67	33.44	24.640	23.92	21.28	HGMD
SOX10	22q13.1	3	10,290	93.22	87.39	71.850	59.57	45.40	HGMD
SYNE1	6q25.2	144	505,896	97.93	92.08	88.410	87.56	84.16	HGMD
TGFB3	14q24.3	7	21,707	84.79	68.64	55.860	52.47	47.91	HGMD
TNNI2	11p15.5	6	1,328	91.76	96.76	96.760	96.76	88.66	HGMD
TNNT3	11p15.5	14	15,618	84.81	98.63	88.570	87.87	79.40	HGMD
TPM2	9p13.3	9	7,737	100.00	100.00	100.000	96.11	87.84	HGMD
TRPV4	12q24.11	15	31,176	100.00	96.59	87.350	85.37	82.31	HGMD
TTN	2q31.2	312	277,631	97.96	99.15	98.600	98.46	97.64	HGMD
UBA1	Xp11.3	25	16,127	90.03	92.01	86.060	82.59	66.73	HGMD
UTRN	6q24.2	74	559,169	100.00	92.71	88.990	88.29	84.57	HGMD
VEGFA	6p21.1	8	13,856	100.00	64.17	46.050	41.77	28.83	HGMD
VIPAS39	14q24.3	19	26,487	92.59	91.12	73.090	70.13	66.83	HGMD
VPS33B	15q26.1	23	23,275	100.00	97.70	86.410	81.93	77.42	HGMD
XRCC1	19q13.31	17	32,067	99.43	100.00	99.790	95.68	89.55	HGMD
XRCC6	22q13.2	12	41,811	99.44	99.11	85.340	81.46	69.44	HGMD
ZC4H2	Xq11.2	5	58,595	82.86	46.83	32.410	29.38	22.94	HGMD
ZMPSTE24	1p34.2	10	34,398	100.00	65.05	55.660	54.18	51.92	HGMD

Table 4.2: Percentage coverage across the 68 genes included in the tiered filtering. **#CE:** Number of coding exons acording to RefSeq database; **CD size:** Size of coding region in bp; **% accessible target:** Percentage of the gene targeted by All Exon V5 capture kit.

Upon variant calling and annotation, 189 variants across the 68 candidate genes were identified in the patient sample. After applying filtering steps described in Figure 4.4, eight variants with possible pathological implication were identified (Table 4.3).

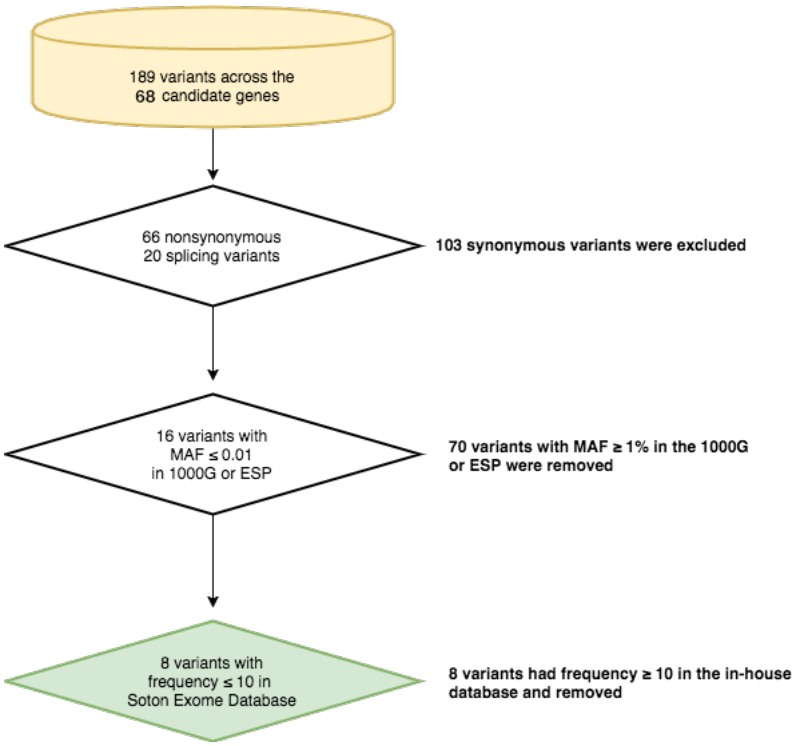


Figure 4.4: Overview of filtering strategy with the number of variants excluded at each step. **1000G:**The 1000 Genomes Project; **ESP:** Exome Server Project.

The clinical significance of candidate variants was investigated in ClinVar. Three variants including *TTN*:p.P23664L, *SYNE1*:p.S3353Y and *ANKRD11*:p.T680S had been previously catalogued in dbSNP with uncertain clinical significance and thus considered unlikely to underlie the condition. Furthermore, MaxEnt analysis of the splice region variant at *ECEL1*:c.1990-10T>G revealed a non-significant ΔMaxEnt score and the variant was subsequently excluded from further follow-up.

Investigation of deleteriousness and conservation scores (Polyphen2-HDVAR and GERP++) for the remaining variants highlighted three possibly damaging variants at *ECEL1*:c.1013T>C, *ECEL1*:155T>C and *ANKRD11*:c.3926C>T (Table 4.3).

ANKRD11 encodes ankyrin repeat-containing cofactors that suppress activation of transcription. Mutations of this gene have been reported to underlie autosomal dominant KBG syndrome (KBGS, OMIM # 148050). KBGS patients typically present with global developmental delay, mental retardation, seizures, short stature, craniofacial and skeletal anomalies and macrodontia[317]. Many characteristics of KBGS patients including intellectual disability, facial dysmorphisms and short stature are inconsistent with the proband's phenotypic features. The *ANKRD11*:c.3926C>T variant has been identified among the European (Non-Finish) subpopulation of the ExAc database with MAF of $1.5e^{-05}$ and it has not been classified as pathogenic in the clinVar database. Given the lack of convincing evidence implicating the *ANKRD11* variant, this variant was deprioritised.

Exome sequencing in proband SD003 identified three heterozygous variants in the *ECEL1* gene that included two non-synonymous variants (*ECEL1*:c.1013T>C, *ECEL1*:155T>C), as well as a splice region T>G transversion (*ECEL1*:c.1990-10T>G) (Table 4.3). Since the ΔMaxEnt score for the splice site variant did not reach the significance threshold of $\geq |3|$ (ΔMaxEnt= 0.22) it considered unlikely to confer clinical

Chr.	g.Position	Gene	Het/Hom	Variant type	Nucleotide	Protein	Polyphen	Gerp++	1000 Genomes	EVS	SED
2	233,349,557	**ECEL1**	HET	ns	c.T1013C	p.M338T	0.901379	3.1	-	-	0,0,0
2	233,351,209	**ECEL1**	HET	ns	c.T155C	p.L52P	0.992739	2.53	-	-	0,1,1
2	233,345,876	**ECEL1**	HET	sp	c.1990-10T>G		MaxEnt= 0.22		-	-	0,0,0
2	179,434,945	**TTN**	HET	ns	c.C70991T	p.P23664L	0.1	4.99	0.0005	0.000242	0,0,0
6	152,686,090	**SYNE1**	HET	ns	c.C10058A	p.S3353Y	0.810343	0.133	0.0027	0.006395	0,6,6
16	89,349,024	**ANKRD11**	HET	ns	c.C3926T	p.T1309M	0.990206	3.74	-	-	0,0,0
16	89,350,911	**ANKRD11**	HET	ns	c.C2039G	p.T680S	0.111165	-4.85	0.01	0.019302	0,10,10
16	2,336,732	**ABCA3**	HET	ns	c.C3241T	p.R1081W	0.999269	4.21	-	0.000116	0,0,0

Table 4.3: Detailed information of eight shortlisted variants identified through WES analysis. g.Position: the genomic position of variant according to the hg19 genome assembly ; 1000 genomes: MAF in the 1000 genomes project, ESP: MAF in NHLBI Exome Sequencing Project; SED: Number of observed genotypes in Soton Exome Database (Frequencies in the SED database is expressed as Hom, Het, Total); ns: Non-synonymous SNV; sp: Splice-site variant; Δ: Maxent splicing score- ΔMaxEnt$\geq |3|$ are considered potentially disrupting to splicing.

impact. The remaining two non-synonymous variants in *ECEL1* were retained for further investigation as they may underlie the patient's phenotype in a compound heterozygous status.

ECEL1 on 2q37.1 encodes a zinc-containing protein that belongs to the M13 family of endopeptidases. This protein plays an important role in the regulation of neuropeptides and peptide hormone activity. Mutations of *ECEL1* have been reported to underlie autosomal recessive distal arthrogryposis, type 5D (DA5D) (OMIM # 615056). DA5D patients present with talipes, hand and wrist joint contracture, extension contracture of lower limbs, micrognathia and short necks. DA5D patients have normal intelligence, and hypoplastic palmar creases have been reported in the context of the disease [318].

Compound heterozygous mutations of *ECEL1* at c.[1470G>A]+[997C>T], p.[Try490*]+[Arg333*] have been identified to underlie DA5D in the affected children of a consanguineous family [319]. The *ECEL1*:c.997C>T variant reported by Dieterich *et al.* maps to exon 5 and is positioned 16 bp (5 amino acids) downstream of the *ECEL1*:c.1013T>C variant identified in the SD003. Interestingly, in both cases patients present with congenital hip dislocation and limited mobility.

Table 4.4 provides a comprehensive summary of major phenotypic complications reported to be associated with mutations of *ECEL1* and *ANKRD11* genes. Given the extensive phenotypic similarity between the patient's phenotype and the reported cases of *ECEL1* mutations, possible implication of the *ECEL1* compound heterozygous mutation could not be excluded, however, based on clinical discretion of genomic medicine multidisciplinary team (GM-MDT), lower priority was assigned to this variant and WGS investigation of reported *de novo* translocation was advised.

				ECEL1 Reported cases											ANKRD11 reported cases					
	phenotypic complications	Proband	ECEL1	Dieterich *et al.*[319] (n= 10)	McMillan *et al.*[318] (n= 9)	Shaheen *et al.*[320] (n= 9)	Patil *et al.*[321] (n= 1)	Barnett *et al.*[322] (n= 2)	Shaaban *et al.*[323] (n= 2)	Hamzeh *et al.*[324] (n= 1)	Stattin *et al.*[325] (n= 1)	Rai *et al.*[326] (n= 1)	Ullmann *et al.*[327] (n= 7)	ANKRD11	Ockeloen *et al.*[328] (n= 19)	Walz *et al.*[329] (n= 11)	Low *et al.*[330] (n= 32)	Goldenberg *et al.*[331] (n= 39)	De Bernardi *et al.*[332] (n= 1)	Miyatake *et al.*[333] (n= 5)
Growth	Developmental delay	-	-	-	-	1/9	+	-	-	+	-	-	-	+	19/19	10/11	-	-	+	3/5
	Intellectual disability	-	-	-	-	-	-	-	-	-	-	-	3/6	+	19/19	10/11	31/32	28/30	+	-
	Delayed bone age	-	-	-	-	-	-	-	-	-	-	-	-	+	3/19	5/5	3/5	8/12	+	3/3
	Short stature	-	+	-	-	4/9	+	-	-	-	+	+	-	+	11/19	8/11	9/32	15/37	+	3/5
Craniofacial anomalies	Microcepahly	-	-	-	-	-	-	-	-	-	-	-	-	+	1/19	-	-	9/33	-	-
	Round/triangular face	-	+	-	6/6	-	-	-	-	-	-	-	-	+	19/19	11/11	+	11/20	-	4/5
	Micrognathia	?	+	-	8/9	1/9	+	2/2	2/2	+	-	-	4/7	+	-	-	-	+	-	-
	Macrodontia and dental abnormalities	-	-	-	-	-	-	-	-	-	-	-	-	-	17/19	11/11	23/32	18/26	-	4/5
	ptosis		+	7/10	8/9	6/9	+	2/2	1/2	+	+	+	5/7	-	-	6/11	-	+/-	-	-
	hypertelorism	?	-	-	-	-	-	-	-	-	-	-	-	+	-	5/11	+	+	+	4/5
	Hairline abnormality	-	-	-	-	-	-	-	-	-	-	-	-	+	5/11	-	+	-	-	4/5
	Eyebrows	?	a	-	a	-	a	a	a	-	-	-	-	b	b	b	b	-	b	b
	Upturned nose	?	+	-	6/6	1/9	-	-	-	-	-	-	-	+	-	-	-	+	-	-
	Prominent nasal bridge	?	-	-	-	-	-	-	-	-	-	-	-	-	19/19	11/11	+	+	-	3/5
	Bulbous Nose	?	+	-	9/9	-	+	2/2	2/2	-	+	+	-	-	-	-	+	-	-	1/5
Skeletal anomalies	Short neck	+	+	10/10	4/7	-	+		2/2	+	-	-	-	-	-	-	+/-	-		-
	Neck contractures	+	+	-	4/4	-	-	2/2	2/2	-	-	-	-	-	-	-	+	-	-	-
	Contractures of shoulders and/or shoulders limited movement	+	+	2/8	6/6	1/9	+	2/2	1/2	+	+	-	4/7	-	-	-	-	-	-	-
	Cervical rib fusion	?	-	-	-	-	-	-	-	-	-	-	-	+	-	2/9	-	-	-	-
	Vertebral body fusion	?	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
	Thoracic kyphosis	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-
	Scoliosis	-	+	7/10	2/9	2/9	+	-	2/2	+	-	+	3/7	-	-	-	-	5/39	-	1/5
	Hyperlordosis	-	+	9/9	-	-	-	-	1/2	-	-	-	-	-	-	-	-	-	-	-
	Contractures of elbows and/or elbows limited movement	+	+	3/7	5/5	1/9	+	2/2	-	+	-	-	3/7	-	-	-	-	-	+	-
	Contractures of wrists	+	+	-	9/9	1/9	-	2/2	2/2	+	+	-	2/7	-	-	-	-	-	-	-
	Camptodactyly	?	+	10/10	9/9	9/9	+	2/2	2/2	+	+	+	7/7	-	-	-	-	2/39	-	-
	Clynodactyly	?	-	-	-	-	-	-	-	-	-	-	-	+	19/19	7/11	+	23/33	-	2/5
	hip dislocation/contractures	+	+	9/9	9/9	6/9	+	2/2	0/2	+	+	-	6/7	-	3/19	-	-	-	-	-
	Contractures of knees	-	+	10/10	8/9	5/9	+	2/2	2/2	+	-	+	7/7	-	-	-	-	-	-	-
	Foot contractures	+	+	-	8/9	-	-	2/2	0/2	-	+	-	6/7	-	-	-	-	-	-	-
	Clubfoot	+	+	9/10	1/9	4/8	+	-	-	+	-	-	-	-	-	-	-	-	-	-
Additional complications	**Skin**																			
	Pterygia of neck	+	+	-	3/8	-	+	2/2	2/2	+	-	+	2/7	-	-	-	-	-	-	-
	Genitourinary																			
	Cryptorchidism	+	-	-	-	-	-	-	-	-	+	-	2/7	+	6/10	5/10	5/16	3/16	-	-
	Sensory impairments																			
	Seizures	-	-	-	-	-	-	-	-	-	-	-	-	+	5/19	1/8	13/32	10/38	-	3/5
	Diminished facial expression	?	+	3/7	1/9	-	-	-	2/2	-	+	-	5/7	-	-	-	-	-	-	-
	hearing loss	-	-	-	-	-	-	-	-	-	-	-	-	+	6/19	-	8/32	11/36	+	3/5
	Heart																			
	Cardiac anomalies	-	-	-	-	-	-	-	-	-	-	-	-	+	3/19	1/8	4/32	10/39	-	-

Table 4.4: Comparison of patients phenotype with reported cases of Distal arthrogryposis type 5D (*ECEL1* mutations) and KBG syndrome (*ANKRD11* mutations).

4.3.2 WGS Results

FastQC analysis of data confirmed the high quality of raw sequence data (Supplementary section 8.3). The mean read-depth across the whole-genome was 34x, and average mapping quality (MQ) was 50. Detailed statistics for coverage analysis is provided in Supplementary section 8.4. In order to increase the sensitivity of SV discovery, called SVs from LUMPY and SVDetect algorithms were merged together. A total of 80,265 SVs including deletions, duplications, inversions and translocations were identified across the whole genome. Only 2,348 SVs were confidently supported by ≥10 independent paired-end (PE) reads. 335 SVs were mapped to contigs outside the primary scaffolding of the hg19 assembly or the mitochondrial genome and therefore excluded from further analysis. In total, 2,013 SVs were identified across chromosomes 1-22 and XY. This number includes common polymorphic markers, private structural differences from the reference genome and potential computational false positive calls at low complexity sequences (Table 4.5).

Chr.	Duplication	Deletion	Inversion	Translocations		Total
				Intra.	Inter.	
1	15	63	5	17	28	128
2	5	60	0	50	21	136
3	8	38	1	12	11	70
4	11	45	0	11	9	76
5	20	46	7	28	10	111
6	19	45	1	10	8	83
7	7	33	1	42	30	113
8	5	34	0	4	5	48
9	6	28	1	22	26	83
10	17	51	0	20	20	108
11	8	36	1	8	8	61
12	10	28	2	12	6	58
13	17	26	2	4	32	81
14	4	19	0	0	7	30
15	6	20	0	30	17	73
16	21	26	2	37	11	97
17	10	38	0	41	40	129
18	5	19	0	8	5	37
19	14	35	0	26	9	84
20	20	32	2	17	81	152
21	3	11	1	4	88	107
22	3	10	0	19	42	74
X	3	11	0	2	8	24
Y	9	17	0	4	20	50
Total	246(**12.2%**)	771(**38.3%**)	26(**1.2%**)	456(**22.6%**)	514(**25.5%**)	**2013**

Table 4.5: Total number of SVs identified by LUMPY and SVDetect across the nuclear genome. Only SVs that are supported by at least ≥10 independent PE reads are considered reliable. The majority of identified SVs belonged to the deletion class. Interchromosomal translocations were more frequent that intrachromosomal translocations and inversions had the least frequency across the genome.

In order to ascertain the breakpoint position related to $t(8,10)$(q22.1; q26.3), previously identified by karyotyping, interchromosomal translocations between chromosomes 8 and 10 were investigated. In total, 11 pairs of reciprocal translocations between chromosome 8 and 10 were identified for which only one pair was supported by ≥10 independent PE reads. This rearrangement disrupts *PLEKHF2* and *MGMT* at 8q22.1 and 10q26.3 and result in a balanced (8,10)(q22.1; q26.3) translocation (Figure 4.5).

PLEKHF2 encodes pleckstrin homology and FYVE domain-containing protein II that regulates receptor trafficking within the cell. It is also suggested that PLEKHF2 enhances cellular sensitivity to TNFα triggered apoptosis[334]. *PLEKHF2* is a small gene with only two coding exons and no Mendelian disorder has been attributed to the mutations of this

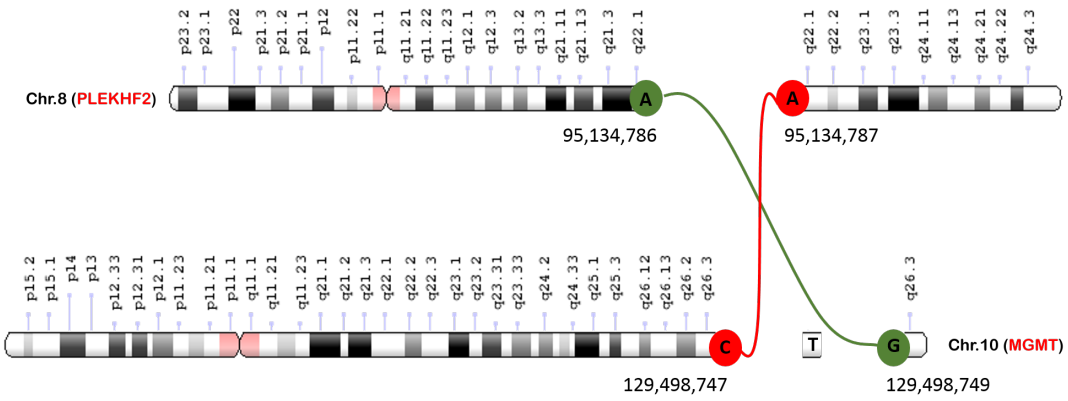


Figure 4.5: Schematic representation of breakpoint for reciprocal translocation identified at t(8,10)(q22.1; q26.3). Corresponding base pair location of breakpoints are given.

gene so far. *MGMT* on 10q26.3 encodes a DNA repair protein that plays an important role in protecting cells against mutagenesis induced by alkylating agents [335]. Methylation of the *MGMT* promoter has been associated with several types of cancers [336]. The fusion gene resulting from this translocation lacks the *MGMT* promoter region and therefore possibly results in reduced activity of the gene. To date, no Mendelian disorder has been attributed to the mutations of this gene.

For the remainder of SVs, the snpEff prediction for putative variant impact was used to prioritise variants. A total of 970 variants were flagged as high impact by snpEff for which only 271 SVs were supported by ≥10 independent reads. Upon filtering variants mapped to the intergenic region of the genome, 222 variants remained. Variants were further excluded if they mapped to non-coding genes (lincRNA, processed pseudogenes, nonsense-mediated decay (NMD) genes) or the immunoglobulin variable region (IgV). Seventeen SVs were recovered that impact protein-coding genes (Table 4.6). Among all shortlisted variants, only the 6Kb deletion on chromosome 4 was identified in the homozygous state. This SV impairs *ZNF718* by deleting exons 2 and 3 (Figure 4.6). Large deletions spanning coding exons of *ZNF718* are classified as pathogenic in the dbVar database and have been reported to underlie congenital anomalies [282]. Furthermore, a 2.02Mb deletions of distal 4p (spanning from chr4:72,447 to 2,094,416) that includes the *ZNF718* gene is classified as "definitely pathogenic" with full penetrance in the DECIPHER (v9.15) database. Although this large deletion encompasses developmentally important genes such as *PIGG* and *IDUA* (enlisted as DD genes) and also haploinsufficient genes including *MAEA* (pLI= 0.84, %HI= 7.12) and *FGFR3* (pLI= 0, %HI= 6.40), similar phenotypic presentations in the DECIPHER patient #341526 that mimic that of patient SD003 renders the investigation of functional impact of *ZNF718* deletion necessary.

A smaller heterozygous deletion of this region (∼298.44 kb) that spans from chr4:37,336 to 335,770 encompassing four protein-coding genes including *ZNF595*, *ZNF718*, *ZNF732* and *ZNF141* (the promoter and first exon only) is also identified as pathogenic in DECIPHER patient #331143. Despite the pathogenic assignment of this heterozygous deletion, no phenotypic information was available for this patient and therefore establishing the correlation between patients' phenotype was not feasible. The fact that this heterozygous CNV has sampling probability >5% (*i.e.* a similar CNV is expected to be identified in greater than 5% of healthy individuals from the Wellcome Trust Case Control Consortium (WTCCC) cohort), undermines the pathogenic involvement of this deletion in the patient's severe dysplastic features.

Chr.	Position	Size	Type	Effect	Zygosity	Gene	PE	SR
1	16,050,023 16,059,762	9,739	Del	EXON_DELETED	Het	CLCNKB	12	19
1	209,761,993 209,762,731	738	Del	EXON_DELETED	Het	TRAF3IP3	11	17
2	178,436,319 178,441,609	5,290	Del	SPLICE_SITE_ACCEPTOR	Het	PRKRA	13	24
2	178,450,414 178,450,965	551	Del	SPLICE_SITE_DONOR	Het	PRKRA	19	10
3	130,044,539 130,087,902	43,363	Del	EXON_DELETED	Het	ALG1L2	10	5
4	**127,126 133,266**	**6,140**	**Del**	**EXON_DELETED**	**Hom**	**ZNF718**	**19**	**9**
6	32,480,809 32,524,614	43,805	Del	EXON_DELETED	Het	HLA-DRB5	14	13
11	62,422,409 62,423,057	648	Del	SPLICE_SITE_ACCEPTOR	Het	SCGB1A1	47	99
11	62,419,150 62,422,219	3,069	Del	SPLICE_SITE_ACCEPTOR	Het	SCGB1A1	58	150
11	99,819,752 99,820,576	824	Inv	FRAME_SHIFT	Het	CNTN5	19	14
13	21,155,151 21,155,692	541	Del	SPLICE_SITE_ACCEPTOR	Het	SKA3	11	13
13	21,155,813 21,157,921	2,108	Del	SPLICE_SITE_DONOR	Het	SKA3	11	15
15	41,573,817 41,577,885		Del	SPLICE_SITE_ACCEPTOR	Het	TYRO3	10	6
19	4,005,350 4,159,440	154,090	Del	EXON_DELETED	Het	PIAS4	13	28
20	45,252,421 45,253,001	580	Del	SPLICE_SITE_DONOR	Het	SLPI	11	23
20	45,253,157 45,253,574	417	Del	SPLICE_SITE_DONOR	Het	SLPI	17	36
20	45,253,733 45,254,458	725	Del	SPLICE_SITE_ACCEPTOR	Het	SLPI	15	31

Table 4.6: Protein-altering SVs supported by >10 independent PE reads (Chr.: chromosome; POS: base-pair position of the SV; Del: Deletion; Inv: Inversion; Het: Heterozygous; Hom: Homozygous; PE: Number of paired-end reads in support of SV; SR: Number of split-reads in support of SV).

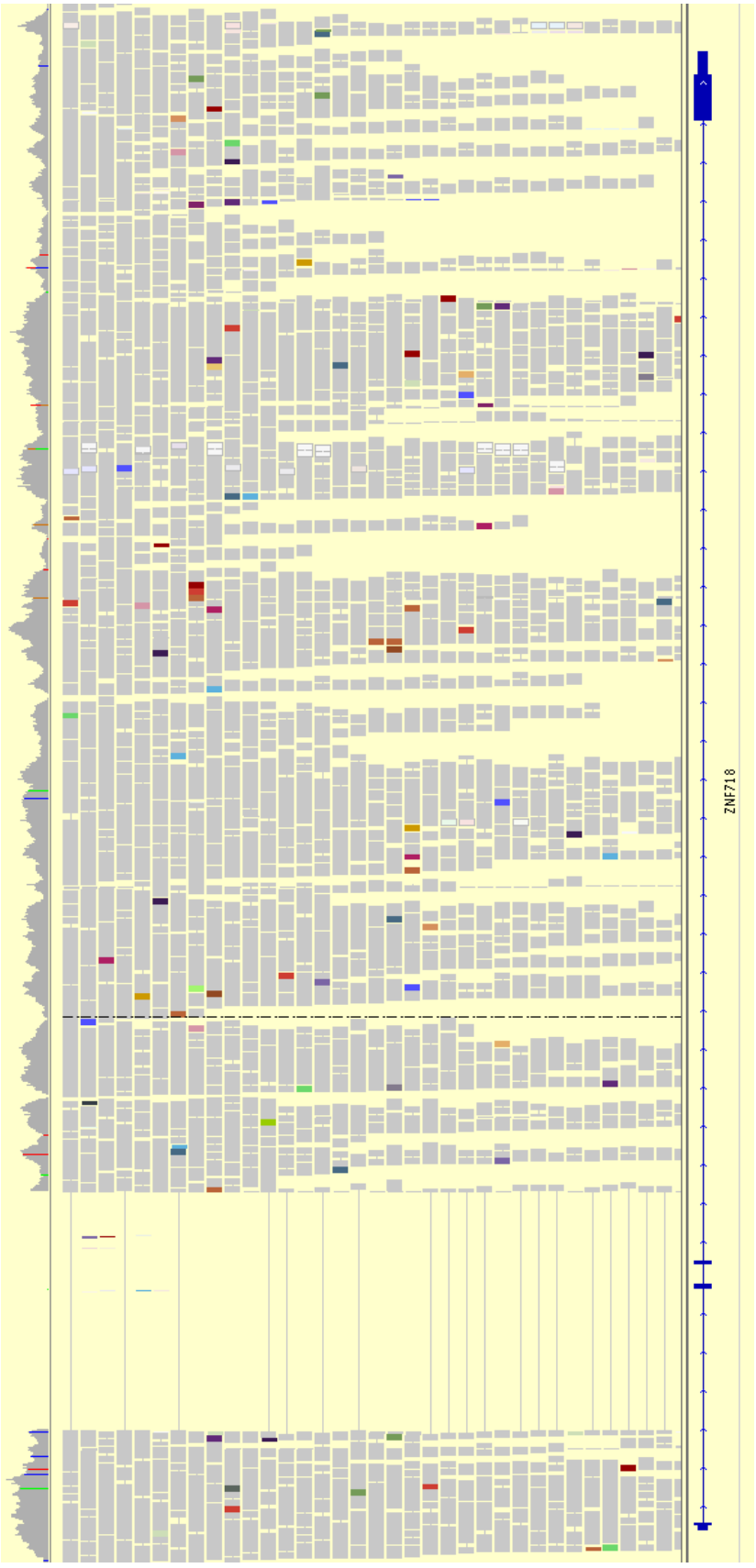


Figure 4.6: Homozygous deletion spanning ~6Kb on chromosome 4. Exons 2 and 3 of *ZNF718* are deleted as a result of this deletion.

4.4 Segregation Analysis

Since variants identified in the *ECEL1* and *ANKRD11* genes were providing the most compelling evidence in favour of pathogenic involvement in the patient's severe condition, segregation of the three variants including *ECEL1*:c.155T>C, *ECEL1*:c.1013T>C and *ANKRD11*:c.3926C>T was followed up in the trio (parents/proband). The raw Sanger sequence traces are provided in the Supplementary Section 8.4.5, 8.4.5 and 8.4.5.

The segregation results (Table 4.7) revealed that the heterozygous *ANKRD11*:c.3926C>T is inherited from the proband's father. Considering that the patient's father is healthy, it is implausible that this variant underlies the patient's phenotype.

On the other hand, segregation analysis for the *ECEL1* variants revealed that the compound heterozygous status in the proband is consistent with maternal transmission of *ECEL1*:c.155T>C and paternal transmission of *ECEL1*:c.1013T>C to the proband. Since the *ECEL1* derived Arthrogryposis 5D (DA5D, OMIM #615065) exhibit an autosomal recessive pattern of inheritance, it is highly likely this the heterozygous compound variants in the *ECEL1* gene underlies the patient's phenotype.

	ECEL1 c.155T>C	**ECEL1 c.1013T>C**	**ANKRD11 c.3926C>T**
Proband (SD003_W1414400)	HET (A:G)	HET (A:G)	HET (G:A)
Mother (SD003_W1510633)	HET (A:G)	HOM (A:A)	HOM (G:G)
Father (SD003_W1510632)	HOM (A:A)	HET (A:G)	HET (G:A)

Table 4.7: Segregation result for the candidate variants tracked through Sanger sequencing in the trio (Parents and the proband).

4.5 Discussion

In this study, we applied WES in combination with WGS to identify the molecular basis of severe congenital anomaly presented by the patient. Genetic diagnosis in multiple congenital anomalies is difficult since many features are unspecific and differential diagnosis includes a range of disorders. In relation to patient SD003, the radiographic findings are highly unusual and suggestive of a novel disorder or an atypical presentation of a known syndrome. The differential diagnosis in the proband included Arthrogryposis Multiplex Congenita with synostosis (AMCN, OMIM #208100), KBG syndrome (KBGS, OMIM #148050) or multiple pterygium syndrome (OMIM #265000). Since skeletal and non-skeletal features of the patient were non-exclusive, genes related to all these syndromes were considered for identification of a single penetrant variant that underlies patient's abnormal phenotype. Furthermore, because the patient presented with a *de novo* balanced translocation, possible gene perturbation as a result of the translocation was investigated through precise breakpoint mapping in WGS analysis. Given the extent of SVs identified through WGS, the potential implication of novel SVs were explored in the context of the patient's phenotype.

Exome sequencing of the proband was suggestive of two independent heterozygous mutations in *ECEL1* gene that may predispose to the disease in the compound heterozygous state. Considering that proband's parents are healthy and unaffected the only plausible scenario for the compound heterozygous mutations to underlie the proband's phenotype is when each variant is singly inherited from each of the parents. Sanger sequencing results

corroborated this scenario and confirmed the compound heterozygous status of the *ECEL1* variants in the patient.

Using WGS analysis I precisely mapped the breakpoints in relation to the *de novo* balanced t(8,10)(q22.1; q26.3) to *PLEKHF2* and *MGMT* genes. The translocation does not impair coding regions of the genes, and none of the genes has been previously reported in the context of congenital anomalies. Balanced translocations that do not involve disruption of genes at the site of translocation are usually benign and do not lead to a pathological condition in the great majority of cases. Given that the identified balanced translocation is in a heterozygous state, it is possible that the translocation originates from the alternative segregation during anaphase I of gametogenesis in one of the asymptomatic parents and therefore the parent of origin for this variant requires further investigation.

In addition to breakpoint mapping of the reported translocation in the patient, a homozygous deletion spanning exons 2 and 3 of the *ZNF718* gene was also identified in the patient. Given the implication of large deletions involving *ZNF718* exons in skeletal phenotypes, the impact of this deletion merits further investigation.

In conclusion, our analysis demonstrates that WES in combination with WGS provides a robust framework for analysing rare congenital anomalies in a comprehensive manner. It is clear that establishing the molecular diagnosis in the present case requires functional follow up of the compound heterozygous *ECEL1* mutations, which is currently underway through CRISPR/Cas9 mutagenesis in *Xenopus tropicalis* by Prof. Matt Guille's group at Portsmouth University.

Chapter 5

Clinical Utility of Exome Gene Panel Sequencing for Molecular Diagnosis in Patients With Unusual Syndromic Cleft Lip/palate Phenotypes

"Some passages in this chapter have been quoted verbatim from Jabalameli *et al.*[337]."

5.1 Introduction

Cleft lip and cleft palate (CLP) is a common congenital orofacial abnormality with a prevalence of between one and two individuals per thousands live births[338]. The aetiology of orofacial clefting ranges from pure genetic aberrations predisposing to clefts to environmental in-utero exposure to teratogens that arrests complete fusion of primordial plates involved in orofacial development. CLP cases are generally classified into two categories: cleft palate only (CPO) cases in which perturbation of primordial plates fusion is restricted to palatine prominences only, and cleft lip with or without cleft palate (CL/P) cases in which failure of fusion of primordial plates also involve maxillary prominence and the medial nasal prominence[339]. Although a range of additional disruptions affecting craniofacial complex is also identified, the majority of orofacial clefts restrict to the upper lip and/or palate. Isolated (non-syndromic) CLP in which orofacial clefting is not associated with additional cognitive or craniofacial structural abnormalities is more common (∼70% of CLP cases[339]), and the disease aetiology in isolated cases usually have a complex origin. The non-syndromic CLP does not cause morbidity among affected individuals, and it has a modest recurrence rate in the general population.

In contrast, syndromic forms of CLP in which additional cognitive and structural abnormalities are also present have a strong genetic component and are generally considered as single gene disorders, and therefore they are amenable to case samples analysis. For over 500 Mendelian syndromes identified to be involved CL with or without CP, the genetic aetiology for only a fraction of them is fully resolved, and the genetic mutations underlying an overwhelming majority of these disorders are largely unknown[339,338]. Given the significant genetic and phenotypic heterogeneity underlying syndromic CLP, exome gene panel sequencing offers a cost-effective solution for molecular diagnosis in these patients.

This chapter describes the genetic analysis of fourteen patients with syndromic CLP

identified at the Operation Smile clinic in Bogota, Colombia. The full description of the cohort ascertained for epidemiological study of orofacial clefts in Colombia is described in the study by Arias Uruena *et al.*[340]. In brief, the Colombian cohort included 311 patients with both syndromic and non-syndromic CLP who have attended the clinic over the period April 2012 to July 2013. About 19% of patients (n= 59) were classified as syndromic cases and were provisionally categorised according to their clinical presentation. Investigation of patients phenotype among syndromic cases was suggestive of an unexpectedly high frequency of cases with a possible diagnosis of Aarskog-Scott syndrome (AAS, OMIM #305400). Amongst the 27 syndromes recorded in this cohort, phenotypic features in 17% of patients (n= 10) revealed to be consistent with a possible diagnosis of AAS.

Aarskog-Scott syndrome (AAS, OMIM #305400) (also known as faciogenital dysplasia) is a complex developmental disorder initially described by Aarskog in a Finnish pedigree[341] and later by Scott[342]. Patients with AAS present with a range of developmental complications including short stature, hypertelorism, ptosis, long philtrum, micrognathia, broad nasal bridge, clinodactyly of the little finger and genitourinary abnormalities including cryptorchidism and shawl scrotum[343]. Causal genetic variants underlying AAS pathophysiology map to the proximal short arm of chromosome X (Xq11.22)[344] and, to date, 61 different mutations across the 18 exons of the *FGD1* gene are reported as pathogenic in the context of the condition. The *FGD1* mutational spectrum includes 32 missense mutations, 16 frameshift variants, 6 nonsense variants, 4 splice site variants, 1 in-frame deletion and 2 out of frame deletions[345]. While only a limited number of the reported cases have been molecularly confirmed (∼35 till November 2018), the population incidence of two to three patients with a proven *FGD1* mutation per year is suggestive of the world population prevalence of 1/25,000[346]. It is notable that the reported mutations only describe 20% of the known cases of AAS[347] and the spectrum of genotype-phenotype correlations is unclear. Failure to identify pathogenic variants in patients who are referred for *FGD1* mutation analysis must reflect extensive clinical and genetic heterogeneity in AAS. In particular, the phenotypic features of the condition overlap with several other developmental disorders. Since it is now possible to undertake cost-effective sequencing of genes of clinical interest, exomes, or indeed whole genomes, there is greatly increased confidence that phenotypic overlaps can be resolved and underlying genetic causal variation understood for AAS and AAS-like syndromes.

A diagnosis of AAS is normally established through the Teebi criteria[348] by evaluating phenotypic features in a mother and an affected son. In this approach the clinical manifestation of the condition is studied in a tiered fashion and diagnosis is established in the presence of all primary and most secondary criteria. A detailed list of the diagnostic criteria and clinical features which differentiate AAS from similar syndromes is provided in Table 5.1. Short stature, hypertelorism and fold of the lower lip are the primary features present in nearly all cases[349]. Brachydactyly, interdigital webbing, shawl scrotum, long philtrum and mild facial hypoplasia are secondary features observed in almost 80% of cases. Additional phenotypic manifestations which include cryptorchidism, inguinal hernia, downward eye slant and ptosis are present in only a fraction of patients and therefore deemed secondary for diagnosis. AAS patients usually present with delayed growth in early childhood but achieve developmental milestones later in life[350]. AAS predominantly influences males and phenotypic complications are attenuated in females. Two separate reports identify impaired executive attentional processes including attention-deficit/hyperactivity disorder (ADHD)[351] and mania[352].

Criteria	AAS	Kuwaiti	Robinow Syndrome	Optiz Syndrome	Naguib-Richieri-Costa Syndrome	Teebi Hypertelorism Syndrome	Craniofrontonosal Dysplasia (CFND)
AAS Primary Diagnostic Criteria							
Hypertelorism	+	+	+	+ (varying degrees)	+	+ (Significant)	+ (varying degrees)
Short nose/ anteverted nares	+	+	+	+ (moderate)	+	+	+/-
Maxillary hypoplasia	+	+	+	+	-	+	+/- (asymmetrical)
Crease below lower lip	+	-	-	-	+	-	-
Short, broad hands	+	+	+	-	-	+/-	+/-
Mild, interdigital webbing	+	+	-	+	+	+	-
Short stature	+	+	+	-	-	-	-
Shawl scrotum	+	+	-	-	+	+	+
Short fifth finger/ clinodactyly	+	+	+	-	-	+	+
AAS Secondary Criteria							
Widow's peak	+	+	-	+	+	+	-
Ptosis	+	-	-	+/-	+	+	-
Downward slant of palpebral fissures	+	-	-	+/-	+	+	-
Abnormal auricle/Fleshy lobules	+	+	+	+	+	+	-
Joint hyperextensibility	+	+	-	+	-	-	+/-
Broad feet with bulbous toes	+	+	+	-	+ (varying degrees)	+	- (elongated toes)
Cryptorchidism/inguinal umbilical hernia	+	-	+	+	-	+	-
Large head/bossing	+/-	-	+	-	-	+	+
Craniosynostosis	-	-	-	-	-	-	+
Broad or bifid nasal tip	-	-	-	+	+	+	-
Thick eyebrows	-	-	-	-	+	+	-
True syndactyly	-	-	-	+ (in toes)	+ (in third and fourth digits)	-	
Nail groove	-	-	-	-	-	-	+/-
Broad thumb/toe, duplications or polydactyly	-	-	-	-	+	-	+
Short limbs	-	-	+	-	-	-	+
Hypoplastic genitalia	-	-	+	-	-/+	-	-
Hypospadias	+/-	-	-	+	+	-	-
Laryngeal complications	-	-	-	+	-	-	-
Vertebral anomalies	+	-	+	-	-	-	-
Intelligence	Usually Normal	Normal	Normal	Variable	Normal	Normal	Variable
Affected sex	Mostly males	Both	Both	Both	Both	Both	Mostly females
Inheritance	Autosomal dominant or X linked	Autosomal recessive	Autosomal dominant or Autosomal recessive	Autosomal dominant or X linked	Autosomal recessive	Autosomal dominant	Probably Autosomal dominant or X linked

Table 5.1: Teebi tiered criteria for the differential diagnosis of Aarskog-Scott Syndrome -Adopted from Teebi diagnostic criteria [348,353]

5.1.1 AAS Phenotypic heterogeneity: related syndromes and differential diagnosis

Arias Uruena *et al.*[340] posited that the unexpectedly high frequency of AAS diagnosis in the Colombian CLP cohort is probably due to the fact that AAS is underdiagnosed. Similarities between the phenotypic faciogenital characteristics of AAS patients and other developmental disorders including Noonan syndrome, SHORT syndrome (Short stature, hyperextensibility, hernia, ocular depression, Rieger anomaly, and teething delay) and Robinow syndrome increase the complexities of diagnosis. Noonan (OMIM #163950) and Noonan-like syndromes, in particular, show significant similarities to AAS. Hypertelorism, genital anomalies and ptosis are present in both conditions; however, additional characteristics of Noonan syndrome including heart abnormalities and lymphatic malformations contribute to unambiguous diagnosis[354]. Similarly, Robinow syndrome (OMIM #268310) presents with many manifestations identical to AAS. While short stature, hypertelorism and facial anomalies are seen in both conditions, shawl scrotum is restricted to AAS patients, and Robinow patients present characteristic shortening of mesomelic limbs[355].

Furthermore, LEOPARD syndrome (OMIM #151100) in which patients present with Lentigines, Electrocardiographic condition defect, Ocular hypertelorism, Pulmonary stenosis, Abnormalities of genitals, Retarded growth and Deafness shares some phenotypic characteristics of AAS. LEOPARD syndrome appears to occur sporadically and therefore differential diagnosis is usually established based on the mode of inheritance[356]. Since related syndromes may share phenotypic features with AAS differential diagnosis, therefore, requires careful consideration of a spectrum of phenotypic features. Establishment of a robust AAS diagnosis requires consideration of the age and gender of the patient[357]. Furthermore, the resolution of any developmental delay issues during later stages of development is indicative of AAS. Female AAS patients, arising through X-linked recessive inheritance, present with attenuated features of the condition. This can be attributed to skewed X-inactivation in which the causal X-linked allele is subject to strong selection to become inactivated[358]. The extent of non-random X inactivation in the context of AAS is poorly understood. Genitourinary abnormalities are almost always restricted to male patients.

Rare AAS cases with additional complications including spina bifida occulta (incomplete closure of the spine and surrounding tissues), cervical spine abnormalities[359], scoliosis (abnormal curve of spine)[360], camptodactyly (fixed flexion deformity of the joints in toes) and lymphoedema[361], macrochidism (abnormally enlarged testis)[362] and ocular complications (optic nerve hypoplasia, retinal vessel tortuosity, deficient ocular elevation, hyperopia and anisometropia)[363,364,365] are also reported.

The very large number of potentially deleterious variants identified by sequencing presents challenges. Even with rigorous filtering and prioritization procedures there may be difficulties in firmly establishing underlying pathogenic variants. In a recent study[93] into the effectiveness of clinical whole-genome sequencing of patients with uncertain molecular diagnosis, the authors established that causal variants could be firmly identified in only 34% of patients with Mendelian disorders. In targeted high-throughput sequencing of heterogeneous disorders the probability of successful identification of causal variants is directly correlated with the number of major clinical features presented by the patient. The latter notion is demonstrated in a study by Redin *et al.*[366] where comparison of clinical phenotypes between patients with more than one detected pathogenic variant and those with either one or no suspicious pathogenic variant clearly demonstrated the importance of precise phenotyping in facilitating accurate identification of underlying causal mutations. Another consideration in the analysis of high-throughput sequencing data is the presence of high frequency pathologic variants in isolated populations that might be otherwise

neglected during filtering and prioritization processes.

It worth noting that the high frequency of AAS cases among syndromic CLP cases in the study by Arias Uruena *et al.* was in sheer contrast with previous studies that found Van der Woude syndrome (VWS1, #119300) as the most frequent condition among CLP patients[367,339,368]. Although the authors suggested that local geographical and ethnic factors might account for the atypical frequency of AAS among syndromic cases, it is likely that these patients are in fact present with other syndromes that due to their extensive phenotypic similarity to AAS are misdiagnosed. The clinical spectrum of AAS phenotypes and overlap with related syndromes is sufficiently broad to complicate diagnosis, and in many cases, a definitive diagnosis can only be established through molecular methods such as targeted sequencing[369].

5.2 Methods

Fourteen patients, ascertained through the Operation Smile Multidisciplinary Centre in the Bogota region of Colombia, with a provisional diagnosis of Aarskog-Scott syndrome were analysed to establish the molecular diagnosis. All patients presented with either unilateral or bilateral cleft lip (CL). Eight patients who were diagnosed with CL were also presented with complete cleft palate (CP). Shawl scrotum (SC) is present in 10 patients and is associated with cryptorchidism in four patients. Available phenotypic information for patients are provided in the Table 5.2.

ID	sex	Cleft Lip	Cleft Palate	Shawl Scrotum	Cryptorchidism	Telecanthus	Hypertelorism	Short Stature	Developmental Delay	Additional Features
CL021	Male	Bilateral	Incomplete	+	-	-	+	+	Delay	
CL022	Male	Bilateral	Complete	+	+	+	+	+	Normal	
CL024	Male	Bilateral	-	-	-	+	+	-	Delay	
CL025	Male	Unilateral	-	+	+	+	+	+	Delay	Anophthalmia, Esophageal atresia
CL027	Male	Unilateral	Complete	+	-	-	+	-	Normal	
CL028	Male	Unilateral	Complete	+	-	-	-	+	Delay	
CL030	Male	Unilateral	Complete	+	-	+	+	-	Normal	
CL033	Male	Unilateral	Complete	+	-	+	+	-	Normal	Clinodactyly
CL035	Male	Unilateral	Complete	+	-	-	-	-	Delay	
CL036	Male	Unilateral	Complete	+	+	+	+	+	Delay	Clinodactyly, Heart murmur
CL037	Male	Unilateral	Complete	+	-	+	+	-	Normal	Left club foot
CL038	Male	Unilateral	Complete	+	-	+	+	-	Delay	
CL039	Male	Unilateral	Complete	+	-	-	+	-	Delay	
CL040	Male	Bilateral	Complete	+	+	-	-	+	Mental Retardation	Lip melanosis

Table 5.2: Phenotypic features of the 14 patients with primary diagnosis of AAS considered for targeted exome sequencing.

5.2.1 Sample processing

Targeted exome enrichment was carried out using the TruSight One Sequencing Panel (Illumina, San Diego, CA, USA). The panel covers exonic regions of 4,813 disease-causing genes that are indexed in the HGMD database[185], OMIM catalogue[1] and the GeneTests[370]. Compared to other enrichment methods, TruSight one effectively reduces the number of enrichment steps by employing a multiplex pre-enrichment sample pooling strategy that eliminates the need for mechanical DNA fragmentation. In brief, genomic DNA is adaptor-tagged using Nextera technology (Illumina, San Diego, CA, USA) and hybridised to biotinylated probes specific to target regions and enriched by streptavidin beads. Sequencing-ready fragments are then magnetically pooled and eluted from beads to give a mean fragment size of 300bp. Sequencing was performed at the Wessex Investigational Sciences Hub laboratory (WISH lab, University of Southampton) using the HiSeq2000 platform (Illumina, San Diego, CA, USA).

Parameter	**Value**
Cumulative Target Region Size	12 Mb
Number of Target Genes	4,813
Number of Target Exons	62,000
Probe Size	80-mer
Number of probes	125,395
Target minimum coverage	20x

Table 5.3: TruSight One sequencing panel coverage details (Adapted from TruSight One sequencing panel technical sheet[371]).

5.2.2 Data analysis

Raw exome sequences were analysed, and quality checked using the WES pipeline as described in Chapter II. Since samples had the provisional diagnosis of Aarskog-Scott syndrome (OMIM #305400), cumulative coverage across the *FGD1* as the primary candidate gene was prioritised in our screening strategy. In order to identify regions with suboptimal depth of coverage in the *FGD1* gene across samples, a heatmap plot of read-depth within 18 exons of *FGD1* was produced using the Bedtools python wrapper (v 0.7.8)[372] and Seaborn statistical data visualization package (v 0.7.1)[373]. Besides, Gviz Bioconductor package[374] was used to plot depth of coverage across exons of *FGD1* gene for all samples. Pathogenicity and conservation scores for each variant computed using prediction algorithms including M-CAP[375], PolyPhen2-HVAR[170] , SIFT[169], LRT[171], FATHMM[173], RadialSVM[376], CADD[174], Phylop[176] and GERP++[175]. For variants outside the exon boundaries, the Ensemble Variant Effect Predictor (VEP)[183] MaxEnt plugin was used to quantify the impact of the mutational event on the splicing process. The splicing variants that fulfilled MaxEnt significance threshold ($|\ \Delta\text{MaxEnt}\ | \geq 3.0$) were further evaluated in the Human Splicing Finder (HSF, v.2.4.1)[377]. Continental and population-specific allele frequencies (AF) from the 1000 Genome Project (2015aug) and the ExAC[5] non-TCGA samples were incorporated to the annotation. The HGMD record for disease-causing genes from the HGMD v.2016.2 VCF file was added to the variants annotation using Pandas data analysis library in python[378]. All positions were defined according to the Human Reference Genome build GRCh37 (hg19).

5.2.3 Filtering and variant analysis

Given the diversity of phenotypic features in each patient, filtering and variant analysis for each individual was carried out separately but in a consistent manner. In order to prioritise putatively causal variants, we adopted a tiered filtering approach focussing on genes previously implicated in the cleft lip and/or cleft palate syndromes (CLP) and also additional genes that underlie shawl scrotum and cryptorchidism. Initially, all samples were screened for the *FGD1* mutations. Next, a collective set of 918 genes prepared in three tiers as follows:

- A primary list of 112 candidate genes prepared from the HGMD v.2016.2 using the following keywords [in phenotype search tab]: Cleft, Shawl Scrotum (HP:0000049) and Cryptorchidism (HP:0000028).
- A complementary list of 20 unique genes (not covered in the tier-1 list) related to Shawl Scrotum (HP:0000049) and Cryptorchidism (HP:0000028) was retrieved from the Harmonizome[379] and prioritised as the second tier in the variant analysis pipeline.
- A comprehensive list of 787 unique genes (not covered by the tier1 and the tier2) prepared by extensive search of OMIM database (January 2017 update), ClinVar[186] (January 2017 update), Orphanet[380] (V4.23.0) and DDG2P[381] (v9.12) database using a set of eleven HPO terms including Mild global developmental delay (HP:0011342), Hyperactivity (HP:0000752), Intellectual disability (HP:0001249), Unilateral or bilateral cleft lip (HP:0100336, HP:0100333), complete or incomplete cleft palate (HP:0000175), telecanthus (HP:0000506), hypertelorism (HP:0000316), shawl scrotum (HP:0000049), cryptorchidism (HP:0000028) and clinodactyly (HP:0000028), and used as third tier panel during variant analysis.

Variants with a minor allele frequency of greater than or equal to 1% were excluded to remove common variants as they are unlikely to contribute to a rare syndromic disease. Variants were further filtered using the following exclusion criteria:

1. All synonymous variants located outside exon sequences were excluded. Variants within 10bp of either donor-splice site or acceptor-splice were retained and analysed as splicing variants;
2. Variants located beyond 10bp of either 5' or 3' of exons were dealt with as non-coding variants and filtered out;
3. Non-frameshift deletion/insertion variants were also excluded as they were unlikely to be causal in the context of AAS; and
4. Variants identified in more than ten individuals in our in-house non-disease database were also excluded.

In order to identify the causal variant in each individual, novel variants with a dominant pattern of inheritance were shortlisted, and a combined score was used to bin variants into a ranked order. The prediction scores from the four functional prediction models (SIFT, Polyphen2-HVAR, LRT and FATHMM), two conservation models (GERP and Phylop) and three ensemble prediction models (M-CAP, CADD and RadialSVM) were transformed to the single combined score (Ψ_i) according to the formula below:

$$\Psi_i = \sum_{i=1}^{9} S_i \tag{5.1}$$

where S_i is the score function and defined as:

$$S_i = \begin{cases} 0 & \text{if } \theta_(x_i) < Q_1 \text{ or } \theta_(x_i) = \text{benign/neutral} \\ 0.5 & \text{if } Q_1 \leq \theta_(x_i) < Q_3 \text{ or } \theta_(x_i) = \text{possibly damaging/uncertain} \\ 1 & \text{if } \theta_(x_i) \geq Q_3 \text{ or } \theta_(x_i) = \text{damaging} \end{cases} \tag{5.2}$$

The $\theta_(x_i)$ is the pathogenicity or conservation score for variant x as defined by model i and Q denotes quartile range for scores from M-CAP, CADD, GERP and Phylop models. Variants were finally ordered from the largest Ψ_i (more pathogenic) to smallest Ψ_i (less pathogenic). Considering that prediction score for stop-gain mutations is not available from the Polyphen2-HVAR, MCAP, FATHMM and RadialSVM models, pathogenicity of these variants evaluated individually.

5.3 Results

5.3.1 QC Results

Quantity and purity of gDNA samples are of paramount importance for target enrichment and library preparation. Contaminants such as organic compounds used for DNA extraction (*i.e.* phenol), RNA and proteins interfere with library preparation and impair sequencing reactions[382]. Sample preparation QC revealed a below-threshold absorbance ratio for 12 samples ($A_{260}/A_{280} < 1.8$) and low DNA concentration (Qubit concentration < 50 μg/μl) for seven samples (Table 5.4). Quality of processed data was carefully inspected at three stages including *pre-alignment*, *post-alignment* and *post variant calling* as per the procedure described in Chapter 2. Mapping quality scores for all samples were consistently above 40 which is indicative of 99.99% base call accuracy (Table 5.4). For all samples $\geq$99% of reads aligned to the reference genome (Table 5.4, the 6th column) and 57-70% of reads aligned in proper read pairs to the TruSight One target region (Table 5.4, the 8th column). The total raw read counts for samples CL024, CL028, CL030 and CL037 were significantly lower than other samples which may be reflective of poor DNA quality. The sequence read lengths in all samples range from ∼30bp to ∼150bp with average length at ∼123-144bp, representing slightly positively skewed distribution (Supplementary Figure 8.14). To verify that samples are not contaminated, sequence read files (BAM files) were checked by the "VerifyBAMID" software[164], and contamination rates for all samples were identified to be consistently well below the 2% threshold. For all cases apart the proband CL038, the gender status of patients inferred from the X chromosome heterozygosity was in accordance with the samples' labels (Table 5.4). Sample CL038 was excluded from further analysis on the basis of gender mismatch.

Sample	Qubit Conc. (ng/ul)	Nanodrop 260/280	DNA Yield (ug)	Number of Reads	Aligned Reads (%)	Unmapped Reads (%)	Mapped to TruSight One Target Region (%)	% Coverage at 20X	Read Mean Length	Mean Mapping Quality	Verify BAMID freemix score (%)	NO. Called Variants	%X Chromosome Heterozygosity	Inferred Gender
CL021	134	1.66	10.7	30010619	29,933,622 (**99.74%**)	76,997 (**0.26%**)	20,080,110 (**67.08%**)	96.58	143.77	44.14	0.01	12101	23.02	Male
CL022	23	1.35	1.6	42409723	42,295,234 (**99.73%**)	114,489 (**0.27%**)	24,502,086 (**57.77%**)	97.13	141.63	44.15	0.019	12312	21.38	Male
CL024	4	1.33	1	2365258	2,354,816 (**99.56%**)	10,442 (**0.44%**)	1,609,730 (**68.05%**)	9.02	144.65	41.28	0.046	5293	75.45	Female
CL025	111	1.65	8.9	29788440	29,715,698 (**99.76%**)	72,742 (**0.24%**)	21,086,178 (**70.78%**)	93.44	137.7	44.32	0.027	11431	19.01	Male
CL027	62	1.56	5	32060201	31,955,822 (**99.67%**)	104,379 (**0.33%**)	19,531,790 (**60.92%**)	96.50	143.98	44.08	0.013	12371	30.72	Male
CL028	46	1.42	4.1	10490498	10,467,454 (**99.78%**)	23,044 (**0.22%**)	6,954,575 (**66.29%**)	63.08	138.87	43.66	0.014	9490	22.44	Male
CL030	48	1.23	4.8	7378595	7,363,223 (**99.79%**)	15,372 (**0.21%**)	5,188,922 (**70.32%**)	45.68	131.8	43.47	0.018	9064	13	Male
CL033	13.5	1.2	1.6	48216232	48,095,299 (**99.75%**)	120,933 (**0.25%**)	31,009,903 (**64.31%**)	97.79	144.03	44.2	0.011	12364	22.56	Male
CL035	128	1.82	10.2	29357251	29,287,572 (**99.76%**)	69,679 (**0.24%**)	19,052,583 (**64.89%**)	96.83	144.45	44.12	0.011	12203	21.42	Male
CL036	97	1.82	7.8	42758969	42,647,420 (**99.74%**)	111,549 (**0.26%**)	27,446,483 (**64.18%**)	97.64	143.31	44.42	0.009	12939	15.43	Male
CL037	99	1.67	8.9	1566231	1,562,592 (**99.77%**)	3,639 (**0.23%**)	1,037,531 (**66.24%**)	2.02	123.12	41.77	0.02	2887	30.01	Male
CL038	9	1.47	1.8	45372198	45,246,719 (**99.72%**)	125,479 (**0.28%**)	28,594,714 (**63.02%**)	97.33	144.17	43.95	0.015	13004	71.32	Female
CL039	18	1.61	0.9	28421558	28,355,860 (**99.77%**)	65,698 (**0.23%**)	18,316,248 (**64.44%**)	95.69	142.72	43.78	0.011	12127	21.76	Male
CL040	61	1.27	6.1	19276303	19,237,589 (**99.8%**)	38,714 (**0.2%**)	13,133,310 (**68.13%**)	81.40	134.98	43.84	0.015	10609	18.12	Male

Table 5.4: Sample preparation, alignment, variant calling and coverage QC results (Mapping quality scores are expressed in Phred scale.)

Depleted read counts for samples CL024, CL028, CL030, CL037 and CL040 were consistent with the poor cumulative depth of coverage (DOC) across the target region (Figure 5.1). *FGD1* coverage for these samples was also insufficient for confident variant calling in this gene (Figures 5.2, 5.3 & Supplementary Figure 8.15) and therefore these samples were excluded from further analysis.

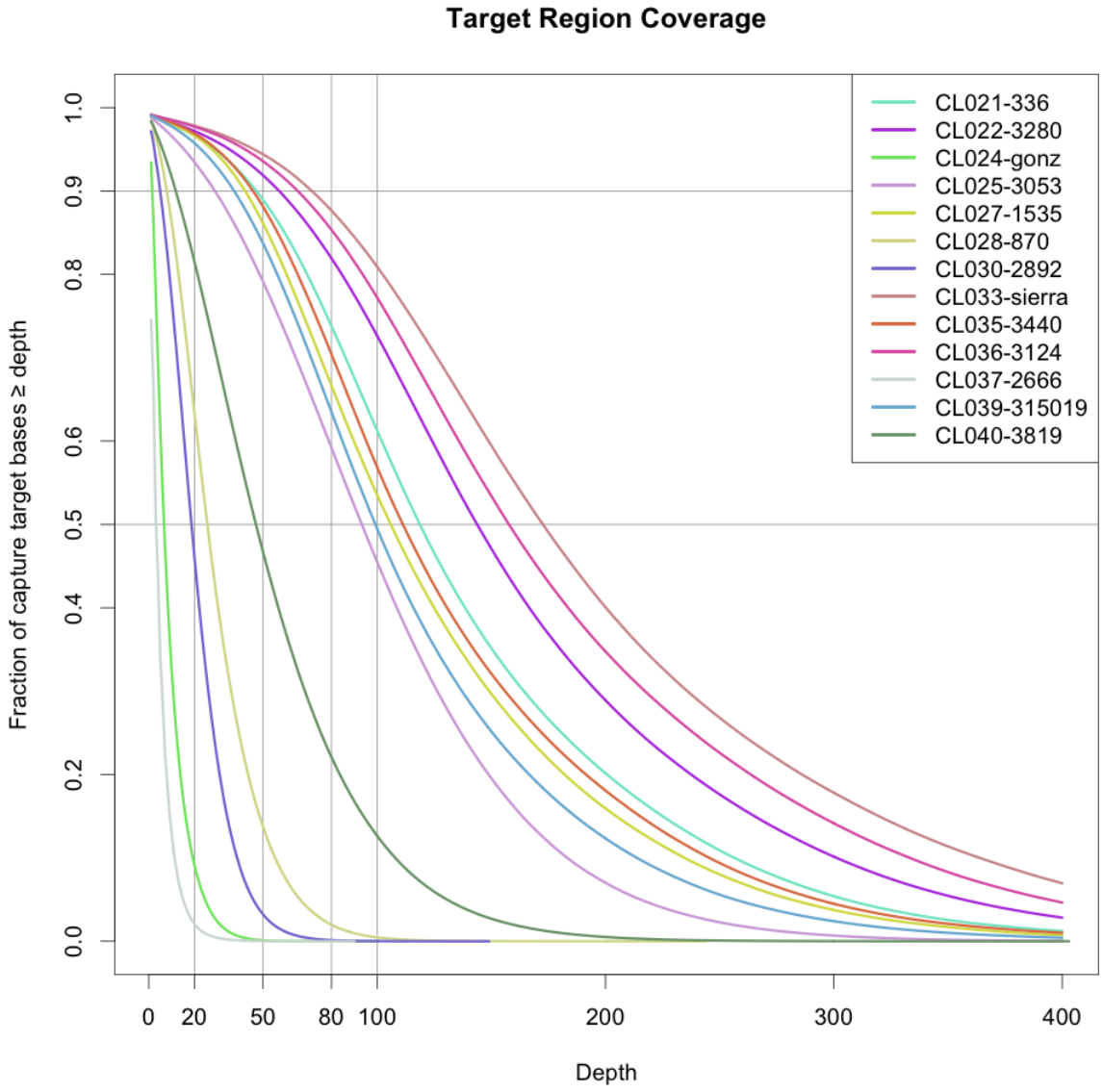


Figure 5.1: The cumulative depth of coverage (DOC) for the 13 samples with correct gender identification across the capture target region (TruSight One sequencing panel). Samples CL037, CL024, CL030, CL028 and CL040 represent the lowest DOC amongst the samples analysed

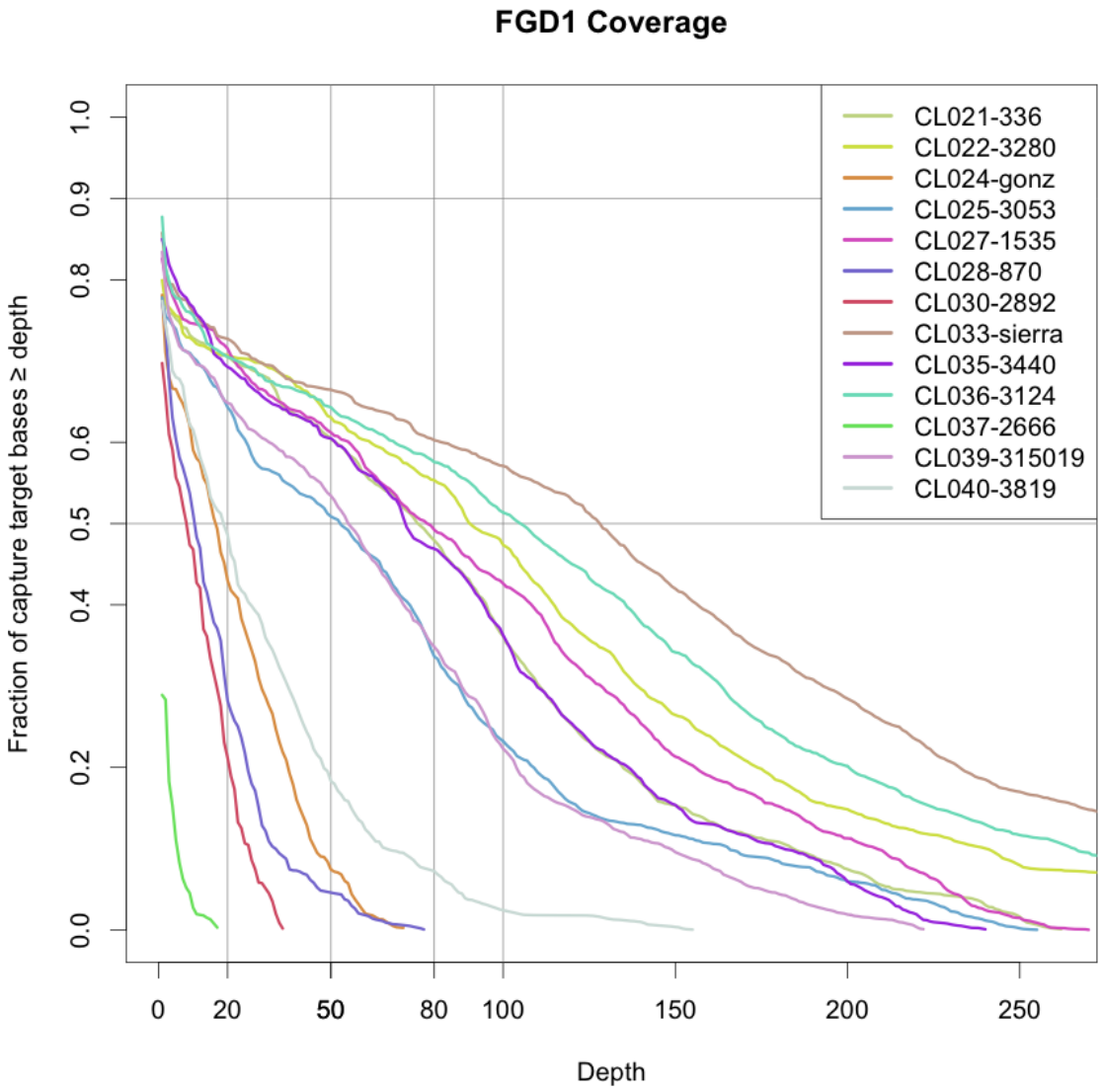


Figure 5.2: The cumulative depth of coverage (DOC) across the coding region of the *FGD1* gene for the 13 samples with correct gender identification. Samples CL037, CL024, CL030, CL028 and CL040 represent poor coverage for the *FGD1* gene.

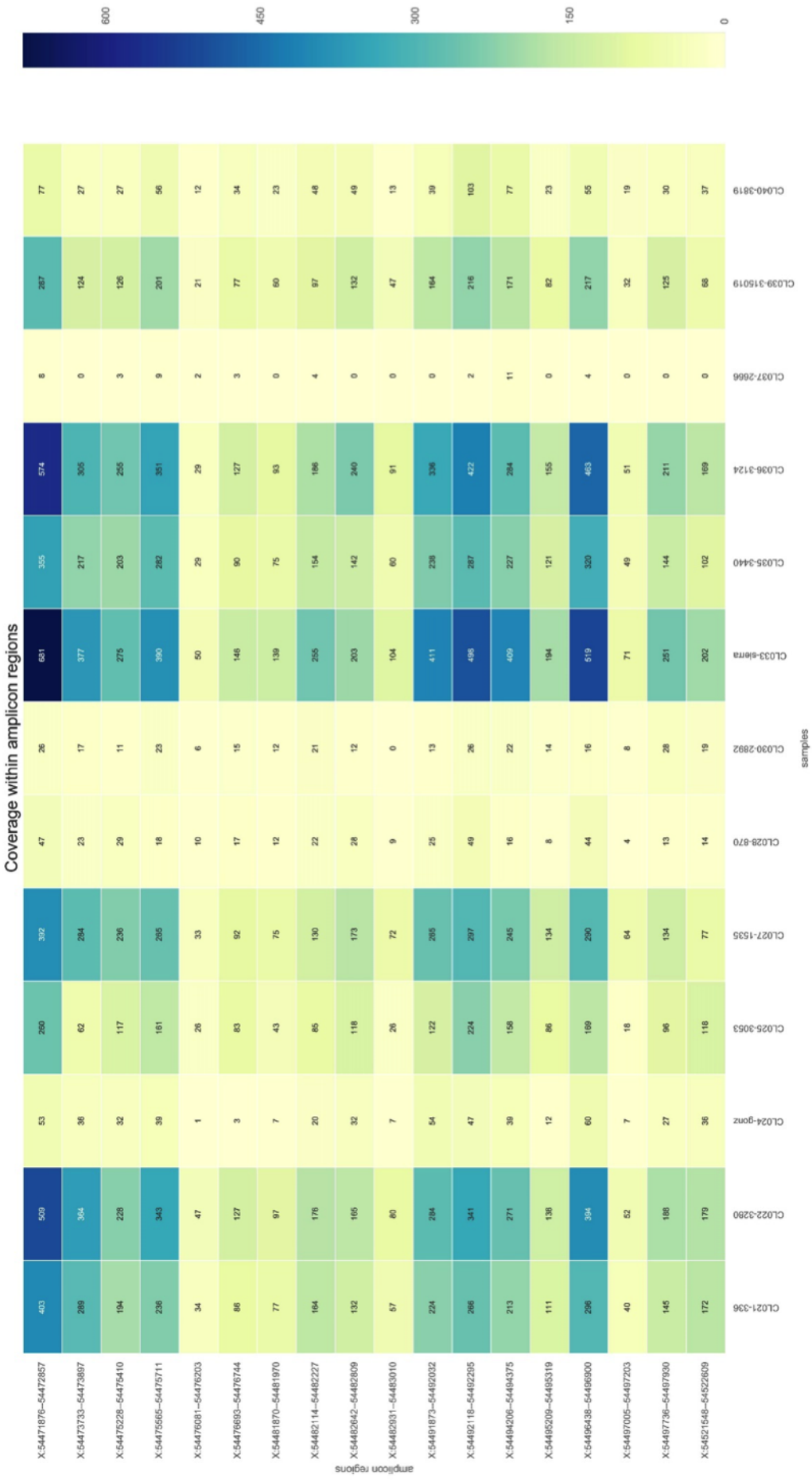


Figure 5.3: Heatmap plot of read depth per exon of *FGD1* gene across the 13 samples with correct gender identification. The vertical axis represents exonic regions of *FGD1* gene and the horizontal axis represents the 13 samples with correct gender identification. Numbers inside the rectangles represent the number of reads covering each exon. Samples CL024, CL028, CL030, CL037 and CL040 are poorly covered at *FGD1* loci and excluded from downstream analysis.

5.3.2 Tiered Analysis

Clinical exome data of good quality with greater than 90% coverage at a depth of 20X from eight samples (CL021, CL022, CL025, CL027, CL033, CL035, CL036 and CL039, (Table 5.4)) were analysed across the three tiers and variants that passed the filtering criteria were inspected individually to establish pathogenicity in each patient. The tiered analysis in each case suggested plausible pathologic variants (Table 5.5) which are thoroughly discussed in the following section.

Gene	Exon	Nucleotide	Protein	Tier	SIFT	PP-2	MCAP	LRT	FATHMM	RadialSVM	CADD	GERP	Phylop	Depth	Ψ_i	CL021	CL022	CL025	CL027	CL033	CL035	CL036	CL039
SOX10	4	c.A718C	p.T240P	III	0.02	0.986	0.689	0	-5.36	1.08	20.3	4.76	7.996	46	8.5		o	o					
COL1A1	5	c.G433C	p.G145R	III	0	0.999	0.916	0	-5.77	0.999	23.6	5.15	5.71	44	8.5						o		
CREBBP	30	c.G5128T	p.G1710C	III	0.01	1	0.328	0	-1.86	0.615	16.95	5.35	7.755	12	8			o					
SALL4	2	c.A1751T	p.H584L	III	0	0.995	0.019	0	-2.18	0.94	22.6	5.59	8.035	11	8			o					
NAGLU	1	c.C311A	p.S104Y	III	0.04	0.961	0.92	0	-3.81	0.985	32	4.94	4.104	15	8				o				
MYO7A	21	c.G2441A	p.R814H	III	0	1	0.472	0	-0.59	0.298	33	5.39	9.863	369	8		o						
CREBBP	30	c.A7037C	p.H2346P	III	0.01	0.986	0.113	0	-2.08	0.466	11.24	5.35	6.08	51	7.5			o					o
KMT2D	42	c.A13885C	p.T4629P	III	0	0.997	0.607	0.028	-2.1	0.7	12.74	5.58	7.969	32	7.5		o						
PTPN11	3	c.C323A	p.T108N	III	0	0.986	0.121	0	2.87	0.264	27.1	5.14	7.811	11	7.5							o	
KMT2D	53	c.C16493T	p.S5498F	III	0	0.999	0.809	0.002	-2.18	0.896	15.74	4.64	9.545	119	7							o	
SOX9	3	c.A715C	p.T239P	I	0.03	0.543	0.177	0	-1.84	0.343	15.95	3.29	8.951	41	7		o						
SOX9	3	c.A706C	p.T236P	I	0.01	0.368	0.322	0	-1.55	0.509	19.36	4.37	8.951	31	7		o						
BRCA2	25	c.A9425T	p.D3142V	III	0	0.99	0.208	0.002	-1.7	0.447	21.7	5.89	6.046	83	7							o	
TSC2	39	c.C5008A	p.P1670T	III	0.31	0.998	0.443	0	-3.45	0.662	19.21	4.21	6.124	58	6.5			o					
SMARCA4	5	c.A889C	p.S297R	III	0.02	0.198	0.231	0	-2.15	0.244	12.37	4.35	6.448	27	6							o	
RYR1	20	c.C2543T	p.T848I	I	0.06	0.986	0.421	0.003	-4.13	0.893	13.7	2.67	2.054	108	5.5		o						
TGFB3	5	c.G785T	p.G262V	I	0.01	0.783	0.095	0.039	-0.47	0.163	20.4	5.25	3.255	47	5.5		o		o				
COMT	4	c.T476G	p.L159R	I	0	0.998	0.111	0	0.94	-0.61	15.26	4.06	7.142	23	5.5						o		
SRCAP	34	c.A8036T	p.E2679V	III	0	0.578	0.22	0.004	-3.03	0.122	7.895	4.88	2.496	166	5	o							
SPAST	1	c.G103C	p.A35P	III	0.33	0.15	0.927	0.014	-3.31	0.069	12.77	3.04	2.946	44	5			o					
PTEN	8	c.G810T	p.M270I	III	0.06	0.437	0.234	0	-1.82	-0.075	18.65	5.13	9.429	71	5					o			
KANSL1	14	c.A2903C	p.Q968P	III	0.15	0.994	0.029	0	0.86	-0.744	25.3	5.2	6.956	23	5	o							
OFD1	20	c.G2635A	p.E879K	I	0.21	0.833	0.232	0.01	-2.01	0.86	34	4.86	3.296	73	5								o
TBX4	8	c.C1106G	p.S369C	III	0	0.991	0.025	0.349	0.58	-0.479	23.6	5.51	7.438	27	5								o
OFD1	20	c.G2610C	p.Q870H	I	0.14	0.799	0.143	0.906	-1.84	0.318	14.98	3.24	0.929	42	4.5			o					
SMARCA4	5	c.G940C	p.A314P	III	0.17	0.041	0.39	0	-2.23	-0.117	15.02	4.36	9.506	38	4.5	o	o						
KAT6B	8	c.C1919A	p.T640N	III	0.05	0.456	0.041	0	-1.04	-0.608	10.04	6.05	5.474	12	4.5	o							
GRIN2A	2	c.G136A	p.V46M	III	0.04	0.12	0.203	0.014	-2.13	-0.614	13.76	3.58	1.062	252	4						o		
TBX1	3	c.C82A	p.P28T	I	0.51	0.033	0.504	0.446	-2.08	-0.743	10.1	1.94	1.946	12	3.5			o					
NIPBL	45	c.A7801G	p.M2601V	I	0.61	0.041	0.04	0	-3.09	-0.112	11.27	4.89	5.462	148	3.5					o			
KIRREL3	5	c.A544T	p.I182F	III	0.03	0.285	0.077	0	-1.03	-0.673	15.24	2.47	1.155	191	3.5					o			
TCOF1	17	c.C2800T	p.Q934X	I	0.24	.	.	0.001	.	.	37	5.8	3.836	11	3.5			o					
KMT2D	39	c.T11759A	p.L3920Q	III	0	0	0.562	0.711	-2.68	-0.567	0.008	-1.73	-0.426	34	3			o					
KMT2D	39	c.T11843A	p.L3948H	III	0	0	0.765		-2.67	-0.644	0.015	-4.77	-0.555	32	3		o						
KMT2D	39	c.T11819A	p.L3940H	III	0	0.09	0.687		-2.76	-0.475	0.036	1.25	-0.121	22	3	o							
KDM6A*	15	c.T2117C	p.M706T	III	0.28	0.884	0.123	0	-1.12	-0.281	6.422	5.08	2.364	241	3	o							
ARID1B	1	c.A293C	p.H98P	III	0	0.144	0.753	.	0.96	-1.051	7.905	2.48	1.012	22	3							o	
SMARCA4	5	c.A892C	p.T298P	III	0.17	0.003	0.183	0.003	-2.15	-0.557	9.496	4.35	0.322	32	3		o						
RPL5	8	c.G838A	p.V280I	I	0.35	0.009	0.01	0	0.57	-1.072	15.6	5.16	6.676	59	3			o					
TSC1	20	c.G2479T	p.E827X	III	0.35	.	.	0	.	.	41	4.88	3.8	22	3					o			
KMT2D	39	c.T11735A	p.L3912H	III	0	0.001	0.314	0.381	-2.77	-0.502	0.006	-0.776	-0.608	37	2.5	o	o						
PIK3R1	4	c.A235G	p.I79V	III	1	0.013	0.009	0	1.04	-0.742	11.02	5.1	6.902	191	2.5		o						
ATR	15	c.G3025A	p.A1009T	III	0.6	0.002	0.005	0	-0.05	-1.008	13.6	4.71	2.912	56	2.5								o
TSC1	20	c.G2484T	p.M828I	III	0.08	0.03	0.07	0	-1.4	-0.119	16.94	4.86	4.556	24	2.5					o			
KMT2D	39	c.A11847T	p.Q3949H	III	0	0.259	0.779		0.93	-1.032	0.05	-1.71	-1.251	68	2					o			
KMT2D	10	c.A1372C	p.T458P	III	0	0.261	0.258		-1.28	-0.799	4.61	4.31	-0.001	17	2				o				
GUCY2D	2	c.G92C	p.R31P	III	0.29	0	0.426	0.402	-1.64	-0.919	10.98	-5.19	-0.378	24	2						o		
KMT2D	10	c.A1923C	p.E641D	III	0	0	0.144		-1.15	-0.927	11.64	-1.53	-0.186	13	2		o						
KMT2D	39	c.T11777A	p.L3926H	III	0	0.083	0.215	0.109	0.96	-1.001	0.008	1.91	-0.724	31	1.5		o						
IQSEC2	15	c.G4195T	p.A1399S	III	0.28	0.008	0.213	0.074	2.84	-0.977	5.868	2.58		10	1.5					o			
RAI1	3	c.G3200A	p.R1067H	III	0.16	0.387	0.132	0.02	0.04	-0.856	9.492	1.91	1.975	11	1.5			o					
KISS1R	5	c.G1046C	p.R349P	III	0.27	0.242	0.154	0.769	0.26	-1.028	12.33	-0.34	0.113	32	1.5			o					
ANKRD11	9	c.A7067C	p.E2356A	III	0.54	0.07	0.056	0.018	1.1	-0.983	13.73	4.99	4.835	32	1.5								o
TCOF1	10	c.G1391A	p.W464X	I	0.09	.	.	0.028	.	.	20.5	-7.81	-2.012	33	1	o							
SALL4	2	c.G429T	p.M143I	III	0.17	0.002	0.009	0	0.52	-1.041	13.36	-0.428	0	11	0.5								o
ANKRD11	9	c.A7084C	p.T2362P	III	0.27	0.03	0.031	0.698	1.17	-1.06	1.363	-0.57	-0.019	40	0			o					
KISS1R	5	c.G1037C	p.R346P	III	0.28	0	0.045	0.911	0.37	-1.024	8.364	-0.733	-0.331	40	0			o					

Table 5.5: Novel variants with the dominant pattern of inheritance identified across the tiered analysis. Variants are ranked according to their respective Ψ_i score (Variants represented in red identify stop-gain mutations; *: hemizygous variant)

CL021

The patient is a 11 year male offspring of a non-consanguineous marriage presented with developmental delay (HP:0011342), incomplete cleft palate (HP:0000175), bilateral cleft lip (HP:0100336), hypertelorism (HP:0000316) and shawl scrotum (HP:0000049). He was delivered at 35 weeks of gestation and had low birth weight ($\sim$ 1100g) at the time of delivery. The proband's parents are healthy, and cleft lip feature is only present in his second-degree cousin. Eight variants were identified as possibly pathogenic in the proband *CL021* (Table 5.5).

The ranking strategy applied to prioritise putatively causal variants revealed the *SRCAP*:c.8036A>T as the top candidate. SRCAP is the activator of *CREBBP* and mutations of *CREBBP* have been identified to underlie Rubinstein-Taybi syndrome 1 (RSTS1, OMIM #180849). Mutations of the *SRCAP* have been described to underlie Floating-Harbor syndrome (FLHS, #136140)[383]. Patients with *SRCAP* mutation present with delayed bone age and speech development, short stature, triangular faces and deep-set eyes with long eyelashes. The ears in FLHS patients are posteriorly rotated, however the facial characteristics of the patients are age-related and change over the time. Unilateral cleft and cryptorchidism have also been reported in the context of the disease[383]. The majority of nonsense and frameshift *SRCAP* mutations that have been identified in FLHS patients map to the exon 34 of the gene[384]. As exon 34 is the last exon of the *SRCAP* gene, these mutations escape the nonsense-mediated mRNA decay and lead to pathogenicity[385]. Given that identified mutation maps to the exon 34 of *SRCAP* and the extent of phenotypic similarity between the proband's phenotype and FLHS features, the functional impact of the variant merits further investigation.

Furthermore, hemizygous mutations of *KDM6A* (rank 6) have been identified to underlie X-linked dominant Kabuki syndrome 2 (KABUK2, OMIM # 300867). Patients with *KDM6A* mutation present with congenital mental retardation short stature, scoliosis and a range of facial dysmorphisms including long eyelashes, arched and sparse eyebrows, broad nasal bridge and cleft palate[386]. Miyake *et al.* demonstrated that *KDM6A* mutations underlie 6.2% of KABUK2 incidence. They also suggested that arched eyebrows are less common among *KDM6A* mutants but short stature and postnatal growth retardation is the consistent feature among all patients with *KDM6A* mutations. Genitourinary anomalies are common among KABUK2 patients but shawl scrotum has never been reported in the context of the disease. Furthermore, congenital mental retardation in combination with multisystemic anomalies involving heart and immune system is common among Kabuki patients. Given that the proband *CL021* does not present with intellectual disability and also considering the extent of phenotypic dissimilarity between the patient and KABUK2 syndrome, it is unlikely that the *KDM6A*:c.2117T>C variant is causal. This has been also reflected in the reduced combined pathogenicity score computed for this variant (Ψ_i=3) and therefore this variant was excluded from further follow up.

The visual inspection of reads using IGV for the remaining variants including *KANSL1*:c.2903A>C (rank 2), *SMARCA4*:c.G940C (rank 3), *KAT6B*:c.1919C>A (rank 4), *KMT2D*:c.11735T>A (rank 5), *KMT2D*:c.11819T>A (rank 7) and *TCOF1*:$c.1391G > A$ (rank 8) revealed low genotype quality calls and erroneous alignment. Based on this evidence, variants identified in these genes are probably spurious and therefore excluded from further follow up.

In conclusion, targeted exome sequencing and tiered filtering identified *SRCAP*:c.8036A>T as a putatively causal variant in the patient *CL021*.

CL022

The patient is a 13 month old male offspring of a non-consanguineous marriage with normal development and no history of CLP in the family. Patient's main features include short stature and low birth weight, bilateral cleft lip (HP:0100336), complete cleft palate (HP:0100336), hypertelorism (HP:0000316), telecanthus (HP:0000506), shawl scrotum (HP:0000049) and cryptorchidism (HP:0000028) as described in the Table 5.2. Fourteen variants were prioritised through the filtering strategy (Table 5.5). The visual inspection of reads in IGV to confirm the authenticity of variants revealed the top three candidates as *MYO7A*:c.2441G>A (rank 2), *RYR1*:c.2543C>T (rank 6) and *PIK3R1*:c.235A>G (rank 12).

MYO7A encodes an unconventional myosin with a very short tail that plays an important role in cellular movements[387]. Heterozygous mutations of *MYO7A* are identified to underlie nonsyndromic progressive hearing loss known as autosomal dominant deafness-11 (DFNA11, OMIM # 601317)[388]. Considering the lack of phenotypic similarity between the DFNA11 and the patient's main features and also absence of vestibular symptoms in the proband *CL022*, it is highly unlikely that *MYO7A*:c.2441G>A is causal in the context of the disease. The variant, therefore, was excluded from further follow-up.

RYR1 encodes a ryanodine receptor which acts as a calcium release channel in the sarcoplasmic reticulum of skeletal muscles[389]. Heterozygous mutations of *RYR1* have been identified to underlie Central core disease (CCD, OMIM # 117000)[390] and a form of malignant hyperthermia known as King-Denborough syndrome (OMIM # 145600)[391]. Facial and genitourinary dysmorphisms have never been reported in the context of CCD or King-Denborough syndrome. Furthermore, patients with pathological *RYR1* usually present with additional muscle and soft tissue involvement including neonatal hypotonia, muscle atrophy and muscle weakness in CCD or muscle rigidity and rhabdomyolysis in King-Denborough syndrome[392]. These phenotypes are absent in the proband *CL022* and therefore, given the inconsistency of symptoms between the patient's main features and the symptoms associated with the *RYR1* mutations, it is unlikely that the *RYR1*:c.2543C>T is aetiological in the context of the disease.

Heterozygous mutations of *PIK3R1* are identified to underlie the autosomal dominant SHORT syndrome (OMIM # 269880). *PIK3R1* encodes the phosphatidylinositol 3-kinase that plays an important role in growth signalling pathways. This lipid kinase phosphorylates the inositol ring of phosphatidylinositol and thereby triggers the second messenger in the insulin signalling pathway[393]. Patients with the SHORT syndrome are present with short stature, triangular faces, micrognathia, telecanthus and deep-set eyes[394]. Bone maturation and teething are usually delayed among SHORT patients, but patients are intellectually normal[395]. CLP and genitourinary abnormalities including shawl scrotum and cryptorchidism have never been reported in the context of the SHORT syndrome. Despite the presence of telecanthus and short stature in the proband *CL022*, the phenotypic similarity between the SHORT syndrome and patients main features are negligible and therefore the *PIK3R1*:c.235A>G is unlikely to be relevant to the patient's phenotype.

As for the remaining variants including variants in *SOX10 KMT2D*, *SOX9*, *SMARCA4* and *IQSEC2* genes, visual inspection of reads in IGV revealed unreliable variant call; therefore, these variants were excluded from further follow up.

In conclusion, neither of the variants shortlisted for the patient *CL022* were proved unequivocally to be causal, and molecular diagnosis in this patient remained unresolved.

CL025

The patient is a 3-year-old male child of a non-consanguineous marriage with a negative history of CLP in the family. The patient's main features include developmental

delay (HP:0011342), unilateral cleft lip (HP:0100333), hypertelorism (HP:0000316), telecanthus (HP:0000506), shawl scrotum (HP:0000049) and cryptorchidism (HP:0000028). The patient is also present with anophthalmia (HP:0000528) and esophageal atresia (HP:0002032) (Table 5.2 and Figure 5.4). Fifteen novel variants were prioritised through the filtering strategy, of which 2 variants including *CREBBP*:c.7037A>C (rank 4) and *OFD1*:c.2610G>C (rank 7) were identified as putatively causal with respect to the patient's phenotype (Table 5.5).

(a) (b)

Figure 5.4: Patient *CL025* dysmorphic facial features; Low-set ears, anophthalmia and unilateral cleft lip are prominent features.

The *CREBBP* encodes the CREB-binding protein with intrinsic histone acetyltransferase activity. This protein act as a stabilising scaffold for other proteins in the transcription complex and plays a vital role in transcriptional co-activation of additional transcript factors involve in homeostasis, growth control and embryonic development [396,397]. Heterozygous mutations of *CREBBP* have been identified to underlie autosomal dominant Rubinstein-Taybi syndrome 1 (RSTS1, # 180849). Patients with pathogenic *CREBBP* mutations present with multiple congenital anomalies including dysmorphic facial features, broad and angulated thumbs, postnatal growth deficiency, shawl scrotum and cryptorchidism [398]. Facial dysmorphisms in RSTS1 patients includes downslanted palpebral fissures and highly arched palate. Occasionally cleft lip present as part of syndrome [399], but anophthalmia has never been reported in the context of the disease. Given the lack of convincing phenotypic similarity between the patient's main features and RSTS1, it seems unlikely that *CREBBP*:c.7037A>C underlies the patient's condition.

Heterozygous mutations of *OFD1* have been identified to underlie the orofaciodigital syndrome 1 (OFD1, # 311200) with an X-linked dominant pattern of inheritance. *OFD1* encodes a centrosomal protein that localise at the basal body of primary cilia and plays an important role in the structure of primary cilia and left-right symmetry [400]. OFD1 patients are typically present with a range of malformations in the face, hand and oral cavity. Main features of the disease include short stature, cleft lip and cleft palate, hypertelorism, telecanthus and polycystic kidney disease [401,402]. Shawl scrotum and cryptorchidism have been rarely reported in the context of the disease as pathogenic *OFD1* mutations lead to

lethality in hemizygous males[402]. Intriguingly, Tsurusaki *et al.*[403] reported a lethal case of a hemizygous male in a family ascertained for the *OFD1* mutation with preterm delivery (33 weeks). The patient presented with cleft lip, soft cleft palate, hypertelorism, microphthalmia, micropenis and cryptorchidism. Tsurusaki *et al.* did not report oesophagal atresia in the patient but reported that the patient died 11 hours after birth due to laryngeal anomalies. Genetic analysis of another male patient from the same family with similar phenotype and oesophagal abnormalities who were also delivered preterm and died one day after birth revealed a pathogenic splicing mutation at *OFD1*:c.2388+1G>C[403]. Given the extensive phenotypic similarity between the patients described by Tsurusaki *et al.* and the proband *CL025* main features it is possible that *OFD1*:c.2610G>C (Ψ_i=4.5) underlies the patient's condition. One possible explanation for the lack of lethality in the proband *CL025* could be the mosaic nature of the mutation in the patient. In fact, the *OFD1*:c.2610G>C variant is supported by 8 reads out of total read depth of 42 suggesting 19% mosaicism. Should mosaicism underlie the reduced penetrance in the patient, a post-zygotic mutation of *OFD1* could be accounted for the patient's phenotype. In view of this, functional impact of *OFD1*:c.2610G>C merits further investigation.

CL027

The patient is a 9-year-old boy of a consanguineous marriage where parents are first cousins and the patient's third-degree cousin present with CLP. The patient is developmentally normal with his weight and height percentiles are in the normal range. The patient's main features include unilateral cleft lip (HP:0100333), complete cleft palate (HP:0000175), hypertelorism (HP:0000316) and shawl scrotum (HP:0000049) (Table 5.2). Three heterozygous novel variants including *NAGLU*:c.311C>A (rank 1), *TGFB3*:c.785G>T (rank 2) and *KMT2D*:c.1372A>C (rank 3) were shortlisted by the filtering strategy (Table 5.5).

Heterozygous mutations of *NAGLU* have been identified to underlie Charcot-Marie-Tooth disease type 2V (CMT2V, OMIM # 616491). Patients with pathogenic *NAGLU* mutations present a range of neurologic conditions including peripheral neuropathy and sensory impairment. The disease has a late age of onset and usually occurs between 18 and 61 years of age[404]. Considering the lack of phenotypic similarity between the patient *CL027* main features and CMT2V, it is unlikely that *NAGLU*:c.311C>A (Ψ_i=8) underlies the patient's condition.

Heterozygous mutations of *TGFB3* have been linked to two autosomal dominant conditions known as Arrhythmogenic right ventricular dysplasia 1 (ARVD1, OMIM # 107970) and Loeys-Dietz syndrome 5 (LDS5, # 615582). Clinical presentation in ARVD1 is restricted to cardiovascular complications[405] whereas, in LDS5, patients present cleft palate in combination with a range of cardiovascular complications including aortic aneurysms[406]. Considering that the proband *CL027* does not present with cardiovascular complications and given the lack of phenotypic similarity to the syndromes described above, it is very unlikely that the *TGFB3*:c.785G>T (Ψ_i=5.5) is causal in the context of the patient's condition.

Visual inspection of reads covering *KMT2D*:c.1372A>C variant revealed an erroneous alignment at this position, and therefore this variant was excluded from further follow-up. Taken together, variant prioritised for the proband *CL027* did not provide convincing evidence for establishing the causal relationship, and therefore molecular diagnosis remained unresolved for this patient.

CL033

The proband is a 10-year-old male patient from a non-consanguineous marriage. The patient present with normal development, unilateral cleft lip (HP:0100333), complete cleft

palate (HP:0000175), hypertelorism (HP:0000316), telecanthus (HP:0000506), shawl scrotum (HP:0000049) and bilateral clinodactyly of the 5th finger (HP:0004209) (Table 5.5). Tiered filtering analysis identified seven novel variants (six non synonymous and one stop-gain) in genes including *PTEN*, *NIPBL*, *KIRREL3*, *TSC1*, *KMT2D* and *IQSEC2* (Table 5.5). Visual inspection of read alignments for shortlisted variants in IGV revealed that only the *NIPBL*:c.7801A>G (rank 2) and *KIRREL3*:c.544A>T (rank 3) appear to be a genuine call. The remaining variants, therefore, were discarded from further investigation.

The *NIPBL* on the short arm of chromosome 5 (5p13) encodes components of enhancer-promoter cohesion complex [407]. Heterozygous mutations of *NIPBL* are identified to underlie autosomal dominant Cornelia de Lange syndrome-1 (CDLS1, OMIM # 122470) [408]. A wide clinical variability for CDLS1 has been described. The CDLS1 patients typically present with multisystem malformations including facial dysmorphisms, upper limb anomalies, developmental delay and mental retardation [409]. Twenty per cent of children with the CDLS1 present with cleft palate [410] and upper limb anomalies and the fifth finger clinodactyly is reported in 74% of cases [411]. Genitourinary anomalies including hypoplastic genitalia and cryptorchidism are fairly common among CDLS1 patients and are reported in 57% and 73% of the male patients respectively [412,411]. Lalatta *et al.* [413] reported hypertelorism in a neonatal case of CDLS1 with novel mutation at exon 35 of *NIPBL*. The wide variability in the clinical presentations of the disease renders its diagnosis challenging. Severe forms of CDLS1 are readily diagnosed at an early age (>90% below the age of 2), but diagnosis becomes more challenging as patients grow older [409]. Establishing diagnosis in the milder form of CDLS1 is even more challenging as patients do not express the characteristic facial appearance of the disease until 2 to 3 years age. The facial features of CDLS1 in patients with milder form gradually fade as patients grow older and it becomes less significant after age 9 [414]. Given the variability in the clinical expression of the disease and extensive phenotypic similarity between the proband's main features and CDLS1, it is highly likely that non-synonymous *NIPBL*:c.7801A>G on exon 45 underlies the patient's condition. To date, eight novel variants on exon 45 of *NIPBL* have been reported in the context of CDLS1. This includes 2 exonic deletions [415,416], 3 duplications [415,416], 1 non-synonymous [417] and 2 splicing variants [418,415]. Phenotypic detail for the patient with non-synonymous *NIPBL*.c.7849C>T is not reported by Bhuiyan *et al.* [417], but through investigation of 39 CDLS1 patients, they concluded that severity of developmental delay is considerably diminished in patients with missense mutations when compared to patients with truncating mutations. This observation is compatible with normal developmental in the patient *CL035*.

Heterozygous mutations of *KIRREL3* have been described to underlie autosomal dominant mental retardation 4 (MRD4, OMIM # 612581) [419]. Given normal development and lack of mental disability in the proband *CL033*, it is implausible that this variant is relevant to the patient phenotype.

Taken together, the variant analysis in the patient *CL033* proposed a highly likely causal variant at the exon 45 of the *NIPBL* gene. Functional impact of *NIPBL*:c.7801A>G merits further investigation.

CL035

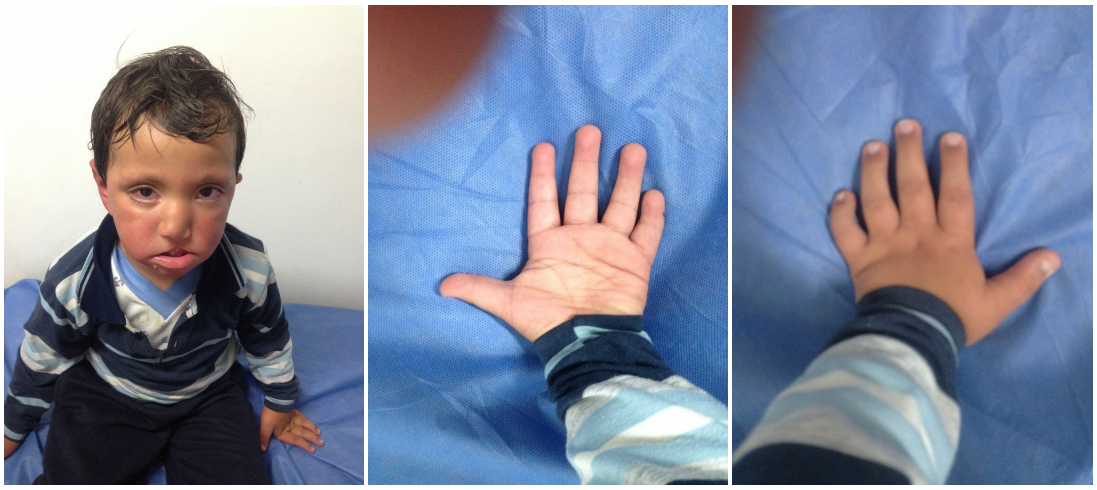
The patient is an 8-month-old infant of a non-consanguineous marriage. The proband's main features include developmental delay (HP:0011342), unilateral cleft lip (HP:0100333) and complete cleft palate (HP:0000175) (Table 5.5). The phenotypic features reported for this patient are very broad and unspecific and therefore establishing a molecular diagnosis is challenging. Four variants, including *COL1A1*:c.433G>C (rank 1), *COMT*:c.476T>G (rank 2), *GRIN2A*:c.136G>A (rank 3) and *GUCY2D*:c.92G>C (rank 4) were identified through variant analysis (Table 5.5). Visual inspection of read alignments in IGV revealed

that only the *GRIN2A*:c.136G>A is a reliable call, concordant with its high genotype quality (GQ= 99) and high read-depth (DP= 249).

The Glutamate Ionotropic Receptor NMDA Type Subunit 2A (*GRIN2A*) encodes a type of glutamate-gated ion channel which is involved in the regulation of synaptic transmission[420]. Mutations in *GRIN2A* have been identified to cause variable neurodevelopmental phenotypes including the autosomal dominant focal epilepsy with speech disorder (FESD, # 245570). Endele *et al.*[421] reported two patients with mutation in *GRIN2A* who present with facial dysmorphisms and neurodevelopmental delay. The facial characteristic of one patient with translocation breakpoint disrupting GRIN2A is restricted to the short nose, but the nature of facial dysmorphism in another patient with non-synonymous *GRIN2A*:c.652C>T mutation is not fully described[421]. Considering incomplete penetrance and variable phenotype reported in the context of *GRIN2A* mutations[422,423], it is possible that identified non-synonymous mutation at *GRIN2A*: c.136G>A (Ψ_i= 4) is causal in the patient. FESD patients usually present childhood-onset seizures that may last until adulthood[424]. Considering that phenotyping in the proband *CL035* was carried out during infancy further neuropsychological assessment is required to establish a diagnosis.

CL036

The patient is a five years old boy of a non-consanguineous marriage present with developmental delay (HP:0011342), unilateral cleft lip (HP:0100333), complete cleft palate (HP:0000175), hypertelorism (HP:0000316), telecanthus (HP:0000506), shawl scrotum (HP:0000049), cryptorchidism (HP:0000028), clinodactyly of the 5th finger (HP:0030084) and heart murmur (HP:0030148) (Table 5.2 and Figure 5.5). The patient had a low height, and weight percentile for his age and proband's maternal cousin was also present with CLP. Five novel variants were identified as putatively pathogenic (Table 5.5). Investigation of genotype quality scores in parallel with visual inspection of read alignments for shortlisted variants revealed that *SMARCA4*:c.889A>C (rank 4 Ψ_i= 6, DP = 27, GQ = 31) and *ARID1B*:c.293A>C (rank 5, Ψ_i= 3, DP = 22, GQ = 28) are probably spurious calls and therefore they were excluded from further analysis. The genotype quality score for the non-synonymous *PTPN11*:c.323C>A (rank 1) variant was also very low (GQ = 28) and considering the low depth of coverage at this site (DP = 11) a reliable basis for a heterozygous call could not be established.



(a) (b) (c)

Figure 5.5: Patient *CL036* dysmorphic features; (a) Unilateral cleft lip, hypertelorism and telecanthus are the prominent dysmorphic facial features; (b & c) Cone shaped fingertips, brachydactyly and clinodactyly of the 5th finger are the salient dysmorphic features in the patient's hand.

Heterozygous mutations of *PTPN11* have been identified to underlie two autosomal dominant conditions known as LEOPARD syndrome 1 (LPRD1) (OMIM # 151100)[425] and Noonan syndrome 1 (NS1) (OMIM # 163950)[426]. Both syndromes exhibit extensive phenotypic similarity with the patient's main features. The identified variant maps to exon 3 of the gene, and several studies reported pulmonic stenosis murmur in association with Noonan main features in the context of *PTPN11* exon three mutations[427,428]. Despite deleteriousness of *PTPN11*:c.323C>A variant (rank 1, Ψ_i = 7.5), read depth for this position is quite low (DP = 11) and therefore a reliable ground for establishing pathogenesis is lacking.

Two variants including *KMT2D*: c.16493C>T (rank 2, Ψ_i = 7) and *BRCA2*: c.9425A>T (rank 3, Ψ_i = 7) were identified to have high read depth and high genotype quality (DP > 80 and GQ = 99) for confident variant analysis.

The *KMT2D* gene encodes a histone methyltransferase that is an important component of ASCOM complex which has been shown to regulate the transcriptional activity of beta-globin and estrogen receptor genes[429,430]. Heterozygous mutations of *KMT2D* are identified to underlie autosomal dominant Kabuki syndrome 1 (KABUK1, OMIM

147920)[431]. Individuals with KABUK1 syndrome present with multiple congenital anomalies and characteristic facial features that typically include highly arched and sparse eyebrows, depressed nasal tip, short columella and large earlobes[432]. Cleft lip and/or cleft palate have been reported in 35% of cases[386]. Genitourinary abnormalities including cryptorchidism and micropenis have been observed in approximately 25% and 10% of male patients, respectively[433,386]. Clinodactyly of the fifth digit is one of the major skeletal anomalies reported in 50% of patients[433]. Additional phenotypes of the patient that are consistent with KABUK1 syndrome include developmental delay and, heart murmur. Developmental delay and variable degrees of mental retardation have been described in the context of KABUK1 syndrome[434]. Mild and moderate developmental delay collectively present in 67% of Kabuki patients[435] and cardiovascular anomalies have been reported in 42% of cases[433]. The cause of heart murmur in the proband *CL036* is not defined, but common heart defects (including Atrial septal defect (ASD), Ventricular septal defect (VSD) and Patent ductus arteriosus (PDA)) that result in childhood heart murmur have been identified in 90.6% of Kabuki patients diagnosed prenatally or at an early age[436]. Given the extensive phenotypic similarity between the patient's main features and KABUK1 syndrome and, also high combined rank score for the *KMT2D*: c.16493C>T mutation (rank 2, $\Psi_i = 7$), it is possible that this variant underlies the patient condition. Evaluation of the phenotypic features, revealed from patients photos, by three independent clinical experts at the Southampton University Hospital (Professor Karen Temple, Dr Nicola Foulds and Dr Katherine Lachlan) strongly supported the diagnosis of Kabuki syndrome.

The *BRCA2* gene was primarily included in our tiered gene list since transcriptional dysregulation of *BRCA2* are identified to associate with susceptibility to DNA damage and non-syndromic CLP[437]. Given the absence of direct link between *BRCA2* mutations and the patient's main features, it is possible that the *BRCA2*: c.9425A>T is an incidental finding without any implication on the patient's phenotype. Although this variant appears unlikely to have any impact on patient's health, because of the *BRCA2* role in familial Breast and ovarian cancer (OMIM #600185), the identified mutation might have a clinical significance for female relatives of the patient. Therefore in compliance with the ACMG guideline for reporting secondary findings[438], it has been brought to the attention of recruiting clinician as an incidental finding.

CL039

The patient is an 8-year-old boy from a non-consanguineous parent. The proband's sister presented with microtia, and there is no history of CLP in his family. The patient's main features include hyperactivity disorder (HP:0000752), unilateral cleft lip (HP:0100333), complete cleft palate (HP:0000175), hypertelorism (HP:0000316) and shawl scrotum (HP:0000049) (Table 5.2). Variant analysis shortlisted six novel mutations in genes including *CREBBP*, *OFD1*, *TBX4*, *ATR*, *ANKRD11* and *SALL4* (Table 5.5). Inspection of read alignments in IGV in parallel with the investigation of genotype quality scores revealed that only *ATR*:c.3025G>A (rank 4, $\Psi_i = 2.5$, DP = 56, GQ= 99) appears to be a reliable call. The remaining variants may reflect alignment errors and therefore were excluded from further analysis.

Heterozygous mutations of the *ATR* gene have been identified to underlie autosomal dominant cutaneous telangiectasia and cancer syndrome (FCTCS, OMIM # 614564)[439]. Patients with pathogenic heterozygous mutations of *ATR* present with cutaneous telangiectasia and dispersed alopecia during infancy[440]. Given the lack of reported telangiectasies complications in the proband *CL039*, it is unlikely that heterozygous *ATR*:c.3025G>A is related to the patient's condition.

Homozygous or compound heterozygous mutations of ATR have been reported to cause

autosomal recessive Seckel syndrome 1 (SCKL1, OMIM # 210600). Among phenotypic features reported for the proband *CL039*, cleft lip/cleft palate and hyperactivity have also been documented in SCKL1 patients[441,442,443]. Considering that no other non-synonymous mutation was identified in the ATR gene, the patient's condition could not be linked to Seckel syndrome 1. Taken together variant analysis in the patient *CL039* failed to identify the causal variant and molecular diagnosis in this patient remained unresolved.

5.4 Discussion

Even though molecularly confirmed cases of AAS have been entirely attributed to the pathogenic mutations of *FGD1*, we did not identify any variant with the pathogenic implication in this gene across the samples analysed. Failure to identify *FGD1* mutations could be attributed to several reasons that are discussed below.

Aarskog-Scott syndrome is a rare disorder with extensive phenotypic similarity with other rare phenotypes (Table 5.1 and section 5.1.1). Genotype-phenotype correlation of the disease is not fully understood, and molecular diagnosis in only 20% of cases is resolved [347]. Differential diagnosis when molecular aetiology of the disease is inconclusive proved to be challenging. Given the overlapping nature of facial characteristics in rare disorders [444], the possibility of alternative diagnoses compatible with molecular findings must be considered. Because of this, it is possible that variants identified in the five patients might genuinely be related to the patients' phenotype and underlie their condition.

Secondly, the TruSight One capture kit does not provide uniform capture across the 18 exons of the *FGD1* gene. To investigate whether the failure to identify *FGD1* mutation in this study is due to sample quality or inherent deficiency of TruSight One kit for capturing *FGD1* coding region, normalised read counts across the gene for the nine samples that passed alignment QC compared to 18 controls from the same sequencing batch. Although coverage efficiency significantly differs between exons of the *FGD1* ($p< 0.001$, One-way ANNOVA), capture coverage of the gene for AAS samples is not significantly different between the AAS cases and controls ($p= 0.9998$, One-tailed t-test) (Figure 5.6). Therefore a possible explanation for the failure to identify *FGD1* mutations can be the limitations of TruSight One kit for capturing *FGD1* exons. In particular exons, 5, 10 and 16 were identified to have low capture coverage in the TruSight One kit.

Investigation of mutational spectrum in the coding region of *FGD1* revealed that the highest number of novel SNVs map to the exon six but the highest density of mutational events (*i.e.* number of SNVs per length of the exon) occurs at the exon 13 (Figure 5.7). The depth of coverage at both exons was sufficient for variant calling in our study. Only one variant have been identified on each of the exon 5, and exon 10 and no variant have been reported in the exon 16 (Figure 5.7). Total number of SNVs identified in these three exons constitute a mere 5.74% of the total mutational density across the coding length of *FGD1* gene, and therefore even with a reduced capture in these exons, we have still been able to robustly screen for pathologic mutations in the remaining exons with theoretical 94.25% mutational density. However, given the modest statistical power of targeted exome sequencing for identifying *de-novo* heterozygous variants[445], the implication of low capture coverage in these exons for detecting *FGD1* variants merits further investigation.

Thirdly, the inherent deficiency of targeted exome sequencing in capturing sequences outside the coding region might also explain the lack of *FGD1* positive cases in our study. The *FGD1* intronic variants (including splicing SNVs) have been reported to underlie AAS[447,347]. Considering that targeted exome sequencing is ill-suited for identifying variants outside the coding region, we might have missed pathogenic variants in the intronic region.

Finally, the mean insert size for the nine samples analysed was ∼143 bp. As discussed

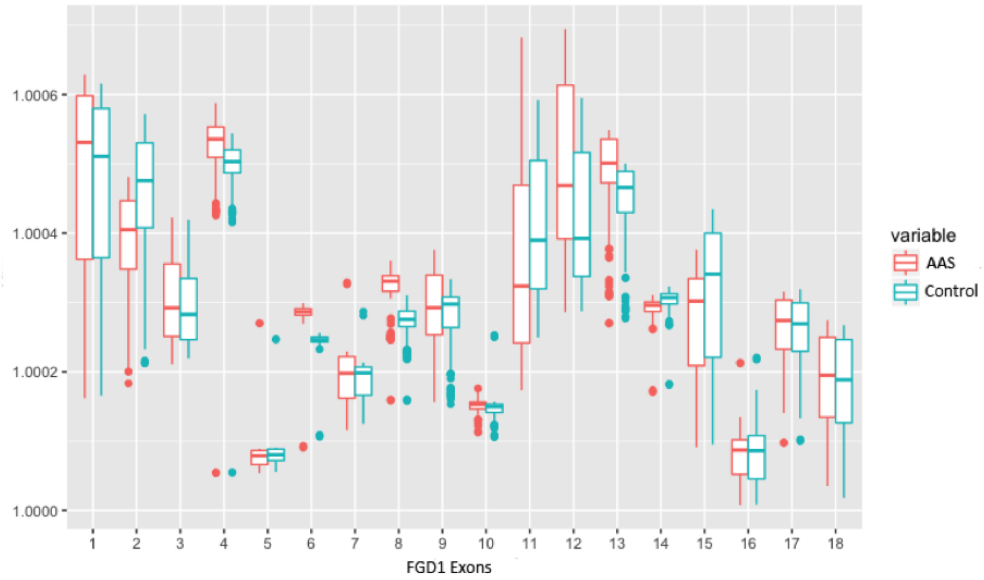


Figure 5.6: Normalised coverage across the *FGD1* in AAS samples (red) compared to 18 controls (cyan). Regardless of variable coverage efficiency across different exons of *FGD1* ((p< 0.001, One-way ANNOVA)), capture coverage among AAS patients is not significantly different from the control samples that were exomed concomitantly on the same dispatch DNA plate (p= 0.9998, One-tailed t-test). In particular, the capture efficiency across exon 5, exon 10 and exon 16 revealed to be significantly lower than the mean coverage for the remaining exons of the gene (P-values for independent groups t-test is consistently <0.01).

in section 5.3.1, the distribution of insert sizes is positively skewed. Low quality DNA or reduced mass input of gDNA changes the distribution of library fragments and result in smaller insert sizes than expected [448]. These smaller fragments are prone to exclusion during the subsequent clean-up step and result in a skewed fragment distribution. In view of this, skewed insert length clearly explains the unevenness in coverage distribution in our samples, especially for the *FGD1* gene. In addition, larger insert lengths ($\geq$ 170) are identified to provide a more uniform coverage distribution across the genes targeted and reduce the false negative rate in the variant detection by two fold [449]. Although larger read length is achieved at the expense of reduced depth, evenness in the coverage dramatically increases the diagnostic power of exome sequencing [449]. The average size of exons in the human genome is 170 bp, and 80-85% of exons are identified to be <200 bp in size [140,450]. Consequently, read lengths above 170bp provide an ideal size distribution for capturing not only exonic variants but also intronic variants that have implications in splicing. Because of this, skewed read length size in our study might provide an explanation for the reduced diagnostic power to identify the *FGD1* mutations in our samples. Given uneven coverage across the exons of *FGD1* in the TruSigth one panel and the reduced insert sizes in this study, screening of the *FGD1* gene in higher uniform coverage must be prioritised.

While rare disorders are believed to be primarily caused by a single penetrant protein truncating mutation, there are cases in which patients with the same genetic defect present different phenotypes [451]. This variable expressivity is generally attributed to the role of modifier genes that collectively impact the penetrance, dominance, and expressivity of the mutation [452]. Besides, many rare disorders share extensive phenotypic similarity. The lack of granularity in phenotypic manifestations of rare disorders is particularly problematic when it comes to molecular diagnosis. Also there are occasions in which mutations in different genes can result in similar phenotypes [453]. The issue pertaining to the lack of consistency in phenotypic manifestations is more pronounced in the context of neurodevelopmental disorders [454], mainly because of a large number of genes identified to implicate

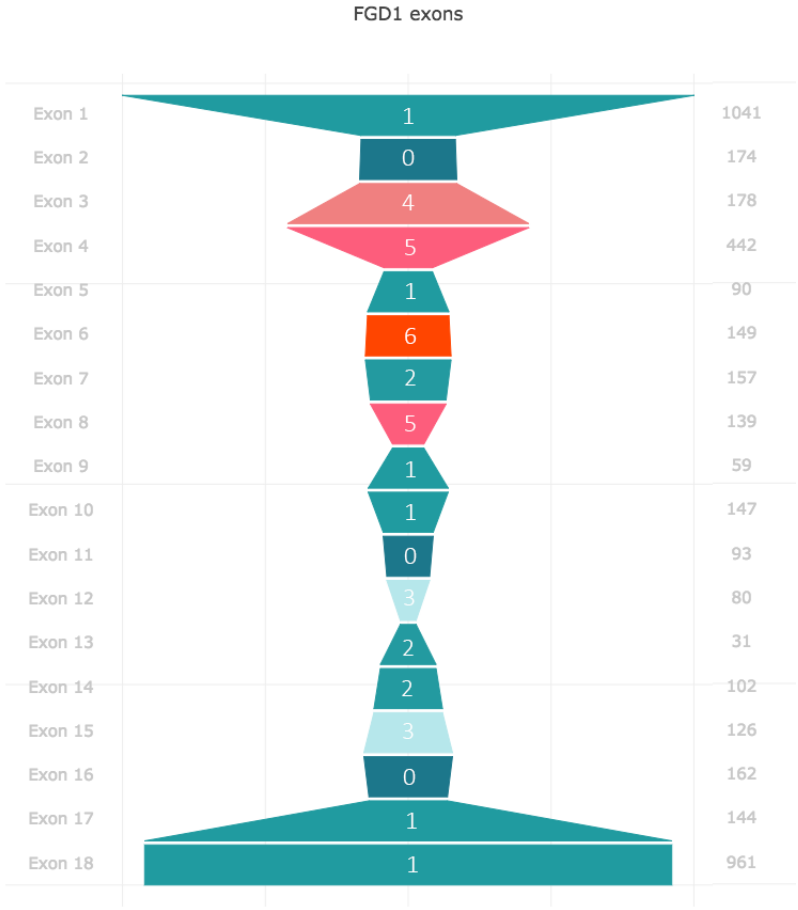


Figure 5.7: Frequency and localisation of 38 novel SNVs identified in the coding region of the *FGD1* gene. The vertical axis on the left identifies the exon number and the numbers on the right axis represent the size of each exon in bp. Trapezoids represent the relative size of exons and the numbers inside them show the total number of novel SNVs identified in each exon. (Data extracted from the LOVD 3.0 FGD1 database [446]). This shows that exons 3,4,6 and eight collectively account for ∼53% novel variants identified across the gene and appear to have sufficient coverage in our analysis.

in the disease pathogenicity [381]. In the absence of detailed phenotypic information, similarities between patients cannot be statistically tested, and the power of NGS analysis is reduced. Obtaining complete and update phenotypic information from patients residing in remote areas in Colombia presented a significant challenge in this study. Indeed, mixed phenotypic information with a variable degree of details and completeness appeared to be the main limitation for molecular diagnosis in these patients. Despite the provisional diagnosis of AAS, many features of the patients are unspecific to AAS. Clinically, this implicates that patients studied here perhaps present a heterogeneous group of disorders that share some phenotypic similarities including CLP, shawl scrotum and hypertelorism (the three most frequent features in the cohort). Interestingly, identification of plausible mutations in five individuals (out of the eight patients satisfied QC criteria) presents a strong case for alternative diagnosis across these patients.

To summarise, we applied targeted exome sequencing to 13 patients with the primary diagnosis of Aarskog-Scott syndrome. Five samples (CL024, CL028, CL030, CL037 and CL040) were discarded from variant analysis early on as they did not comply with the

quality standards required for confident variant analysis. Investigation of novel variants in the remaining eight samples resulted in the identification of putative causal variants for five samples including CL021, CL025, CL033, CL035 and CL036 (Table 5.6).

An important finding from this analysis is the Kabuki syndrome (KS) diagnosis in the patient *CL036*. Having reviewed the findings with our local clinical experts, we confirmed the Kabuki diagnosis in this patient. While KS was believed to have a higher incidence in Japanese populations, recent findings in other ethnic groups have extended the phenotypic spectrum of the disease and revealed new pathological mutations [455,456].

Establishing molecular diagnosis in the remaining patients requires further follow-up including in-depth investigation of phenotypic details that may corroborate the molecular finding from WES analysis.

Sample	**Variant**	**Exon**	Ψ_i	**Depth**	**GQ**
CL021	SRCAP:c.A8036T	34	5	166	99
CL025	OFD1:c.G2610C	20	4.5	42	36
CL033	NIPBL:c.A7801G	45	3.5	148	99
CL035	GRIN2A:c.G136A	2	4	252	99
CL036	KMT2D:c.C16493T	53	7	119	99

Table 5.6: Summary of putatively causal variants identified in five samples (ψ_i: Combined rank score; GQ: Genotype quality score)

Chapter 6

Fine-scale characterisation of LD-structure in sub-Saharan African population: Functional overrepresentation analysis and relationship to the disease genome

6.1 Introduction

One of the fundamental aspects of sexual reproduction is recombination in which breakdown of linked loci enables haplotype shuffling throughout the genome. Activation of a homologous recombination repair mechanism upon a double-strand break in DNA, during meiosis I, leads to the breakdown of linked loci and enables new combinations of alleles to be produced. Conversely, recombination reduces the level of disequilibrium between linked loci and results in rapid shrinkage of linkage disequilibrium (LD) in genomic regions near recombination hotspots. As a result, the genome-wide pattern of LD is primarily determined by the local recombination rate[457]; however, it is well established that other evolutionary forces such as selection, mutation and genetic drift also play an important role in shaping the LD pattern at the genome level[458,459]. It has been shown that characteristics of the neighbouring sequence features are also influencing the extent of LD at local domains. For example, it is suggested that factors such as GC content, gene density and the presence and the type of interspersed repeats correlates with the extent of LD at genome level[460,461,457]. The higher recombination rate across the GC rich domain of the genome leads to the breakdown of LD whereas selective sweeps generally increase the extent of LD. Sequence changes in the coding part of the genome, brought about by recombination and random mutation, are more likely to influence the fitness of the species and therefore are subject to a more intense impact of natural selection. Consequently selective forces, by limiting the frequency of mutational events at the coding regions, result in elevation of LD in functional parts of the genome[462]. Non-genomic features including population size and historical patterns of migration have been also suggested to influence the LD pattern. While population bottlenecks and genetic drifts generally result in increase in LD, rapid population expansion and non-selective mating reduces the extent of LD[458,463].

According to Hill–Robertson interference (HRI), recombination plays a substantial role in fixation of individual advantageous alleles arising from the random mutation in the population[464,465]. HRI suggests that in the absence of initial linkage disequilibrium

(LD) between two alleles arising from new mutations in a finite population, fixation of either allele is effectively hampered by the competitive fitness of the other allele. In situations where there is no recombination, co-inheritance of both alleles is only feasible when the second allele arises in a progeny inheriting the first allele. In the absence of recombination, the selection at one locus effectively reduces the fitness of the competing allele at another locus (on a separate chromosome) and neither alleles can reach fixation in the population. Given that efficacy of selection at independent loci is largely defined by the population effective size (N_e) and the selection coefficient, in a finite population, loci under selection inevitably interfere with each other's fixation. Recombination, by creating haplotypes carrying both favourable alleles, enables fixation of advantageous mutations in the population. Conversely, in weakly recombining regions of the genome, such haplotype diversity is less frequent, and therefore the efficacy of selection is diminished. As suggested by the Muller's ratchet effect [466,467], non-recombining regions of the genome tend to accumulate deleterious mutations. In such regions, like the mammalian Y chromosome, the absence of recombination hampers regeneration of mutation-free haplotypes and consequently results in excessive accumulation of deleterious mutations, which ultimately lead to gene inactivation. Given the importance of recombination in maintaining genome plasticity against selective sweeps and also in view of highly variable recombination rate across the genome, delineation of the pattern of recombination may reflect the underlying distribution of disease-associated regions across the genome.

Historical recombination events are predominantly reflected in LD patterns. While other evolutionary forces such as selection and mutation are also involved in shaping the LD structures, recombination is known as the primary force in shaping the LD pattern across the genome [59]. Recombination and mutation, by breaking-up the LD across the linked region of the genome, act in favour of increased haplotype diversity, whereas selective sweeps act toward increasing LD across the genome. Delineation of LD patterns directly demonstrates the historical impact of recombination and provides insights into the regions of the genome that have undergone selection. Deciphering selected regions across the genome is of particular interest for evolutionary studies and disease gene identification [468].

The striking concordance between the linkage maps generated from quantification of meiotic recombination over a few generations and LD maps underlies the impact of historical recombination as the major force in the determination of LD structure. The fact that contours of LD maps across different populations demonstrate highly concordant patterns confirms that LD structures are broadly conserved across different populations.

Previous efforts into delineation of LD structures across different populations made cost-effective genome-wide association studies feasible through the development of genotyping arrays consisting of 'tag' SNPs [469]. LD studies across different populations reveal underlying population structure and migration histories involved in shaping the LD patterns [470,471] and identify the sequence context underlying recombination hotspots [472,473]. Myers *et al.* [473] showed that recombination hotspots are enriched around 13-mer degenerate motifs, which play a key role in recruiting crossover machinery across the human genome. These motifs are identified to define binding arrays for PRDM9 [474] which regulates activation of recombination hotspots. Binding of PRDM9 to specific DNA sequence targets initiates methylation of lysine 4 at histone 3 (H3K4me3) and thereby triggers recombination by recruiting homologous recombination repair machinery [475].

LD maps generated from the whole-genome sequence (WGS) data achieve an unprecedented resolution for delineation of LD structures. Using WGS data from the HapMap project (Phase 3) Pengelly *et al.* demonstrated a ∼2.8 fold increase in the resolution of recombination hotspots across human chromosome 22 [82]. High-resolution maps of the LD structure across individual genes provide a deeper insight into the evolutionary forces that operate at the gene level.

6.1.1 Linkage disequilibrium maps

Efforts into delineation of LD patterns in the human genome to aid association mapping, identification of recombination hotspots and infer selective sweeps, resulted in a flurry of diverse metrics to measure LD across single-nucleotide polymorphisms (SNPs). Among these metrics (including covariance (D), correlation (r), regression (b) and frequency difference) association was demonstrated to be the most effective metric for modelling SNPs relationships across the length of the linked loci[476]. LD maps based on association are derived from the Malécot-Morton model in which the probability of association between SNPs (ρ) is computed according to Equation 6.1.

$$\rho = (1 - L)Me^{-\varepsilon d} + L \tag{6.1}$$

In this equation 'L' denotes the residual association at a large distance, 'M' denotes the probability of association at zero distance, 'ϵ' describes the exponential decay of association with physical distance and 'd' denotes the distance between each pair of SNPs in kilobases[477]. Collins *et al.* demonstrated that LD hotspots are discernible by estimating ε in the Malécot model[478]. The first LD map based on this model demonstrated that haplotype blocks across the human genome are broken by sharp steps that align with recombination hotspots. In these maps, additive physical distances across the chromosome arms are expressed in linkage disequilibrium units (LDU) in which the LDU distance between the ιth pair of SNPs is equated by $\epsilon_\iota d_\iota$[479]. The product ϵd also demonstrates the number of generations after an effective bottleneck over which LD has accumulated[480]. The ϵd is also equivalent to θt (θ denotes the frequency of recombination and t is the number of generation over which recombination has accumulated). Although ϵd is primarily a function of recombination and time, it is also informed by the effect of mutation, selection, and other evolutionary forces. Therefore in contrast to coalescent based maps that exclude the impact of processes other than recombination, LDU maps reflect the collective impact of mutation, selection and population history[481,482].

Zhang *et al.*[479] demonstrated a close correspondence between the steps in LDU map with recombination hotspots identified by Jeffreys *et al.* through sperm typing of class-II HLA complex[472]. At the same time, they identified that plateau regions in LDU maps correspond closely to the regions of low haplotype diversity identified by Daly *et al.*[60].

The LDU scale provides a framework that enables investigation of LD extent in different genomic regions irrespective of physical distance (*i.e.* the Kb scale). In LDU maps, one LDU corresponds to a variable physical distance across the chromosome over which LD declines to its 'background' levels. In another term, the distance in Kb over which LD has declined to e^{-1} ~0.37 of its starting value corresponds to one LDU. In that sense genomic locations that are greater than one LDU apart are effectively LD independent.

As discussed earlier, since the LDU framework accounts for additional evolutionary forces other than recombination (including genetic drift, gene conversion, selection and mutation), although contours of population-specific LD maps are highly concordant[483], the magnitude of the LDU scale is different among different populations. In fact the LDU scale across different populations is directly correlated with the duration of respective population since an effective bottleneck. In contrast to isolated populations that have a more extensive LD, the sub-Saharan African population (SSA) due to their extended population history have less extensive LD. Erosion of LD during the time in SSA population enables delineation of LD structures at a higher resolution which directly result in an enhanced identification of genomic regions under selection. Furthermore, fully saturated LD maps from the WGS data of SSA population improve our understanding about human population differences and reveals key regions of the genome that have been under selection during evolution.

6.1.2 Population structure in SSA

Recent studies into the population structure across the SSA population has identified a modest difference amongst SSA groups and revealed that population differentiation across the sub-Saharan region is predominantly influenced by the ethnolinguistic grouping across this region[87]. Comparison of fixation index (F_{ST}) between different SSA populations revealed a minimal interpolation difference among the Niger-Congo language groups (F_{ST}= 0.009)[87]. This demonstrates that recent population expansion and migration across the sub-Saharan region started around 3,000 to 5,000 years ago from West Africa and had a deterministic impact in shaping the population structure across SSA region[484].

In addition to continental dispersion and ethnolinguistic subdivisions that had a primary role in population differentiation across SSA, Eurasiatic gene flow has been revealed to be a dominant factor in residual divergence between the SSA populations[87]. Population admixture analysis identified a widespread Eurasian and hunter-gatherer admixture events across all SSA populations revealing the highest proportion of Eurasian ancestry in East African populations and the greatest proportion of hunter-gatherer ancestry across Southern populations (*i.e.* Zulu and Sotho). Notably, the Eurasian admixture events across the West African and the Ethiopian populations predates the recent complex admixture events across some Southeastern populations consistent with the recent colonial admixture events into Africa[485]. Similarly, hunter-gatherer proportions in Western populations are indicative of ancient admixture events (around 9,000 years ago), whereas hunter-gatherer proportions in the Eastern and Southern populations represent a more recent gene flow through multiple admixture events which have been occurred 100 to 3,000 years ago. This specific admixture pattern is consistent with the 'back-to-Africa' gene-flow that is assumed to have taken place within the last 10,000 years[486].

Since most of the sampled populations in our study belong to the Bantu linguistic group (Table 6.1), haplotype diversity in each population has to be assessed in the context of a broad 'Bantu expansion' model. According to this model, population differentiation across the SSA is derived by the millennia-long series of migrations starting from West Africa around 3,000 years ago that ultimately led to the establishment of regional populations[484]. The majority of ethnolinguistic groups south of the latitude between Nigeria/Cameroon and Somalia speak a derivative of the Bantu language that appears to have been spread quickly across the SSA[487] (Figure 6.1). In view of this, the language spoken by Zulu populations in the south coast of Africa is highly similar to the language spoken by the Baganda population (situated in the east coast of Africa) underlying the shared population history between these two populations, despite their geographical distance[485].

The extensive population history resulted from the more distant effective population bottleneck time across the SSA populations enable delineation of LD patterns at an unprecedented resolution. Increased haplotype diversity across the SSA populations reflect the historical impact recombination, selection and mutation in the accumulation of LD across different genomic regions.

6.2 Materials and Methods

6.2.1 Samples and populations

The whole genome sequence (WGS) data of 320 healthy individuals from seven SSA ethnolinguistic groups who were recruited to the African Genome Variation Project (AGVP)[87] were obtained and processed (Table 6.1). Since all samples were sequenced at low coverage (∼4X), to assess the concordance of variant calls across populations for whom BeadChip array data was also available (Zulu and Baganda populations), concordance of variant calls

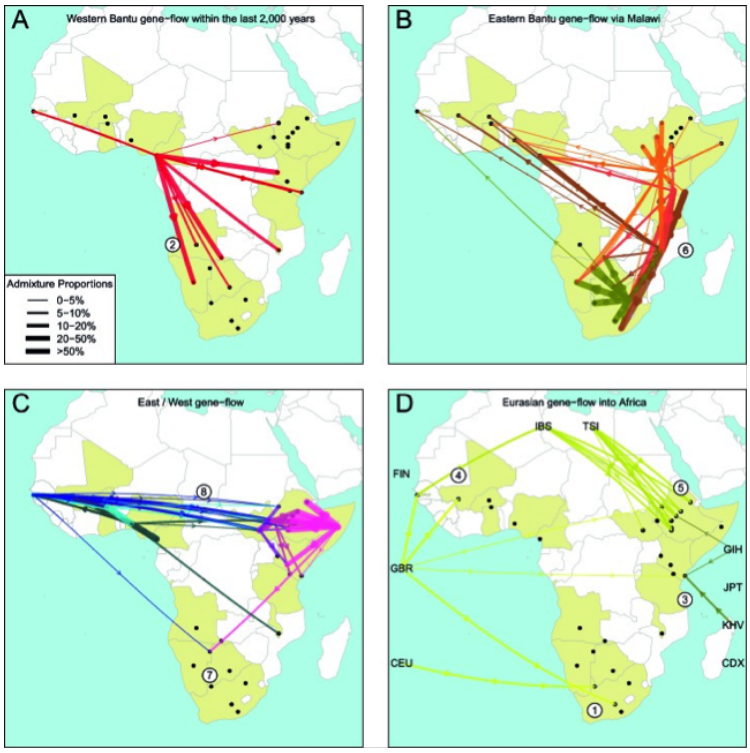


Figure 6.1: Schematic representation of Bantu expansion and recent Eurasian gene flow into SSA; (A) The broad 'Bantu expansion' believed to have taken place in two major waves between 3,000 and 2,000 years ago. This resulted in rapid spread of Niger-Congo language across South and East coast of Africa; (B) Admixture events between Niger-Congo speaking populations across the South and East coast of Africa; (C) Admixture events between Western Niger-Congo speaking populations with Eastern Afro-Asiatic and Nilo-Saharan speaking population in the Eastern coast of Africa; (D) Extensive recent Euro-Asian gene flow into the Eastern coast of Africa; (Figure adopted from Busby *et al.*[485]).

between the array data and WGS data was checked and confirmed across a panel of 24 SNPs[143]. To ensure that maps are representative of major ethnic populations, principal component analysis (PCA) was carried out using SNPRelate package in R[488].

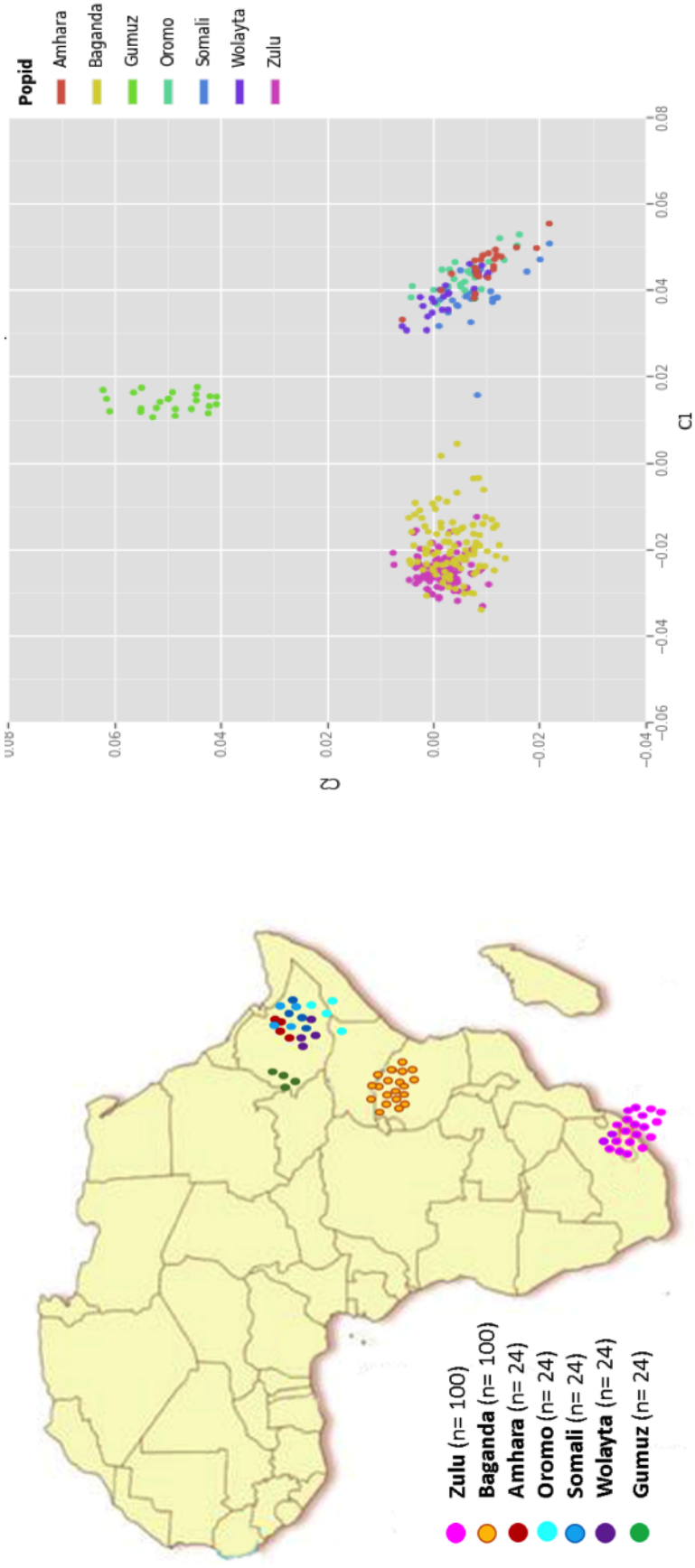
For PCA analysis, individuals with >10% missing genotypes, polymorphic markers with >10% missing genotypes across all individuals and those who deviate from Hardy-Weinberg proportions (HWE) at the significance level of $p < 10^{-7}$ were removed. Only polymorphic markers with minor allele frequency (MAF) of greater than 5% were retained for PCA analysis. Identity by descent (IBD) was measured within each population and related individuals (IBD > 0.05) were removed. Furthermore, squared correlation coefficient between allele counts were used to prune correlated SNPs at r^2 <0.5 thresholds. Consistent with the pattern of admixture across these populations, PCA revealed three major clusters mirroring distinct linguistic grouping across these populations. Furthermore, one Somali individual was excluded from further analysis since he did not cluster with any of the major groups (Figure 6.2 & Supplementary figure 8.16).

Given the sensitivity of LDU map length to sample size, samples from the Gumuz population (n= 24), with distinct clustering pattern separate from the rest of SSA sub-populations, were excluded from downstream LD analysis. To ensure sample size > 90 for the purpose of LDU map construction, samples from Amhara (n=24), Oromo (n= 24), Somali (n= 23) and Wolayta (n=24) sub-populations were grouped together to represent a broader Ethiopian sub-population (n= 95). Ultimately, samples from the three major populations including Zulu (n= 100), Baganda (n=100) and Ethiopian (n=95) were

Population	Region	Country	Language	Language subgroup	Size
Baganda	Eastern Africa	Uganda	Niger-Congo	Bantoid	100
Zulu	Southern Africa	South Africa	Niger-Congo	Bantoid	100
Amhara	Eastern Africa	Ethiopia	Afro-Asiatic	Bantoid	24
Oromo	Eastern Africa	Ethiopia	Afro-Asiatic	Cushitic	24
Somali	Eastern Africa	Ethiopia/Somalia	Afro-Asiatic	Cushitic	24
Wolayta	Eastern Africa	Eastern Africa	Omotic	-	24
Gumuz	Eastern Africa	Ethiopia/Sudan	Nilo-Saharan	-	24

Table 6.1: Details of populations selected for LD map construction.

selected for LD analysis.



(a) (b)

Figure 6.2: (*a*) The geographical origin of SSA samples recruited for LD map analysis; (*b*) SSA samples represented along PC1 and PC2.

6.2.2 Variant pre-processing

The whole-genome VCF files from chromosomes 1-22 were used to compute metrics of LD maps across 295 individuals. Population-specific multivariant VCF files were generated according to the PCA clustering pattern, and the ethnicity assignment for each individual was cross-checked with the information provided by the AGVP.

In order to investigate the resolution of LDU maps under different thresholds for the exclusion of rare variants, maps for chromosome 22 at two alternative MAF threshold (less than 5% and less than 1%) were constructed, and the LDU map lengths were consulted to decide about the optimal MAF threshold. Prior map construction, the raw data from each population underwent a pre-processing procedure to exclude: 1) individuals with > 5% missing genotypes; 2) markers with > 5% missing genotypes across all individuals; 3) markers with MAF less than 1% in the respective sub-population; and 4) markers that deviate from the HWE proportions at the significance threshold of $p < 10^{-3}$.

Since the relaxation of MAF threshold from 5% to 1% led to a ∼3.35%-21.18% increase in the LDU scale and therefore a better resolution of LD structure in highly recombining regions of the chromosome 22, minor allele frequency of 1% was used as the general threshold for construction of maps across all autosomes (Supplementary Table 8.19).

6.2.3 LD map construction

Following data pre-processing, construction of LDU maps was carried out in LDMAP program with default parameters (maximum distance between any marker pairs= 500 kb, the maximum number of intervals between marker pairs= 100, the maximum number of iteration= 20,000 and stop iterating in an interval when ϵ_i=0)[480]. Initially, an intermediate file containing information about pairwise association probabilities (ρ) of marker intervals was computed according to Equation 6.2:

$$\rho = \frac{D}{Q \times (1 - R)} \tag{6.2}$$

Where 'D' denotes the absolute value of the difference between observed haplotype frequency and its expected value from allele frequencies in the population, 'Q' denotes reference allele frequency and 'R' denotes the alternative allele frequency[489,490]. Upon computation of the pairwise association between markers, the Malécot model given by Equation 6.1 is fitted to data to compute ϵ iteratively across SNP intervals using a maximum likelihood approach. Ultimately, the pairwise value of ϵd score is used to compute LDU score across each interval.

Since the construction of LD maps are computationally intensive and requires estimation of ϵ for each marker interval for a maximum window size of 100 SNPs, VCF files from each chromosome were split to smaller segments with the maximum number of 12,000 SNPs in each segment. In order to compute ϵ for markers at the boundaries of adjacent segments, an overlapping region comprised of 200 markers was set at the end of each segment. Finally, 25 terminal markers from each end of segments were trimmed and respective map for each chromosome was assembled by joining the segments.

6.2.4 Interpolation of LD at different genomic regions

SNPs intervals with LDU locations were used to interpolate LDU size at different genomic regions. Genomic locations of all human genes were determined according to the Refseq assembly (version GRCH37/hg19), and genes with a curated status record of 'VALIDATED' or 'REVIEWED' were selected for downstream analysis. In order to avoid spurious inflation of cumulative LDU scale across the genic region, overlapping genes were merged to

contain only the longest interval. This resulted in a total of 19,455 genes across chromosome 1-22 (Supplementary Table 8.20). The RefSeq accession number for genes was used to group genes into protein-coding (NM) and non-protein-coding (NR) categories and the LD structure across coding genes, ncRNA genes and the intergenic regions was investigated. Since recombination is depressed in centromeric regions, coordinates of centromeric heterochromatin across autosomal chromosomes were defined according to GRCH37/hg19 assembly and excluded from the calculation of LDU sizes for intergenic regions (Supplementary Table 8.21).

In order to relate the gene-specific LDU size to measures of gene essentiality, gene groups were defined according to the categories described by Spataro *et al.*[491]. This classified a total of 17,927 genes across five categories including NDNE: non-disease non-essential genes (n= 13,080); END: essential non-disease genes (n= 1,572); CNM: complex non-Mendelian genes (n= 2,388); MNC: Mendelian non-complex genes (n= 684) and CM: complex-Mendelian genes (n= 203).

Since LDU size is influenced by the gene size, a regression approach was used to correct for the gene size. Given that the relationship between LDU size and gene length is essentially non-linear and given the highly skewed distribution of the gene lengths and LDU size across the five categories of genes (overall LDU skewness= 9.06, kurtosis= 156.08, Supplementary Figure 8.17), a natural logarithm transformation was used to transform values for LDU size and gene length (Supplementary Figures 8.17c & 8.17d). To investigate the nature of the relationship between the transformed LDU size (LDU_T) and the transformed gene length (L_T), a symbolic regression model was used. Since the primary focus here is to remove the correlation component attributed to the gene length, the possibility of fitting more sophisticated regression model was also explored.

The performance of fitted models was benchmarked against residuals diagnostic plots and ultimately scaled residuals (e_{LDU}) from the best-fit model (with the least heteroscedasticity) were used to investigate the relationship between the e_{LDU} size and the gene-grouping pattern. Finally, a Kruskal-Wallis H test was used to analyse for statistically significant differences among the mean e_{LDU} size between the five gene groups.

6.2.5 Functional clustering and overrepresentation analysis

In order to check whether enrichment of particular cell function or metabolic pathway is related to gene-specific LDU size, normalised residuals (e_{LDU}) were converted to percentile rank scores. For the purpose of over-representation analysis, we assumed that gene-specific LDU sizes across each functional cluster are independent of each other. Genes were divided into four groups according to their quartile range and functional overrepresentation analysis of 'Biological processes' was carried out in PANTHER (v.13.1)[492]. To inspect the relationship between e_{LDU} and the essentiality of genes in the functional clusters, pre-tabulated essentiality scores from 17 different essentiality metrics adopted from Bartha *et al.*[493]. Scores were scaled and the sum of values across the 17 metrics was used as the single essentiality score for each gene. Missing values for each essentiality metric were replaced by the arithmetic mean of scaled values for that specific score.

In order to investigate the relationship between the extent of LD and the evolutionary divergence of a gene, overrepresentation of homologous genes in each quartile was investigated in Enrichr[494]. To further delineate the specific pattern of homologous genes overrepresentation across e_{LDU} quartiles, the average gene age for overrepresented gene groups in each quartile was calculated according to the human protein-coding gene-age annotation[495]. Based on this framework, the evolutionary age of each gene is estimated according to the presence or absence of orthologs in the vertebrate phylogeny, and a score in the scale of 0 (representing the oldest genes that appeared before the divergence of Ze-

brafish) to 12 (for human-specific genes) is assigned to each gene. For a small proportion of the genes with missing age score (n=1,783, ∼10%), the estimated age of the most significant functional partner (identified through the search of STRING 10.5 database[496]) was used as a proxy for the evolutionary origin of the gene.

6.3 Results

6.3.1 Characteristics of LD maps across different populations

The LD maps generated from the three SSA populations show highly concordant contours although overall map lengths are different (Figures 6.3 & 6.4 and Tables 6.2 and 6.3). The characteristic 'step and plateau' pattern indicating regions of high and low recombination is highly concordant across the three populations. The pairwise Spearman rank correlation coefficient between Mb/LDU rates for chromosome 22 in the three populations shows a very strong correlation ($\rho > 0.9$, p-value $<$ 2.2e-16 and Supplementary Figure 8.18), however, the relationship between the LDU size and the physical distance of the bins from the centromeres is not monotonic and therefore the correlation between the maps is perhaps overestimated.

The Zulu and the Baganda populations represent a relatively similar LDU scale across all autosomes, but the LDU magnitude is significantly reduced for the Ethiopian population (Figures 6.3 & 6.4). Based on these results, the longer map length for the Zulu population presumably provides evidence in support of the 'late-split hypothesis' for Bantu expansion toward the East coast of Africa. It is a matter of debate whether the Eastern Bantu groups were split off before the departure of Southern Bantu populations from West Africa or the Eastern Bantu spread occurred after the Bantu expansion toward the South. In fact, the total autosomal LDU map length across the Zulu population revealed to be ∼0.4% longer than the Baganda population, and both maps were identified to be ∼17% longer than the Ethiopian map (Tables 6.2 & 6.3).

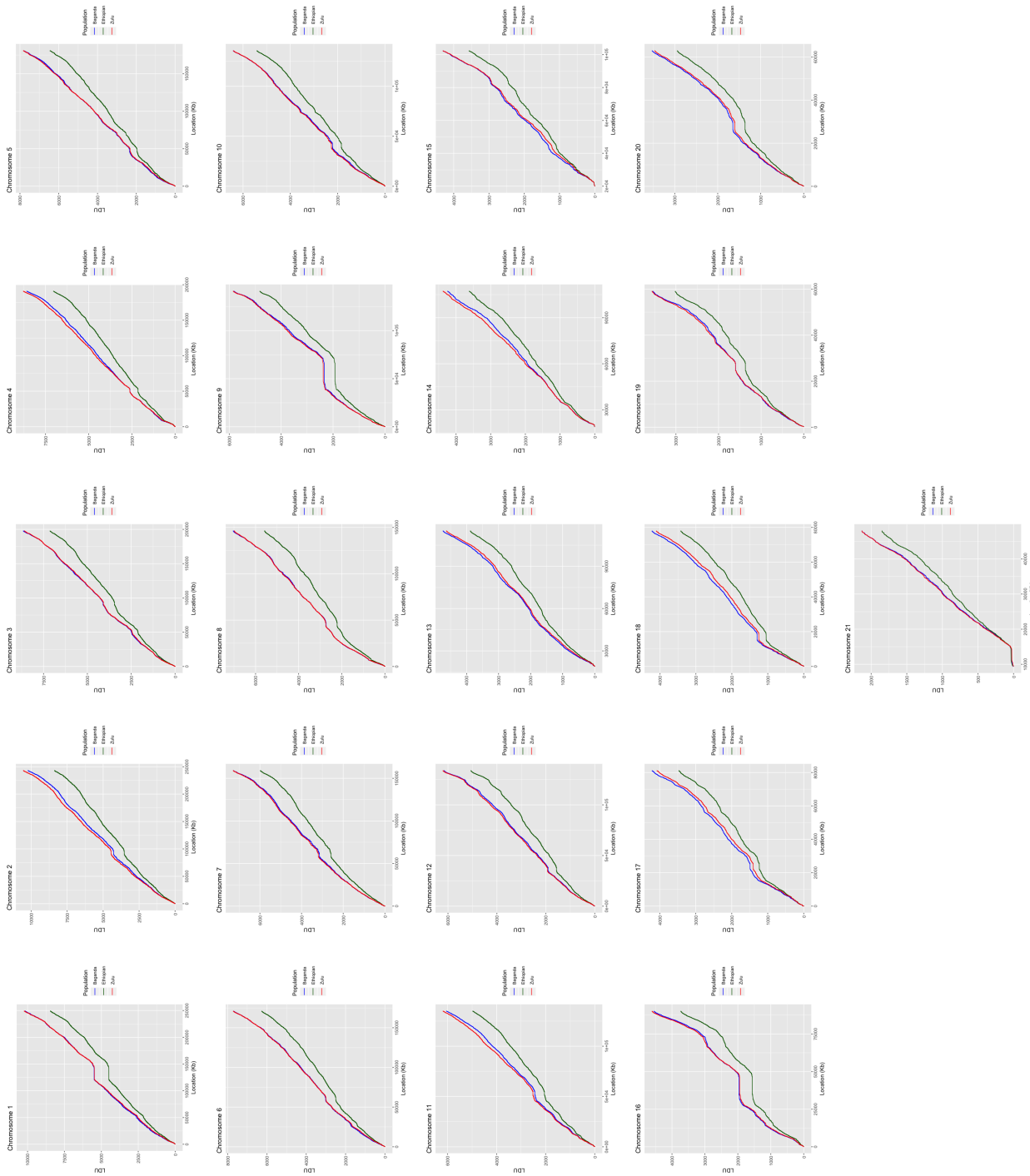


Figure 6.3: LD map plots for chromosomes 1-21 for the three major SSA populations. The contours of LD maps across the three populations are highly concordant although the overall map lengths are different across the three populations. Overall the Zulu and Baganda populations represent a more extensive LDU length consistent with a more extensive population history across these two populations

Table 6.2: Details of overall LDU map length across different SAA populations for chromosomes 1-12.

Chr.	population	Failed missingness (GENO > 0.05)	Failed frequency test	Failed HWE test	Total number of remaining markers	Number of markers after LD construction	Physical length (Kb)	Map length (LDU)
1	Buganda	105,191	1,422,792	40,342	1,232,089	1,055,438	249,208.93	10,166.92
	Zulu	87,406	1,365,281	40,048	1,290,150	1,118,373	249,208.93	10,232.35
	Ethiopian	83,399	1,547,636	38,155	1,111,508	1,097,822	249,208.93	8,453.55
2	Buganda	112,263	1,577,133	41,669	1,334,195	1,139,730	243,167.81	10,222.70
	Zulu	92,458	1,495,564	41,121	1,416,970	1,220,123	243,167.81	10,550.32
	Ethiopian	87,877	1,714,522	39,005	1,201,366	1,183,401	243,167.73	8,379.45
3	Buganda	94,617	1,292,806	35,875	1,136,341	975,075	197,840.11	8,652.03
	Zulu	78,299	1,247,771	36,258	1,181,916	1,023,841	197,840.11	8,592.47
	Ethiopian	74,145	1,408,441	34,520	1,024,289	1,008,464	197,840.11	7,133.60
4	Buganda	98,284	1,249,129	39,422	1,144,310	985,643	190,920.29	8,581.00
	Zulu	81,917	1,188,147	39,715	1,206,129	1,046,827	190,920.29	8,784.09
	Ethiopian	78,622	1,369,558	37,422	1,027,984	1,013,830	190,920.29	7,033.83
5	Buganda	87,149	1,169,617	33,118	1,034,785	887,187	180,730.86	7,811.35
	Zulu	71,919	1,129,908	32,424	1,075,826	933,217	180,769.37	7,825.82
	Ethiopian	68,212	1,284,337	30,730	924,378	913,305	180,769.37	6,448.32
6	Buganda	89,439	1,113,726	37,316	1,016,020	872,316	170,875.10	7,698.63
	Zulu	73,720	1,071,094	37,900	1,058,727	919,776	170,875.10	7,717.59
	Ethiopian	71,148	1,218,286	36,391	914,569	903,271	170,875.10	6,268.16
7	Buganda	78,555	1,040,028	29,759	935,850	802,132	159,112.37	7,272.00
	Zulu	64,848	1,005,322	29,919	971,327	839,044	159,112.39	7,284.13
	Ethiopian	62,118	1,147,682	28,439	831,628	822,286	159,112.37	5,979.92
8	Buganda	68,801	1,024,638	26,560	911,772	780,359	146,211.02	7,040.27
	Zulu	57,277	989,159	26,729	947,457	818,877	146,211.02	6,991.71
	Ethiopian	53,827	1,133,031	25,406	806,386	795,075	146,234.56	5,584.68
9	Buganda	55,186	791,991	21,223	681,338	582,606	141,069.82	5,825.25
	Zulu	45,536	753,288	21,146	720,485	623,917	141,069.82	5,852.42
	Ethiopian	43,926	861,471	20,085	613,753	605,025	141,069.82	4,824.72
10	Buganda	65,102	893,417	24,688	784,041	671,952	135,451.23	6,418.79
	Zulu	53,285	864,305	24,472	813,910	703,875	135,444.57	6,409.52
	Ethiopian	51,925	959,108	23,684	720,593	711,933	135,456.05	5,413.39
11	Buganda	63,861	904,066	24,140	787,500	670,444	134,795.08	6,040.50
	Zulu	53,023	865,020	24,259	827,051	716,867	134,766.76	6,184.26
	Ethiopian	50,320	987,321	23,228	706,794	698,268	134,771.21	4,970.81
12	Buganda	65,101	857,102	24,744	755,903	646,565	133,749.97	6,157.24
	Zulu	53,549	827,436	24,768	786,173	679,337	133,749.53	6,203.88
	Ethiopian	50,978	946,713	23,404	668,972	659,882	133,695.23	5,060.85

Table 6.3: Details of overall LDU map length across different SAA populations for chromosomes 13-22.

Chr.	population	Failed missingness (GENO > 0.05)	Failed frequency test	Failed HWE test	Total number of remaining markers	Number of markers after LD construction	Physical length (Kb)	Map length (LDU)
13	Buganda	51,167	625,517	19,495	569,797	490,190	96,089.53	4,734.58
	Zulu	42,287	605,848	19,523	589,791	512,552	96,089.53	4,629.56
	Ethiopian	40,383	686,862	18,525	510,378	503,470	96,089.53	3,882.26
14	Buganda	44,832	598,650	17,178	517,870	442,302	88,289.30	4,242.93
	Zulu	36,813	570,908	17,324	545,724	472,127	88,289.30	4,375.81
	Ethiopian	35,448	648,015	16,224	470,490	465,741	88,289.22	3,618.30
15	Buganda	39,566	538,957	14,826	470,615	402,244	82,468.18	4,304.27
	Zulu	32,632	513,176	14,589	497,039	428,946	82,468.18	4,304.06
	Ethiopian	31,215	590,712	14,211	420,463	413,527	82,464.48	3,564.66
16	Buganda	35,988	599,353	14,211	504,723	430,997	90,093.89	4,548.90
	Zulu	30,064	570,317	14,138	534,167	460,119	90,093.94	4,607.21
	Ethiopian	28,187	653,828	13,389	451,713	444,727	90,194.82	3,733.21
17	Buganda	37,505	501,814	14,685	435,199	371,164	81,187.57	4,215.68
	Zulu	30,858	486,145	14,358	451,332	389,757	81,187.57	4,080.27
	Ethiopian	29,448	547,721	13,730	391,200	386,786	81,188.22	3,472.50
18	Buganda	38,824	501,360	14,495	459,736	394,313	77,989.03	4,222.90
	Zulu	32,093	486,146	14,339	475,547	412,631	77,989.07	4,102.97
	Ethiopian	30,442	552,901	14,103	409,573	403,919	77,989.04	3,420.51
19	Buganda	32,318	389,770	14,802	363,074	313,363	58,880.86	3,541.08
	Zulu	27,393	374,661	14,504	378,657	329,827	58,880.86	3,561.55
	Ethiopian	25,741	431,307	13,467	323,504	318,734	58,880.86	3,017.19
20	Buganda	28,549	411,063	10,967	361,035	309,093	62,903.88	3,528.43
	Zulu	23,669	396,894	10,870	375,653	325,305	62,904.06	3,467.01
	Ethiopian	22,340	447,488	10,559	325,775	320,876	62,904.44	2,941.58
21	Buganda	19,254	237,218	7,927	220,698	190,638	38,587.85	2,131.71
	Zulu	16,114	225,956	7,835	232,087	202,296	38,587.85	2,131.18
	Ethiopian	15,445	258,023	7,526	200,595	197,627	38,587.88	1,849.53
22	Buganda	17,722	244,146	7,103	215,457	184,354	35,186.78	2,282.84
	Zulu	14,795	233,990	6,973	225,851	196,211	35,186.88	2,267.67
	Ethiopian	14,795	233,990	6,973	225,851	197,052	35,192.05	1,950.29

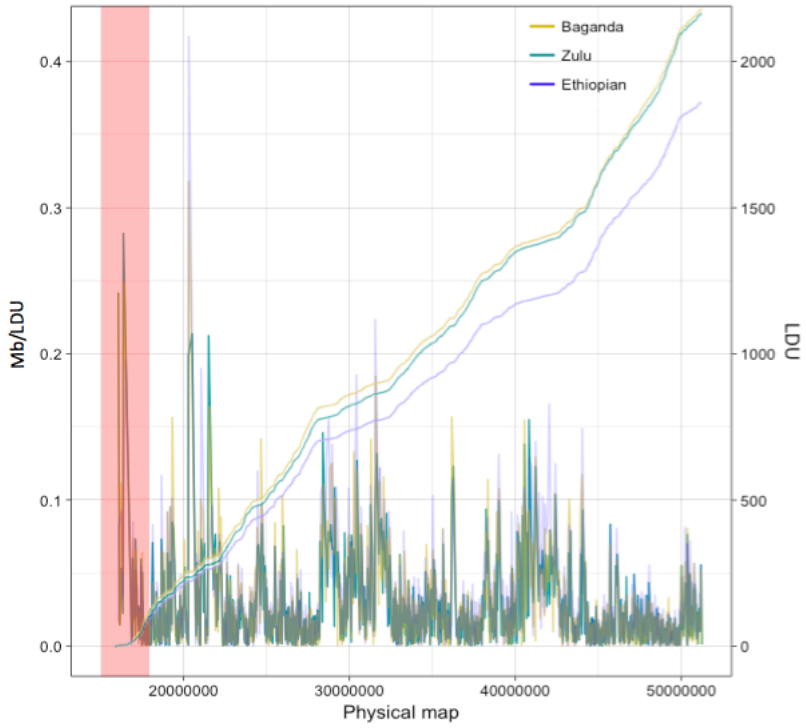


Figure 6.4: Population-specific LD map for chromosome 22; The physical length of chromosome 22 is represented along the x-axis, The extent of chromosome over which LDU increases by one unit is identified by the Mb/LDU rate along the y-axis to the left and cumulative LDU extent is identified by the y-axis to the right. The red block shows the centromeric heterochromatin. The plateau regions of LDU map align with intensity picks of Mb/LDU rate trend line.

Based on the framework proposed by Zhang *et al.*[480], assuming a generation time of 25 years, the effective bottleneck time for these populations can be estimated from the LDU/Morgan ratio (the sex-averaged morgan rations for human autosomes were adopted from the study by Kong *et al.*[497]). The effective population bottleneck time for the Zulu population revealed to be around 92,975 years (range for autosomal chromosomes: 75,538-109,974), whereas the bottleneck time for the Baganda and Ethiopian populations are estimated to be around 92,794 (range 78,045- 110,738) and 76,818 (range 64,287- 87,843) years. Given these results, it is possible that the residual difference in the LD map length between the Zulu and Baganda population is indicative of an earlier southward spread from the Bantu homeland. It is noteworthy that these estimates are more compatible with the more recent models of human dispersal out of Africa and in fact demonstrate that 'Bantu Expansion' may have occurred before the modern human dispersal out of Africa[498,499].

The shorter map length in the Ethiopian population, by contrast, demonstrates the impact of the Eurasian admixture events that have led to the retention of extended LD structures across individuals of this population. Most importantly, the attenuated LD decay with genomic distance in Ethiopian population (reflected as extended LD haplotypes), suggests a more recent bottleneck event for this population. Consistent with this observation, the presence of non-Bantu linguistic groups in this region (including Semitic and Cushitic) perhaps reflects the distinctive population history for the Ethiopian population.

Despite nearly perfect correlation of maps at Mb scale (Figure 6.4), cross correlation comparison of maps at Kb scale reveals more subtle differences between the maps. For example for chromosome 22, a positive cross correlation covariance between the populations

implies that LD sizes in all three population tend to vary together, albeit the magnitude of this syntonic behaviour at zero lag is more intense between the Bantoid populations (Zulu and Baganda, r_k= 0.098) than Bantoid versus Ethiopian population (Zulu versus Ethiopian: r_k= 0.066, Baganda versus Ethiopian: r_k= 0.055) (Figure 6.5).

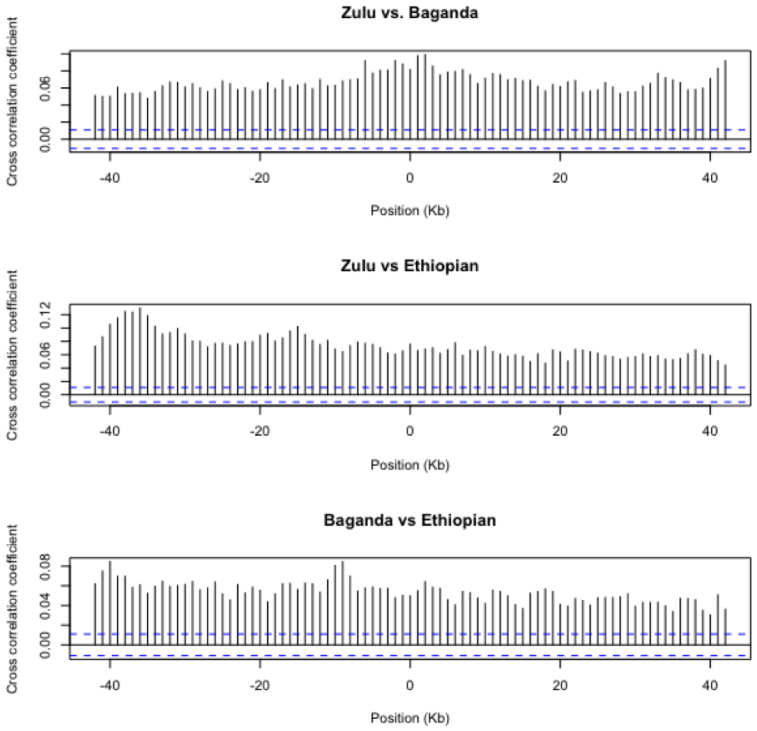


Figure 6.5: Cross-correlation comparison of SSA LD maps at Kb scale.

Decreased cross-correlation function (CCF) covariance between the Ethiopian and Bantoid populations (Zulu or Baganda) is most likely representative of a distinctive localisation pattern for LD hotspots in this population. Consistent with this hypothesis, it has been revealed that East African populations possess lower frequency of active *PRDM9* alleles in comparison to West African populations with Bantoid origin [500].

6.3.2 Extent of LD in different genomic regions

Analysis of LD across different genomic regions revealed a highly variable size for the coding part of the genome versus non-coding regions. In total, the extent of LD across the 19,455 non-overlapping genic regions were compared against the 19,224 non-overlapping intergenic regions (Supplementary Table 8.20). The genic region was identified to constitute ∼45% of the length of the autosomes while the intergenic regions and centromeric heterochromatin constitute 51% and ∼4.8% of the genome respectively (Supplementary Table 8.21). The genic LDU size accounts for 43.5% of total LDU size across all autosomes and the LDU extent across the intergenic region constitutes 55.8% of the entire LDU length (Table 6.4). The remainder 0.7% of the LDU size is attributable to the heterochromatic regions in which severely depressed recombination has resulted in the accumulation of LD.

The average LDU extent (KB/LDU rate) across genic regions identified to be ∼5.7% more extensive compared to whole chromosome scale (21.89 and 20.71 respectively)(Supplementary Table 8.22). Furthermore, the average LDU extent across the intergenic regions is ∼10% shorter than the genic regions which probably reflects the higher frequency of recombination outside the gene boundaries and/or attenuated intensity of

selection across these regions (Supplementary Table 8.22). We also identified that there is a close correlation between the centimorgan length of the chromosome and LDU length with relatively consistent ratio across all autosomes (30.21-43.98, SD=3.50) (Table 6.4).

To further delineate the extent of LD across sub-genic levels, the LDU size across exonic and intronic regions was investigated. The extent of LD across exonic regions of coding genes that constitute only ∼5% of total length of the genic region is dramatically higher than intronic regions (34.13 Kb versus 22.39 Kb) which spans ∼84% of the length of the genic region (Supplementary Table 8.22). This pattern was also consistent across the exonic and intronic regions of the non-coding genes, albeit with a reduced magnitude of the difference (22.39 Kb versus 20.85 Kb). It is important to note that these two regions collectively constitute ∼10% of the total length of the genic region (0.69% and 9.77% for non-coding exonic and non-coding Intronic regions respectively). The overall difference between the averaged extent of LD across exonic and intronic regions of non-coding genes is small (One-tailed t-test p-value= 0.09), but increased length of LD across exonic regions of non-coding genes is consistent across all chromosomes (Supplementary Table 8.22).

The results also demonstrate that LD is noticeably more extensive across genic regions compared to intergenic segments (t = 4.05, df = 21, p-value = 2.14e-4, Figure 6.6). The extent of LD at a chromosomal level is significantly different from the LD extent across intergenic region (t = -2.6224, df = 21, p-value = 0.01), but it does not reach significance when compared to the LD extent at genic level (t = 1.6347, df = 21, p-value = 0.11). This probably indicates that the pattern of LD at chromosomal level is predominantly influenced by the extent of LD at the genic level. While this pattern is generally conserved across all autosomes, three chromosomes including chromosomes 9, 16 and 21 present exceptional cases in which the extent of LD in the coding region is less than the LD extent at the chromosomal level. This perhaps can be attributed to the extended region of "gene desert" immediate to centromeric regions of chromosome 9 and 16 in which overrepresentation of LINE elements immediate to centromeric region leads to extended LD size.

In agreement with recombination maps, we identified that the extent of LD in large chromosomes that have lower sex averaged recombination rates is greater and LD is more frequently broken down in small chromosomes where the sex averaged rate of recombination is higher (Figure 6.6).

Taken together, our result demonstrate that genes themselves represent high degree of linkage disequilibrium because they are subject to a more intense effect of selective forces, but it appears that LD is more readily broken down across intronic and intergenic regions of the genome.

Table 6.4: Total LDU size across different genomic regions (The sex-averaged genetic lengths of chr.1-22 (cM ratios) are adopted from Kong *et al.*[497])

Chr.	**Whole Chr.**	**Genic**	**Coding Genes**		**ncRNA genes**		**Intergenic**	**cM length**	**LDU/cM**
			Exonic	*Intronic*	*Exonic*	*Intronic*			
1	10232.35	4667.52	137.53	3913.1	26.2	528.15	5514.88	270.27	37.86
2	10550.32	4384.82	118.15	3451.47	30.73	651.26	6125.46	257.48	40.98
3	8592.47	4035.66	101.37	3412.31	20.95	409.19	4486.3	218.17	39.38
4	8784.09	3354.36	76.06	2665.91	13.56	407.06	5400.17	202.8	43.31
5	7825.82	3091.64	79.73	2384.46	16.82	471.93	4686.23	205.69	38.05
6	7717.59	3210.16	98.88	2660.2	23.99	335.53	4473.41	189.6	40.70
7	7284.13	3388.81	94.37	2905.06	22.77	233.79	3867.88	179.34	40.62
8	6991.71	3036.87	71.93	2511.72	22.75	343.35	3907.55	158.94	43.99
9	5852.42	2342.17	83.85	2026.67	10.73	174.54	3480.04	157.73	37.10
10	6409.52	2972.04	75.06	2558.27	23.78	215.53	3412.54	176.01	36.42
11	6184.26	2798.03	112.47	2377.23	20.15	216.2	3343.51	152.45	40.57
12	6203.88	2834.57	111.4	2320.29	17.7	242.68	3301.55	171.09	36.26
13	4629.56	1794.3	34.16	1394.99	19.46	312.84	2808.68	128.60	36.00
14	4375.81	1718.91	68.45	1369.75	12.8	156.31	2638.54	118.49	36.93
15	4304.06	2014.89	58.68	1629.94	16.3	236.18	2263	128.76	33.43
16	4607.21	2078.97	78.34	1775.4	13.4	150	2498.99	128.86	35.75
17	4080.27	2103.9	114.87	1718.02	17.31	145.06	1930.05	135.04	30.22
18	4102.97	1567.09	44.01	1318.24	10.9	163.69	2523.29	120.59	34.02
19	3561.55	1840.12	150.66	1443.57	15.35	101.49	1661.89	109.73	32.46
20	3467.01	1475.29	59.52	1245.31	14.66	109.15	1971.96	98.35	35.25
21	2131.18	854.33	24.24	629.98	9.19	164.72	1258.26	61.86	34.45
22	2267.67	1113.06	52.05	896.8	12.75	103.05	1087.06	65.86	34.43
Total	**130155.84**	**56677.5**	**1845.78**	**46608.71**	**392.23**	**5871.67**	**72641.26**	**3435.71**	**37.88**

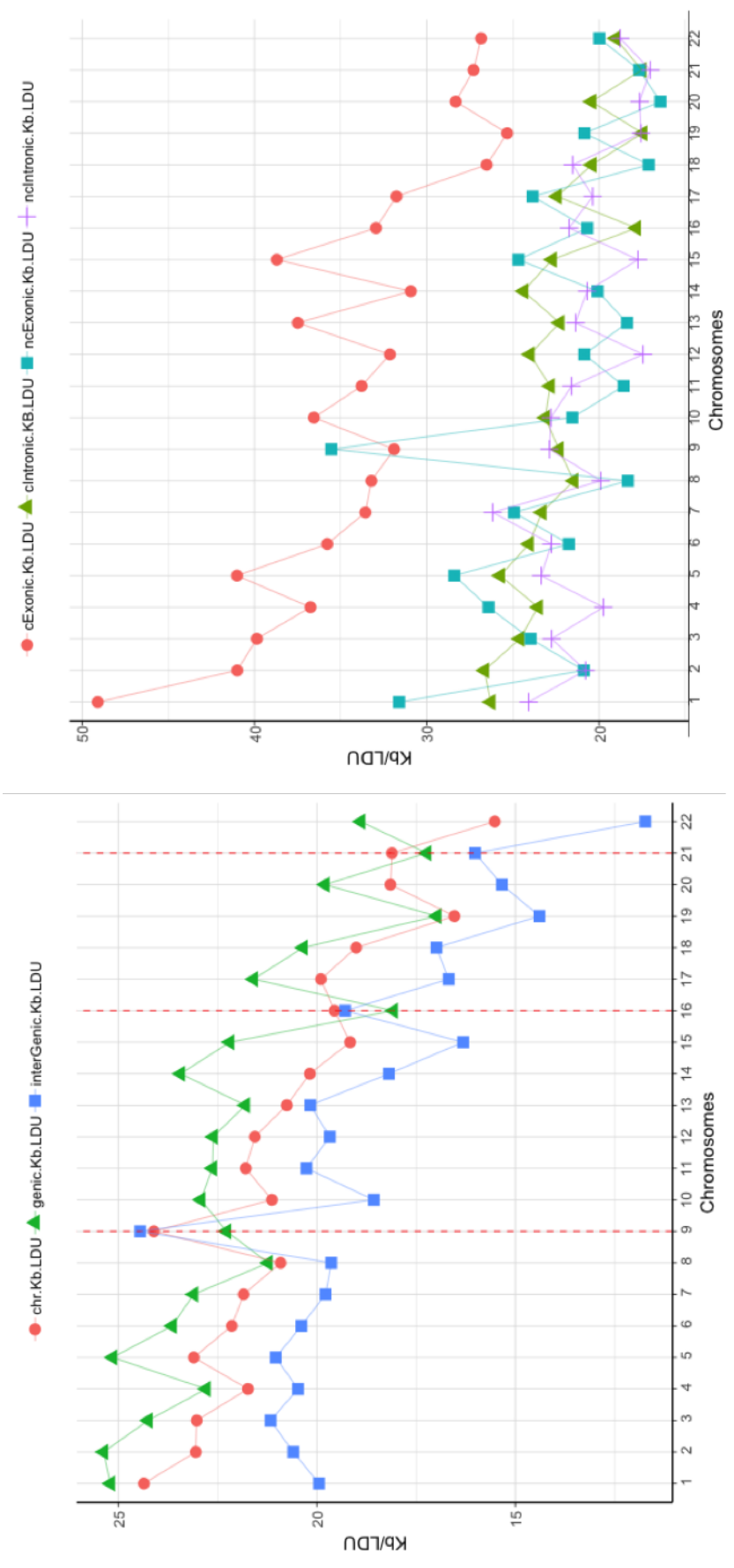


Figure 6.6: (a) . Total LDU extent across genic, intergenic and chromosomal level; (b) Total LDU extent across exonic and Intronic regions of protein coding and non-protein coding genes.

6.3.3 Regression model selection for correcting LDU variance for gene size

As mentioned in the Section 6.2.4, in order to correct the LDU size for gene length a regression approach was used. Initially, the correlation between the LDU size and the gene length was modeled with a linear function. Although the linear model using non-transformed data explains a fair amount of LDU variance at the gene level (adjusted R^2= 0.72, p-value< 2.2e-16), diagnostic plots for residuals are indicative of strong heteroscedasticity (Supplementary Figure 8.19). The inconsistency of the covariance matrix of the estimated regression coefficients in the presence of heteroscedasticity renders the residuals of the linear-fit unsuitable for downstream analysis. From the biological prospect, there is also a number of plausible reasons that explain why regression residuals from the linear model are unsuitable. It is well established that the intensity of recombination has a strong association with *PRDM9* motifs in the genome and distribution of these 13-mer motifs across the genome is strictly non-linear [474]. Furthermore, there might be additional genomic features (such as Alu repeats or GC contents) that contribute to the variance in LD extent, but a linear model overlooks their significance at gene levels. It is noteworthy that the possibility of applying weighted least square regression was also explored, but it failed to correct for the observed heteroscedasticity of residuals from the linear model. Given the demonstrated skewness in the distribution of gene lengths and LDU sizes (Supplementary Figures 8.17a & 8.17b), the subsequent regression analysis carried out using the natural logarithm transformed data (Supplementary Figures 8.17c & 8.17d).

The emulator function derived from a symbolic regression model was suggestive of a cubic relationship between the transformed LDU size (LDU_T) and the transformed gene length (L_T). The model generated using a genetic programming (GP)-run simulation with a primary space of 5000 possible functions that evolve during 100 generations with 0.7 probability of random crossover and 0.1 probability of mutation of functions. The initial function set for model construction included, addition, subtraction, multiplication, and division. Based on the symbolic regression model, the transformed size of LDU (Y_{LDU}) for each gene is calculated according to formula 6.3:

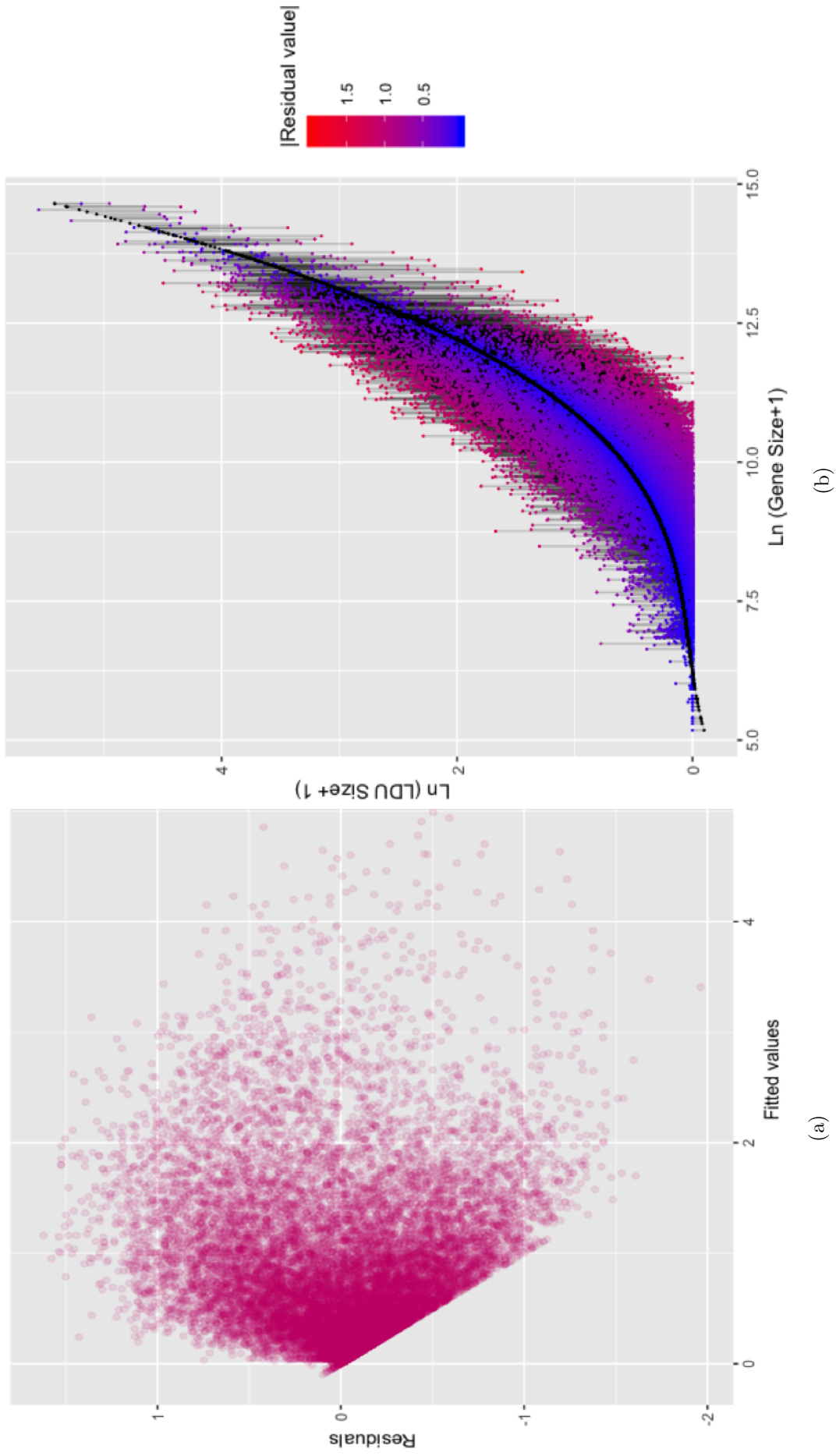
$$Y_{LDU} = 0.03\chi^3 - 0.05\chi^2 - 0.35\chi + 0.6 \tag{6.3}$$

In which χ represents the transformed value for respective gene size. Despite the mediocre performance of the model in explaining the variance (R^2= 0.64), it is suggestive of a third order linear relationship between transformed LDU size and gene size.

Investigation of alternative regression models confirmed the superior fitness of the cubic model for explaining the observed LDU_T variance at the gene level (adjusted R^2= 0.70, Table 6.5, Supplementary Figure 8.20). Since the sole purpose of fitting a regression model was to account for gene size, residuals plots from the fitted models were also assessed to identify the most appropriate model. Residuals derived from the cubic model revealed to be more randomly distributed (Figure 6.a) and therefore provide better estimates for LDU size at the gene level. In other words, by replacing residuals from the cubic model, gene-specific LDU size was corrected by removing the linear component attributable to gene size (Figure 6.b).

Regression model	AIC	BIC	Adjusted R2	p- value	Residual SE	Formula	Convergence tolerance	Number of iterations to convergence
Linear	29054	29077	0.5919	< 2e-16	0.5441	y ∼ x	-	-
Second order polynomial	23340	23371	0.7033	< 2.2e-16	0.4639	y ∼ x2	-	-
Cubic	23029	23068	0.7084	< 2.2e-16	0.4599	y ∼ x3	-	-
Logarithmic	31333	31357	0.5366	< 2.2e-16	0.5798	y ∼ log(x)	-	-
Inverse	33608	33631	0.4739	< 2.2e-16	0.6177	y ∼ (1/x)	-	-
Non-linear exponential inverse	30936	30960	-	< 2e-16	0.5734	y ∼ (1/(x *a))+ b* x	1.65E-08	1
Non-linear Log	31333	31357	-	< 2e-16	0.5798	y ∼ (a + b * log(x))	1.05E-09	0
Non-linear Exponential	23583	23607	-	< 2e-16	0.4671	y ∼ I(exp(1)$\hat{}$(a + b * x))	1.13E-06	6
Non-linear inverse	33608	33631	-	< 2e-16	0.6177	y ∼ I(1/x * a) + b	3.41E-09	1

Table 6.5: Details of regression model statistics fitted to transformed data; The cubic model represents the lowest AIC/BIC values and has the highest adjusted R^2 in comparison to other models fitted to the transformed data.

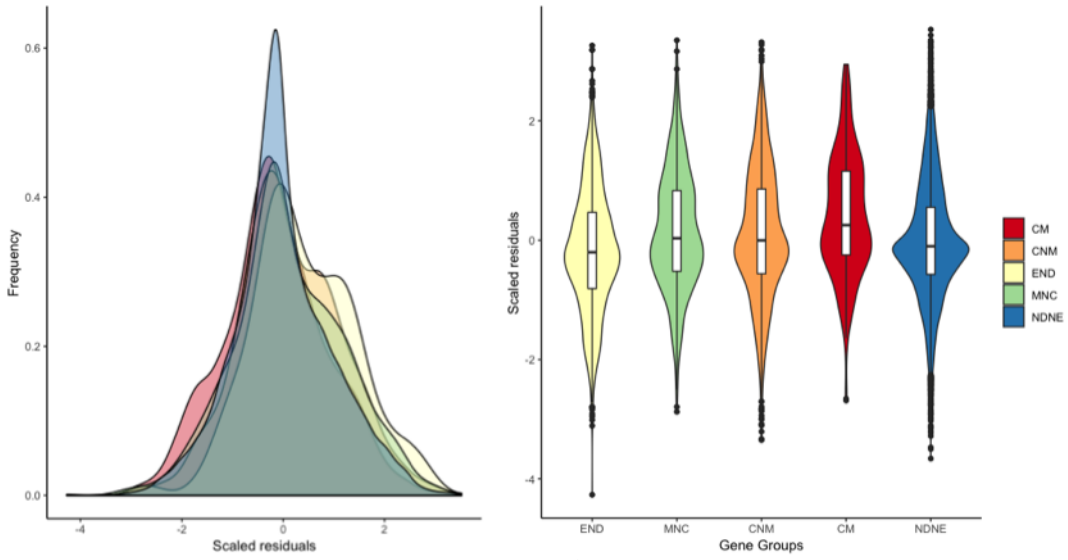


(a) (b)

Figure 6.7: (a) Scatterplot showing distribution of residuals versus fitted values; (b) Schematic representation of absolute residual distance from the fitted cubic curve for the 17,927 genes.

6.3.4 LD characteristics across different gene groups

Scaled residuals (e_{LDU}) from the fitted cubic model were used as the gene-size corrected measures for comparison of LDU across different gene groups. A Kruskal-Wallis H test between five categories of genes revealed that there is a statistically significant difference in the mean e_{LDU} size between different classes of genes (Kruskal-Wallis H= 114.35, p-value= 8.56e-24, df= 4) (Figure 6.8).



(a) (b)

Figure 6.8: (a) Histogram illustration of scaled residuals (e_{LDU}) distribution across the five categories of gene groups; (b) Distribution and probability density of scaled residuals across the five gene categories.

Post-hoc analysis using pairwise Wilcoxon comparison of ranked sum e_{LDU} size across the five categories of genes revealed that ranked sum e_{LDU} score across the END genes is significantly different from other gene groups (with sum rank e_{LDU} score of 8,136.73 for END genes versus 9,510.62 for CNM, 11,049.72 for CM, 9,656.62 for MNC and 8,895.04 for NDNE genes)(Table 6.6 & Supplementary Table 8.23). The observed difference was most significant when END genes were compared against the three categories of genes involved in human diseases (*i.e.*, CNM, CM, MNC). Similarly, the ranked sum e_{LDU} scores across the NDNE genes were identified to be most significantly different from that of CM, END and CNM groups and to the lesser extent with MNC group. Although pairwise comparison of ranked sum e_{LDU} scores between MNC group versus NDNE, CM and CNM groups reached the nominal significance (p-value$<$ 0.05), it did not surpass the Bonferroni corrected threshold for multiple testing (p-value$<$ 0.005).

	Gene Group				
	END	**MNC**	**CNM**	**CM**	**NDNE**
END					
MNC	1.40E-08				
CNM	6.30E-13	0.96659			
CM	4.80E-12	0.0679	0.0046		
NDNE	4.00E-06	0.0165	9.00E-06	3.90E-07	

Table 6.6: Post-hoc comparison of ranked sum e_{LDU} size across the five gene groups. P-values for the Wilcoxon paired test is provided. END genes represent the most significant difference when compared to other groups. While both the END and NDNE groups represent a significant ranked sum e_{LDU} difference with human disease (HD) genes (MNC, CNM and CM), the difference among the HD group is least significant.

6.3.5 Functional overrepresentation analysis result

As discussed in section 6.3.2, we observed that recombination is consistently suppressed across genic regions. Given this, it would be interesting to investigate whether the extent of LD at the gene level is influenced by the functional relevance of the genes. Previous studies on Caucasian populations have revealed that the variation in the magnitude of LD is influenced by the biological function of the gene product [462,501]. To assess whether a similar pattern is discernible from the LD structures in SSA populations, the gene-level e_{LDU} size across the Zulu population was used to investigate overrepresentation pattern of different gene ontology groups across quartile ranges of e_{LDU}. The Zulu genome was preferably used for the purpose of overrepresentation analysis (ORA) since the largest LDU maps in this population enables capturing of residual e_{LDU} differences at a superior resolution.

Analysis of gene sets in quartile range of normalised ranked e_{LDU} revealed that genes located in the first quartile (strong LD tail) are statistically overrepresented in essential intracellular functions that mainly involve in transcription, translational modifications and intracellular transport. Significant GO terms were ordered according to the combined score implemented in Enrichr method. Amongst the top 10 Genome Ontology (GO) terms that reached significance across genes in the first quartile, RNA metabolic process (GO:0016070) achieved the highest score (adjusted p= 5.59e-11, Z-score= -1.99, combined score= 47.03) followed by protein polyubiquitination (GO:0000209) (adjusted p= 9.42e-14, Z-score= -1.56, combined score= 46.73) and protein modification by small protein conjugation (GO:0032446) adjusted p= 6.34e-17, Z-score= -1.14, combined score= 42.57). Intriguingly, gene clusters related to Golgi vesicle transport (GO:0048193) and cell cycle G2/M phase transition (GO:0044839) were among the overrepresented clusters in the first quartile gene sets (Figure 6.9).

Despite gene sets in the first quartile where ORA revealed a specific pattern of enrichment for gene clusters in the RNA polymerase II network (primarily responsible for mRNA synthesis), gene sets across the second quartile revealed another interesting pattern of functional overrepresentation. Across the genes, in this quartile, the top 10 overrepresented functional clusters appeared to be generally related to the transcriptional activity of RNA polymerases I and III that involve transcriptional processing of non-coding RNAs. Across genes in this quartile, the functional cluster related to protein targeting to Endoplasmic Reticulum (ER) (GO:0045047) achieved the highest combined score (adjusted p= 1.83e-7, Z-score= -2.08, combined score= 46.60) followed by the functional cluster related to nonsense-mediated mRNA decay (GO:0000184) (adjusted p= 1.34e-7, Z-score= -2:00, combined score= 46.39) and SRP-dependent co-translational protein targeting to mem-

brane (GO:0006614) (adjusted p= 1.25e-7, Z-score= -1.84, combined score= 43.53). The remaining functional clusters across this quartile are indicative of enrichment for biological processes involved in the transcriptional activity of RNA Pol. I and III (Figure 6.9).

The biological processes overrepresented across the third quartile included many sensory receptors and immune-related genes that are known to have high allelic diversity. ORA across this quartile revealed that genes involved in detection of chemical stimulus (GO:0007608) and perception of smell (GO:0007608) (adjusted p= 3.46e-22, Z-score= -1.69, combined score= 96.53) achieve the highest combined scores. Furthermore, gene clusters involved in the activation and regulation of immune responses were also overrepresented across this quartile (Figure 6.9). Across the top 10 overrepresented biological processes in this quartile, six clusters were related to immune functions but after adjustment for multiple testing only three clusters including regulation of peptidyl-serine phosphorylation of STAT protein (GO:0033139) (adjusted p= 0.02, Z-score= -2.26, combined score= 23,41), T cell activation involved in immune response (GO:0002286) (adjusted p= 9.32e-3, Z-score= -2.20, combined score=23.72) and positive regulation of peptidyl-serine phosphorylation of STAT protein (GO:0033141) (adjusted p= 0.02, Z-score= -2.13, combined score= 22.50) surpassed Bonferroni correction for multiple testing. Overrepresentation of such molecular functions in the third quartile of LD reflects high recombination rate across these genes involved in sensory and immune-related functions.

Gene sets across the last quartile (weak LD tail) were specifically enriched for GO terms related to extracellular matrix organisation and functions related to extracellular mechanisms. Across the top 10 biological processes, which surpassed the Bonferroni correction, gene clusters involved in extracellular matrix organization (GO:0030198) achieved the highest combined score (adjusted p= 6.81e-9, Z-score= -1.65, combined score= 44.90) followed by biological processes involved in flavonoid glucuronidation (GO:0052696) (adjusted p= 2.25e-3, Z-score= -3.59, combined score= 40.59) and sensory perception of mechanical stimulus (GO:0050954) (adjusted p= 3.22e-6, Z-score= -1.73, combined score= 33.68). The biological processes overrepresented across the Q4 included many genes that involve in cell-cell communication which are known to have high allelic diversity (Figure 6.9).

Investigation of gene essentiality scores across the top 40 overrepresented clusters revealed a decreasing trend in essentiality with the decrease in the extent of LD (adjusted p= 3.94e-6, df= 38) (Figure 6.10). Genes in the first quartile (strong LD tail) represent a broader distribution of essentiality scores with more outliers toward the high end of essentiality spectrum. In contrast, gene clusters in the weak LD tail (third and fourth quartiles) represent low variability in essentiality distribution and are skewed toward the lower extreme of essentiality.

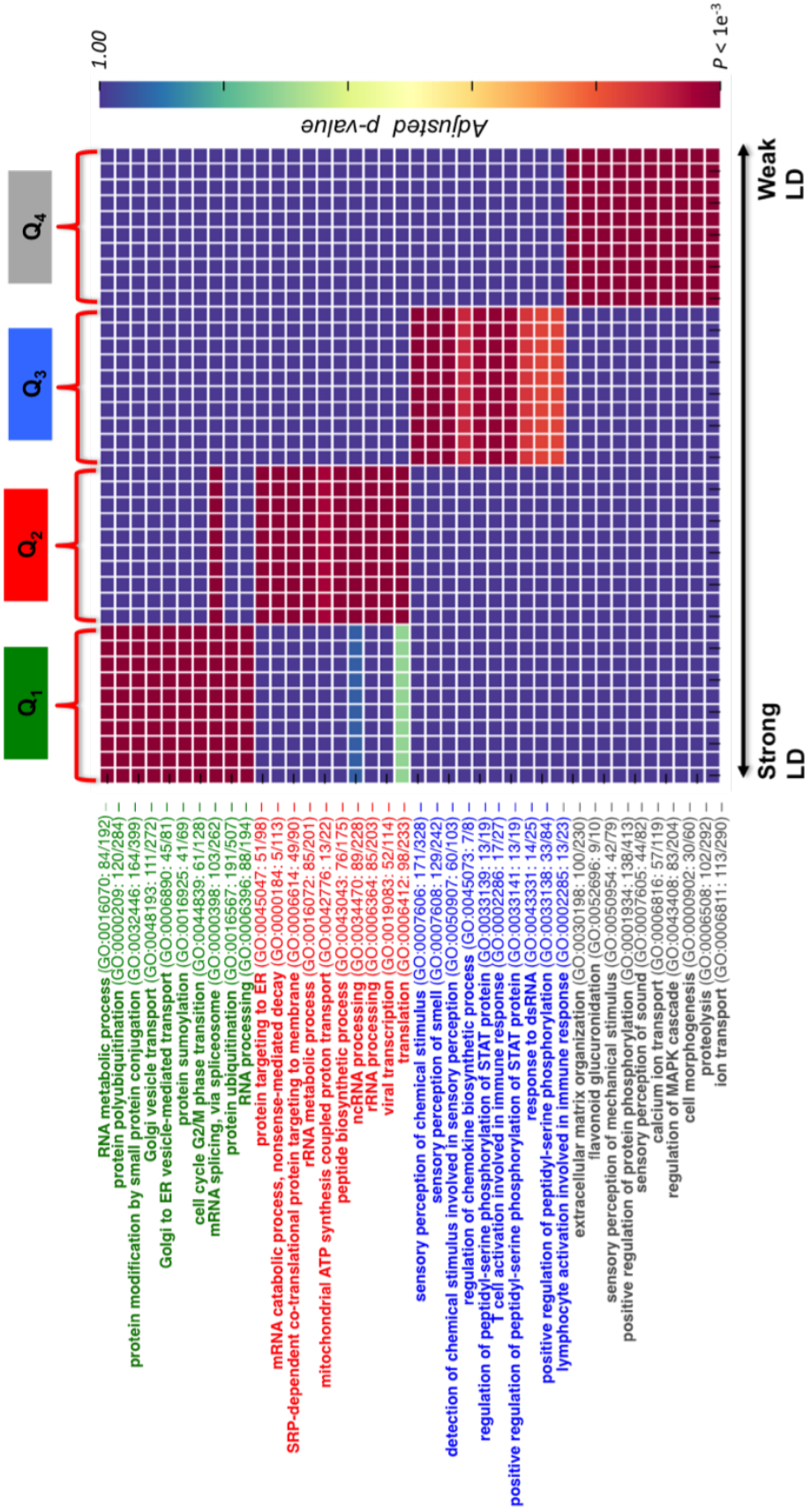


Figure 6.9: Heatmap plot of top 40 overrepresented Gene Ontology (GO) biological processes across the quartile range of LDU size.

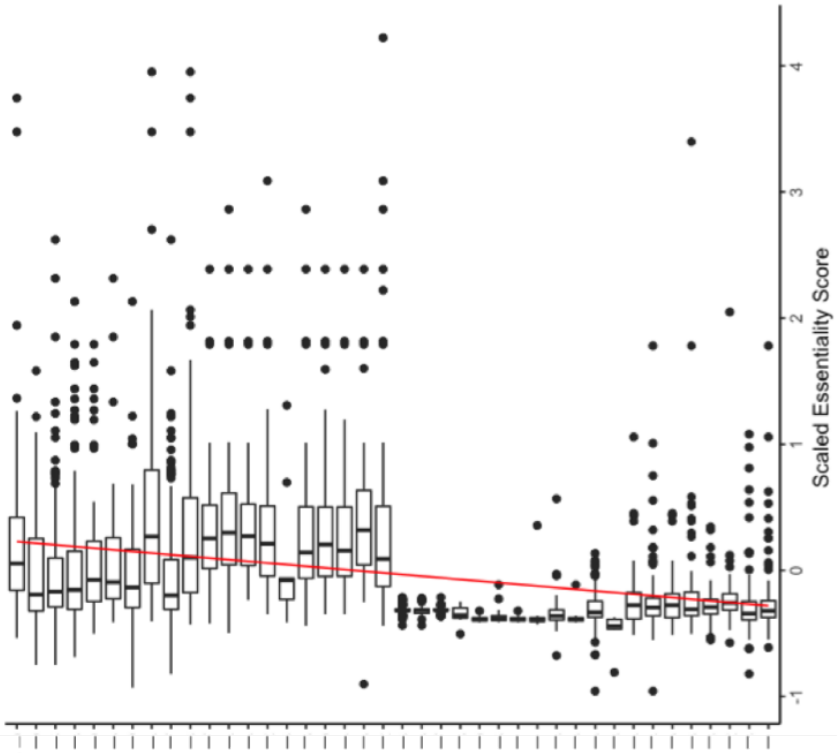


Figure 6.10: Scaled essentiality scores across the genes of top 40 overrepresented GO terms across the LD quartile ranges. The linear fit line shows the general negative trend in essentiality as we move from overrepresented GO clusters in the strong LD tail to weak LD tail(Essentiality scores were adopted from Bartha *et al.*[493] and scaled according to the method described in Section 6.2.5.)

Taken together this result indicates that the extent of LD varies between different gene ontology groups. The fact that biological functions occurring inside the cell are overrepresented across the strong LD tail (Q1 and Q2) whereas overrepresentation of functions related to the external region of the cell across the weak LD tail (Q3 and Q4) perhaps reflects that a greater intensity of purifying selection inside the cell. The decreasing essentiality trend across the top 40 overrepresented gene classes corroborates the reduced intensity of purifying selection at the extracellular environment.

Investigation of the relationship between the essentiality status of the gene (essential versus non-essential according to the OGEE framework [502] and the extent of LD using a multinomial regression model revealed that e_{LDU} scores in the Q1 range are significant positive predictors for essential genes (p= 8.60e-3, OR= 1.96, $CI_{95\%}$= 1.18- 3.25). In contrast, for non-essential genes, e_{LDU} scores within the strong LD tail (Q1 and Q2 range) appear to have negative association; however this negative association do not reach statistical significance (Supplementary Table 8.24).

6.3.6 Orthologous gene analysis

Given that the extent of LD is influenced by the intensity and duration of selection we sought to investigate the relationship between the gene age and the extent of LD. Previous studies have shown that gene age is associated with the strength of purifying selection and the majority of human disease genes have an ancient evolutionary origin [503]. It has been suggested that younger genes specific to mammalian lineage are underrepresented in human disease genome and are subject to a more relaxed coding constraints leading to a faster and more variable sequence evolution [504]. In order to identify the relationship between the extent of LD and the evolutionary divergence of the gene, overrepresentation analysis of homologous gene clusters was carried out in Enrichr [494].

Analysis of genes across the first quartile (genes with strong LD, Q1) revealed significant over representation of three homoloGene clusters. Homologous genes to Euteleostomi clade (bony vertebrates) achieved the most significant overrepresentation p-value (adjusted p-value= 8.56e-37, Z-score= -1.28, combined score= 110.17) followed by homologous genes to Tetrapoda superclass (first four limbed vertebrates and descendants) (adjusted p-value= 1.14-9, Z-score= -1.18, combined score= 26.37) and Amniota (four limbed vertebrates that can lay eggs on land) (adjusted p-value= 2.47e-3, Z-score= -0.98, combined score= 7.26).

Across the genes in the second quartile a significant overrepresentation of homologous clusters was identified across genes common to Eukaryota (adjusted p-value= 9.98e-9, Z-score= -1.38, combined score= 28.98) followed by gene clusters homologous to Boreoeutheria kingdom (placental Mammals) (adjusted p-value= 5.66e-7, Z-score= -1.19, combined score= 19.35).

Two homologous clusters across the Q3 gene achieved significant overrepresentation p-value. Genes homologous to Boreoeutheria (Placental Mammals) revealed the most significant overrepresentation p-value (adjusted p-value= 1.01e-19, Z-score= -1.50, combined score= 69.41) followed by gene clusters homologous to Euarchontoglires Superorder (superorder of mammals including rodents and primates) (adjusted p-value= 0.04, Z-score= -1.37, combined score= 6.60).

Ultimately, homologous clustering pattern across the last quartile (weak LD tail) revealed a significant overrepresentation p-values across three gene sets homologous to Euteleostomi (adjusted p-value= 8.56e-37, Z-score= -1.28, combined score= 110.17) followed by genes common to Tetrapoda (adjusted p-value= 1.14e-9, Z-score= -1.17, combined score= 26.37) and Amniota (four limbed vertebrates that can lay eggs on land) (adjusted p-value= 2.47e-3, Z-score= -0.98, combined score= 7.26).

Overrepresentation analysis of orthologous genes across the quartile range of e_{LDU}

score revealed an interesting relationship between the taxonomic age of the gene sets and the extent of LD. The relative overrepresentation percentage across the orthologous gene sets that reached significance indicates that genes with e_{LDU} scores at the strong LD tail (Q1) generally appear to have a more distant evolutionary origin whereas gene at the weak LD tail seems to have been evolved more recently (Q4) (Figure 6.11).

The relative frequency percentage across the first quartile demonstrated extensive overrepresentation of Q1 genes in lower taxa which gradually decreases toward the higher primates (Figure 6.11). Conversely, gene sets across the fourth quartile (weak LD tail) represent the highest relative frequency across homologous gene groups with intermediate evolutionary origin. Finally, gene sets with e_{LDU} size in the interquartile range (Q2 and Q3) have higher relative frequency across younger homologous gene sets.

The higher relative frequency of intermediate e_{LDU} size genes across the higher taxa perhaps indicates that primate-specific genes in human are under moderate selective pressures while genes with vertebrate-specific origins are under weaker levels of purifying selection. Interestingly, previous studies into the rates of selective evolutionary pressure across the human disease genes have revealed an attenuated selective pressure across the homologous genes common to vertebrates[505]. This observation may explain why modelling of some immune-related disorders (in which the causal genes have an e_{LDU} score within the interquartile range) in other vertebrates including rodents poorly mimic the phenotype in human[506]. In other words, the reduced intensity of selective pressures on vertebrate-specific homologous genes may explain the erratic overrepresentation of weak LD genes across these clusters. Similarly, overrepresentation of intermediate e_{LDU} genes across the homologous clusters specific to higher primates indicates the impact of active selective sweeps on these genes.

Furthermore, given the role of gene duplication in mammalian speciation the reduced extent of LD across more recent genes is consistent with the impact of gene conversion in the erosion of LD. In fact, the impact of GC-biased gene conversion in the evolution of mammalian genome is extensively documented, and it has been estimated to occur 10 times more frequently than crossover in the human genome[507]. The higher relative frequency of intermediate e_{LDU} size genes across the mammalian-specific homologous cluster probably reflects the collective impact of gene conversion and balancing selection acting on these genes. Conversely, overrepresentation of genes with weak LD across taxa with intermediate age perhaps reflects a more intricate process underlying species divergence.

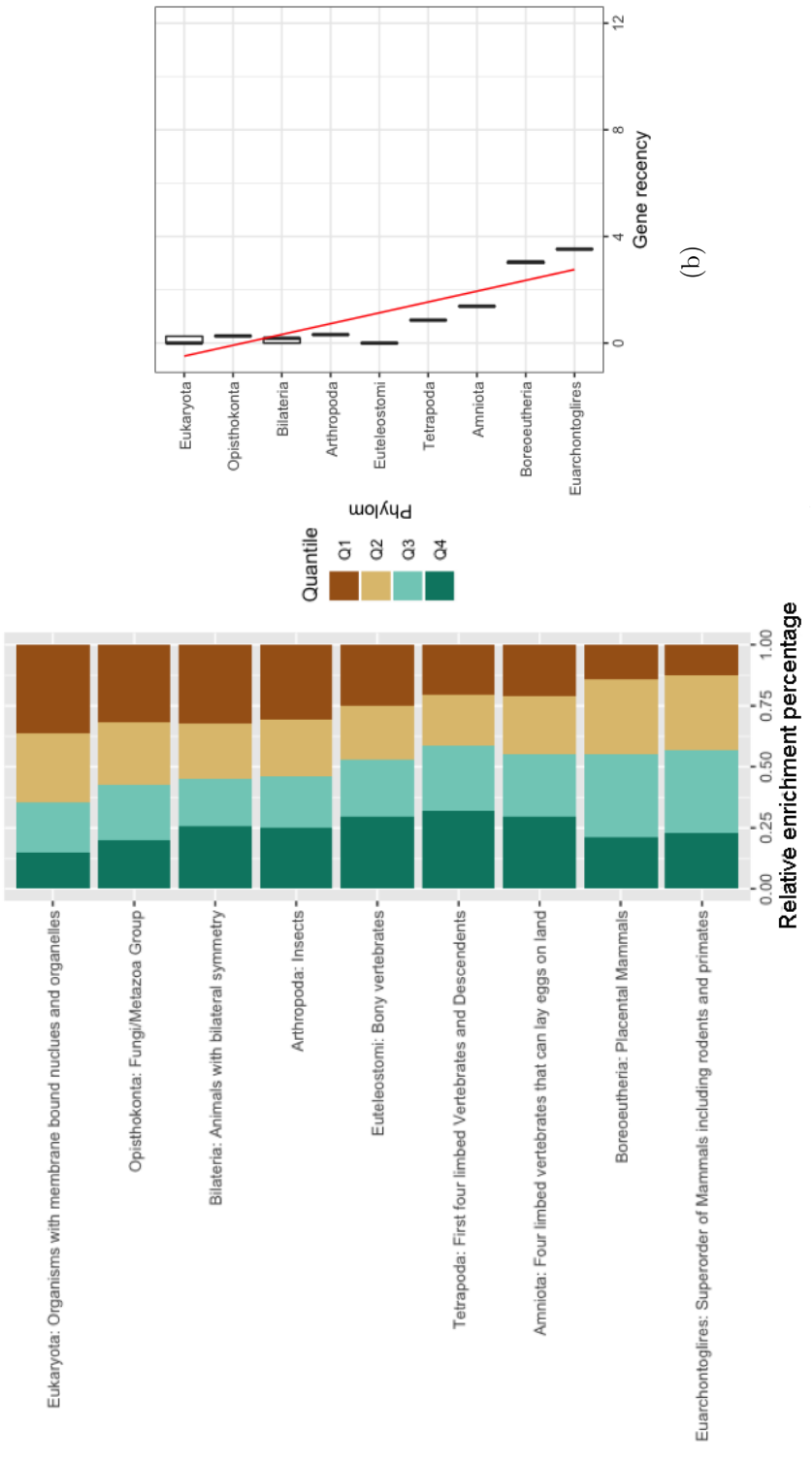


Figure 6.11: (a).Relative overrepresentation percentage of quartile gene sets across different taxa; (b) Representation of mean age for overrepresented homologous gene sets across different phyla based on the framework proposed by Zhang *et al.*[495]. The higher numbers identify genes that diverged more recently.

Homologous genes are generally categorised as orthologous and paralogouse genes. Despite orthologous genes that evolve before speciation and maintain their ancestral function, paralogous genes are derived from gene duplication and usually lead to new function within the species. Delineation of orthologous genes from paralogous genes may provide further insight into the relationship between the extent of LD and gene age.

Taken together our result demonstrates that LD is more extensive across ancient genes, and it has the most reduced extent across the vertebrate-specific homolog genes.

6.3.7 Relationship between gene groups and e_{LDU} quartile range

The Pearson chi-square test of association between disease gene groups and gene specific e_{LDU} quartile range revealed that the group assignment of a gene is significantly associated with its e_{LDU} quartile range (χ^2= 177.19, df= 12, p-value= 1.60e-31, *Cramer's* V= 0.06).

To check whether group assignment of a gene can be predicted from its respective e_{LDU} quartile, gene age, GC content and gene essentiality score, a multinomial logistic regression model according to Equation 6.4 was fitted to the data:

$$log\frac{Prob(category_i)}{Prob(category_{END})} = \beta_{i0} + \beta_{i1}(geneage) + \beta_{i2}(GCcontent) + \beta_{i3}(Essentialityscore) + \beta_{i4}(e_{LDU}quartilerank) \quad (6.4)$$

Given that ranked sum e_{LDU} score across END group is most significantly different from the other groups, END was selected as the reference category in the model and the probability odd ratios of other categories (*i.e.* MNC, CM, CNM and NDNE) were calculated against the reference category. The fitness of the model was investigated using a chi-square statistic measuring the difference in log-likelihood ratios between the final model and a reduced model. Model fitting information confirmed the statistical significance of the fitted model (χ^2= 990.75, df= 24, p-value< 2.2e-16).

Inspection of model covariates (Table 6.7) revealed that gene age and essentiality score are significant predictors of group assignment across all categories. Multinomial logit estimates (B) for gene age across all groups was positive and significant revealing that higher gene age scores (*i.e.* newer genes) are positive predictors of group assignment as non-essential. The magnitude of this positive association was highest across NDNE genes (p-value= 2.25e-47, OR= 1.57) and lowest across CM genes (p-value= 0.081, OR= 1.13). Given that CM genes represent a very specific subgroup of HD genes with extended protein network and higher expression level[491], reduced odds ratio for gene age association is not surprising across this group.

Evaluation of logit estimates for essentiality score revealed a significant negative association between higher essentiality scores and group assignment as either human disease (MNC, CM, CNM) or NDNE. This negative association is most significant across NDNE genes (p-value= 1.37e-28) and least significant across CM group (p-value= 2.47e-04). Despite this, inspection of odds ratios across the four groups revealed that higher essentiality scores are more protective against group assignment as either MNC or CM (both with OR= 0.35) and have attenuated protective impact for group assignment as CNM (OR= 0.56) or NDNE (OR= 0.51).

The logit estimate for the GC content reached significance only for CM, CNM and NDNE groups, but it did not reveal a significant directional association with group assignment across these groups when compared to END genes ($OR\sim$ 1).

Perhaps, the most interesting insight from the multinomial logit model was the relationship between gene groups and the e_{LDU} quartile rank. Inspection of model estimates

Table 6.7: Model covariates for the fitted multinomial logistic regression predicting the relationship between gene groups and e_{LDU} quartile range.

Parameter Estimates									
Group Code		**B**	**Std. Error**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95% CI for Exp(B)**	
								Lower Bound	**Upper Bound**
MNC	Intercept	4.473	0.88	25.816	1	3.76E-07			
	Gene Age	0.148	0.046	10.463	1	0.001	1.159	1.06	1.268
	GC Content	0.006	0.006	0.781	1	0.377	1.006	0.993	1.018
	Essentiality	-1.062	0.156	46.574	1	8.82E-12	0.346	0.255	0.469
	Q4	0.185	0.145	1.642	1	0.2	1.203	0.907	1.598
	Q3	0.163	0.146	1.246	1	0.264	1.177	0.884	1.567
	Q2	0.546	0.137	15.911	1	0.000066	1.727	1.32	2.258
	Q1	0†	.	.	0	.	.	.	.
CM	Intercept	4.04	1.618	6.23	1	0.013			
	Gene Age	0.125	0.072	3.035	1	0.081	1.133	0.985	1.304
	GC Content	-0.024	0.011	5.024	1	0.025	0.976	0.956	0.997
	Essentiality	-1.062	0.29	13.437	1	0.000247	0.346	0.196	0.61
	Q4	1.406	0.245	33.032	1	9.06E-09	4.078	2.525	6.586
	Q3	0.996	0.259	14.806	1	0.000119	2.707	1.63	4.495
	Q2	0.486	0.282	2.969	1	0.085	1.626	0.935	2.828
	Q1	0†	.	.	0	.	.	.	.
CNM	Intercept	4.699	0.494	90.352	1	1.99E-21			
	Gene Age	0.34	0.033	103.646	1	2.42E-24	1.405	1.316	1.5
	GC Content	-0.034	0.005	51.576	1	6.89E-13	0.967	0.958	0.976
	Essentiality	-0.584	0.083	48.982	1	2.58E-12	0.557	0.473	0.657
	Q4	0.669	0.098	46.802	1	7.85E-12	1.953	1.612	2.366
	Q3	0.275	0.101	7.369	1	0.007	1.316	1.079	1.605
	Q2	0.256	0.1	6.546	1	0.011	1.292	1.062	1.572
	Q1	0†	.	.	0	.	.	.	.
NDNE	Intercept	5.715	0.368	241.031	1	2.34E-54			
	Gene Age	0.454	0.031	209.019	1	2.25E-47	1.575	1.481	1.675
	GC Content	-0.008	0.004	4.717	1	0.029862	0.992	0.985	0.999
	Essentiality	-0.672	0.061	123.029	1	1.37E-28	0.511	0.454	0.575
	Q4	0.2	0.082	5.944	1	0.015	1.221	1.04	1.434
	Q3	0.121	0.082	2.16	1	0.142	1.129	0.96	1.327
	Q2	0.198	0.081	6.036	1	0.014	1.219	1.041	1.429
	Q1	0†	.	.	0	.	.	.	.

The reference category is: END

†This parameter is set to zero because it is redundant.

for e_{LDU} quartiles revealed a significant association for at least one quartile in each category of genes (Table 6.7).

Across MNC genes, e_{LDU} scores in the second quartile revealed to be a positive predictor for MNC genes (p-value= 6.603-05, OR= 1.72). Even though parameter estimates for the third and the fourth quartile was also showing a positive association with group assignment as MNC, their associated p-value was not significant, and therefore their positive association with MNC genes could not be established.

Inspection of parameter estimates for CM genes, revealed a significant positive association for e_{LDU} scores in the third (Q3) and fourth quartile (Q4) with CM genes. The magnitude and the significance of this positive association identified to be higher for e_{LDU} scores in the Q4 range (p-value= 9.06e-09, OR= 4.07).

Similarly, across the CNM genes, e_{LDU} quartile rank in the Q4 range revealed the most significant positive correlation (p-value=7.85E-12, OR=1.2), although the e_{LDU} scores in the Q2 and Q3 range were also showing a significant positive association but with dramatically reduced significance.

Finally, the logit estimates for e_{LDU} quartile rank across the NDNE group was indicative of two equally sized positive associations between the e_{LDU} scores and genes in this group. The e_{LDU} scores in the Q2 and Q4 range both revealed to increase the odds of group assignment as NDNE by a factor of 1.2 (p-value $\sim$ 1.50e-02). Perhaps one possi-

ble explanation for this observation is the fact that NDNE genes comprise a substantial heterogeneous group of genes (13,080 genes), in which a proportion of them might be in fact involved in human disorders, but due to our limited knowledge of aetiological factors outside the coding region of these genes, they have not been attributed to any Mendelian or complex disorders yet.

6.3.8 Discussion

In this chapter, I investigated the LD structure across three SSA populations and described the characteristics of LDU variation across different genomic features. Delineation of LD structure both at the population level and the gene annotation level revealed several interesting insights that I summarise here.

At a population level, LDU magnitude is revealed to be directly correlated with the population history. The larger LDU map length for Bantoid populations (Zulu and Baganda) in comparison to Ethiopian population was consistent with the proposed model of 'Bantu expansion' across the sub-Saharan region [485]. Furthermore, the relatively shorter map length for the Baganda population compared to Zulu was consistent with the 'late-split' hypothesis for Bantu expansion. Given that decline of LD is directly related to duration since the last effective bottleneck, the increased LDU scale across the Zulu population supports the hypothesis of primary southward migration before secondary split toward East coast. This insight revealed from the LDU maps is supported by the recent analysis that confirms linguistic and genetic similarities between the two populations better match the 'late-split' hypothesis [485]. On the other hand, reduced LDU length across the Ethiopian population reflects the significance of this region as the principal migratory route out of Africa. Admixture events are extensively documented across the Ethiopian populations and provide the most compelling evidence for the observed higher LD extent (Mb/Kb) across this population [508] (Supplementary Figure 8.18).

Despite differences in the LDU magnitude, the broad characteristics of the maps revealed to be similar across the three populations. Nearly perfect correlation of Mb/LDU rates across the three populations indicates that regions of high and low recombination are broadly co-localised across the three populations, whereas reduced cross-correlation coefficients between the maps at fine scales corroborate differential intensity of active recombination hotspots in different populations. The previous studies into the evolution of recombination hotspots have revealed that *PRDM9* motifs across the African populations have higher allelic diversity and the probability of recombination across the genome of these populations is fully dictated by *PRDM9* hotspots [509,500]. Given population-specific differences in the allelic diversity and frequency of *PRDM9* motifs, a great proportion of intra-population LD variation at fine scale can be attributed to differential localisation of these hotspots. This indicates that the longer LDU length across the Bantoid populations (Zulu and Baganda), compared to Ethiopian population, is not due to differences in population history alone and in fact, the ethnic-specific frequency of hotspots need to be considered. As discussed earlier, populations with West African ancestry tend to show higher frequency of 17-mer PRDM9 motifs across their genome [500] and therefore it would be reasonable to assume that the longer LDU length across the Zulu and Baganda populations primarily reflects the higher frequency of recombination hotspots across the genome of Bantoid populations. Despite this, regions that demonstrate above average Mb/LDU rate (recombination cold-spots) are common across different populations and are usually centred at gene-desert regions of the genome such as 9q12.

Taking advantage of WGS data in the present study, we achieved the highest resolution for delineation of LD structures at the fine scale. Consistent with the greater sex-averaged recombination rate across small chromosomes [497], we identified a shorter extent of LD in

small chromosomes. For example, while LD extends 24.36 Kb on chromosome 1, it only extends 15.52 kb on chromosome 22 (Supplementary Table 8.22). Given that identified pattern of LD in SSA populations broadly mirrors recombination rate variations across autosomes[497], it would be reasonable to conclude that LD magnitude across SSA populations is predominantly determined by the intensity of recombination. In view of this, the role of other evolutionary forces such as population history, selection and drift in defining the pattern of LD appear secondary to the role of recombination.

Having established the prominent role of recombination in shaping the LD pattern at the population level, we sought to investigate the LD variation across different genomic regions. Here we demonstrated that LD is not only ∼17% more extensive across genic regions compared to non-genic regions, but also it is more extensive across exonic regions of both coding and non-coding genes. Comparison of LD extent across exonic and intronic regions of coding genes revealed that LD is ∼53% more extensive across exonic regions, whereas the difference in LD extension between the exonic and intronic regions of non-coding genes is dramatically reduced (∼7%)(Supplementary Table 8.22). This result compared quite closely with earlier findings from the array-based LD maps, in which autosomal genic regions are identified to have ∼14% more extensive LD[510]. The variation observed in the LD profile of exonic/intronic regions for both coding and non-coding genes confirm the greater intensity of selective constraints across exonic regions[511].

The observed heterogeneity in LD pattern at the gene level also has medical implications. Given that disease-associated recombination errors that lead to insertions or deletions are more frequent near the recombination hotspots[512], genic regions with essential biological functions are depleted for recombination hotspots[510]. In fact, purifying selection by limiting the evolution of *PRDM9* motifs across genic regions reduces the disruption of haplotypes that have an essential biological function. Despite the generally suppressed recombination rate across genic regions, our result indicates a large degree of heterogeneity in the LD profile of different gene ontology groups. Strikingly we identified that genes expressed in the outer areas of the cell tend to have shorter LD extent, whereas genes with essential functions in the cytoplasmic or nuclear areas represent more extensive LD. This result is of particular importance especially given the role of recombination in generating diversity across loci that are under varying intensity of selection over time. It may be that there is a selective advantage for overrepresented ontology clusters in the weak LD tail to undergo more frequent recombinations, as a more diverse reservoir of haplotypes is better suited for the ever-changing environment outside the cell. In contrast, overrepresented ontology groups at the strong LD tail are usually involved in essential biological processes inside the cell where the environment is extensively stable and strong purifying selection favours retention of ancestral haplotypes. Consistent with this speculation, Hussin *et al.* showed that regions of low recombination are highly enriched for genes with essential functions[513].

Despite the suppressed recombination rate across essential genes, the evidence as to whether recombination is mutagenic is conflicting. While more extensive LD across exonic regions is suggested to reflect selection against recombination[481], higher proportions of rare ($MAF < 0.01$) non-synonymous variants across low recombining regions of the genome reflect the significance of Muller's ratchet effect in the build-up of damaging mutations across these regions[513]. Given the complex relationship between recombination, mutation and selection, the establishment of the relationship between the recombination rate and disease pathogenicity across different regions of the genome remains challenging. The high dimensionality of clinical NGS data complicates molecular diagnosis and confounds the establishment of the disease-gene relationship. This problem is more pronounced when genetic pleiotropy is in effect, and therefore LD inspired gene-specific information that shows the interplay between recombination, mutation and selection might be useful for

establishing the molecular diagnosis in clinical NGS applications [501,514].

Using LD-specific gene structure across five categories of genes we have shown that essentiality of a gene has a significant correlation with the extent of LD. Furthermore, using a multinomial regression model we showed that genes associated with Mendelian disorders are significantly associated with e_{LDU} scores in the second quartile, whereas genes involved in complex disorders are more associated with e_{LDU} scores in third and fourth quartile (Table 6.7). Taken together, using LD-specific structure (reflecting the combined effect of selection and recombination) and gene essentiality scores (as a proxy for functional importance) we show that disease-containing genes can be identified through their characteristic LD structure.

The findings presented in this chapter demonstrate that LD structure not only at chromosome level but also at genic and sub-genic level provides a reliable framework for delineation of the interplay between selection, recombination and mutation.

Chapter 7

Conclusion

In this thesis, I have explored applications of next-generation sequencing (NGS) at both the individual and population level and attempted to investigate the possibility of applying population-informed LD maps for a more precise delineation of the human disease genome. In this final chapter, I reflect on the successes and the limitations encountered over the course of my research and discuss the significance of findings for future research.

The primary objective of this thesis is to explore applications of NGS technology for molecular diagnosis in rare disorders. To this aim, different aspects of analytical workflow for investigating whole exome/genome data has been discussed, and strategies for improving the diagnostic credibility of WES/WGS results have been explored in Chapter 2. I have shown that designing and implementing an optimal pipeline requires meticulous attention to numerous sequencing and variant calling parameters. When I started my research, the best practice as to the choice of the optimal variant caller for short-read sequencing was rather obscure, but active development in probabilistic models for variant calling along with efforts for benchmarking pipelines enabled a standard frameworks to be adopted in the field [515]. Despite this, the best practice in relation to variant discovery is subject to change. My approach for implementing the optimal pipeline for analysing NGS data throughout my research was informed by careful benchmarking against the reference set obtained from the Genome in a Bottle Consortium [191]. I have shown that utilising the GATK variant caller enables better precision and lower recall rate for identification of SNVs, however recall rates for detection of INDELs still benefit from the use of joint callers (e.g. Samtools and GATK). Identification of complex variants such as nearby SNPs and INDELs are extremely challenging, and while GATK base-quality score recalibration and realignment methods mitigate the rate of false positives in such regions, it might overlook true variants in low-complexity regions. While merging callsets from different variant callers improves false detection rate (FDR) for INDELs discovery [516], the greater number of variants pose another challenge for delineation of causal mutation. Methods for improved identification of INDELs and copy-number variants are under active development, and indeed availability of long-read sequencing have already started to improve CNV detection [517,518,519].

As has been shown, even trivial details such as software version have a great impact on the total number and type of variants identified and therefore once a decision about the optimal design of pipeline made, samples must be processed using a single stable pipeline. As the performance of diagnostic pipelines continues to improve new insights from previous unresolved cases are inevitable. However, there persist major limitations that continue to dominate the field of NGS analysis in rare disorders. I discuss these limitations in accordance with their relevance to each chapter.

From the methodological point of view, algorithms for annotating variants outside coding regions are underdeveloped. Due to the popularity of WES in clinical diagnosis a

great majority of annotation efforts have been concentrated on the coding region, and our knowledge of annotation attributes for non-coding regulatory elements is limited. Despite active and ongoing efforts for annotating regulatory elements (*e.g.* ENCODE project [520]), the function of nearly all *cis*-regulatory elements in human health and development is still unknown [132]. Causal variants in these regions might be easily overlooked due to the lack of clear relevance to the phenotype at the time of analysis. Moreover, the accuracy of methods that allegedly provide annotation to non-coding variants, such as CADD [174] and Eigen [521], is not comparable to their accuracy for coding regions and their performance in the non-coding region of the genome is open for debate [521]. This limitation is reflected by the modest discovery rate (less than 4%) of *de novo* mutations in regulatory elements across the 7,903 patients with severe undiagnosed neurodevelopmental disorders ascertained from the DDD project [522]. Based on this, identification of disease-associated mutations in regulatory elements pose a greater challenge and require robust annotation tools to be developed.

In Chapter 3, the potential of WES to provide a molecular diagnosis in three kindreds ascertained for multiple cases of nephrolithiasis has been explored and segregation of candidate variants proposed by WES analysis investigated in an extended number of individuals from the families. While comprehensive exome analysis has been suggestive of putative variants in at least two pedigrees, validation of these variants proved to be challenging. Variants involved in late-onset disorders, such as adult-onset nephrolithiasis, tend to escape the impact of purifying selection and therefore reach higher frequencies in the general population. For analysing nephrolithiasis cases, we adopted a 2% threshold for exclusion of common variants. While this threshold appears reasonably non-stringent, there is a chance that we have missed more common variants that may underlie the disease with reduced penetrance. There is emerging evidence that even in rare neurodevelopmental disorders that are primarily thought to be monogenic, at least 7.7% of risk variance is attributable to common variants [451] hence the polygenic transmission disequilibrium test [278] appears to be a more robust method for studying nephrolithiasis. Applying this method, however, requires a larger sample size, not only from affected patients but also their parents and unaffected siblings. Given the biochemical heterogeneity underlying nephrolithiasis ascertaining a granular cohort of patients with the same biochemical impairment is practically challenging and requires a multi-centre effort. Besides, availability of detailed clinical and biological information is key to a successful diagnosis. Cohorts such as the one studies in this chapter are generally ascertained over many years, and obtaining detailed and up-to-date clinical information from all patients across multiple generations is difficult. Phenotypes like adult-onset nephrolithiasis would greatly benefit from the availability of deep phenotyping and genomic data from the UK biobank cohort [523]. In particular, availability of urine, saliva and blood samples from the 500,000 participants and also their electronic health-record enables in-depth analysis of genetic aetiology underlying nephrolithiasis.

Nevertheless, variants identified through WES analysis of kidney-stone patients implicate novel pathways for disease pathology that was not predictable from biochemical results. Recently, Hofherr *et al.* [524] demonstrated the imperative role of *SLC25A25* in kidney development. This new insight may strengthen the case for causal role of *SLC25A25* mutations in nephrolithiasis.

The application of WES in combination with WGS for molecular diagnosis of a severe case of skeletal dysplasia has been explored in Chapter 4. Through the comprehensive analysis of SNVs, INDELs and structural variants, we established a possible molecular diagnosis of distal arthrogryposis 5D in the proband. The patient presents an interesting case for the compound impact of recessive mutations in rare disorders. Due to the restrictive nature of short-read sequencing for identification of variants phasing statues, the

index cases design is usually underpowered for detection of compound heterozygous mutations. Although methods such as GATK Read-back Phasing are proposed for delineation of variants phasing statues, they are restrictively applicable to variants that occur within the same sequencing reads. Hence, resolving the phasing statues for variants that map too far apart (longer than one read length) cannot be reliably established. In ambiguous cases where analysis is primarily focused on the identification of a single penetrant mutation, the compounding impact of heterozygous mutations can be easily overlooked. This perhaps explains why the proportion of cases attributable to recessive coding variants across the DDD patients (∼3.6%) is significantly lower than the proportion of patients identified to be influenced by dominant *de novo* mutations (∼49.9%)[454]. One may speculate that the high proportion of unexplained recessive disorders even across families in the DDD database indicates the possibility of under-diagnosed cases for compound mutations.

The patients studied in Chapter 4 and Chapter 5 present a typical dilemma in analysing rare disorders. Rare disorders are notoriously known to exhibit an extensive phenotypic heterogeneity even among patients who carry the same disease-causing variants[451]. A portion of differences in the phenotypic presentations of the disease can be attributed to the role of modifying variants[451] and the patients ethnicity[454]. In the absence of a consistent framework for documenting phenotypic details, establishing phenotypic similarity between cases is difficult. Besides, many rare disorders share extensive phenotypic similarities that complicates clinical diagnosis. More recently, the introduction of the Human Phenotype Ontology (HPO)[525] enabled standardization of phenotypic descriptions in rare disorders. Methods developed for comparing HPO-coded profiles of patients against known disease facilitate differential diagnosis and enable a systematic evaluation of rare disorders[526,527,528]. Despite the successful application of HPO terms in large international projects such as the Matchmaker Exchange consortium[529] and the DDD project, its widespread use is still limited. This limitation specially applies to patients from outside the UK who were phenotyped before the introduction of HPO terminology. Wherever possible, I attempted to convert descriptive phenotypic informations to standardised HPO terminology, however the lack of detailed phenotypic information, especially in the Colombian cohort, rendered this task difficult. For example, the broad term of intellectual disability and developmental delay (ID/DD) comprises a large subcategory of the motor and neurological deficits that range from mild to severe forms, each with individual HPO designation. The absence of in-depth descriptive information about the nature of ID/DD in patients hampered the application of HPO similarity tests[530,453] in this cohort.

In Chapter 5, I demonstrated that how the adoption of the proprietary Ψ_i score for ranking variants, has facilitated diagnosis in some exome-negative cases. The large number of variants identified through NGS analysis is prohibitive for molecular diagnosis and efficient filtering and prioritisation strategies are required for identification of causal variant(s). The difficulty in the delineation of causal variants from a large background of noise (unrelated variants) is specifically exacerbated when a definite clinical diagnosis is missing. In such cases, inaccurate clinical diagnosis can effectively mislead the genome analysis and result in negative reports. This situation demonstrates yet another challenge involved in the NGS analysis of rare disorders. A typical genome from a healthy individual harbours at least ∼100 LoF variants which are not related to any disorders[531]. Moreover, the performance of annotation tools in prioritising pathogenic variants is highly context-dependent and widely varies depending on the disease phenotype[532]. To overcome this issue, I devised a hypothesis-free framework for ranking pathogenic variants based on the cumulative pathogenicity burden of novel variants identified in each patient. This practice is fully in compliance with the American College of Medical Genetics and Genomics (ACMG) guidelines that demands pathogenicity predictions be consulted from multiple tools[533]. The successful validation of Kabuki syndrome diagnosis in a patient with avail-

able photographs demonstrates the robustness of this method for molecular diagnosis. It is worth reiterating that establishing the molecular diagnosis, without access to up to date and in-depth phenotypic information, is practically impossible.

While variant-level attributes such as conservation and pathogenicity scores help to identify disease-associated variants, it does not provide any insight into the spatial distribution of these putatively causal mutations across the genome. Recent population expansions resulted in an excess load of rare variants across the human genome [534], however only a fraction of these mutations are pathogenic. Delineating the *disease genome*, which is defined as set of genes containing coding and regulatory variation that contribute to disorders, relies on better understanding of the interplay between selection, recombination and mutation. This complex relationship is reflected by the pattern of linkage disequilibrium (LD) and may have a close relationship with the disease genome [513,514].

In Chapter 6, I have used fine-scale LD maps to investigate the relationship between genome-wide LD patterns and the disease-genome. Extended population history in the sub-Saharan African populations enables delineation of LD patterns at an unprecedented resolution. Using WGS data from the 295 individuals from the three major SSA populations I constructed the LD maps according to the Malecot model and demonstrated that map lengths in the SSA populations correlate quite closely with the proposed model for Bantu expansion across Africa.

I also investigated the extent of LD at various genomic levels and demonstrated that LD is about ∼17% more extensive in genic regions. Fine-scale comparison of the LD extent at sub-genic level revealed that LD is also more extensive in the exonic regions (up to 52% for protein-coding genes and about 7% in ncRNA genes). The general trend revealed from the African genome compares quite closely with scale-corrected LD extent in the Caucasian population [510] and demonstrates the greater intensity of selection [535,482] and suppressed recombination in genic regions [513].

Even though LD is generally more extensive in genic regions, I demonstrated that there is substantial heterogeneity between gene ontology groups. It appears that ontology groups overrepresented in the strong LD tail are enriched for genes with nuclear functions (e.g. RNA synthesis and processing), whereas ontology groups in the weak LD tail are comprised of genes that primarily function in outer areas of the cell (e.g. extracellular matrix organisation). Since LD is primarily influenced by recombination, one may speculate that higher recombination rates are favourably selected among genes that experience varying degrees of selective pressures over time. Higher haplotype diversity rendered by an increased recombination rate across, for example, sensory or immune-related genes provides a more diverse population that is better adapted to cope with the ever-changing environment.

I also attempted to investigate the relationship between the gene-age and the extent of LD in genes identified to have a homologous counterpart in different taxa. I observed that genes with strong LD are overrepresented in lower taxa, whereas the weak LD genes tend to be overrepresented in vertebrate-specific genes. The fact that younger genes represent an intermediate LD extent perhaps underlies the varying intensity of selection that acts on these genes. This is an interesting result, as one may speculate that genes with intermediate LD signify recently evolved functions that are specific to higher taxa. It is worth mentioning that my approach to investigating the relationship between the gene-age and LD extent was quite exploratory and further analysis through simulation, is required to corroborate this relationship. Since a great deal of primate evolution is driven by *Alu* repeats [536,537], perhaps a more interesting project would be investigation of the relationship between the LD extent and evolutionary history of dispersed *Alu* repeats in the genome.

In the final analysis, I attempted to study the association between the gene-specific LDU size and gene-group assignment as either essential non-disease (END), Mendelian

non-complex (MNC), Complex non-Mendelian (CNM), Complex-Mendelian (CM) and Non-disease non-essential (NDNE). To do this, I adopted a multinomial regression method that models gene-group assignment as a covariate of gene-level annotation attributes such as gene age, GC content, gene essentiality score and e_{LDU} quartile rank. Using this model, I demonstrated that gene age and essentiality scores are significant predictors of group assignment across all categories. Inspection of model estimates for e_{LDU} quartile rank revealed that e_{LDU} scores in the lower band are generally positive predictors for END and MNC genes, whereas e_{LDU} scores in the upper band are generally associated with CM,CNM and NDNE genes. Taken together, this suggests that LD might be useful for the identification of disease-conferring genes. While the work described in Chapter 6 has revealed interesting insights about the relationship between the LD extent and gene attributes such as gene age, ontology essentiality and involvement in the human disorders, a number of questions remained to be answered. The most important question perhaps relates to the annotation attributes that explain the extent of LD at fine-scale. While I used regression residuals to account for the LDU variance attributable to the gene-size, it is clear that gene-length alone only partially explains the variance in the LDU size. Given the non-random distribution of recombination hotspots throughout the genome, probably a model that accounts for additional annotation attributes such as GC-content, distance to nearest recombination hotspot and distance from the centromeres to name a few, would generalise better and provide a better explanation for LDU variance at fine scale.

In conclusion, research detailed herein demonstrates the current status of NGS analysis in the field of rare disorders. Rapid advances in the sequencing technology along with developments in computational methods are changing the foreground of genomics. Many limitations of NGS technology attributable to short-read lengths are being resolved by new methods such as the 10X linked-read technology [538] or long-read sequencing [539]. Adoption of these methods by a greater number of diagnostic and research laboratories around the world promises a hopeful future for delineation of human disease genome. Soon, multi-layered frameworks that utilise the wealth of data from multiple high-throughput resources will be commonplace in the field of clinical genetics. It is an exciting time for studying genetic disorders since the impact of plausible variants revealed from NGS analysis can be tracked down through multiple layers of data including patients' transcriptome, microbiome and metabolome. This, in combination with advances in genome engineering technology and improved drug design, holds promise for enhanced diagnostic precision and treatment refinement for many genetic disorders.

Chapter 8

Appendices

8.1 Supplementary Data for Chapter 3

Family member	1*	23*	30*	20	36	37	49	26	27	28	39	40	43
Stones	Y multiple	Y	Y multiple	Y	N	N	N	N	N	N	N	N	N
Sex	F	M	M	M	M	F	M	F	M	F	F	M	F
Age	79	67	49	86	57	54	43	39	36	29	31	29	32
Relation to No.1	/	nephew	nephew	brother	nephew	niece	nephew	gt niece fath 23	gt neph fath 23	gt niece fath 23	gt niece fath 36	gt neph fath 36	gt niece
24h calcium	4.22	1.65 L	6.87	4.31	5.32	5.88	3.93	4.44	9.36 H	3.59	4.33	4.64	4.45
24h oxalate	0.31	0.27	0.34	0.42	0.44	0.52	0.45	0.28	0.29	0.49	0.28	0.56 H	0.34
24h citrate	2.15	3.47	3.76	1.09	2.64	3.82	3.72	5.74	7.16	3.93	5.33	3.79	5.33
Calcium/creat	0.53	0.12	0.48	0.42	0.12	0.84 H	0.18	0.44	0.43	0.16	0.16	0.26	0.34
Citrate/creat	0.33	0.24	0.23	0.10 L	0.14 L	0.4	0.19	0.67	0.34	0.19	0.35	0.16	0.37
TmPO4/GFR	0.70 L	0.64 L	0.78 L	0.60 L	0.8	/	0.91	0.82	0.72 L	0.48 L	0.95	1.08	1.21
Urine pH	6.8	/	6.2	6.2	/	/	5	/	5.2	5	/	/	/
1,25 vitamin D	70	67	145 H	/	111 H	/	56	107	93	102	107	63	94
PTH	2	3.2	7.5 H	10.6 H	5.9	/	1.9	1.2	2.6	2.2	2.7	3.8	1.8
Plasma phos	0.87	0.9	1.03	0.94	1.01	/	1.1	0.91	1.02	0.71	1.14	1.3	1.14
Plasma bicarb	31	29	26	27	25	/	28	28	29	26	26	29	29
Plasma calcium	2.59 H	2.31	2.43	2.35	2.24	/	2.26	2.32	2.43	2.36	2.34	2.29	2.39
Pl. creatinine	69	101	85	121 H	88	/	82	69	90	77	72	82	57
Pl. urate	0.22	0.2	0.32	0.31	0.29	/	0.23	0.11 L	0.24	0.2	0.09 L	0.26	0.13
Creat clear	93.9	117.4	194.4	75.0 L	148.8	/	190.5	139.1	202.0 H	154.8	157.4	186.5	167.5

Table 8.1: Results of biochemical assays (plasma, 24h urine and random urine tests) on 13 members of family-A; Results which are outside the reference ranges are highlighted in red; H indicates high & L indicates low values;*DNA from stone formers 1,23 and 30 were analysed by exome sequencing.
Ref. Ranges:
Plasma: calcium 2.15-2.55 mmol/l; phosphate 0.70-1.50 mmol/l; bicarbonate 24-31 mmol/l; PTH <7.3 pmol/l; 1,25 OH vitamin D 48-110 pmol/l; urate: men 0.15-0.45 mmol/l; women 0.12-0.36 mmol/l; creatinine men 80-115 μmol/l; women 53-97 μmol/l. ***24H urine***: calcium: men 2.00-7.50 mmol/324h; women 2.00-6.25 mmol/24h; oxalate $\leq$ 0.50 mmol/24h; citrate $\geq$ 1.60 mmol/24h; creatinine clearance men 135-200 l/24h; women 120-180 l/24h. ***Random urine (adults)***: calcium $\leq$ 0.59 mmol/mmol creat; citrate $\geq$ 0.16 mmol/mmol creat; TmPO4/GFR 0.80-1.35 mmol/l (low values indicate phosphaturia); PTH = parathyroid hormone; Y = yes; N = no

Family member	1*†	71*†	32*†	2*	21*	49	75	24	61	62	77	27	25*	105	23
Stones	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	N	N	N
Sex	M	F	F	M	F	M	M	F	M	M	M	M	F	M	F
Age	41	49	39	63	65	36	54	60	25	24	35	54	58	29	62
Relation to No.1	/	cousin	cousin	father	aunt	cousin	cousin	aunt	cousin	cousin	cousin's son	uncle	aunt	cousin's son	aunt
24h calcium	7.53 H	7.50 H	6.65 H	4.86	4.57	2.09	6.19	11.99 H	6.93	6.55	6.1	5.95	5.89	5.8	3.37
24h oxalate	0.49	0.36	0.32	0.34	0.47	0.44	0.41	0.4	0.25	0.34	0.32	0.25	0.28	0.27	0.53 H
24h citrate	2.57	3.38	2.74	0.05 L	2.61	4.06	4.87	6.49	2.4	1.37 L	/	1.6	3.12	3.78	3.98
Calcium/creat	0.42	0.65 H	0.34	0.25	0.36	0.11	0.82 H	0.66 H	0.51	0.3	0.5	0.54	0.68 H	0.34	0.19
Citrate/creat	0.16	0.2	0.09 L	0.16	0.2	0.19	0.45	0.35	0.08 L	0.09 L	0.08 L	0.14 L	0.32	0.2	0.31
TmPO4/GFR	0.08 L	0.58 L	1.18	0.70 L	0.88	1.2	0.70 L	0.95	1.02	0.82	/	0.98	0.63 L	1	0.96
Urine pH	5.6	5.6	7	6	6.6	5.8	5.5	6.3	5.3	6.2	/	5.2	5.3	5.5	5.8
1,25 vit D	161 H	99	147 H	134 H	81	99	88	112 H	102	58	/	70	97	84	112 H
PTH	1.4	4.1	4.7	4.8	4.5	7.4 H	4.6	3.6	3.3	1.5	/	2.5	11.4 H	5.7	4.6
Plasma phos	0.87	0.8	1.27	0.89	1.06	1.28	1	1.17	1.4	1.14	/	1.19	0.85	1.16	1.06
Plasma bicarb	28	25	25	27	25	25	31	28	27	29	/	22 L	26	32	27
Plasma calcium	2.4	2.25	2.32	2.33	2.31	2.22	2.43	2.38	2.44	2.45	/	2.35	2.67 H	2.38	2.33
Pl. creatinine	107	80	67	92	82	82	98	98	101	112	/	99	81	79	73
Pl. urate	0.22	0.27	0.14	0.29	0.25	/	0.28	0.26	0.29	0.33	/	0.35	0.33	0.3	0.25
Creat clearance	145.4	163.8	137.4	146.7	136.9	195.3	113.1	176.8	205.8 H	210.4 H	/	124.8	109	150.8	153

Table 8.2: Results of biochemical assays (plasma, 24h urine and random urine tests) on 28 members of family-B; Results outside the reference range are highlighted in red; H indicates high & L indicates low values; Highlighted columns represent individuals present with kidney stone; *DNA from stone formers 1,23 and 30 were analysed by exome sequencing.

Ref. Ranges:

Plasma: calcium 2.15-2.55 mmol/l; phosphate 0.70-1.50 mmol/l; bicarbonate 24-31 mmol/l; PTH <7.3 pmol/l; 1,25 OH vitamin D 48-110 pmol/l; urate: men 0.15-0.45 mmol/l; women 0.12-0.36 mmol/l; creatinine men 80-115 μmol/l; women 53-97 μmol/l. ***24H urine***: calcium: men 2.00-7.50 mmol/324h; women 2.00-6.25 mmol/24h; oxalate $\leq$ 0.50 mmol/24h; citrate $\geq$ 1.60 mmol/24h; creatinine clearance men 135-200 l/24h; women 120-180 l/24h. ***Random urine (adults)***: calcium $\leq$ 0.59 mmol/mmol creat; citrate $\geq$ 0.16 mmol/mmol creat; TmPO4/GFR 0.80-1.35 mmol/l (low values indicate phosphaturia); PTH = parathyroid hormone; Y = yes; N = no

Family member	48	74	34	89	88	78	26	103	101	43	42	41	20	3 "married-in"
Stones	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Sex	F	F	M	M	F	M	F	F	M	F	F	F	M	F
Age	38	51	?	20	25	34	55	27	31	35	43	47	75	62
Relation to No.1	cousin	cousin	cousin	cousin's son	cousin's daughter	cousin's son	aunt	cousin's daughter	cousin's son	cousin	cousin	cousin	uncle	mother
24h calcium	3.27	2.83	1.66 L	2.66	2.4	3.34	1.93 L	1.92 L	1.89 L	1.73 L	1.67 L	1.63 L	/	1.58 L
24h oxalate	0.42	0.45	/	0.21	0.18	0.22	0.51 H	0.14	0.26	0.32	0.4	0.22		0.65 H
24h citrate	2.88	3.48	/	0.99 L	2.58	1.16 L	2.49	1.71	2.12	0.78 L	1.09 L	/	/	2.49
Calcium/creat	0.2	0.41	/	0.12	0.09	0.31	0.21	0.15	0.1	0.15	0.18	0.22	0.46	0.59
Citrate/creat	0.19	0.23	/	0.06 L	0.17	0.06 L	0.3	0.17	0.21	0.10 L	0.11 L	0.04 L	0.04 L	0.28
TmPO4/GFR	0.94	0.83	/	1.46	0.77 L	1.22	1.2	0.73 L	0.74 L	0.99	0.93	0.82	/	0.98
Urine pH	5	7	/	5.3	5.5	7	8.6	6.1	7.1	5.8	7	5.5	/	6.4
1,25 vit D	98	44 L	74	60	66	52	97	67	56	38 L	52	54	64	149 H
PTH	6.1	5.3	/	1.6	1.5	1.3	1.6	1.1	2.4	2.7	3.8	2	2	4.8
Plasma phos	0.99	1.15	/	1.68 H	0.98	1.4	1.19	0.97	0.8	1.2	1.13	1.12	/	1.25
Plasma bicarb	26	26	/	29	25	32	24	25	29	30	21	29	/	28
Plasma calcium	2.37	2.48	/	2.59 H	2.28	2.45	2.37	2.42	2.52	2.2	2.38	2.34	/	2.53
Pl. creatinine	80	83	/	106	77	76	89	81	91	80	76	79	/	75
Pl. urate	0.17	0.40 H	/	0.22	0.13	0.29	0.22	0.16	0.32	0.27	0.27	0.24	/	0.18
Creat clearance	160.6	128.9	/	116.2	91.4	113	88.1	86.9	118.4	118.8	97.1	91.7	/	113.5

Table 8.3: Continued results of biochemical assays (plasma, 24h urine and random urine tests) for members of family-B; Results which are outside the reference ranges are highlighted in red; H indicates high & L indicates low values;*DNA from stone formers 1,23 and 30 were analysed by exome sequencing.
Ref. Ranges:
Plasma: calcium 2.15-2.55 mmol/l; phosphate 0.70-1.50 mmol/l; bicarbonate 24-31 mmol/l; PTH <7.3 pmol/l; 1,25 OH vitamin D 48-110 pmol/l; urate: men 0.15-0.45 mmol/l; women 0.12-0.36 mmol/l; creatinine men 80-115 μmol/l; women 53-97 μmol/l. ***24H urine***: calcium: men 2.00-7.50 mmol/324h; women 2.00-6.25 mmol/24h; oxalate $\leq$ 0.50 mmol/24h; citrate $\geq$ 1.60 mmol/24h; creatinine clearance men 135-200 l/24h; women 120-180 l/24h. ***Random urine (adults)***: calcium $\leq$ 0.59 mmol/mmol creat; citrate $\geq$ 0.16 mmol/mmol creat; TmPO4/GFR 0.80-1.35 mmol/l (low values indicate phosphaturia); PTH = parathyroid hormone; Y = yes; N = no

Family member	1	3*	8*	34*	13	4	10	15	23	26	33	35	37	38	39
Stones	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no
Sex	F	M	M	M	F	M	F	M	F	F	M	M	F	F	M
Age	50	24	54	52	46	19	22	19	83	85	59	49	22	19	17
24h calcium	7.31 H	9.94 H	6.36	7.32	6.76 H	7.67 H	6.44 H	-	-	-	-	8.20 H	5.12	4.47	12.18 H
24h oxalate	0.35	0.37	0.48	0.67 H	0.3	0.57 H	0.3	-	-	-	-	0.46	0.3	0.27	0.37
24h citrate	3.48	1.92	0.70 L	3.11	1.77	2.08	2	-	-	-	-	6.44	4.17	2.8	1.52 L
Calcium/creat	-	0.62 H	0.48	0.35	0.66	0.47	0.73 H	0.33	1.08 H	0.26	0.62	0.28	0.42	0.44	0.73 H
Citrate/creat	-	0.16	0.06 L	0.24	0.59	0.18	0.22	0.03 L	0.35	0.42	0.3	0.21	0.25	0.32	0.07 L
TmPO4/GFR	0.70 L	0.9	0.95	0.96	0.83	-	1.1	-	-	0.96	-	0.70 L	1.03	0.78 L	1.24
Urine pH	5.5	5.4	8.3	6.7	6.8	-	5.6	5.5	6.5	5.9	5	5.6	5.8	5.5	5.3
1,25 vit D	95	119 H	-	140 H	151 H	-	126 H	-	-	80	-	90	-	-	-
PTH	2.7	3.3	2.5	3.2	1.5	-	1.8	-	-	3.5	-	3.2	1.8	2.6	2.5
Plasma phos	0.85	1.03	1.25	1.15	1	-	1.19	-	-	1.1	-	0.86	1.18	0.97	1.36
Plasma bicarb	25	27	27	28	29	-	27	-	-	27	-	29	28	24	29
Plasma calcium	2.34	2.43	2.52	2.37	2.52	-	2.27	-	-	2.36	-	2.39	2.48	2.27	2.44
Pl. creatinine	77	88	105	92	78	-	67	-	-	83	-	97	61	66	72
Pl. urate	0.35	0.39	0.48 H	0.32	0.22	-	0.25	-	-	0.18	-	0.22	0.18	0.18	0.19
Creat clearance	163.1	155.4	125.7 L	152.4	104.0 L	-	175.4	-	-	-	-	219.7 H	151.9	127	202.7 H
Comments	No blood stored for DNA					No blood		No blood or 24h urine	No blood or 24h urine	No 24h urine	No blood or 24h urine				

Table 8.4: Results of biochemical assays (plasma, 24h urine and random urine tests) for 15 members of family-C; Results outside the reference range are highlighted in red; H indicates high & L indicates low values; Highlighted columns represent individuals present with kidney stone; *DNA from stone formers 3,08 and 34 were analysed by exome sequencing.

Ref. Ranges:

Plasma: calcium 2.15-2.55 mmol/l; phosphate 0.70-1.50 mmol/l; bicarbonate 24-31 mmol/l; PTH <7.3 pmol/l; 1,25 OH vitamin D 48-110 pmol/l; urate: men 0.15-0.45 mmol/l; women 0.12-0.36 mmol/l; creatinine men 80-115 μmol/l; women 53-97 μmol/l. ***24H urine***: calcium: men 2.00-7.50 mmol/324h; women 2.00-6.25 mmol/24h; oxalate $\leq$ 0.50 mmol/24h; citrate $\geq$ 1.60 mmol/24h; creatinine clearance men 135-200 l/24h; women 120-180 l/24h. ***Random urine (adults)***: calcium $\leq$ 0.59 mmol/mmol creat; citrate $\geq$ 0.16 mmol/mmol creat; TmPO4/GFR 0.80-1.35 mmol/l (low values indicate phosphaturia); PTH = parathyroid hormone; Y = yes; N = no

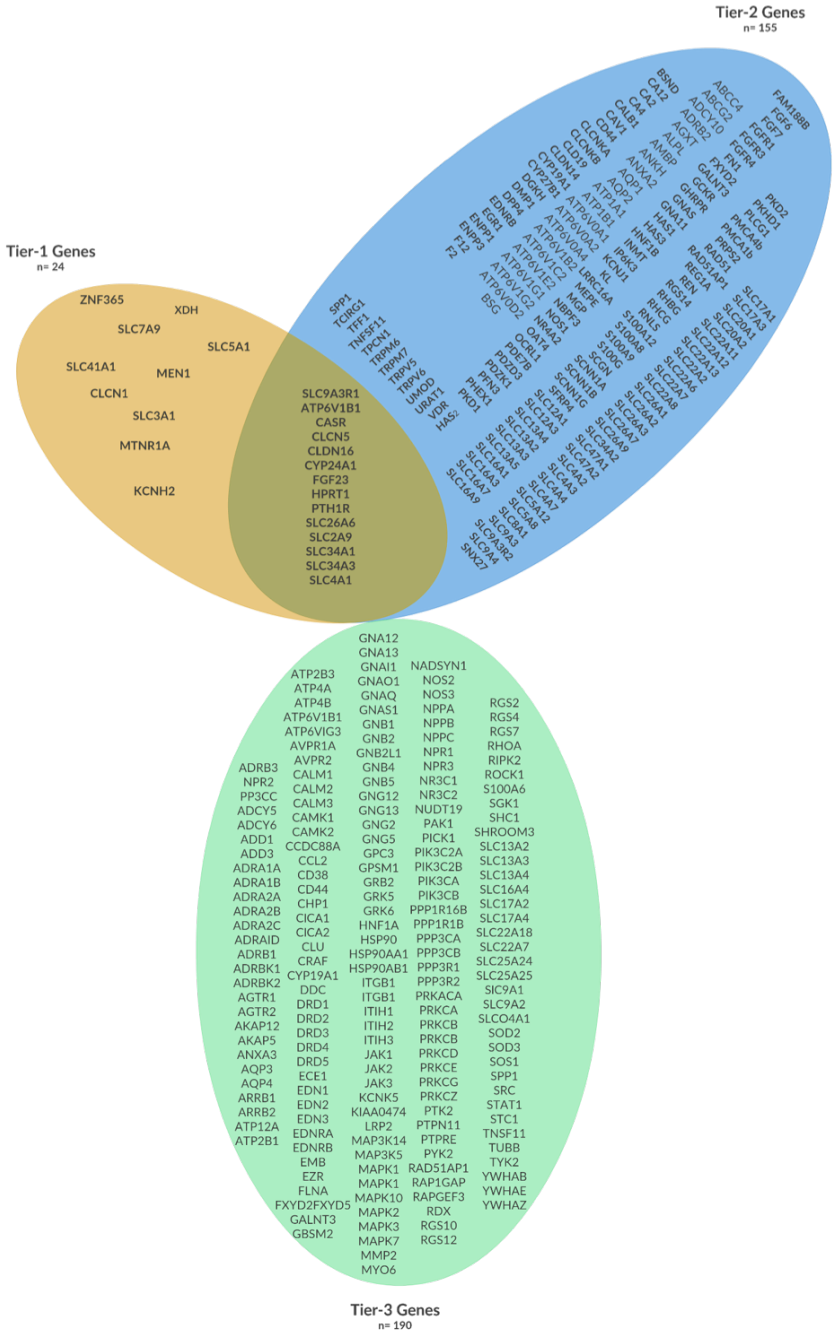


Figure 8.1: Venn representation of prioritised genes for variant analysis. Genes that are covered by both HGMD and OMIM database in the context of nephrolithiasis is represented in the overlapping region and considered among the tier-1 gene list

Table 8.5: Agilent SureSelect Human All Exon V.5 coverage efficiency for 367 genes considered in tiered filtering.

	Genes	Tier	Gene Size	Base Coverage	%Coverage
1	SLC9A3R1	Tier-1	2016	1561	77.4
2	ATP6V1B1	Tier-1	1939	1939	100.0
3	CASR	Tier-1	5265	4939	93.8
4	CLCN5	Tier-1	10385	9549	91.9
5	CLDN16	Tier-1	3273	3182	97.2
6	CYP24A1	Tier-1	3266	1953	59.8
7	FGF23	Tier-1	3018	3018	100.0
8	HPRT1	Tier-1	1415	1415	100.0
9	PTH1R	Tier-1	2146	1973	91.9
10	SLC26A6	Tier-1	3514	3406	96.9
11	SLC2A9	Tier-1	2079	1960	94.3
12	SLC34A1	Tier-1	3461	3404	98.4
13	SLC34A3	Tier-1	2433	2104	86.5
14	SLC4A1	Tier-1	4953	4820	97.3
15	SLC41A1	Tier-1	4854	4628	95.3
16	XDH	Tier-1	5717	5717	100.0
17	SLC3A1	Tier-1	2320	2320	100.0
18	MTNR1A	Tier-1	1105	1105	100.0
19	CLCN1	Tier-1	3175	3175	100.0
20	KCNH2	Tier-1	5717	5390	94.3
21	ZNF365	Tier-1	7522	7070	94.0
22	MEN1	Tier-1	4642	3791	81.7
23	SLC7A9	Tier-1	1987	1794	90.3
24	SLC5A1	Tier-1	5231	5056	96.7
25	ABCC4	Tier-2	6252	5978	95.6
26	ABCG2	Tier-2	4840	4429	91.5
27	ADCY10	Tier-2	5421	5104	94.2
28	ADRB2	Tier-2	2042	2042	100.0
29	AGXT	Tier-2	1598	1598	100.0
30	ALPL	Tier-2	2596	2443	94.1
31	AMBP	Tier-2	1433	1433	100.0
32	ANKH	Tier-2	8207	8091	98.6
33	ANXA2	Tier-2	1856	1670	90.0
34	AQP1	Tier-2	3451	3451	100.0
35	AQP2	Tier-2	4179	4179	100.0
36	ATP1A1	Tier-2	4300	3990	92.8
37	ATP1B1	Tier-2	2200	2200	100.0
38	ATP6V0A1	Tier-2	4244	2769	65.2
39	ATP6V0A2	Tier-2	6541	5696	87.1
40	ATP6V0A4	Tier-2	3357	3033	90.3
41	ATP6V1B2	Tier-2	3044	3018	99.1
42	ATP6V1C2	Tier-2	3240	3157	97.4
43	ATP6V1E2	Tier-2	1979	966	48.8
44	ATP6V1G1	Tier-2	1611	1244	77.2
45	ATP6V1G2	Tier-2	20524	2932	14.3
46	ATP6V0D2	Tier-2	2370	2307	97.3
47	BSG	Tier-2	2391	1832	76.6
48	BSND	Tier-2	1370	1370	100.0
49	CA12	Tier-2	3975	3975	100.0
50	CA2	Tier-2	1666	1666	100.0
51	CA4	Tier-2	1154	1079	93.5
52	CALB1	Tier-2	2531	2531	100.0
53	CAV1	Tier-2	3938	3601	91.4
54	CD44	Tier-2	6706	6706	100.0
55	CLCNKA	Tier-2	2617	2521	96.3
56	CLCNKB	Tier-2	3191	3005	94.2
57	CLD19	Tier-2	3191	3005	94.2
58	CLDN14	Tier-2	2568	1155	45.0
59	CYP19A1	Tier-2	4514	4301	95.3
60	CYP27B1	Tier-2	2487	2487	100.0
61	DGKH	Tier-2	4748	4549	95.8
62	DMP1	Tier-2	2686	2603	96.9
63	DPP4	Tier-2	3900	3900	100.0
64	EDNRB	Tier-2	6473	6290	97.2
65	EGR1	Tier-2	3136	3136	100.0
66	ENPP1	Tier-2	7442	7348	98.7
67	ENPP3	Tier-2	3165	3165	100.0
68	F12	Tier-2	2051	2051	100.0
69	F2	Tier-2	2010	2010	100.0
70	FAM188B	Tier-2	2733	2733	100.0
71	FGF6	Tier-2	744	744	100.0
72	FGF7	Tier-2	3936	932	23.7
73	FGFR1	Tier-2	6584	5500	83.5
74	FGFR3	Tier-2	4438	4438	100.0
75	FGFR4	Tier-2	3403	3131	92.0
76	FN1	Tier-2	9663	9663	100.0
77	FXYD2	Tier-2	669	371	55.5
78	GALNT3	Tier-2	3267	2988	91.5
79	GCKR	Tier-2	2188	2188	100.0
80	GHRPR	Tier-2	1328	152	11.4
81	GNAS	Tier-2	6204	5874	94.7
82	GNA11	Tier-2	1599	1293	80.9
83	HAS1	Tier-2	2087	1814	86.9
84	HAS3	Tier-2	4896	4425	90.4
85	HNF1B	Tier-2	3023	3023	100.0
86	INMT	Tier-2	2764	2562	92.7
87	IP6K3	Tier-2	2823	2466	87.4
88	KCNJ1	Tier-2	2890	2386	82.6
89	KL	Tier-2	5006	5006	100.0
90	LRRC16A	Tier-2	5626	5626	100.0
91	MEPE	Tier-2	2394	2287	95.5
92	MGP	Tier-2	1473	1473	100.0
93	NBPF3	Tier-2	4433	3442	77.6
94	NOS1	Tier-2	12534	11762	93.8
95	NR4A2	Tier-2	3531	3276	92.8
96	OAT4	Tier-2	2092	1979	94.6
97	OCRL1	Tier-2	5154	5154	100.0
98	PDE7B	Tier-2	5387	5286	98.1
99	PDZD3	Tier-2	2532	2532	100.0
100	PDZK1	Tier-2	2300	618	26.9
101	PFN3	Tier-2	530	530	100.0
102	PHEX1	Tier-2	2861	2861	100.0
103	PKD1	Tier-2	14188	6849	48.3
104	PKD2	Tier-2	5079	5079	100.0
105	PKHD1	Tier-2	17407	17215	98.9
106	PLCG1	Tier-2	5390	5390	100.0
107	PMCA4b	Tier-2	1936	1936	100.0
108	PMCA1b	Tier-2	1936	1936	100.0
109	PRPS2	Tier-2	2724	2724	100.0
110	RAD51	Tier-2	2562	2002	78.1
111	RAD51AP1	Tier-2	2269	2269	100.0
112	REG1A	Tier-2	808	751	92.9
113	REN	Tier-2	1465	1465	100.0
114	RGS14	Tier-2	2406	2406	100.0
115	RHBG	Tier-2	2247	2051	91.3
116	RHCG	Tier-2	1937	1488	76.8
117	RNLS	Tier-2	3585	2856	79.7
118	S100A12	Tier-2	466	418	89.7
119	S100A8	Tier-2	523	375	71.7
120	S100A9	Tier-2	577	549	95.1
121	S100G	Tier-2	453	407	89.8
122	SCGN	Tier-2	1479	1479	100.0
123	SCNN1A	Tier-2	4292	4292	100.0
124	SCNN1B	Tier-2	2597	2429	93.5
125	SCNN1G	Tier-2	3499	3404	97.3
126	SFRP4	Tier-2	2973	2973	100.0
127	SLC12A1	Tier-2	4805	4775	99.4
128	SLC12A3	Tier-2	5762	4556	79.1
129	SLC13A4	Tier-2	2894	2894	100.0
130	SLC13A2	Tier-2	2782	2641	94.9
131	SLC13A3	Tier-2	4371	4274	97.8
132	SLC13A5	Tier-2	3282	3282	100.0
133	SLC16A1	Tier-2	4697	3590	76.4
134	SLC16A3	Tier-2	2606	1524	58.5
135	SLC16A7	Tier-2	3541	3415	96.4
136	SLC16A9	Tier-2	3987	3574	89.6
137	SLC17A1	Tier-2	1843	1456	79.0
138	SLC17A3	Tier-2	1808	1598	88.4
139	SLC20A1	Tier-2	3377	3104	91.9
140	SLC20A2	Tier-2	4448	3564	80.1
141	SLC22A11	Tier-2	2512	2512	100.0
142	SLC22A12	Tier-2	4244	4244	100.0
143	SLC22A13	Tier-2	2560	2560	100.0
144	SLC22A2	Tier-2	2512	2512	100.0
145	SLC22A6	Tier-2	2609	2609	100.0
146	SLC22A7	Tier-2	2663	2663	100.0
147	SLC22A8	Tier-2	2606	2489	95.5
148	SLC26A1	Tier-2	4185	3739	89.3
149	SLC26A2	Tier-2	8075	7749	96.0
150	SLC26A3	Tier-2	2894	2771	95.7

	Genes	Tier	Gene Size	Base Coverage	%Coverage
151	SLC26A7	Tier-2	5764	5638	97.8
152	SLC26A9	Tier-2	6977	6881	98.6
153	SLC34A2	Tier-2	4387	4208	95.9
154	SLC47A1	Tier-2	3279	3273	99.8
155	SLC47A2	Tier-2	2495	2495	100.0
156	SLC4A2	Tier-2	5774	5027	87.1
157	SLC4A3	Tier-2	4710	4534	96.3
158	SLC4A4	Tier-2	8146	4001	49.1
159	SLC4A7	Tier-2	8209	8209	100.0
160	SLC5A12	Tier-2	6253	6253	100.0
161	SLC5A8	Tier-2	3286	3286	100.0
162	SLC8A1	Tier-2	6244	6157	98.6
163	SLC9A3	Tier-2	2777	2777	100.0
164	SLC9A3R2	Tier-2	3694	2630	71.2
165	SLC9A4	Tier-2	4138	3734	90.2
166	SNX27	Tier-2	7085	6975	98.4
167	SPP1	Tier-2	1814	1663	91.7
168	TCIRG1	Tier-2	3220	3005	93.3
169	TFF1	Tier-2	492	492	100.0
170	TNFSF11	Tier-2	2727	2198	80.6
171	TPCN1	Tier-2	5349	5177	96.8
172	TRPM6	Tier-2	8632	8632	100.0
173	TRPM7	Tier-2	10403	7377	70.9
174	TRPV5	Tier-2	2486	2486	100.0
175	TRPV6	Tier-2	2907	2907	100.0
176	UMOD	Tier-2	2442	2415	98.9
177	URAT1	Tier-2	1585	817	51.5
178	VDR	Tier-2	5243	5078	96.9
179	HAS2	Tier-2	3275	2737	83.6
180	EDN1	Tier-3	2266	2266	100
181	EDN2	Tier-3	1243	1243	100
182	EDN3	Tier-3	5807	5807	100
183	EDNRA	Tier-3	4150	3689.35	88.9
184	EDNRB	Tier-3	6473	6291.756	97.2
185	ECE1	Tier-3	5966	5870.544	98.4
186	DRD1	Tier-3	3373	2600.583	77.1
187	DRD2	Tier-3	2699	2493.876	92.4
188	DRD5	Tier-3	2376	2376	100
189	DRD3	Tier-3	1824	1428.192	78.3
190	DRD4	Tier-3	1362	1362	100
191	DDC	Tier-3	2487	1847.841	74.3
192	ADRA1A	Tier-3	3946	3736.862	94.7
193	ADRA1B	Tier-3	2254	1886.598	83.7
194	ADRA1D	Tier-3	2666	2666	100
195	ADRA2A	Tier-3	3873	3873	100
196	ADRA2B	Tier-3	3266	3266	100
197	ADRA2C	Tier-3	1958	1958	100
198	ADRB3	Tier-3	2646	2646	100
199	ADRB1	Tier-3	2862	2862	100
200	AGTR1	Tier-3	2571	2015.664	78.4
201	AGTR2	Tier-3	2899	2727.959	94.1
202	AVPR2	Tier-3	4217	3854.338	91.4
203	AVPR1A	Tier-3	7831	2224.004	28.4
204	NR3C1	Tier-3	11755	8169.725	69.5
205	NR3C2	Tier-3	5898	5538.222	93.9
206	NPPA	Tier-3	854	854	100
207	NPPB	Tier-3	698	698	100
208	NPPC	Tier-3	701	480.185	68.5
209	NPR1	Tier-3	4185	4185	100
210	NPR2	Tier-3	3430	3430	100
211	NPR3	Tier-3	7682	7682	100
212	CYP19A1	Tier-3	4514	4301.842	95.3
213	GNAO1	Tier-3	7923	2709.666	34.2
214	GNAI1	Tier-3	3636	3319.668	91.3
215	GNAS	Tier-3	6204	5875.188	94.7
216	GNAQ	Tier-3	2196	2180.628	99.3
217	GNA12	Tier-3	4384	4384	100
218	GNA13	Tier-3	6311	1325.31	21
219	GNB1	Tier-3	3163	1587.826	50.2
220	GNB2	Tier-3	1664	1479.296	88.9
221	GNB4	Tier-3	6428	2886.172	44.9
222	GNB5	Tier-3	3282	3137.592	95.6
223	GNG2	Tier-3	3923	3436.548	87.6
224	GNG5	Tier-3	806	580.32	72
225	GNG12	Tier-3	4409	4228.231	95.9
226	GNG13	Tier-3	984	916.104	93.1
227	MAPK1	Tier-3	6206	1625.972	26.2
228	LAMTOR3	Tier-3	4393	4138.206	94.2
229	CCDC88A	Tier-3	10193	6870.082	67.4
230	PIK3CA	Tier-3	3709	3627.402	97.8
231	PIK3CB	Tier-3	6383	5004.272	78.4
232	PIK3C2B	Tier-3	7606	7088.792	93.2
233	PIK3C2A	Tier-3	8288	8188.544	98.8
234	RHOA	Tier-3	1921	1661.665	86.5
235	ROCK1	Tier-3	6648	6648	100
236	CD38	Tier-3	1491	1491	100
237	MAPK10	Tier-3	7070	6129.69	86.7
238	SRC	Tier-3	4201	3654.87	87
239	PTK2B	Tier-3	4923	3948.246	80.2
240	ADRBK1	Tier-3	3444	3240.804	94.1
241	ADRBK2	Tier-3	9055	2553.51	28.2
242	GRK5	Tier-3	2558	2558	100
243	GRK6	Tier-3	4174	4174	100
244	PICK1	Tier-3	2624	1897.152	72.3
245	PRKCA	Tier-3	8770	2166.19	24.7
246	PRKCZ	Tier-3	3079	2761.863	89.7
247	PRKCB	Tier-3	8665	8665	100
248	PRKCD	Tier-3	2835	2500.47	88.2
249	PRKCE	Tier-3	5520	5520	100
250	RAF1	Tier-3	3275	2885.275	88.1
251	RAP1GAP	Tier-3	3676	2624.664	71.4
252	RAP1GAP	Tier-3	3676	2624.664	71.4
253	SHC1	Tier-3	4017	3844.269	95.7
254	GRB2	Tier-3	3524	2963.684	84.1
255	SOS1	Tier-3	8318	8318	100
256	GPSM1	Tier-3	4917	4597.395	93.5
257	PAK1	Tier-3	3479	2967.587	85.3
258	MAP3K5	Tier-3	5195	5195	100
259	MAPK3	Tier-3	2244	1640.364	73.1
260	MAPK1	Tier-3	6206	1625.972	26.2
261	MAPK7	Tier-3	3402	2769.228	81.4
262	SGK1	Tier-3	3965	3965	100
263	ITGB1	Tier-3	4087	3731.431	91.3
264	RIPK2	Tier-3	2585	2585	100
265	GNB2L1	Tier-3	1109	1109	100
266	CALM1	Tier-3	4256	4256	100
267	CALM2	Tier-3	1302	1302	100
268	CALM3	Tier-3	2259	2259	100
269	PPP3CB	Tier-3	3232	3232	100
270	PRCC	Tier-3	2123	2123	100
271	PPP3R1	Tier-3	3011	3011	100
272	PPP3R2	Tier-3	3387	3387	100
273	NOS3	Tier-3	6185	5844.825	94.5
274	ADCY5	Tier-3	6257	6257	100
275	ADCY6	Tier-3	10085	9883.3	98
276	CAMK1	Tier-3	1492	1338.324	89.7
277	CAMK2N2	Tier-3	1360	1360	100
278	RAPGEF3	Tier-3	6842	6520.426	95.3
279	ATP6V1B1	Tier-3	1939	1939	100
280	ATP4A	Tier-3	3559	3559	100
281	ATP4B	Tier-3	1497	1497	100
282	ATP12A	Tier-3	4003	4003	100
283	PRKACA	Tier-3	2720	1256.64	46.2
284	PRKCB	Tier-3	8665	8665	100
285	PRKCG	Tier-3	3128	3128	100
286	PTPN11	Tier-3	6732	2531.232	37.6
287	LRP2	Tier-3	15735	15735	100
288	AKAP5	Tier-3	6487	1732.029	26.7
289	AKAP12	Tier-3	8789	5818.318	66.2
290	ARRB1	Tier-3	7522	3572.95	47.5
291	ARRB2	Tier-3	2361	2361	100
292	FLNA	Tier-3	8510	8373.84	98.4
293	YWHAB	Tier-3	3105	2806.92	90.4
294	YWHAE	Tier-3	1860	1826.52	98.2
295	YWHAZ	Tier-3	3749	1372.134	36.6
296	JAK1	Tier-3	5053	4881.198	96.6
297	JAK2	Tier-3	5285	4814.635	91.1
298	JAK3	Tier-3	5434	3923.348	72.2
299	TYK2	Tier-3	4248	3891.168	91.6
300	EZR	Tier-3	3342	3231.714	96.7

	Genes	Tier	Gene Size	Base Coverage	%Coverage
301	PTK2	Tier-3	4710	4432.11	94.1
302	MMP2	Tier-3	3843	3758.454	97.8
303	GPC3	Tier-3	2499	2499	100
304	TUBB	Tier-3	19865	2423.53	12.2
305	RDX	Tier-3	5874	4634.586	78.9
306	SPP1	Tier-3	1814	1663.438	91.7
307	SOD2	Tier-3	1940	1940	100
308	SOD3	Tier-3	1529	1339.404	87.6
309	CLU	Tier-3	3303	3303	100
310	S100A6	Tier-3	683	389.993	57.1
311	PPP1R1B	Tier-3	2178	1829.52	84
312	FXYD5	Tier-3	1930	1930	100
313	FXYD2	Tier-3	669	371.295	55.5
314	PPP1R16B	Tier-3	6251	6163.486	98.6
315	RGS2	Tier-3	1350	1350	100
316	RGS4	Tier-3	4549	4212.374	92.6
317	RGS7	Tier-3	2440	2159.4	88.5
318	RGS10	Tier-3	974	961.338	98.7
319	RGS12	Tier-3	7573	6770.262	89.4
320	HSP90AA1	Tier-3	4074	3874.374	95.1
321	HSP90AA1	Tier-3	4074	3874.374	95.1
322	HSP90AB1	Tier-3	2538	2441.556	96.2
323	NUDT19	Tier-3	2967	1166.031	39.3
324	NADSYN1	Tier-3	2430	2430	100
325	PTPRE	Tier-3	5511	5263.005	95.5
326	ANXA3	Tier-3	1618	1346.176	83.2
327	ITGB1	Tier-3	4087	3731.431	91.3
328	SLCO4A1	Tier-3	2776	2667.736	96.1
329	AQP3	Tier-3	1871	1871	100
330	AQP4	Tier-3	5632	5632	100
331	SLC13A3	Tier-3	4371	4274.838	97.8
332	SLC17A2	Tier-3	2375	1933.25	81.4
333	SLC17A4	Tier-3	3600	3517.2	97.7
334	HNF1A	Tier-3	3238	3238	100
335	SLC25A24	Tier-3	3899	3899	100
336	SLC25A25	Tier-3	4639	4639	100
337	SLC22A7	Tier-3	2663	2663	100
338	SLC16A4	Tier-3	3692	3260.036	88.3
339	SLC13A2	Tier-3	2782	2640.118	94.9
340	SLC22A18	Tier-3	1625	1542.125	94.9
341	SLC13A4	Tier-3	2894	2894	100
342	SIC9A1	Tier-3	4742	4599.74	97
343	SLC9A2	Tier-3	5442	5442	100
344	CICA1	Tier-3	3120	3120	100
345	CICA2	Tier-3	4024	4024	100
346	KCNK5	Tier-3	3783	3783	100
347	EMB	Tier-3	4309	4309	100
348	ATP2B1	Tier-3	6933	6933	100
349	ATP2B3	Tier-3	6574	6574	100
350	ADD3	Tier-3	4628	4095.78	88.5
351	ADD1	Tier-3	4333	4164.013	96.1
352	SHROOM3	Tier-3	11019	11019	100
353	STC1	Tier-3	3877	3877	100
354	CHP1	Tier-3	3228	2905.2	90
355	MAP3K14	Tier-3	4467	2961.621	66.3
356	MYO6	Tier-3	8662	5604.314	64.7
357	ATP6V1G3	Tier-3	678	631.218	93.1
358	STAT1	Tier-3	4659	4272.303	91.7
359	GALNT3	Tier-3	3267	2989.305	91.5
360	TNFSF11	Tier-3	2727	2197.962	80.6
361	NOS2	Tier-3	4206	4016.73	95.5
362	RAD51AP1	Tier-3	2269	2269	100
363	ITIH1	Tier-3	3329	3032.719	91.1
364	ITIH2	Tier-3	3189	3189	100
365	ITIH3	Tier-3	3037	3037	100
366	CD44	Tier-3	6706	6706	100
367	CCL2	Tier-3	747	747	100

Table 8.6: Rare variants identified across tier analyses in family A.

Chr.	LBP	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.A-1	p.A-23	p.A-30
1	16355728	CLCNKA	s	c.T1161G	p.L387L	-		0.01	-	-	0,3,3	* (R.D.=22)	het (R.D.=17)	* (R.D.=16)
1	16354590	CLCNKA	ns	c.A944T	p.Y315F	0.893102	2.11	0.18	0.215	0.207	9.58.167	het (R.D.=78)	* (R.D.=95)	het (R.D.=82)
1	16355287	CLCNKA	ns	c.C1000G	p.L334V	0.216257	-1.21	0.01	0.03	0.022	0,22,22	* (R.D.=109)	* (R.D.=122)	het (R.D.=135)
1	16378739	CLCNKB	s	c.G1455A	p.A485A	-		0.022	0.0301	0.027	0,19,19	* (R.D.=30)	* (R.D.=35)	het (R.D.=24)
1	21809750	*NBPF3*[y]	ns	c.G1563C	p.Q521H	-		-	-	-	162,11,173	hom (R.D.=8)	hom (R.D.=8)	hom (R.D.=8)
1	156354355	RHBG	fsd	c.1271delC	p.P424fs	-		-	-	0.00003	23,10,33	hom	* (R.D.=117)	* (R.D.=128)
1	154938954	SHC1	s	c.G693A	p.E231E	.		0.02	0.032442	0.03151	0,26,26	het (R.D.=28)	*	*
1	21571475	ECE1	sp	c.1278+7C¿T		.		0.01	0.031163	0.02344	2,29,31	het (R.D.=54)	*	het (R.D.=52)
1	154938214	SHC1	s	c.G1098C	p.S366S	.		0.01	0.043488	8.33E-06	1,27,28	*	het (R.D.=30)	*
2	71187104	ATP6V1B1	ns	c.G481A	p.E161K	0.999172	4.12	0.02	0.021	0.0306	1,16,17	* (R.D.=89)	het (R.D.=90)	* (R.D.=108)
2	216251538	FN1	ns	c.C4213T	p.R1405W	0.999392	1.82	0.0046	0.00814	0.0049	0,7,7	het (R.D.=36)	* (R.D.=65)	* (R.D.=63)
2	220499233	SLC4A3	s	c.G1653A	p.P551P	-		-	0.00104	0.00108	0,0,0	het (R.D.=74)	* (R.D.=47)	* (R.D.=72)
2	40392095	SLC8A1	ns	c.A2053G	p.S685G	0.997982	5.28	-	-	-	0,0,0	* (R.D.=67)	het (R.D.=66)	* (R.D.=70)
2	96780997	ADRA2B	nfi	c.891_892ins GAAGAGGAG	p.E298delinsEEEE	.		.	.		143,142,285	het (R.D.=32)	het (R.D.=33)	het (R.D.=28)
2	170089934	LRP2	ns	c.G5085A	p.S1695S	.		0.01	0.016047		1,11,12	het (R.D.=81)	*	*
2	170100011	LRP2	ns	c.C3452T	p.P1151L	0.994071	4.67	0.01	0.006279	0.009917	0,3,3	het (R.D.=111)	*	*
2	191862937	STAT1	sp	c.633+6T¿A		.		0.0005	0.004538	0.002994	0,1,1	* (R.D.=23)	het (R.D.=26)	het (R.D.=21)
2	170145661	LRP2	sp	c.923-6G¿A		.		0.0046	0.008023	0.01159	0,6,6	*	het (R.D.=71)	*
3	190106074	CLDN16	fsd	c.166delG	p.A56fs	-		-	-	0.000008	20,136,156	* (R.D.=77)	het (R.D.=56)	het (R.D.=64)
3	38316517	SLC22A13	s	c.C675T	p.A225A	-		0.0032	0.0037	-	0,1,1	het (R.D.=52)	* (R.D.=47)	* (R.D.=47)
3	52833881	ITIH3	ns	c.C1019T	p.T340M	.		0.01	0.044418	0.02665	2,37.39	het (R.D.=22)	het (R.D.=37)	*
4	10027519	SLC2A9	sp	c.63+9¿ TTTTTTTTC		-		-	-	-	2,0,2	* (R.D.=54)	* (R.D.=65)	hom (R.D.=76)
4	187455399	MTNR1A	ns	c.G497A	p.G166E	0.997169	4.39	0.01	0.016	-	0,6,6	* (R.D.=24)	* (R.D.=23)	* (R.D.=32)
4	983612	SLC26A1	ns	c.G1115A	p.R372H	0.998371	3.82	0.01	0.0086	0.0251	0,8,8	het (R.D.=21)	het (R.D.=17)	* (R.D.=19)
4	2877619	ADD1	sp	.		.		.	.	.	.	*	het (R.D.=16)	*
5	32710896	NPR3	sp	c.121+7G¿A		.		0.01	.	0.01616	0,13,13	*	*	het (R.D.=126)
5	142779317	NR3C1	ns	c.A83G	p.N28S	0.976987	3.5	0.01	0.030581	0.02076	0,27,27	*	*	het (R.D.=145)
5	176831085	F12	ns	c.C1025T	p.P342L	0.007516	-10.1	0.0014	0.00154	0.00112	0,0,0	* (R.D.=54)	het (R.D.=56)	* (R.D.=45)
6	25689391	SCGN	sp	c.528-9-¿ ATTTTTTTTTTT		-		-	-	-	57,0,57	* (R.D.=66)	hom (R.D.=60)	hom (R.D.=69)
6	25653667	SCGN	ns	c.A140G	p.K47R	0.973395	3.87	-	-	-	0,0,0	* (R.D.=49)	het (R.D.=53)	* (R.D.=69)
6	25556962	LRRC16A	ns	c.C2626T	p.P876S	-		0.0014	0.00072	0.001	0,2,2	* (R.D.=43)	* (R.D.=41)	het (R.D.=37)
6	151672185	AKAP12	ns	c.A2659T	p.S887C	0.975398	3.07	0.01	0.033372	0.02319	2,28,30	*	het (R.D.=43)	*
6	151674121	AKAP12	nfi	c.4595_4596 insGGA	p.D1532delinsED	.		.	.		279,145,424	HOM (R.D.=109)	HOM (R.D.=128)	HOM (R.D.=146)
6	151671747	AKAP12	ns	c.A2221C	p.S741R	0.917402	3.18	0.0032	0.003256	0.005734	0,1,1	*	*	het (R.D.=85)
6	136882715	MAP3K5	ns	c.G3943A	p.D1315N	0.998946	4.37	0.0037	0.009302	0.006569	0,9,9	het (R.D.=153)	het (R.D.=168)	het (R.D.=167)
6	39159407	KCNK5	s	c.C759T	p.H253H	.		0.01	0.019767	0.0166	0,21,21	het (R.D.=53)	*	*
6	44216483	HSP90AB1	s	c.C117T	p.F39F	.		0.01	0.007907	0.008319	0,6,6	het (R.D.=85)	* (R.D.=110)	het (R.D.=88)
7	30795466	INMT	stp_los	c.G791C	p.X264S	0.919282	0.413	0.02	0.036	0.02492	1,20,21	* (R.D.=9)	het (R.D.=17)	* (R.D.=15)
7	30951658	AQP1	ns	c.C134T	p.A45V	0.78671	-0.0543	0.02	0.04	0.026	0,24,24	* (R.D.=47)	* (R.D.=62)	het (R.D.=62)
7	138453908	ATP6V0A4	sp	c.196+10-¿ TTTTTTTTTTTTT		-		-	-	-	38,0,38	* (R.D.=35)	* (R.D.=45)	hom (R.D.=27)
8	92346523	SLC26A7	ns	c.A643G	p.I215V	0.982422	4.49	0.02	0.03	0.025	1,24,25	* (R.D.=15)	het (R.D.=37)	* (R.D.=28)
8	90796369	RIPK2	sp	.		.		.	.	0.03694	0,2,2	*	*	het (R.D.=4)
9	140127725	SLC34A3	ns	c.C625T	p.L209L	-		-	0.006	0.00003	0,5,5	het (R.D.=37)	* (R.D.=43)	* (R.D.=42)
9	140128858	SLC34A3	sp	c.1094-10T¿A		-		0.14	0.11	0.1343	5,69,74	het (R.D.=78)	* (R.D.=68)	* (R.D.=71)
9	139235485	GPSM1	fd	c.1242delC	p.P414fs	.		.	.	.	2,101,103	het (R.D.=12)	*	het (R.D.=10)
9	130868670	SLC25A25	ns	c.G1047C	p.Q349H	0.999532	5.08	0.0018	0.002209	0.002637	0,2,2	het (R.D.=48)	het (R.D.=57)	het (R.D.=52)
10	7759650	ITIH2	ns	c.G529A	p.V177M	0.999479	4.99	0.0009	0.005116	0.002908	0,5,5	*	*	het (R.D.=12)

Chr.	LBP	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.A-1	p.A-23	p.A-30
11	64577147	**MEN1**	s	c.C435T	p.S145S	-		0.02	0.035	0.02862	0,39,39	* (R.D.=57)	* (R.D.=77)	het (R.D.=65)
11	46751059	**F2**	s	c.G1602A	p.P534P	-		0.01	0.024	0.01612	0,16,16	het (R.D.=52)	het (R.D.=51)	* (R.D.=55)
11	67811321	**TCIRG1**	ns	c.G754C	p.A252P	0.999414	4.6	-	-	-	0,13,13	* (R.D.=31)	* (R.D.=28)	het (R.D.=30)
11	77034400	**PAK1**	ns	c.A1606G	p.I536V	.		0.0005	.	8.33E-06	0,0,0	het (R.D.=29)	*	*
11	17150917	**PIK3C2A**	ns	c.G2329A	p.D777N	0.999392	5.82	.	0.000815	0.000833	0,0,0	*	het (R.D.=34)	*
11	637391	**DRD4**	nfd	c.76_87del	p.26_29del	.		.	.	.	0,30,30	*	*	het (R.D.=6)
12	48240233	**VDR**	s	c.C1059T	p.A353A	-		0.0009	0.00325	0.00339	0,3,3	* (R.D.=40)	* (R.D.=45)	het (R.D.=65)
12	50344652	**AQP2**	s	c.G39A	p.V13V	-		0.0027	0.0069	0.0058	0,5,5	* (R.D.=52)	* (R.D.=42)	het (R.D.=42)
12	6472600	**SCNN1A**	sp	c.861+9C¿T		-		0.0014	0.0057	0.0145	0,0,0	* (R.D.=11)	* (R.D.=6)	het (R.D.=8)
12	101573812	**SLC5A8**	s	c.T1228C	p.L410L	-		0.01	0.0356	0.027	0,21,21	* (R.D.=58)	het (R.D.=54)	* (R.D.=56)
12	117768346	**NOS1**	ns	c.G529A	p.G177S	0.897205	1.99	-	0.00012	0.000008	0,0,0	* (R.D.=39)	* (R.D.=43)	het (R.D.=36)
12	117705934	**NOS1**	ns	c.A1855T	p.M619L	0.996509	4.18	0.0014	0.0048	0.0026	0,2,2	het (R.D.=31)	* (R.D.=22)	het (R.D.=52)
15	52433397	**GNB5**	s	c.G567A	p.K189K	.		0.02	0.036921	0.02666	2,23,25	*	het (R.D.=84)	*
15	52446260	**GNB5**	s	c.G252A	p.A84A	.		0.02	0.036921	0.02703	2,24,26	*	het (R.D.=27)	*
16	2160622	**PKD1**	ns	c.G4546A	p.A1516T	4.24E-04	-5.82	0.0023	0.0071	0.00608	0,2,2	* (R.D.=6)	het (R.D.=4)	* (R.D.=4)
16	2143865	**PKD1**	s	c.C10765T	p.L3589L	-		0.01	0.03344	0.04416	2,24,26	* (R.D.=49)	het (R.D.=53)	* (R.D.=41)
16	2140972	**PKD1**	s	c.C11913T	p.R3971R	-		0.01	0.0276	0.02488	2,20,22	* (R.D.=51)	het (R.D.=48)	* (R.D.=52)
16	56901062	**SLC12A3**	ns	c.G363C	p.E121D	0.999612	3.94	-	0.0011	0.00077	0,0,0	het (R.D.=42)	het (R.D.=44)	* (R.D.=36)
17	26816369	**SLC13A2**	sp	c.231+9-¿C		-		-	-	0.000016	73,211,284	hom (R.D.=50)	hom (R.D.=47)	hom (R.D.=61)
18	18533532	**ROCK1**	sp	c.4061+7A¿T		.		.	0.029516	0.009133	0,9,9	het (R.D.=56)	*	*
19	580710	**BSG**	s	c.G180A	p.L60L	-		0.02	0.0372	0.025	0,28,28	het (R.D.=114)	* (R.D.=124)	* (R.D.=129)
20	43535058	**YWHAB**	s	c.C720T	p.D240D	.		0.0009	0.000233	0.000256	0,0,0	*	*	het (R.D.=8)
20	57415812	**GNAS**	s	c.T651A	p.R217R	-		0.01	0.0082	0.0059	0,7,7	* (R.D.=66)	het (R.D.=50)	* (R.D.=41)
X	153587448	**FLNA**	ns	c.C4378T	p.P1460S	0.96666	2.88	.	.	.	0,0,0	het (R.D.=61)	*	*

The traffic light colour code corresponds to zygosity of variants. Homozygous reference variants are identified in green, heterozygous variants in red and homozygous alternative variants in orange.

Table 8.7: Rare variants identified across tier analyses in family B.

Chr.	LBP	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.B-1	p.B-32	p.B-71
1	21809750	NBPF3	ns	c.G1563C	p.Q521H	-		-	-	-	162,11,173	het (R.D.=44)	hom (R.D.= 10)	* (R.D.=36)
1	156347121	RHBG	ns	c.G217A	p.V73M	-		-	0.0008	0.0008	0,0,0	het (R.D.=102)	* (R.D.=78)	* (R.D.=103)
1	16355728	CLCNKA	s	c.T1161G	p.L387L	-		0.01	-	-	0,3,3	* (R.D.=20)	het (R.D.=12)	het (R.D.=24)
1	167792356	ADCY10	ns	c.C3599T	p.P1200L	0.158114,	-1.31	0.02	0.04	0.029	0,29,29	het (R.D.=71)	het (R.D.=55)	het (R.D.=62)
1	167787307	ADCY10	sp	c.4482+3G¿A		-		0.01	0.003	0.006	0,1,1	* (R.D.=77)	het (R.D.=42)	* (R.D.=54)
1	86891073	CLCA2	ns	c.G238A	p.V80I	0.926891	3.37	0.02	0.042134	0.03036	1,25,26	het (R.D.=26)	*	*
1	86961291	CLCA1	ns	c.C2046G	p.N682K	0.001815	-10.1	0.0046	0.018605	0.012	0,8,8	het (R.D.=64)	*	*
1	204423811	PIK3C2B	s	c.C2052T	p.I684I	.		0.02	0.034186	0.0308	0,19,19	*	het (R.D.=4)	*
1	21571475	ECE1	sp	c.1278+7C¿T		.		0.01	0.031163		2,29,31	*	*	het (R.D.=41)
2	44528256	SLC3A1	ns	c.G1126A	p.G376S	0.999422,	4.91	0.0009	0.0005	0.0004	0,0,0	het (R.D.= 64)	* (R.D.=59)	* (R.D.=55)
2	113404974	SLC20A1	s	c.A408C	p.A136A	-		0.004	0.01	0.007	0,4,4	* (R.D.=151)	het (R.D.=99)	* (R.D.=107)
2	216272025	FN1	s	c.G2538A	p.S846S	-		-	-	0.00004	0,0,0	* (R.D.=89)	hom (R.D.= 59)	* (R.D.=55)
2	216251538	FN1	ns	c.C4213T	p.R1405W	0.999392,	1.82	0.0046	0.0081	0.0049	0,7,7	* (R.D.=58)	* (R.D.=41)	het (R.D.=43)
2	216236712	FN1	ns	c.A6091G	p.I2031V	0.999155,	4.4	0.02	0.02	0.016	0,18,18	het (R.D.=94)	* (R.D.=67)	* (R.D.=70)
2	220505237	SLC4A3	s	c.C3363T	p.S1121S	-		0.0009	0.001	0.00074	0,3,3	* (R.D.=103)	* (R.D.=79)	het (R.D.=86)
2	96780997	ADRA2B	nfi	c.891_892ins GAAGAGGAG	p.E298del insEEEE	.	.	.	.	.	143,142,285	het (R.D.=103)	het (R.D.=91)	het (R.D.=92)
2	45879305	PRKCE	s	c.C66T	p.A22A	.	.	0.02	0.014425	0.0139	0,4,4	het (R.D.=130)	*	het (R.D.=105)
3	190106074	CLDN16	fsd	c.166delG	p.A56fs	-		-	-	0.000008	20,136,156	het (R.D.=66)	* (R.D.=62)	* (R.D.=50)
3	52833881	ITIH3	ns	c.C1019T	p.T340M	.	.	0.01	0.044418	0.02665	2,37,39	*	het (R.D.=26)	het (R.D.=75)
4	72433545	SLC4A4	ns	c.A3220C	p.I1074L	-		0.16	0.186	0.26	23,123,146	het (R.D.=20)	het (R.D.=13)	het (R.D.=13)
5	176825069	SLC34A1	ns	c.C1702T:	p.H568Y	0.105189,	-2.88	0.01	0.033	0.022	1,28,29	* (R.D.=100)	* (R.D.=76)	het (R.D.=80)
5	176813234	SLC34A1	nfsd	c.272_292del	p.91_98del	-		-	-	-	1,23,24	* (R.D.=122)	* (R.D.=74)	het (R.D.=59)
5	137801600	EGR1	s	c.C150A	p.G50G	-		-	-	-	0,0,0	het (R.D.=101)	* (R.D.=67)	* (R.D.=81)
5	142779317	NR3C1	ns	c.A83G	p.N28S	0.976987	3.5	0.01	0.030581	0.02076	0,27,27	*	*	het (R.D.=127)
6	25689391	SCGN	sp	c.528-9-¿ ATTTTTTTTTTTT		-		-	-	-	57,0,57	hom (R.D.=56)	* (R.D.=60)	hom (R.D.=58)
6	160679518	SLC22A2	ns	c.G272A	p.R91H	0.239867,	-0.609	-	0.0001	0.00003	0,0,0	* (R.D.=56)	* (R.D.=67)	het (R.D.=61)
6	44220889	HSP90AB1	s	c.C1695T	p.D565D	.		0.0018	0.006628		0,4,4	het (R.D.=43)	*	het (R.D.=30)
6	39162068	KCNK5	ns	c.G511A	p.V171I	0.998817	5.2	0.0009	0.001047	0.001023	0,1,1	*	het (R.D.=53)	*
6	134583262	SGK1	ns	c.A94G	p.M32V	.		0.01	0.010329		0,17,17	*	*	het (R.D.=37)
6	151670656	AKAP12	ns	c.T1130C	p.V377A	0.95695	3.8	0.0027	0.008256	0.006506	0,6,6	*	*	het (R.D.=38)
6	76550348	MYO6	s	c.C600T	p.N200N	.		0.0005	0.000581	0.00009962	0,1,1	*	*	het (R.D.=43)
6	151672573	AKAP12	ns	c.C3047T	p.T1016I	0.998412	3.6	0.0023	0.001744	0.001363	0,0,0,	*	*	het (R.D.=71)
6	151674121	AKAP12	nfi	c.4595_4596 insGGA	p.D1532 delinsED	.	.	.	.	.	279,145,427	hom (R.D.=167)	hom (R.D.=102)	hom (R.D.=125)
7	107423773	SLC26A3	s	c.C996T	p.D332D	-		0.0046	0.0029	0.009	0,4,4	* (R.D.=66)	* (R.D.=47)	het (R.D.=68)
7	107427322	SLC26A3	ns	c.T921G	p.C307W	0.999107,	5.08	0.02	0.0456	0.04251	1,54,55	* (R.D.=105)	* (R.D.=65)	het (R.D.=81)
7	138453908	ATP6V0A4	sp	c.196+10-¿ TTTTTTTTTTTTT		-		-	-	-	38,0,38	* (R.D.=20)	hom (R.D.=20)	* (R.D.=30)
7	79840384	GNAI1	s	c.C690T	p.Y230Y	.		0.0005	0.002209	0.00147	0,0,0	*	het (R.D.=87)	*
7	150698349	NOS3	ns	c.G1264A	p.A422T	0.999084	4.58	0.0018	0.004884	0.003881	0,6,6	*	*	het (R.D.=19)
8	38287213	FGFR1	s	c.C345T	p.S115S	-		0.01	0.006	0.008	0,0,0	het (R.D.=40)	* (R.D.=25)	het (R.D.=50)
8	26721888	ADRA1A	ns	c.T599G	p.I200S	0.998437	5.07	0.02	0.024535	0.02458	0,20,20	*	het (R.D.=92)	het (R.D.=75)
10	64416221	ZNF365	ns	c.G457A	p.V153I	0.850042,	0.323	-	-	0.00004	0,0,0	* (R.D.=71)	het (R.D.= 56)	* (R.D.=81)
11	67810414	TCIRG1	ns	c.T419C	p.L140P	0.05649,	-0.428	-	-	-	0,0,0	het (R.D.=32)	het (R.D.=19)	* (R.D.=21)
11	17112977	PIK3C2A	s	c.A4782G	p.P1594P	.		0.0032	0.007685	0.004556	0,2,2	het (R.D.=82)	*	*
11	637380	DRD4	nfd	c.76_87del	p.26_29del	.	.	.	.	.	0,30,30	het (R.D.=11)	*	*
11	113287694	DRD2	s	c.G423A	p.L141L	.	.	0.01	0.039223	0.0286	0,27,27	*	het (R.D.=36)	het (R.D.=33)
12	50344652	AQP2	s	c.G39A	p.V13V	-		0.0027	0.006	0.0058	0,5,5	* (R.D.=38)	het (R.D.=30)	* (R.D.=35)
12	6458350	SCNN1A	ns	c.T1546C	p.W516R	0.998129,	4.88	0.01	0.02	0.018	1,15,16	* (R.D.=25)	* (R.D.=12)	het (R.D.=16)
12	113727960	TPCN1	s	c.C2040T	p.R680R	-		0.0023	0.007	0.009	0,2,2	* (R.D.=7)	het (R.D.=7)	* (R.D.=6)
12	58158558	CYP27B1	s	c.G942A	p.L314L	-		0.02	0.03	0.023	0,26,26	* (R.D.=110)	* (R.D.=77)	het (R.D.=78)
12	49168506	ADCY6	ns	c.G2134T	p.V712L	0.977146	2.55	.	.	.	0,0,0	*	*	het (R.D.=17)
13	42733399	DGKH	sp	c.623-3-¿ TTTTTTTTTTTTTT		-		-	-	-	0,0,0	* (R.D.=19)	hom (R.D.=18)	* (R.D.=25)
15	52433397	GNB5	s	c.G567A	p.K189K	.	.	0.02	0.036921	0.02666	2,23,25	*	het (R.D.=81)	*
15	52446260	GNB5	s	c.G252A	p.A84A	.	.	0.02	0.036921	0.02703	2,24,26	*	het (R.D.=25)	*
16	2147371	PKD1	ns	c.G10351C	p.G3451R	0.004436,	-6.39	-	-	-	0,0,0	* (R.D.=53)	het (R.D.=37)	* (R.D.=64)
16	2143865	PKD1	s	c.C10765T	p.L3589L	-		0.01	0.03	-	2,24,26	* (R.D.=45)	het (R.D.=32)	* (R.D.=47)
16	2140972	PKD1	s	c.C11913T	p.R3971R	-		0.01	0.02	0.024	2,20,22	* (R.D.=60)	het (R.D.=32)	* (R.D.=47)
17	26816369	SLC13A2	sp	c.231+9-¿C		-		-	-	0.00001	73,211,284	het (R.D.=71)	hom (R.D.=50)	* (R.D.=55)
17	80193938	SLC16A3	s	c.C54A	p.G18G	-		-	-	0.00002	0,0,0	* (R.D.=73)	het (R.D.=33)	* (R.D.=33)
17	43364294	MAP3K14	u	UNKNOWN		.		.	.		423,0,423	hom (R.D.=101)	hom (R.D.=64)	hom (R.D.=75)
17	19284899	MAPK7	s	c.C1377T	p.V459V	.	.	.	0.000698	0.0002315	0,1,1	*	*	het (R.D.=62)
17	42333071	SLC4A1	s	c.G1770A	p.K590K	-		0.01	0.021	0.015	0,13,13	* (R.D.=23)	* (R.D.=21)	het (R.D.=16)
19	580710	BSG	s	c.G180A	p.L60L	-		0.02	0.03	0.025	0,28,28	het (R.D.=161)	* (R.D.=102)	* (R.D.=99)
19	33183575	NUDT19	ns	c.T709A	p.Y237N	0.940848	2.69	.	0.001093	0.000649	0,3,3	*	*	het (R.D.=8)
X	49854843	CLCN5	s	c.C1815T	p.I605I	-		-	-	0.00001	0,0,0	hom (R.D.= 40)	* (R.D.=62)	* (R.D.=79)
X	132826404	GPC3	ns	c.G1285A	p.V429M	0.998019	4.72	0.01	0.003716	0.005827	0,0,0	*	het (R.D.=25)	*

Table 8.8: Rare variants identified across tier analyses in family C.

Chr.	LBP	Gene	Variant Type	Nucleotide	Protein	PolyPhen2	Gerp++	MAF in 1000 Genomes	MAF in EVS	MAF in ExAC (n=60,706)	SED (n=460)	p.C-3	p.C-8	P.C-34
1	86961291	CLCA1	ns	c.C2046G	p.N682K	0.041	-10.3	0.0026	0.013	0.012	0.0181	*	het (R.D.= 33)	*
1	86891073	CLCA2	ns	c.G238A	p.V80I	0.015	3.08	0.016	0.031	0.0304	0.0327	*	het (R.D.= 11)	*
1	16355287	CLCNKA	ns	c.C871G	p.L291V	0.015	-1.2	0.01	0.021	0.0224	0	*	het (R.D.= 107)	*
1	43204109	CLDN19	ns	c.C371T	p.A124V	0	0.671	.	.	.	.	het (R.D.= 121)	*	het (R.D.= 105)
1	21806667	NBPF3	ns	c.C1122G	p.D374E	0.001	-1.32	.	0.13	0.0851	0.1514	hom (R.D.= 4)	*	*
1	21806710	NBPF3	ns	c.T1165G	p.L389V	0	-1.32	.	0.14	0.2165	0.2953	hom (R.D.= 5)	*	*
1	204434438	PIK3C2B	ns	c.C943T	p.R315W	0.711	3.35	0.018	0.039	0.0393	0.0404	het (R.D.= 28)	*	*
1	205890810	SLC26A9	ns	c.G1939A	p.G647S	0	-4.27	.	.	0.0001		*	*	het (R.D.= 38)
2	96781257	ADRA2B	ns	c.G632C	p.G211A	0.005	1.06	0.017	0.041	0.0362	0	*	het (R.D.= 93)	*
2	170129547	LRP2	ns	c.G2006A	p.G669D	0.899	1	0.015	0.025	0.0285	0.0269	*	het (R.D.= 26)	*
2	113416643	SLC20A1	ns	c.G1020C	p.E340D	0.004	3.69	.	.	.	1.22E-05	het (R.D.= 20)	het (D.P.= 10)	het (D.P.= 27)
2	39249741	SOS1	ns	c.A1828G	p.I610V	0.98	5.78	.	.	8.28E-06	4.08E-06	het (R.D.= 35)	*	*
2	31593265	XDH	ns	c.A1936G	p.I646V	0.026	0	0.019	0.025	0.0249	0.0058	*	het (R.D.= 17)	*
3	179143945	GNB4	ns	c.G44A	p.R15Q	0.027	5.26	.	0.0002	6.60E-05		*	*	het (R.D.= 27)
4	983612	SLC26A1	ns	c.G1115A	p.R372H	0.082	3.36	0.0072	0.0063	0.0251	0	*	het (R.D.= 28)	*
5	149360877	SLC26A2	ns	c.T1721C	p.I574T	0	4.37	.	0.99	0.9928	0.9931	hom (R.D.= 106)	hom (D.P.= 55)	hom (D.P.= 67)
5	176824011	SLC34A1	ns	c.C1352T	p.T451I	0.934	5.39	.	0.0001	7.49E-05	.	*	het (R.D.= 34)	het (R.D.= 32)
5	176825069	SLC34A1	ns	c.C1702T	p.H568Y	0	0.517	0.0088	0.025	0.0227	0.0364	*	het (R.D.= 41)	hom (R.D.= 151)
5	475104	SLC9A3	ns	c.T2368C	p.C790R	0	3.84	.	0.87	0.8253	0.8144	hom (R.D.= 94)	het (D.P.= 50)	hom (R.D.= 115)
6	51920485	PKHD1	ns	c.C1736T	p.T579M	0.999	3.74	0.013	0.026	0.0246	0.0245	het (R.D.= 37)	*	*
6	51774082	PKHD1	ns	c.A6681T	p.R2227S	0.006	4.63	.	.	.	0	*	het (R.D.= 13)	*
6	134583262	SGK1	ns	c.A94G	p.M32V	0.016	0.683	0.0064	0.0079	0.0125	0.0105	*	het (R.D.= 40)	*
7	150693567	NOS3	ns	c.G346C	p.G116R	0.571	-0.454	.	.	.	0	*	het (R.D.= 14)	*
8	26636918	ADRA1A	ns	c.A913G	p.N305D	.	-1.55	0.0068	.	0.0081		*	*	*
9	117354865	ATP6V1G1	ns	c.A116C	p.E39A	0.004	4.54	0.0012	0.0027	0.0046		*	*	het (R.D.= 17)
9	139222174	GPSM1	ns	c.T23C	p.V8A	.	.	.	.	1	0.9986	hom (R.D.=15)	hom (D.P.= 19)	hom (D.P.= 16)
9	140128582	SLC34A3	ns	c.C947T	p.T316M	0.475	-0.567	.	.	0.0003	0.0002	het (R.D.= 19)	*	*
12	48132962	RAPGEF3	ns	c.C2299T	p.P767S	0.556	4.66	.	0.0001	1.67E-05	3.31E-05	het (R.D.= 32)	*	*
12	6463637	SCNN1A	ns	c.G1504C	p.E502Q	0.422	4.95	.	.	.	.	*	het (R.D.= 18)	*
12	6472752	SCNN1A	ns	c.C718T	p.R240W	0.764	0.872	0.0072	0.012	0.0202		*	*	het (R.D.= 39)
12	113715138	TPCN1	ns	c.G949T	p.A317S	0.1	5.25	.	0.0008	0.001		*	*	het (R.D.= 56)
16	23360165	SCNN1B	ns	c.C245G	p.S82C	0.905	5.01	0.0008	0.0064	0.0048	0.0048	het (R.D.= 61)	*	*
17	42338993	SLC4A1	ns	c.G118A	p.E40K	0.252	0.246	0.0058	0.012	0.0108	0	*	het (R.D.= 40)	*
18	24436192	AQP4	ns	c.G874A	p.V292I	0.256	0	0.0002	0.0004	0.0006	0.0032	*	het (R.D.= 24)	het (R.D.= 23)
18	18534948	ROCK1	ns	c.C3649G	p.Q1217E	0.139	5.43	.	.	0.0734	0.1241	*	het (R.D.= 11)	*
19	33182993	NUDT19	ns	c.C127T	p.R43W	0.986	3.14	0.0012	0.0077	0.0236	0.0079	*	het (R.D.= 42)	*
19	10463118	TYK2	ns	c.C3310G	p.P1104A	0.899	0.73	0.01	0.029	0.0273		*	*	het (R.D.= 36)

The traffic light colour code corresponds to zygosity of variants. Homozygous reference variants are identified in green, heterozygous variants in red and homozygous alternative variants in orange.

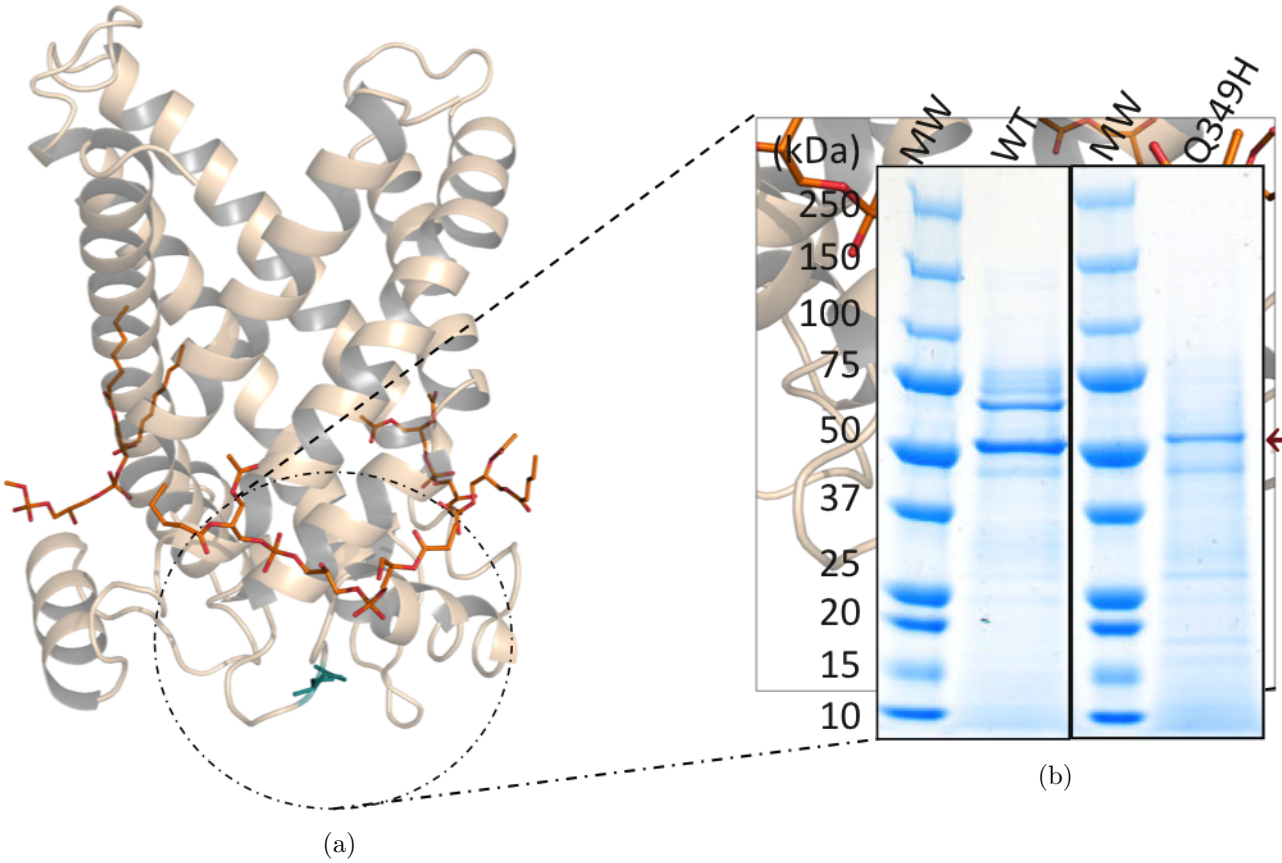
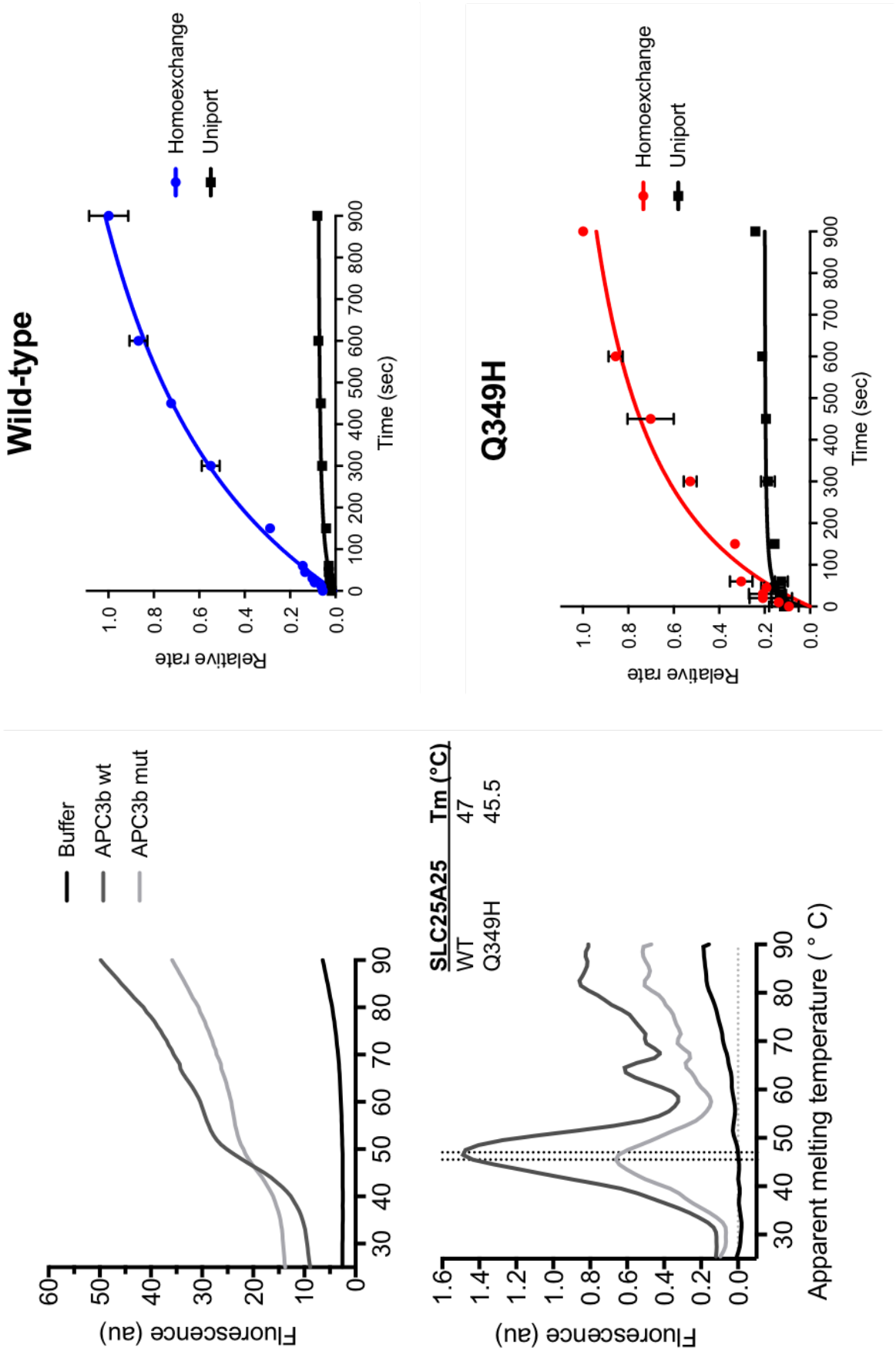


Figure 8.2: **(a):**3D representation of Q349H mutation on *SLC25A25* (APC3), The mutation is in the loop between transmembrane helix H3 and matrix helix h34; (b): Purification yield for the wild-type and mutant protein as shown on SDS-PAGE ($SLC25A25_{WT}$: 0.8 mg, $SLC25A25_{Q349H}$:0.5 mg)(Figures courtesy of Fiona Fitzpatrick of MRC Mitochondrial Biology Unit, University of Cambridge)



(a) (b)

Figure 8.3: (a): Thermal stability assay for $SLC25A25_{Q349H}$ versus wild-type protein; (b): Transport assay with the two $SLC25A25$ isoforms. The assay measures the uptake of radiolabelled ATP($[^{14}C]$-ATP) into proteoliposomes.(Figures courtesy of Fiona Fitzpatrick of MRC Mitochondrial Biology Unit, University of Cambridge)

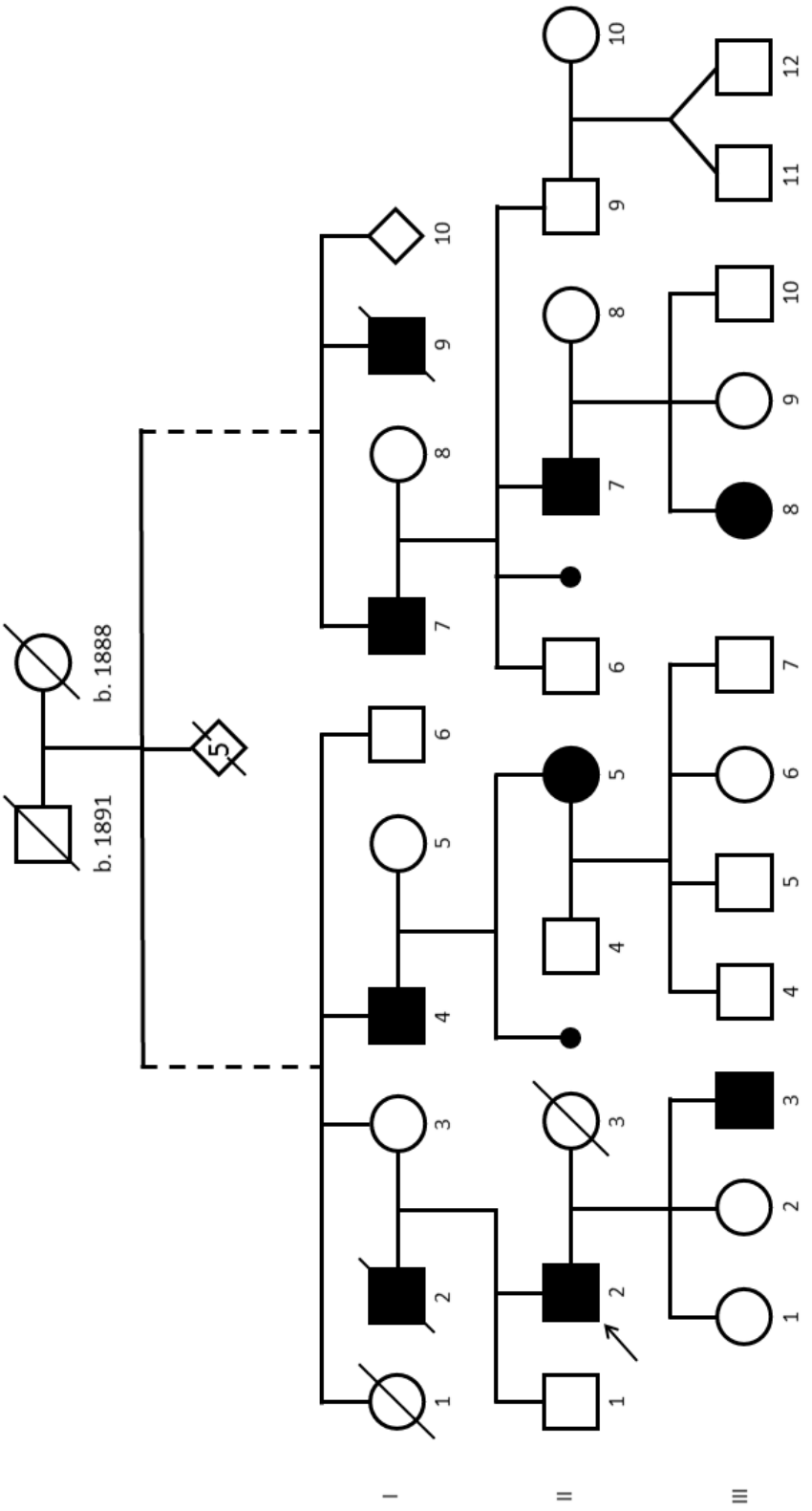


Figure 8.4: The Italian pedigree with recurrent stone phenotype. The *SLC25A25:NM* − 001006641 : *exon*8 : *c.G*1047*C* mutation consistently segregates among affected individuals of this kindred (Pedigree courtesy of Professor Roberto Colombo of Università Cattolica del Sacro Cuore, Italy)

8.2 Supplementary Data for Chapter 4

8.3 WGS FastQC Results

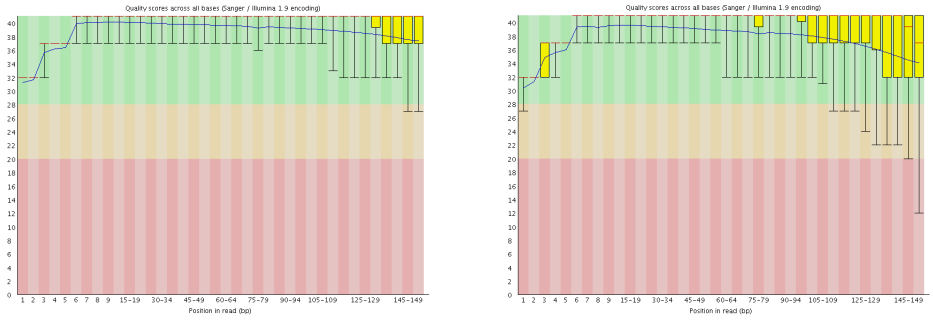
Quality of cleaned raw data for each pair of lane investigated in FastQC software and results for each lane is provided separately:

8.3.1 Lane 1

Basic Statistics

Table 8.9: Basic Statistics for Lane 1 forward & reverse reads

File name	SD003_DHG07669_HWNCVCCXX_L4_1(&2).clean.fq.gz
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	73,343,137
Sequence flagged as poor quality	0
Read Length	150 bp
%GC	43



(a) (b)

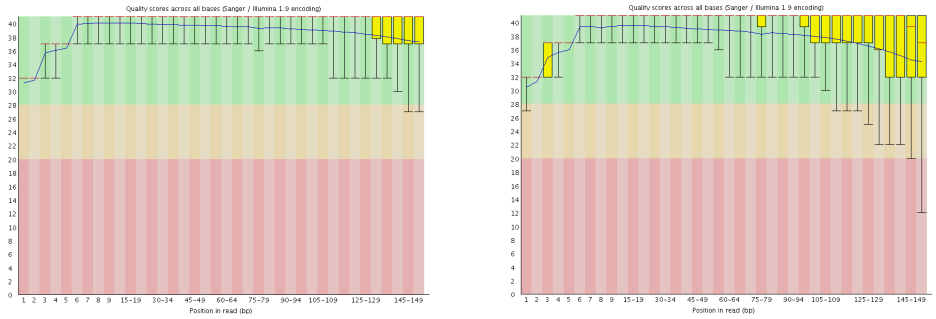
Figure 8.5: Per Base Sequence Quality plot of trimmed WGS reads from Lane-01; For each position a BoxWhisker type plot is drawn; The mean quality across the length of the read is represented by the blue line.

8.3.2 Lane 2

Basic Statistics

Table 8.10: Basic Statistics for Lane 2 forward & reverse reads

File name	SD003_DHG07669_HWNCVCCXX_L5_1(&2).clean.fq.gz
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	69,400,196
Sequence flagged as poor quality	0
Read Length	150 bp
%GC	43



(a) (b)

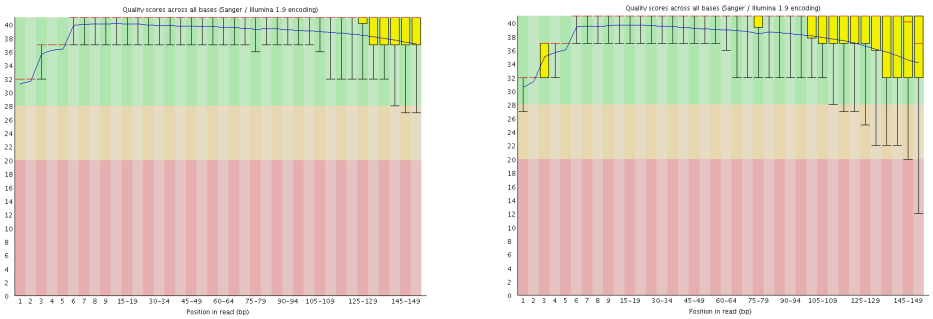
Figure 8.6: Per Base Sequence Quality plot of trimmed WGS reads from Lane-02; For each position a BoxWhisker type plot is drawn; The mean quality across the length of the read is represented by the blue line.

8.3.3 Lane 3

Basic Statistics

Table 8.11: Basic Statistics for Lane 2 forward & reverse reads

File name	SD003_DHG07669_HWNCVCCXX_L6_1(&2).clean.fq.gz
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	70,575,832
Sequence flagged as poor quality	0
Read Length	150 bp
%GC	43



(a) (b)

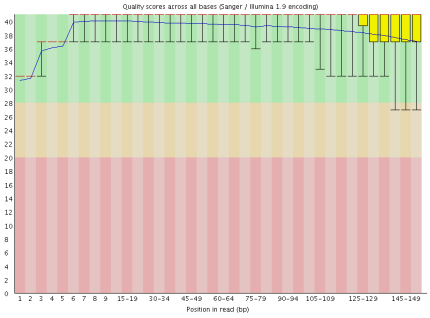
Figure 8.7: Per Base Sequence Quality plot of trimmed WGS reads from Lane-03; For each position a BoxWhisker type plot is drawn; The mean quality across the length of the read is represented by the blue line.

8.3.4 Lane 4

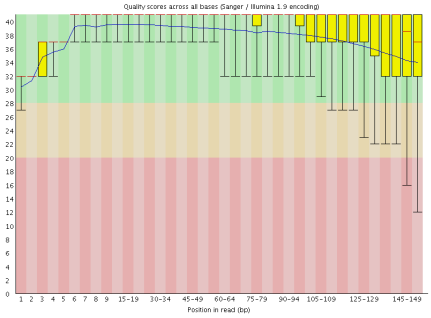
Basic Statistics

Table 8.12: Basic Statistics for Lane 4 forward & reverse reads

File name	SD003_DHG07669_HWNCVCCXX_L7_1(&2).clean.fq.gz
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	71,472,977
Sequence flagged as poor quality	0
Read Length	150 bp
%GC	43



(a)



(b)

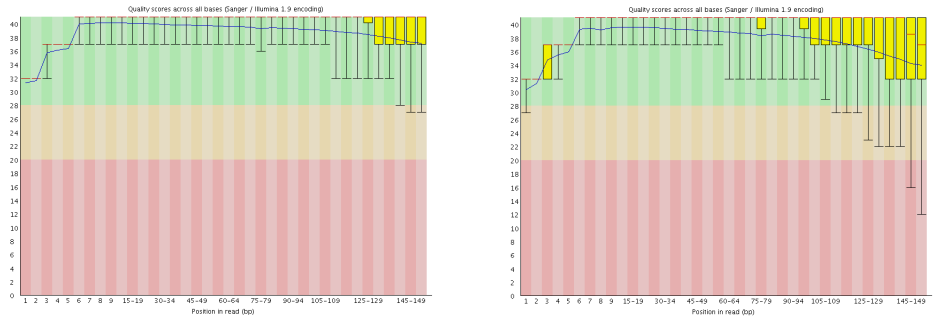
Figure 8.8: Per Base Sequence Quality plot of trimmed WGS reads from Lane-04; For each position a BoxWhisker type plot is drawn; The mean quality across the length of the read is represented by the blue line.

8.3.5 Lane 5

Basic Statistics

Table 8.13: Basic Statistics for Lane 5 forward & reverse reads

File name	SD003_DHG07669_HWNCVCCXX_L8_1(&2).clean.fq.gz
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	70,280,629
Sequence flagged as poor quality	0
Read Length	150 bp
%GC	43



(a) (b)

Figure 8.9: Per Base Sequence Quality plot of trimmed WGS reads from Lane-05; For each position a BoxWhisker type plot is drawn; The mean quality across the length of the read is represented by the blue line.

8.4 WGS Coverage Analyses

8.4.1 Global Statistics

Table 8.14: SD003 WGS global statistics

Reference size	3,099,922,541
Number of reads	727,228,800
Mapped reads	725,858,397 / 99.81%
Unmapped reads	1,370,403 / 0.19%
Mapped paired reads	725,858,397 / 99.81%
Mapped reads, first in pair	363,184,333 / 49.94%
Mapped reads, second in pair	362,674,064 / 49.87%
Mapped reads, both in pair	725,457,874 / 99.76%
Mapped reads, singletons	400,523 / 0.06%
Read min/max/mean length	30 / 150 / 147.76
Duplicated reads (flagged)	73,914,402 / 10.16%
Clipped reads	58,223,508 / 8.01%

8.4.2 ACGT content

Table 8.15: SD003 WGS ACGT contents

Number/percentage of A's	29,733,995,705 / 28.16%
Number/percentage of C's	22,928,062,103 / 21.72%
Number/percentage of T's	29,723,739,471 / 28.15%
Number/percentage of G's	23,191,840,491 / 21.97%
Number/percentage of N's	0 / 0%
GC Percentage	43.68%

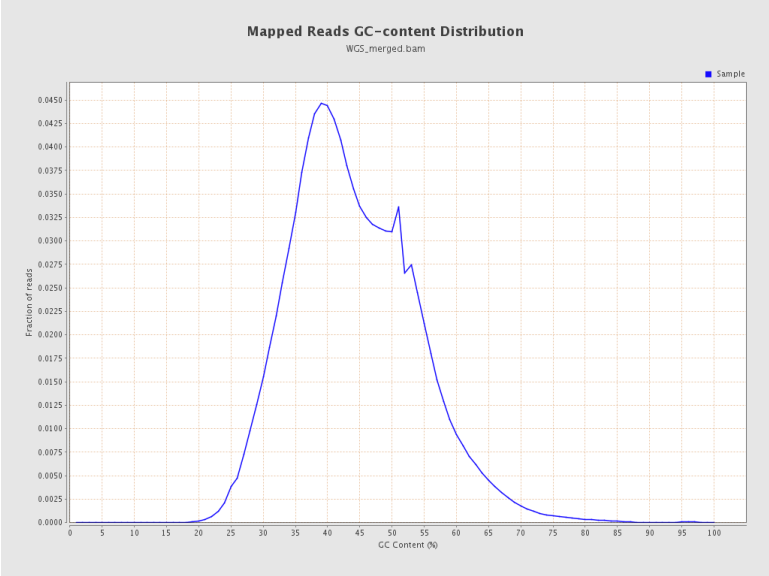


Figure 8.10: SD003 WGS GC-content distribution

8.4.3 Coverage & Mapping Quality

Mapping quality of a read is phred-scaled posterior probability that the read is aligned to the wrong place. The probability is calculated as:

$$p = 10^{-}q/10 \tag{8.1}$$

Table 8.16: SD003 WGS mean coverage.

Mean Coverage	34.0618
Standard Deviation	189.1737
Mean Mapping Quality	44.9

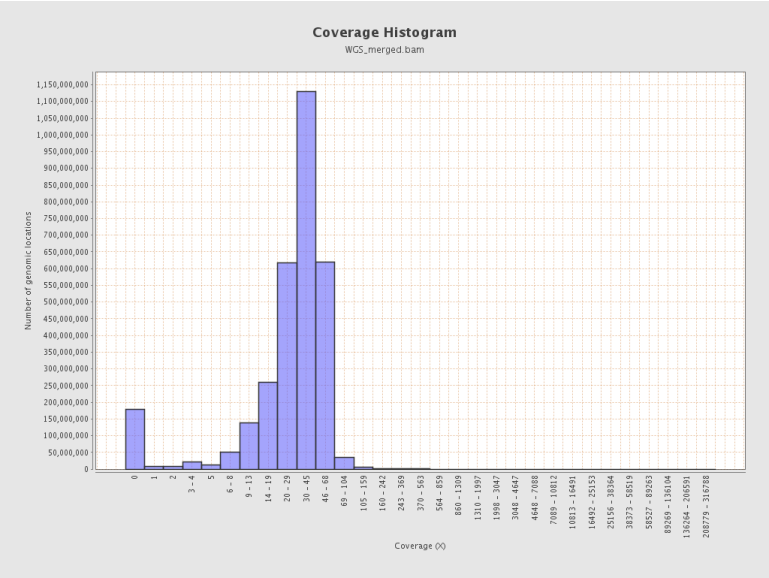


Figure 8.11: **Coverage Histogram:** Number of genomic regions covered at X fold

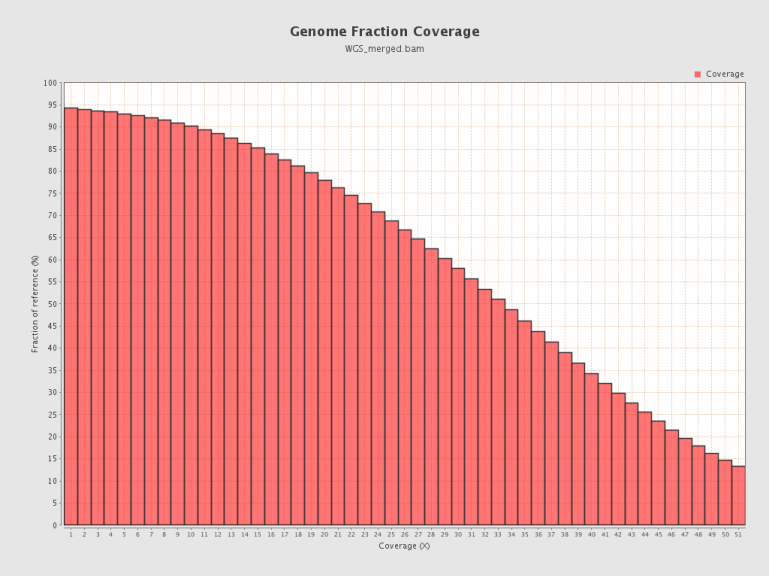


Figure 8.12: Genome fraction coverage

8.4.4 Insert Size

Table 8.17: SD003 WGS insert size statistics.

Mean Insert Size	275.82bp
Standard Deviation	2,728,455.33
P25/Median/P75	254 / 289 / 326

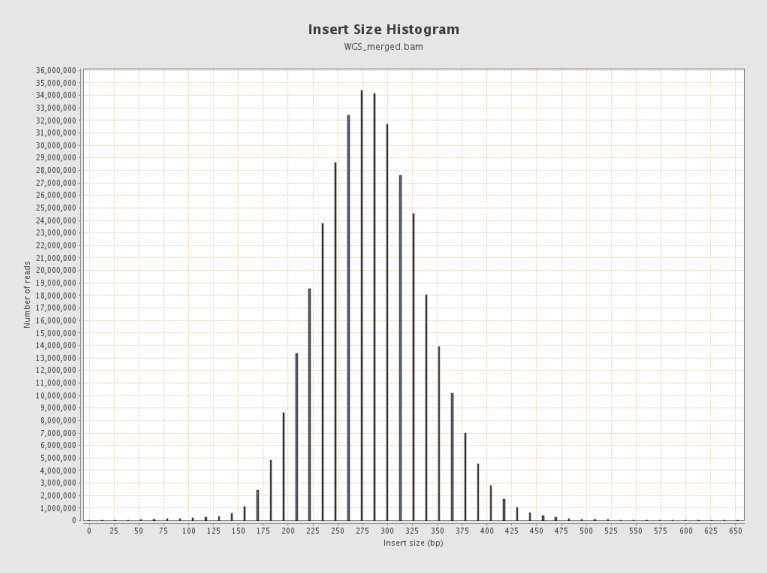


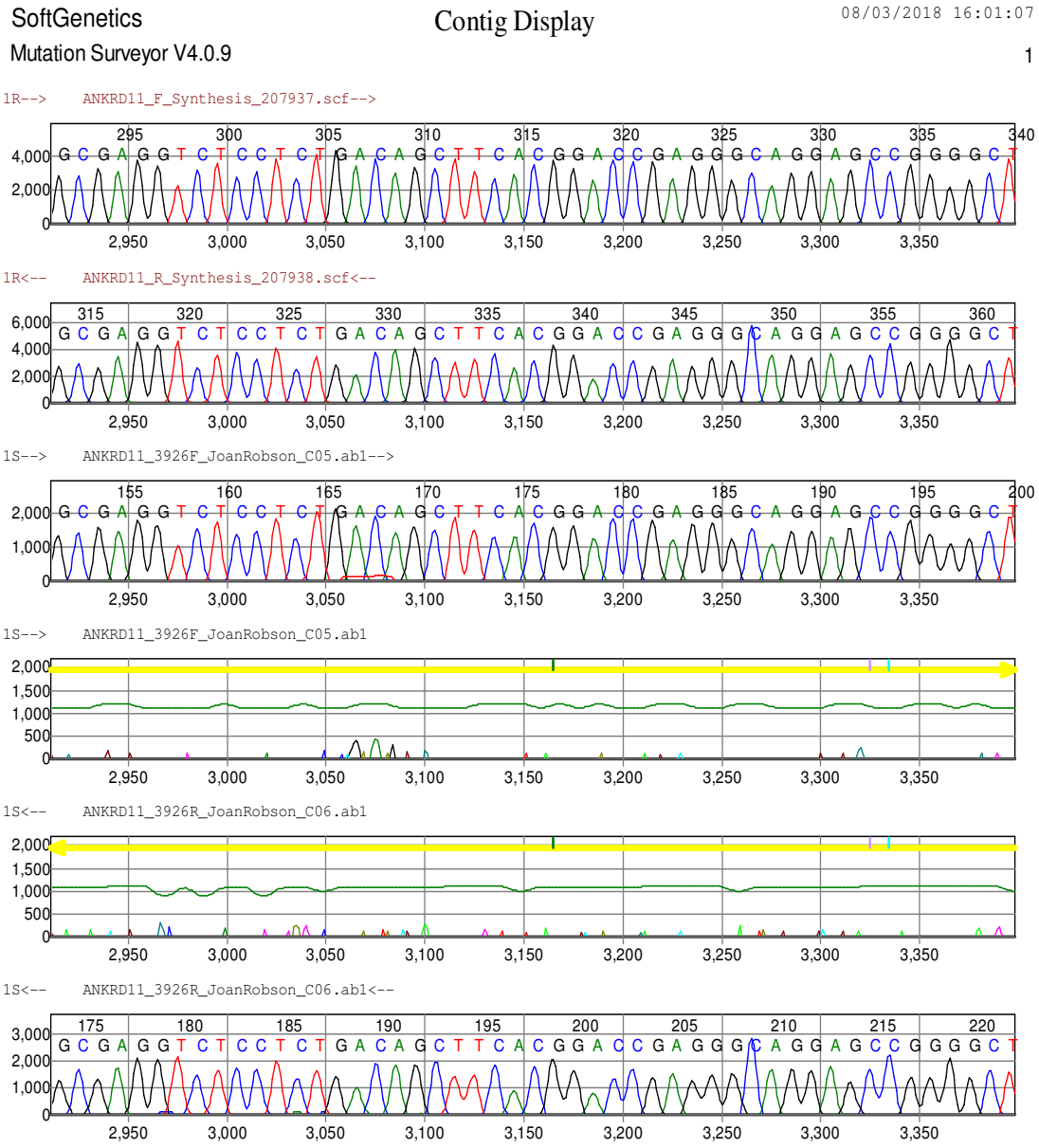
Figure 8.13: Insert Size Histogram.

8.4.5 Mismatches and indels

Table 8.18: SD003 WGS mismatches & indels statistics

General error rate	0.48%
Mismatches	505,748,606
Insertions	103,192
Mapped reads with at least one insertion	0.01%
Deletions	130,279
Mapped reads with at least one deletion	0.02%
Homopolymer indels	64.3%

8.4.6 Sanger sequencing traces for *ANKRD11*:c.3926C>T

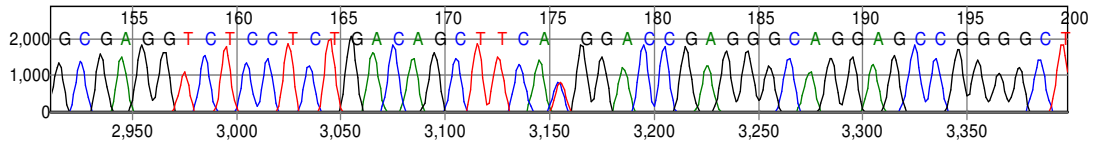


SoftGenetics
Mutation Surveyor V4.0.9

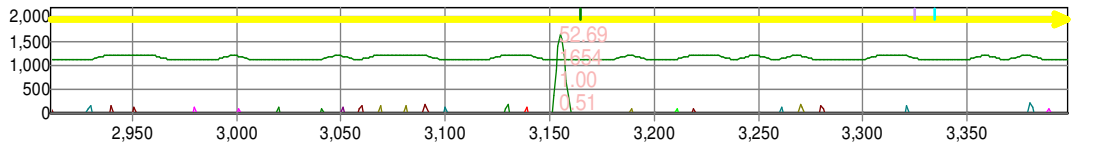
Contig Display

08/03/2018 16:01:07
2

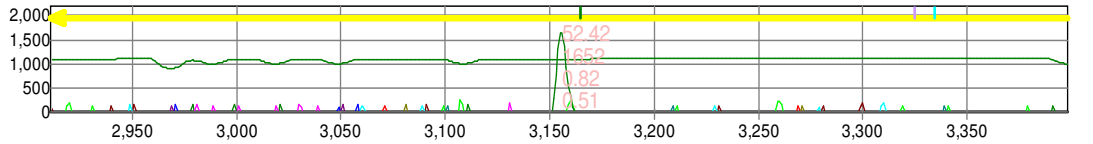
1S--> ANKRD11_3926F_NicholasRobson_A05.ab1-->



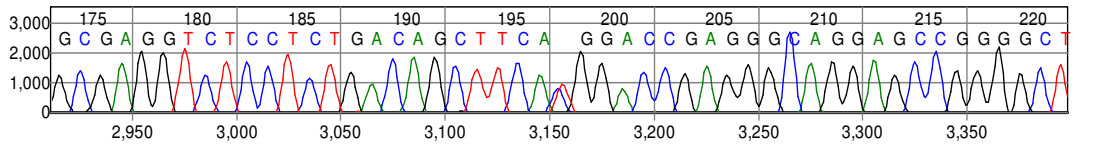
1S--> ANKRD11_3926F_NicholasRobson_A05.ab1 Mutations: 207946C>CT$52.7



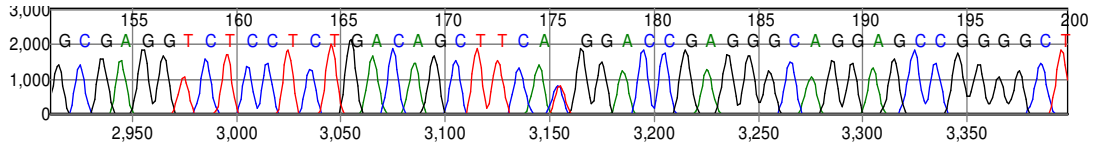
1S<-- ANKRD11_3926R_NicholasRobson_A06.ab1 Mutations: 207946C>CT$52.4



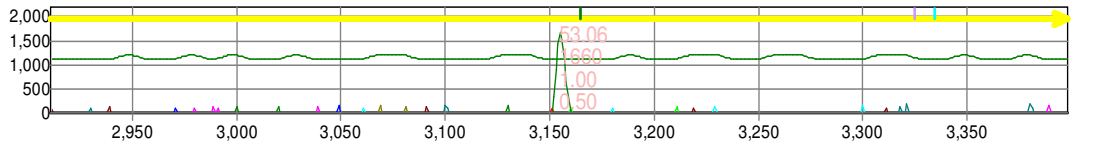
1S<-- ANKRD11_3926R_NicholasRobson_A06.ab1<--



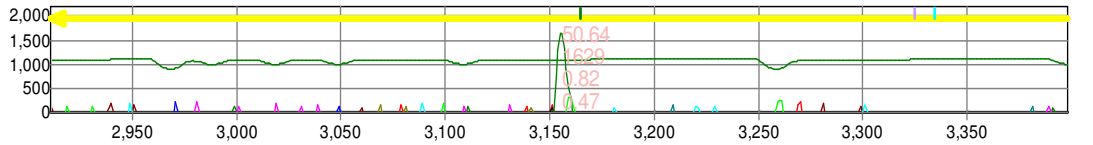
1S--> ANKRD11_3926F_NigelRobson_B05.ab1-->



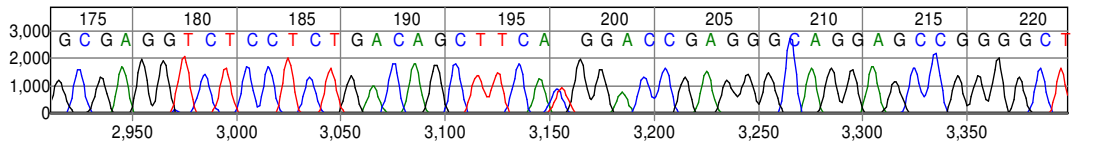
1S--> ANKRD11_3926F_NigelRobson_B05.ab1 Mutations: 207946C>CT$53.1



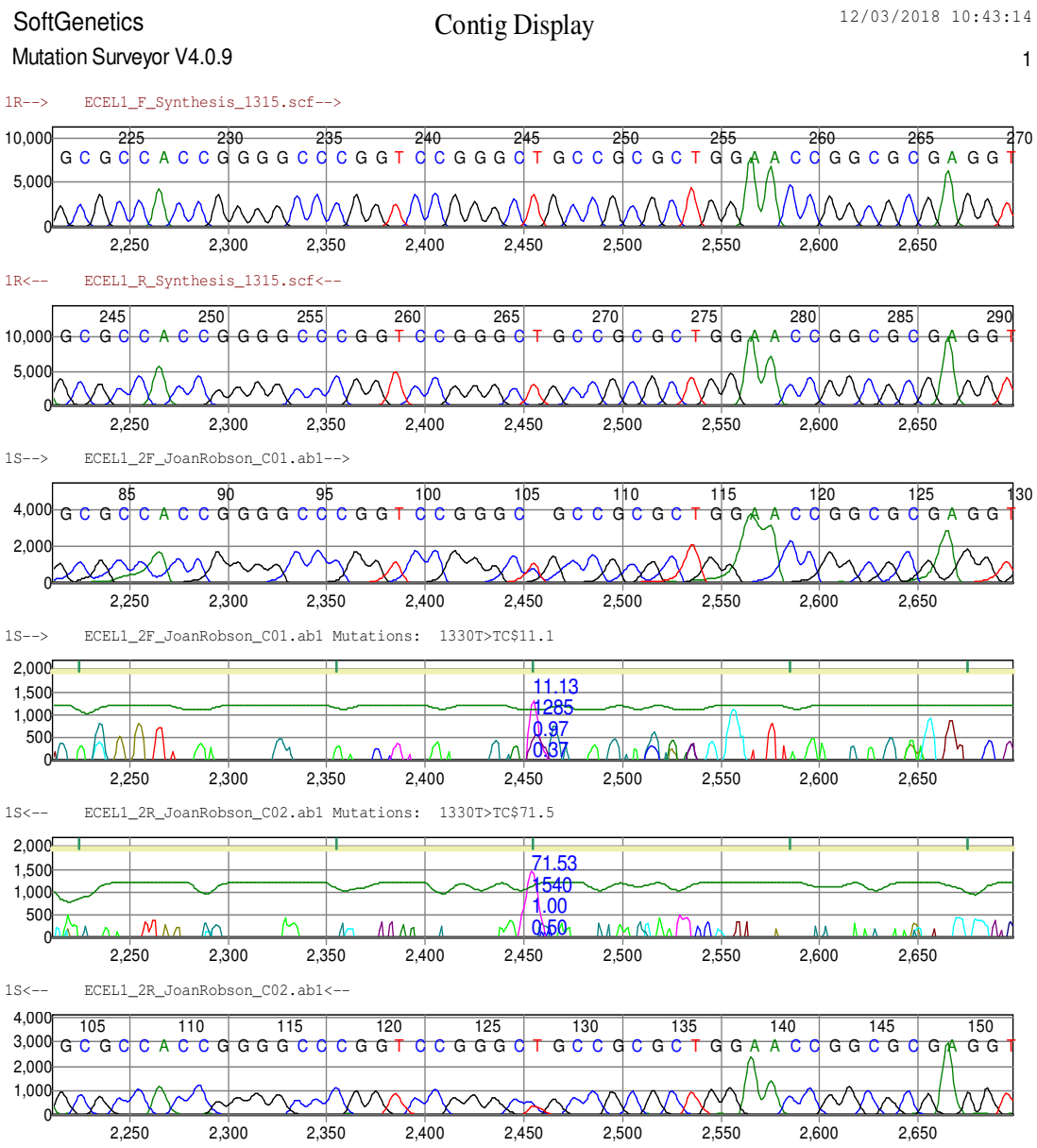
1S<-- ANKRD11_3926R_NigelRobson_B06.ab1 Mutations: 207946C>CT$50.6



1S<-- ANKRD11_3926R_NigelRobson_B06.ab1<--



8.4.7 Sanger sequencing traces for *ECEL1*:c.155T>C

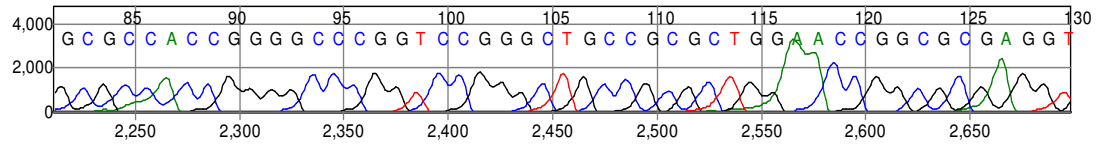


SoftGenetics
Mutation Surveyor V4.0.9

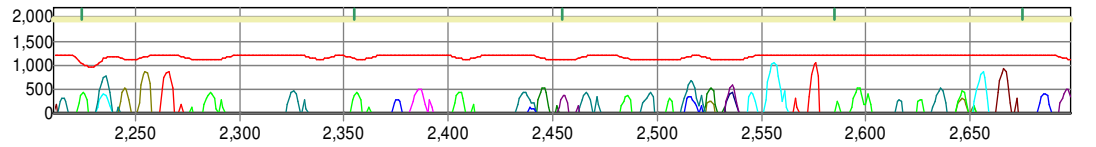
Contig Display

12/03/2018 10:43:14
2

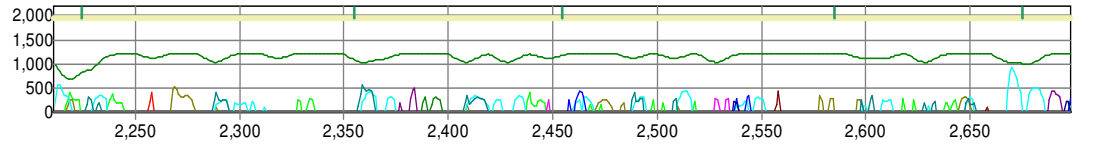
1S--> ECEL1_2F_NigelRobson_B01.ab1-->



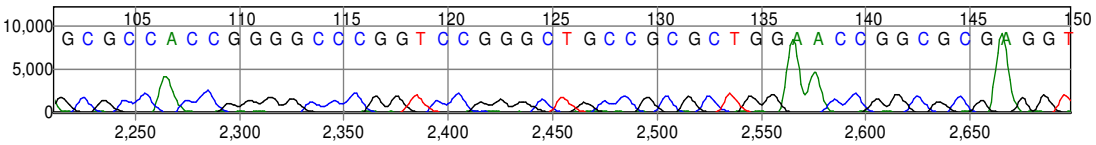
1S--> ECEL1_2F_NigelRobson_B01.ab1



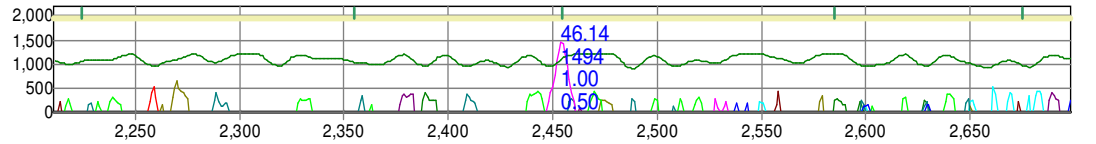
1S<-- ECEL1_2R_NigelRobson_B02.ab1



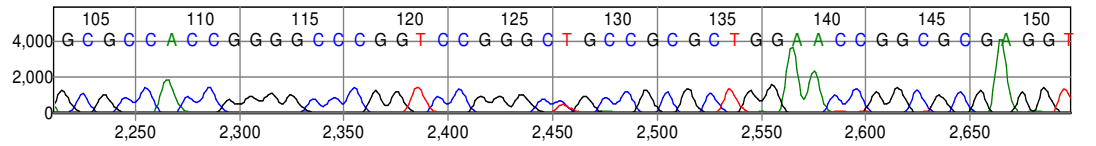
1S<-- ECEL1_2R_NigelRobson_B02.ab1<--



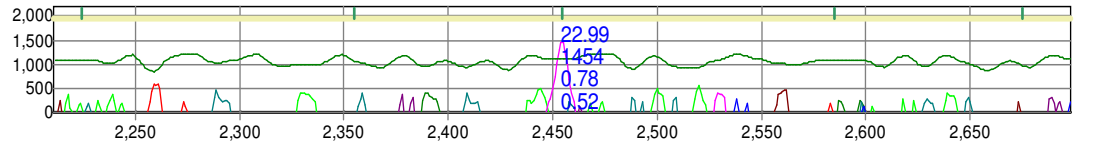
1S<-- ECEL1_2R_NicholasRobson_G12.ab1 Mutations: 1330T>TC$46.1



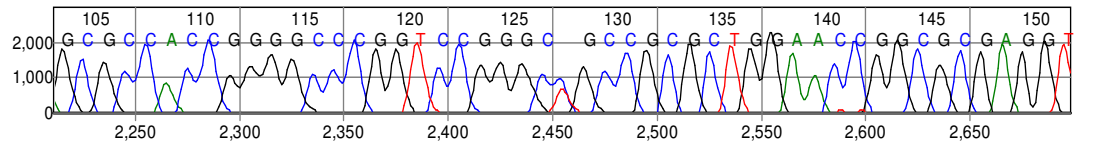
1S<-- ECEL1_2R_NicholasRobson_G12.ab1<--



1S<-- ECEL1_2R_NicholasRobson_H12.ab1 Mutations: 1330T>TC$23.0



1S<-- ECEL1_2R_NicholasRobson_H12.ab1<--



8.4.8 Sanger sequencing traces for *ECEL1*:c.1013T>C

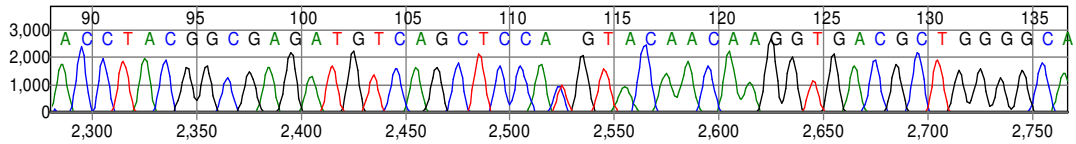
SoftGenetics
Mutation Surveyor V4.0.9

Contig Display

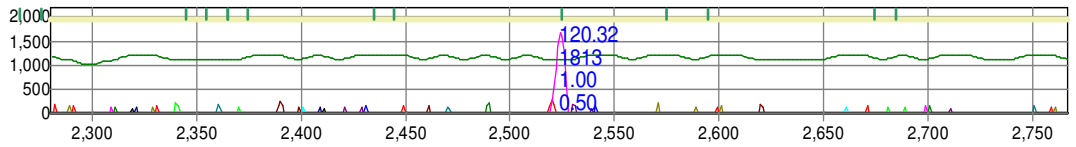
08/03/2018 15:55:09

2

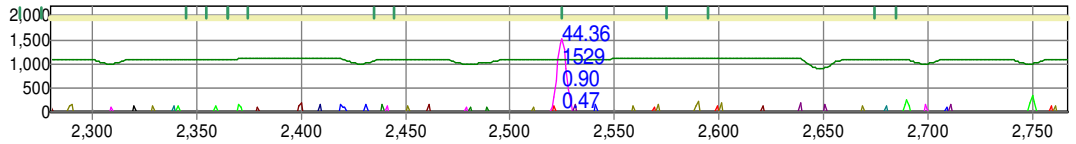
1S--> ECEL1_5F_NicholasRobson_A03.ab1-->



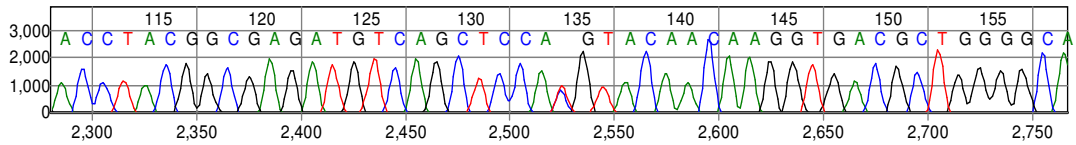
1S--> ECEL1_5F_NicholasRobson_A03.ab1 Mutations: 2982T>TC$120.3



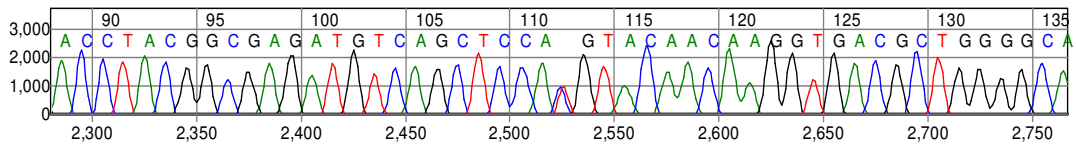
1S<-- ECEL1_5R_NicholasRobson_A04.ab1 Mutations: 2982T>TC$44.4



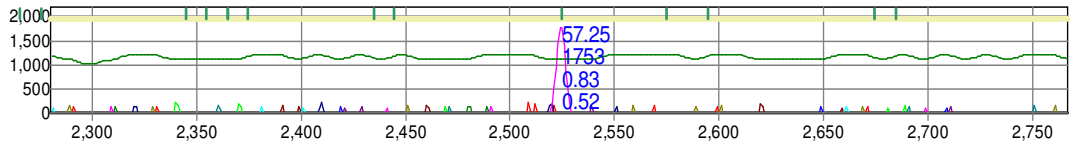
1S<-- ECEL1_5R_NicholasRobson_A04.ab1<--



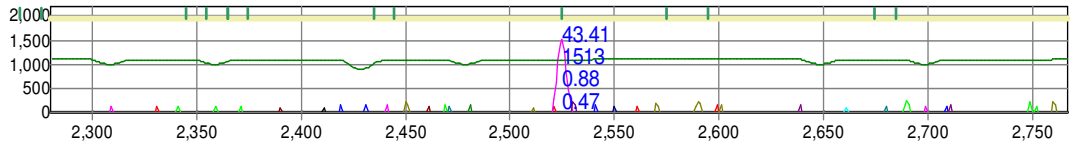
1S--> ECEL1_5F_NigelRobson_B03.ab1-->



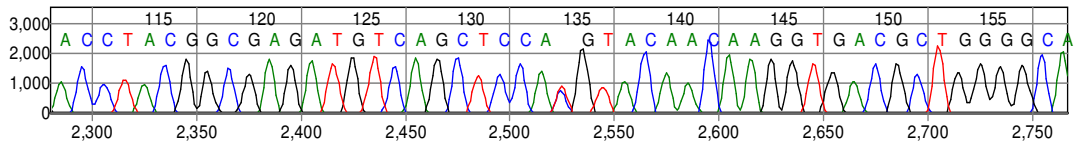
1S--> ECEL1_5F_NigelRobson_B03.ab1 Mutations: 2982T>TC$57.3



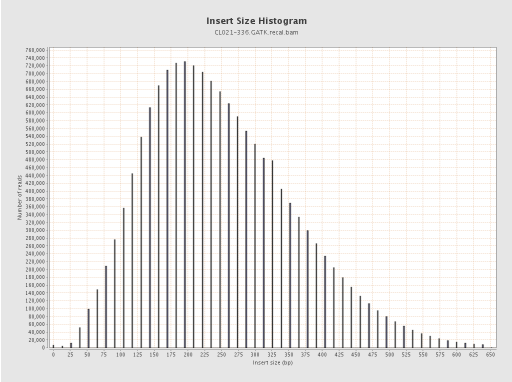
1S<-- ECEL1_5R_NigelRobson_B04.ab1 Mutations: 2982T>TC$43.4



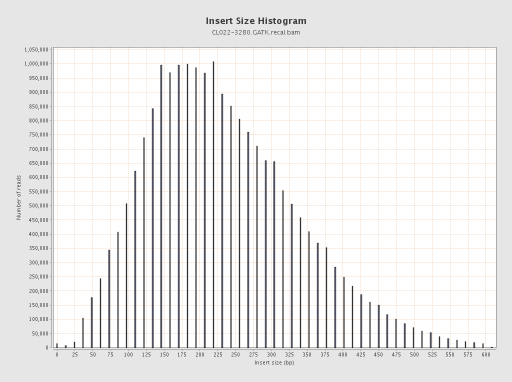
1S<-- ECEL1_5R_NigelRobson_B04.ab1<--



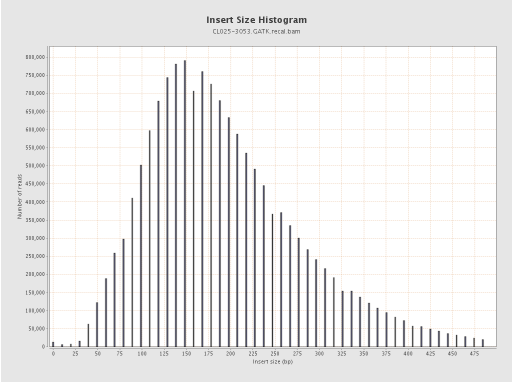
8.5 Supplementary Data for Chapter 5



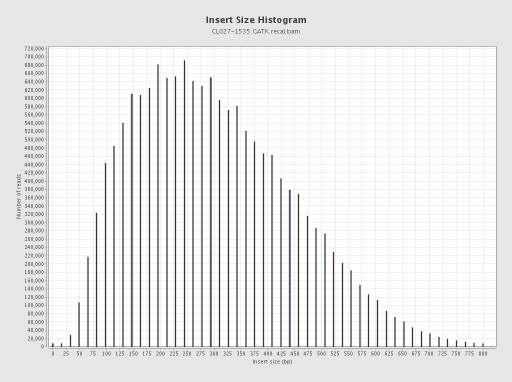
(a) CL021 insert size histogram



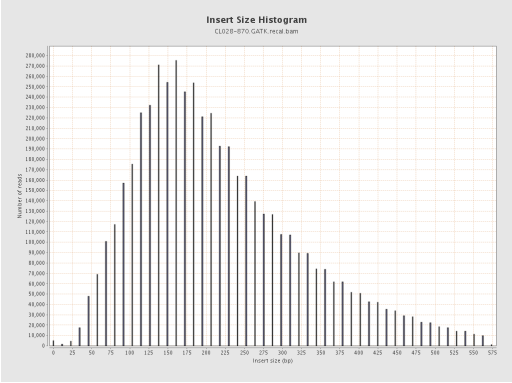
(b) CL022 insert size histogram



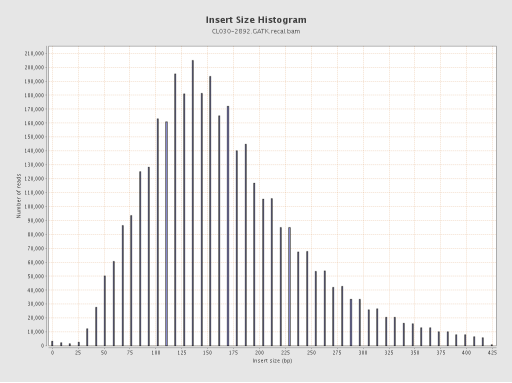
(c) CL025 insert size histogram



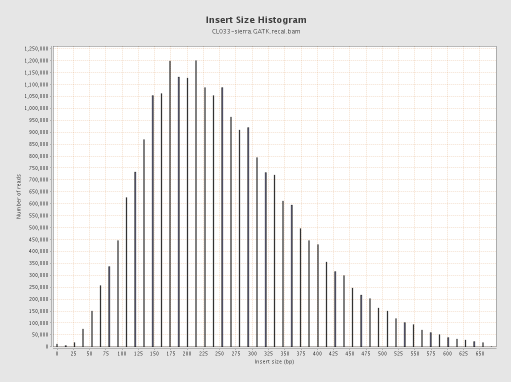
(d) CL027 insert size histogram



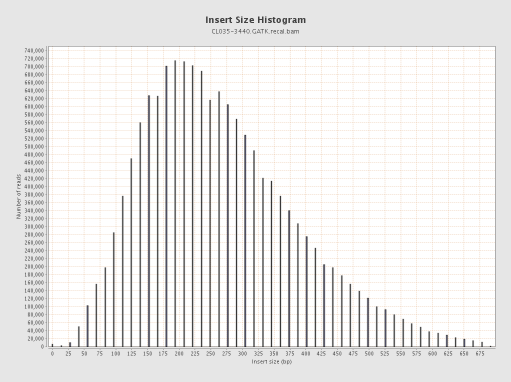
(e) CL028 insert size histogram



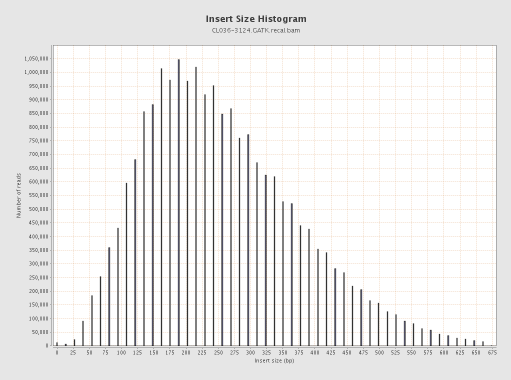
(f) CL030 insert size histogram



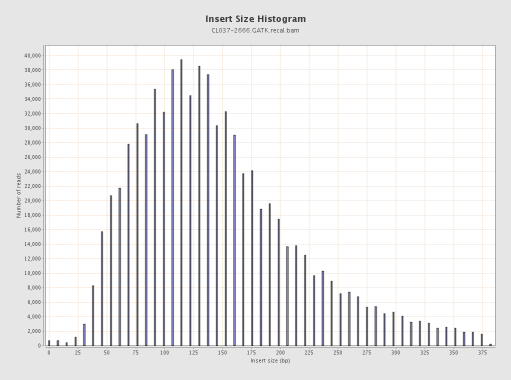
(g) CL033 insert size histogram



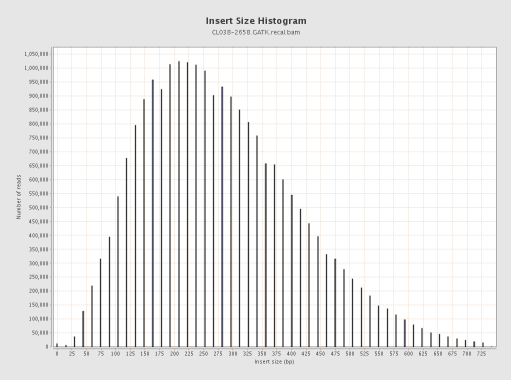
(h) CL035 insert size histogram



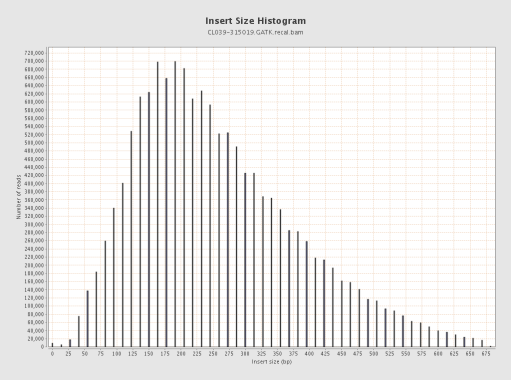
(i) CL036 insert size histogram



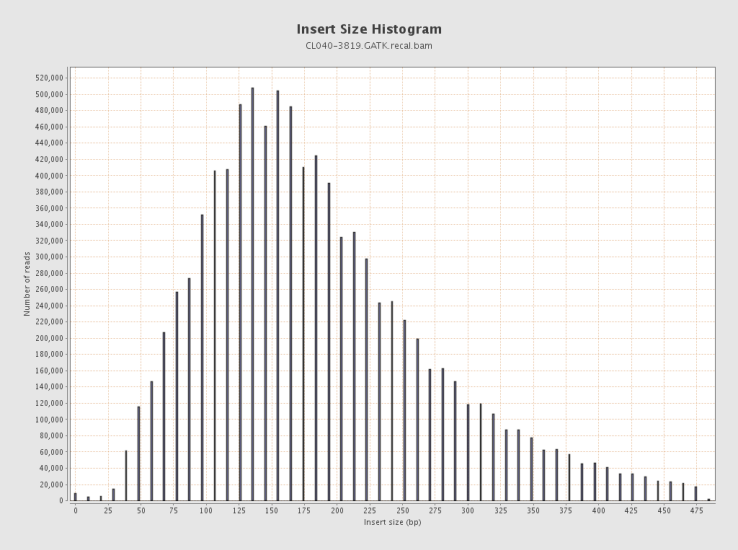
(j) CL037 insert size histogram



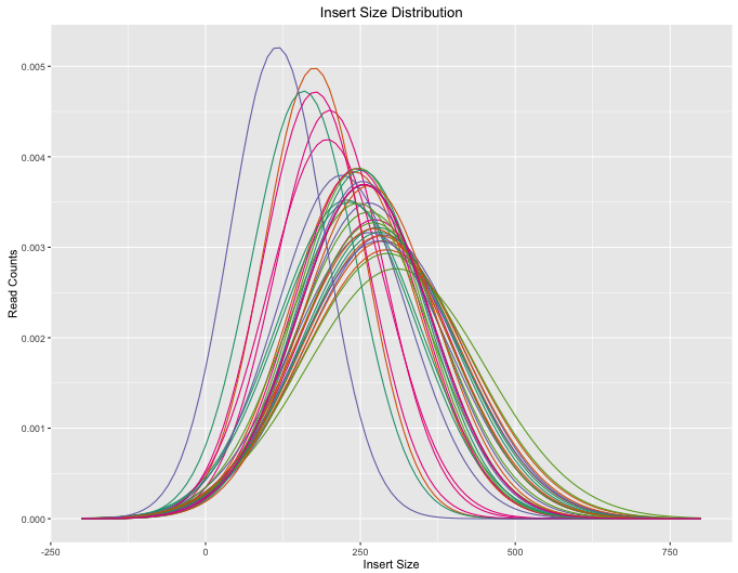
(k) CL038 insert size histogram



(l) CL039 insert size histogram

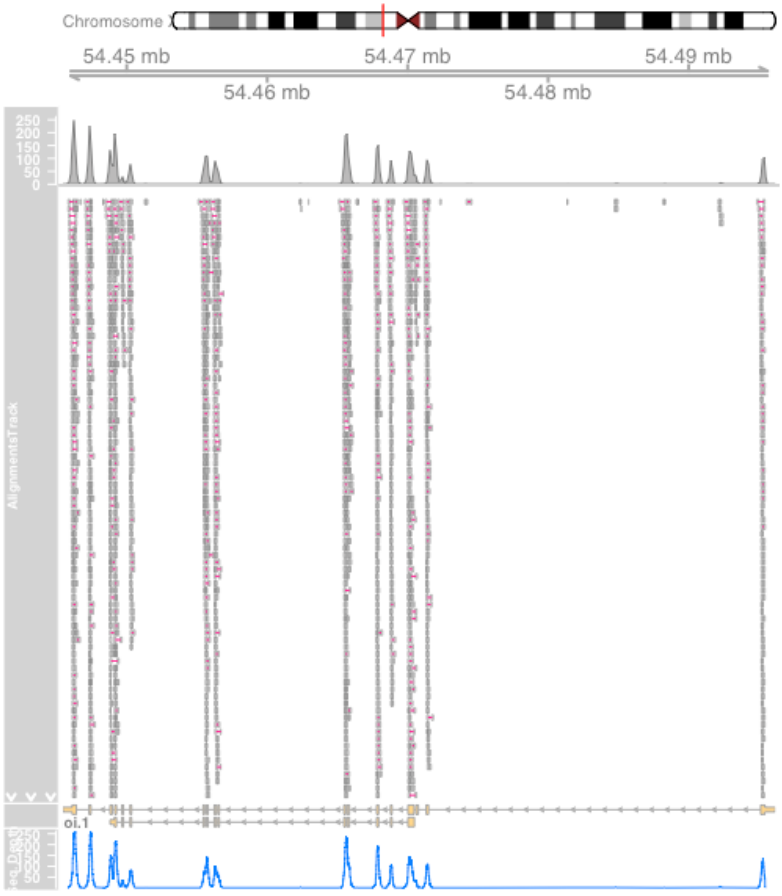


(m) CL040 insert size histogram



(n) Theoretical distribution of insert-size across all samples on the same dispatch DNA plate

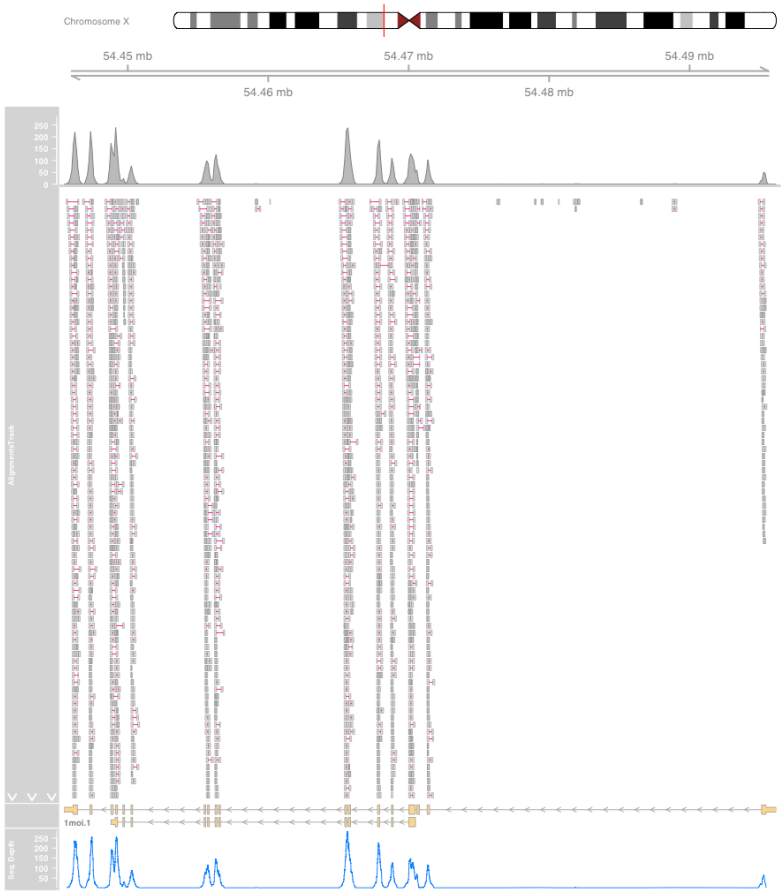
Figure 8.14: Histogram of insert sizes of read pairs for a single library from the 14 samples considered for targeted exome sequencing; (Figures 8.14a to 8.14m depict positively skewed distribution of insert sizes for the 14 samples with primary diagnosis of AAS and figure 8.14n depicts theoretical distribution of insert sizes for all samples on the same dispatch DNA plate).



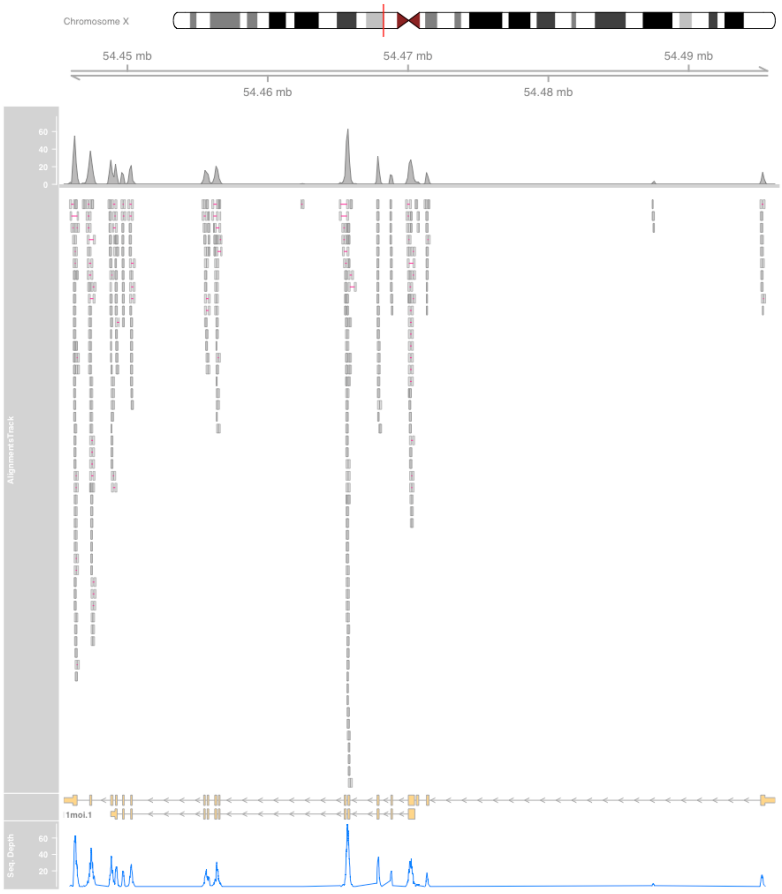
(a) Depth of coverage across *FGD1* gene for sample CL021

(b) Depth of coverage across *FGD1* gene for sample CL022

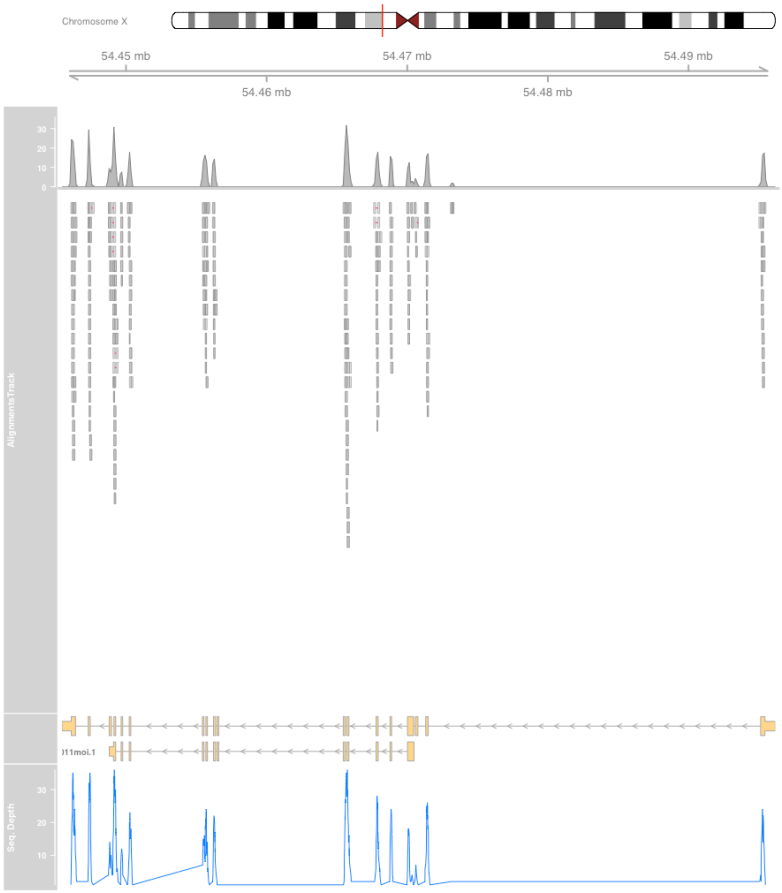
(c) Depth of coverage across *FGD1* gene for sample CL025



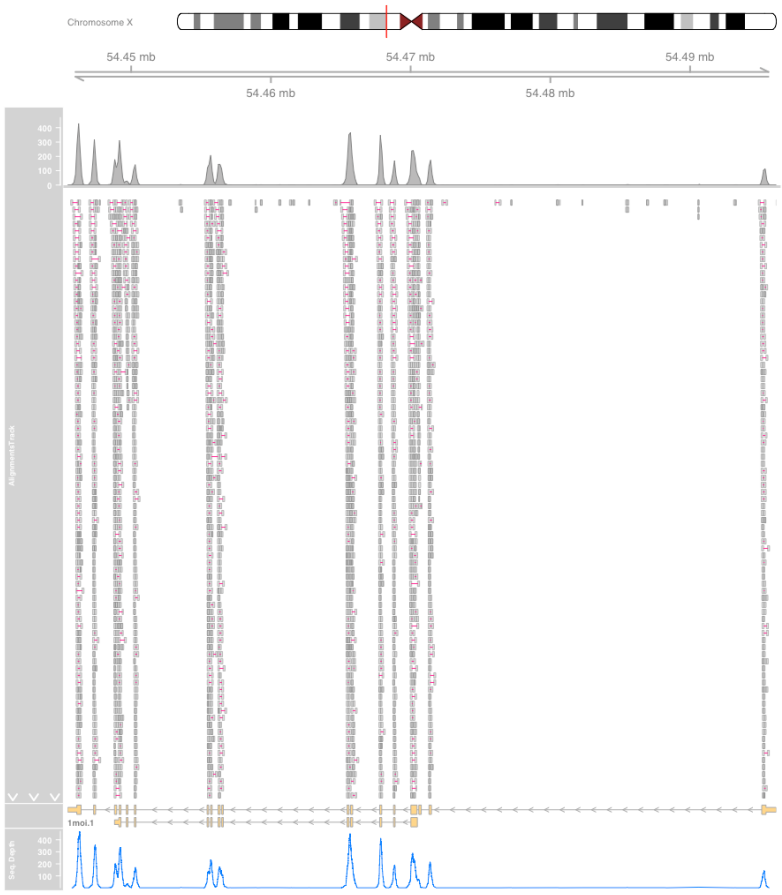
(d) Depth of coverage across *FGD1* gene for sample CL027



(e) Depth of coverage across *FGD1* gene for sample CL028



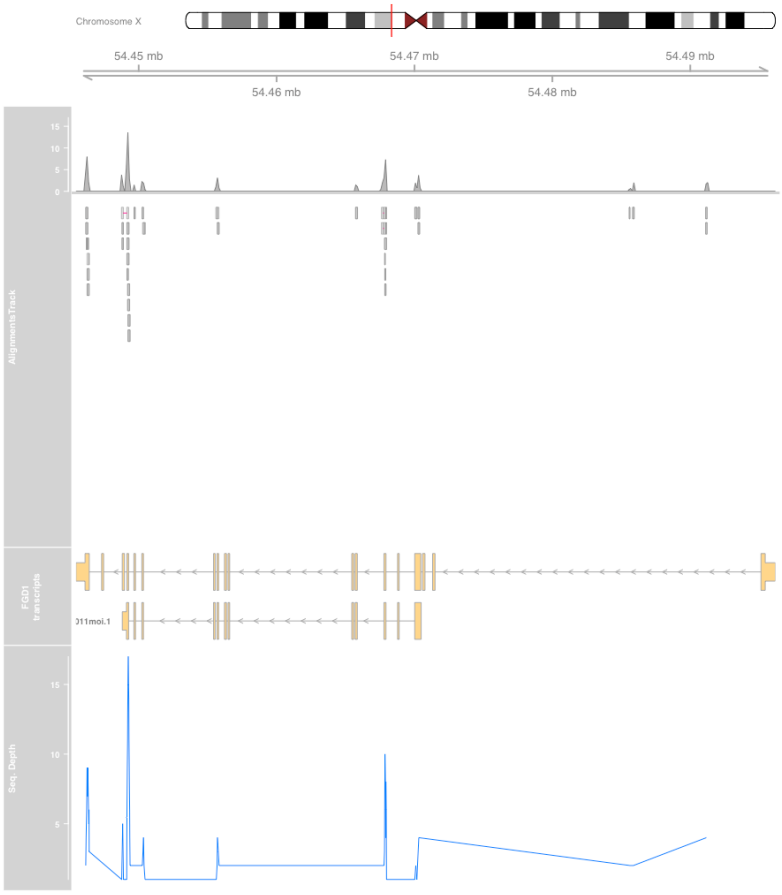
(f) Depth of coverage across *FGD1* gene for sample CL030



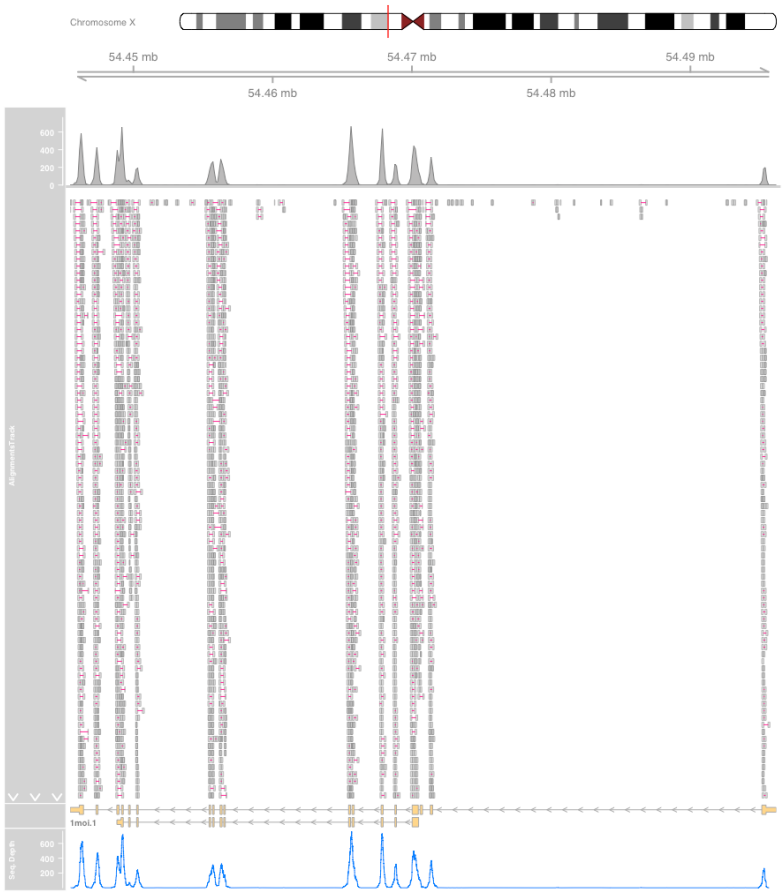
(g) Depth of coverage across *FGD1* gene for sample CL033

(h) Depth of coverage across *FGD1* gene for sample CL035

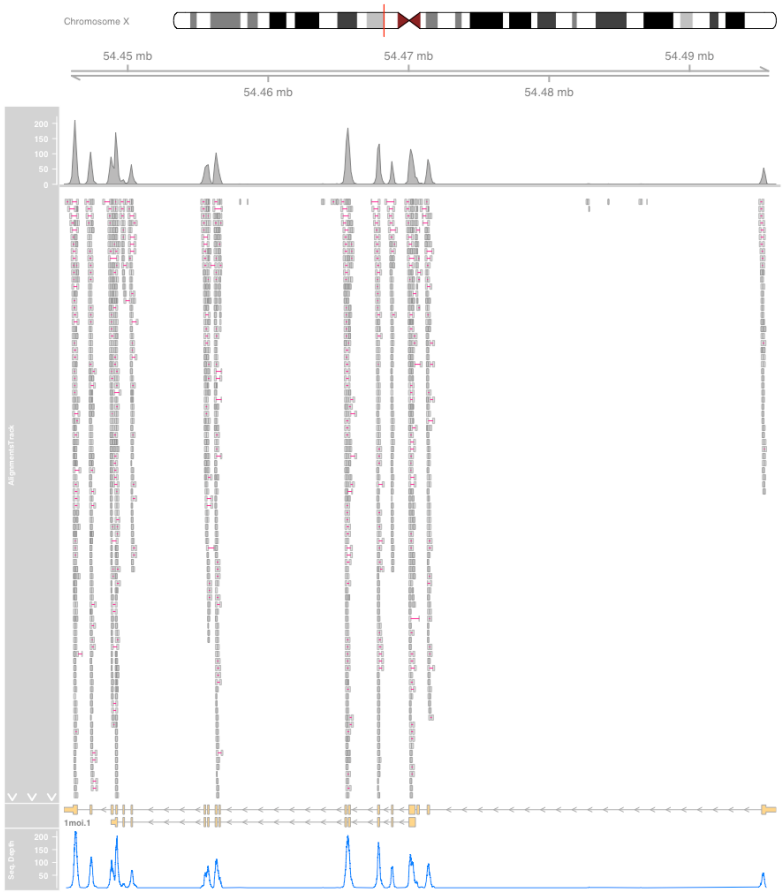
(i) Depth of coverage across *FGD1* gene for sample CL036



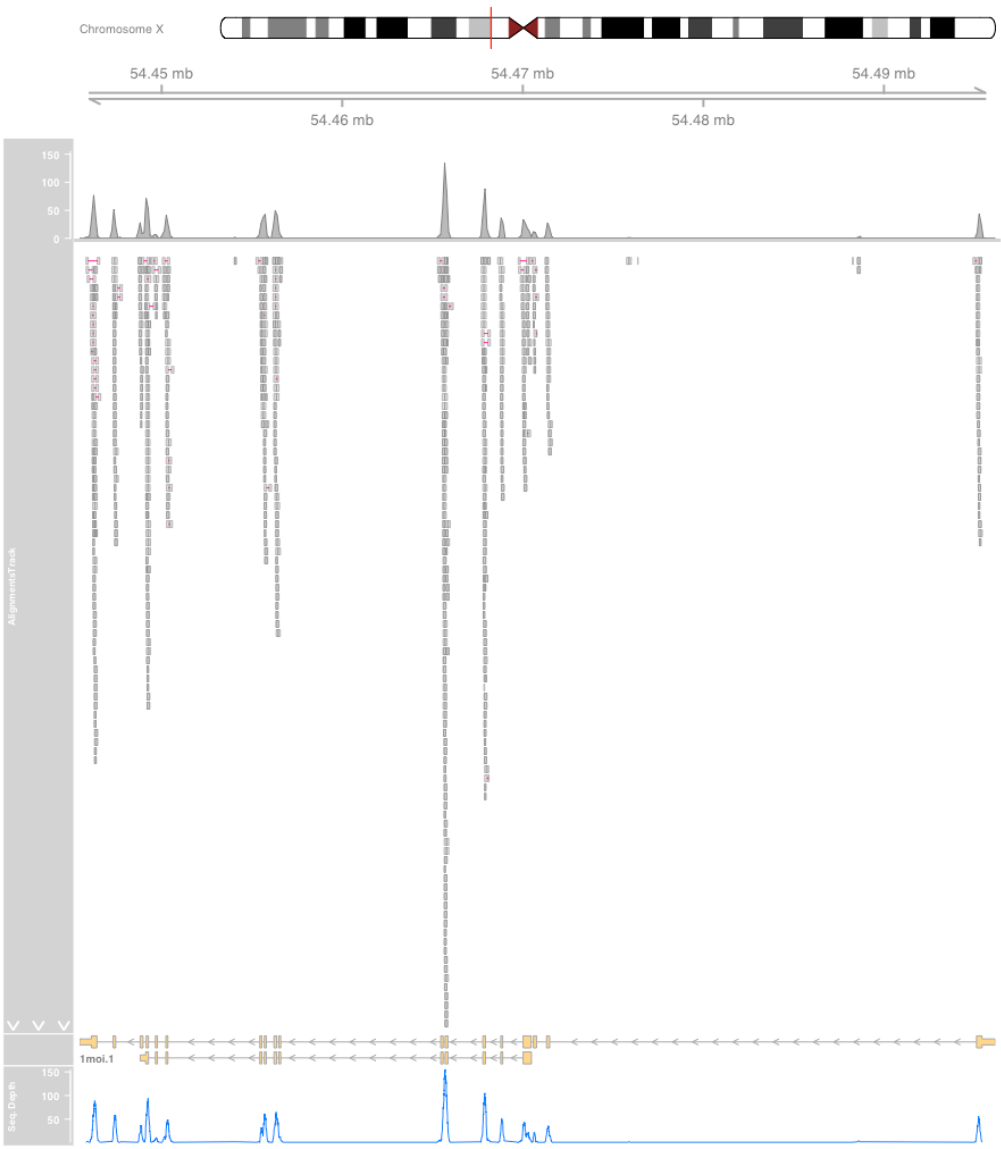
(j) Depth of coverage across *FGD1* gene for sample CL037



(k) Depth of coverage across *FGD1* gene for sample CL038



(l) Depth of coverage across *FGD1* gene for sample CL039



(m) Depth of coverage across *FGD1* gene for sample CL040

Figure 8.15: Depth of coverage across *FGD1* gene for the 14 samples considered for WES analysis

8.6 Supplementary Data for Chapter 6

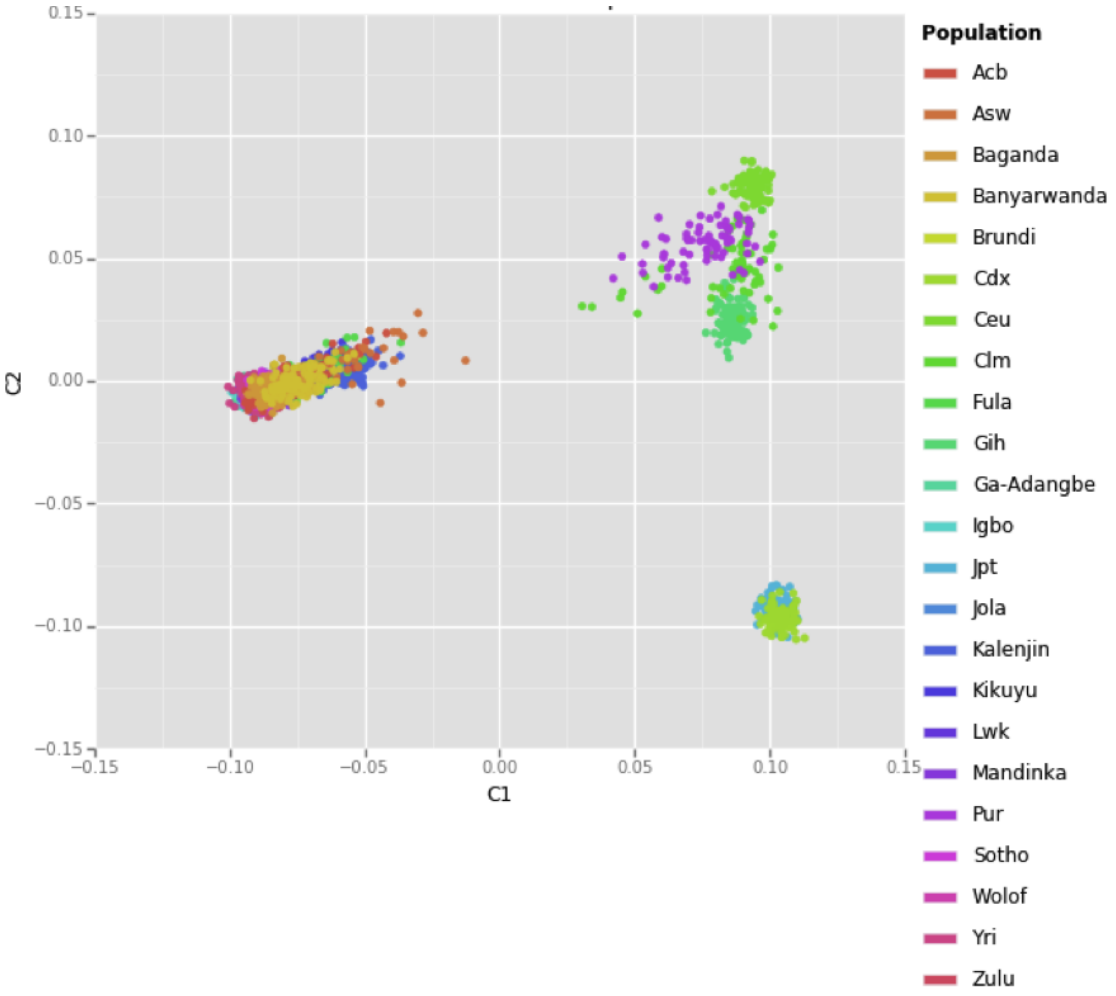


Figure 8.16: All SSA and HapMap populations for whom genotype data was available are presented across PC1 and PC2; ACB:African Caribbeans in Barbados, ASW: Americans of African Ancestry in Southwestern USA, CDX: Chinese Dai in Xishuangbanna, CEU: Utah Residents (CEPH) with Northern and Western European Ancestry, CLM: Colombians from Medellin- Colombia, GIH: Gujarati Indian from Houston- Texas, JPT: Japanese in Tokyo- Japan, LWK: Luhya in Webuye-Kenya, PUR: Puerto Ricans from Puerto Rico, YRI: Yoruba in Ibadan- Nigeria.

Table 8.19: Metric LD map lengths and the total number of markers remained after each filtering step for two alternative MAF thresholds (1% vs. 5%)

Population	Frequency	Failed missingness (GENO > 0.05)	Failed frequency test	Failed HWE	Total number of remaining markers	Number of markers after LD construction	KBmap length (Kb)	Map length (LDU)	Number of samples
Baganda	0.01	17,722	244,146	7,103	215,457	184,304	35,186.78	2,282.84	100
	0.05	17,722	346,431	7,103	116,838	111,889	35,186.78	1,912.33	
Zulu	0.01	14,795	233,990	6,973	225,851	196,211	35,186.88	2,267.67	100
	0.05	14,795	342,146	6,973	121,282	115,564	35,186.78	1,871.40	
Ethiopian	0.01	14,393	260,872	6,533	199,614	197,052	35,192.05	1,950.29	95
	0.05	14,393	360,327	6,533	103,340	102,990	35,186.78	1,887.08	

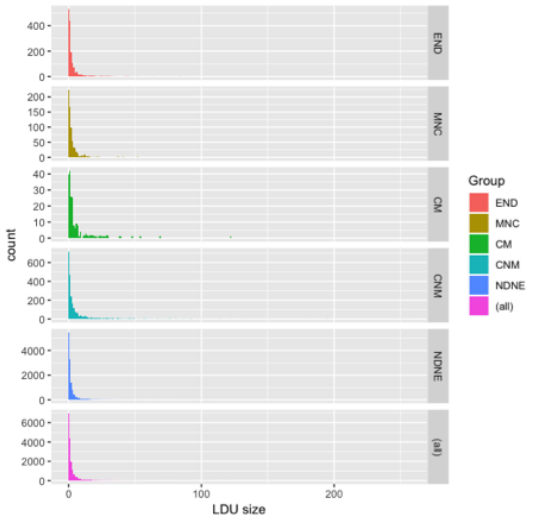
Relaxation of the MAF threshold from 5% to 1% results in 19.37%, 21.18% & 3.35% map length increase in the Baganda, Zulu ,and Ethiopian populations respectively.

Table 8.20: Total number of genomic regions across chromosome 1-22.

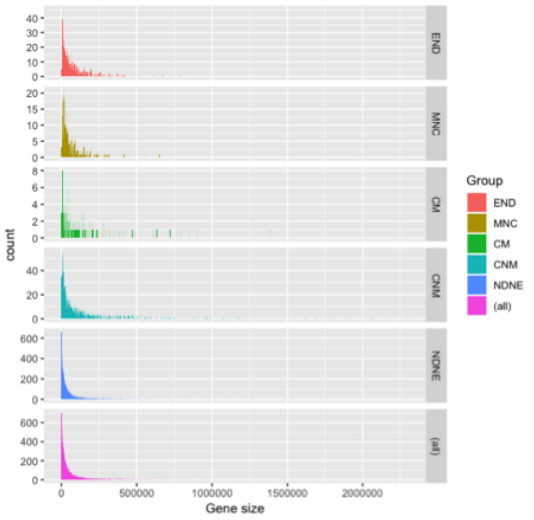
Chr.	**Genic**	**Coding Genes**		**ncRNA genes**		**Intergenic regions**
		Exonic	*Intronic*	*Exonic*	*Intronic*	
1	2,002	19,196	16,775	2,078	1,237	2,003
2	1,341	15,913	12,848	1,757	906	1,314
3	1,081	12,451	10,059	1,366	628	1,072
4	845	8,180	6,575	999	515	836
5	964	9,163	7,480	1,279	659	955
6	1,087	10,452	8,331	1,194	575	1,081
7	938	9,832	7,993	1,485	815	927
8	769	7,125	5,713	999	506	750
9	822	8,390	6,931	991	570	807
10	804	8,650	6,728	1,210	636	791
11	1,161	11,736	9,413	929	454	1,144
12	1,037	11,901	9,837	1,075	539	1,034
13	449	3,696	3,013	994	558	445
14	623	6,252	5,129	711	302	622
15	625	7,328	6,015	1,207	701	609
16	834	8,758	7,261	1,019	627	813
17	1,106	12,256	10,005	971	485	1,093
18	319	3,232	2,610	455	220	318
19	1,368	12,185	9,878	873	439	1,363
20	548	5,196	4,163	676	345	538
21	272	2,137	1,672	426	216	268
22	460	4,367	3,557	657	415	441
Total	19,455	198,396	161,986	23,351	12,348	19,224

Table 8.21: Physical size of different genomic regions in Kb.

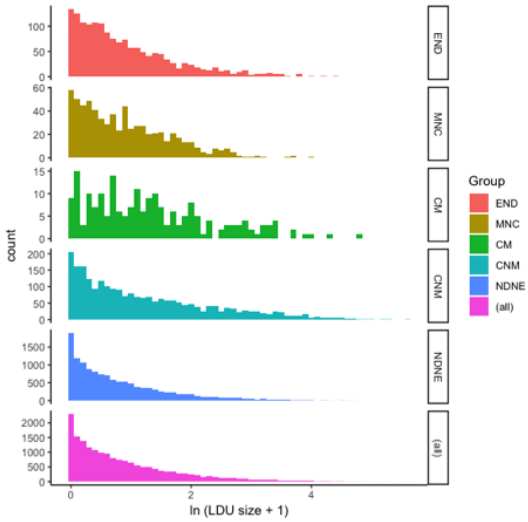
Chr.	**Whole Chr.**	**Whole Chr. LDMAP**	**Genic**	**Coding Genes**		**ncRNA genes**		**Intergenic**	**Centromeric heterochromatin**
				Exonic	*Intronic*	*Exonic*	*Intronic*		
1	249,250.62	249,208.93	117,612.81	6,752.56	102,877.18	827.99	12,722.37	109,978.17	7,400.00
2	243,199.37	243,167.81	111,240.33	4,844.49	91,995.71	641.76	13,540.27	126,152.02	6,300.00
3	198,022.43	197,840.11	97,867.20	4,041.29	83,941.11	502.3	9,311.18	94,955.73	6,000.00
4	191,154.28	190,920.29	76,442.85	2,795.36	62,756.52	358.09	8,039.88	110,570.32	4,500.00
5	180,915.26	180,769.37	77,732.10	3,270.10	61,370.82	477.66	11,023.23	98,563.08	4,600.00
6	171,115.07	170,875.10	75,899.56	3,537.44	64,001.74	521.93	7,638.59	91,210.24	4,600.00
7	159,138.66	159,112.39	78,286.82	3,166.68	67,759.65	567.78	6,118.99	76,514.91	3,700.00
8	146,364.02	146,211.02	64,448.34	2,389.30	53,936.47	416.91	6,832.65	76,725.77	5,000.00
9	141,213.43	141,069.82	52,164.61	2,674.29	45,265.41	381.27	3,993.40	85,071.61	3,400.00
10	135,534.75	135,444.57	68,121.75	2,743.18	59,160.09	512.58	4,913.52	63,358.33	4,300.00
11	135,006.52	134,766.76	63,315.26	3,799.76	54,297.02	374.22	4,670.01	67,741.60	4,100.00
12	133,851.90	133,749.53	64,084.27	3,580.47	55,717.39	369.09	4,233.84	64,937.76	4,900.00
13	115,169.88	96,089.53	39,112.55	1,280.50	31,096.35	357.48	6,684.92	56,640.42	3,200.00
14	107,349.54	88,289.30	40,274.35	2,117.27	33,398.50	257.17	3,234.83	47,974.72	3,000.00
15	102,531.39	82,468.18	44,710.52	2,271.14	37,048.06	402.34	4,190.51	36,916.69	4,900.00
16	90,354.75	90,093.94	37,569.52	2,580.96	31,675.24	277.11	3,261.56	48,206.77	4,000.00
17	81,195.21	81,187.57	45,431.57	3,647.96	38,550.52	412.88	2,954.60	32,190.94	3,600.00
18	78,077.25	77,989.07	31,881.65	1,167.27	26,926.02	186.54	3,524.70	42,863.65	3,600.00
19	59,128.98	58,880.86	31,240.97	3,816.20	25,231.03	319.96	1,783.02	23,916.76	4,200.00
20	63,025.52	62,904.06	29,202.41	1,685.56	25,455.86	240.91	1,926.31	30,244.68	3,800.00
21	48,129.90	38,587.85	14,722.57	661.58	11,041.08	162.39	2,804.57	20,151.52	3,400.00
22	51,304.57	35,186.88	21,031.11	1,397.51	17,072.41	254.73	1,934.64	12,743.47	5,700.00
Total	2,881,033.29	2,794,812.95	1,282,393.10	64,220.86	1,080,574.16	8,823.10	125,337.57	1,417,629.16	98,200.00



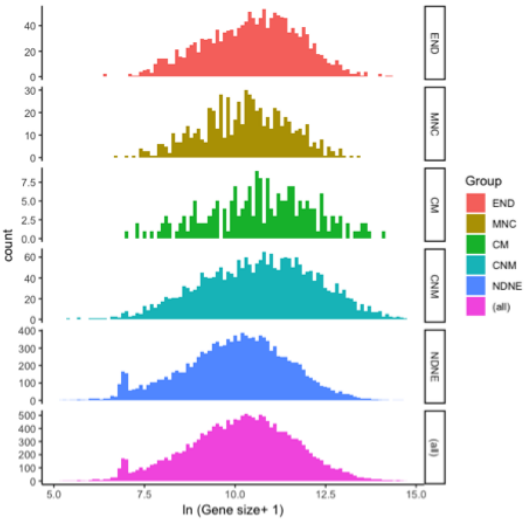
(a) Raw gene-size distribution



(b) Raw LDU size distribution



(c) Log transformed LDU size



(d) Log transformed gene size

Figure 8.17: a. Raw LDU size across the five categories of gene groups ($END_{Skewness}$= 5.40, $END_{kurtosis}$= 46.37, $MNC_{Skewness}$= 5.44, $MNC_{Kurtosis}$= 46.41, $CM_{Skewness}$= 5.07, $CM_{Kurtosis}$= 38.80; $CNM_{Skewness}$= 6.28, $CNM_{Kurtosis}$= 66.54; $NDNE_{Skewness}$= 6.93, $NDNE_{Kurtosis}$= 78.06) b. Distribution of gene size across the five category of gene groups. c. Distribution of natural logarithm transformed LDU size across five gene groups. d. Distribution of natural logarithm transformed gene size across five groups.

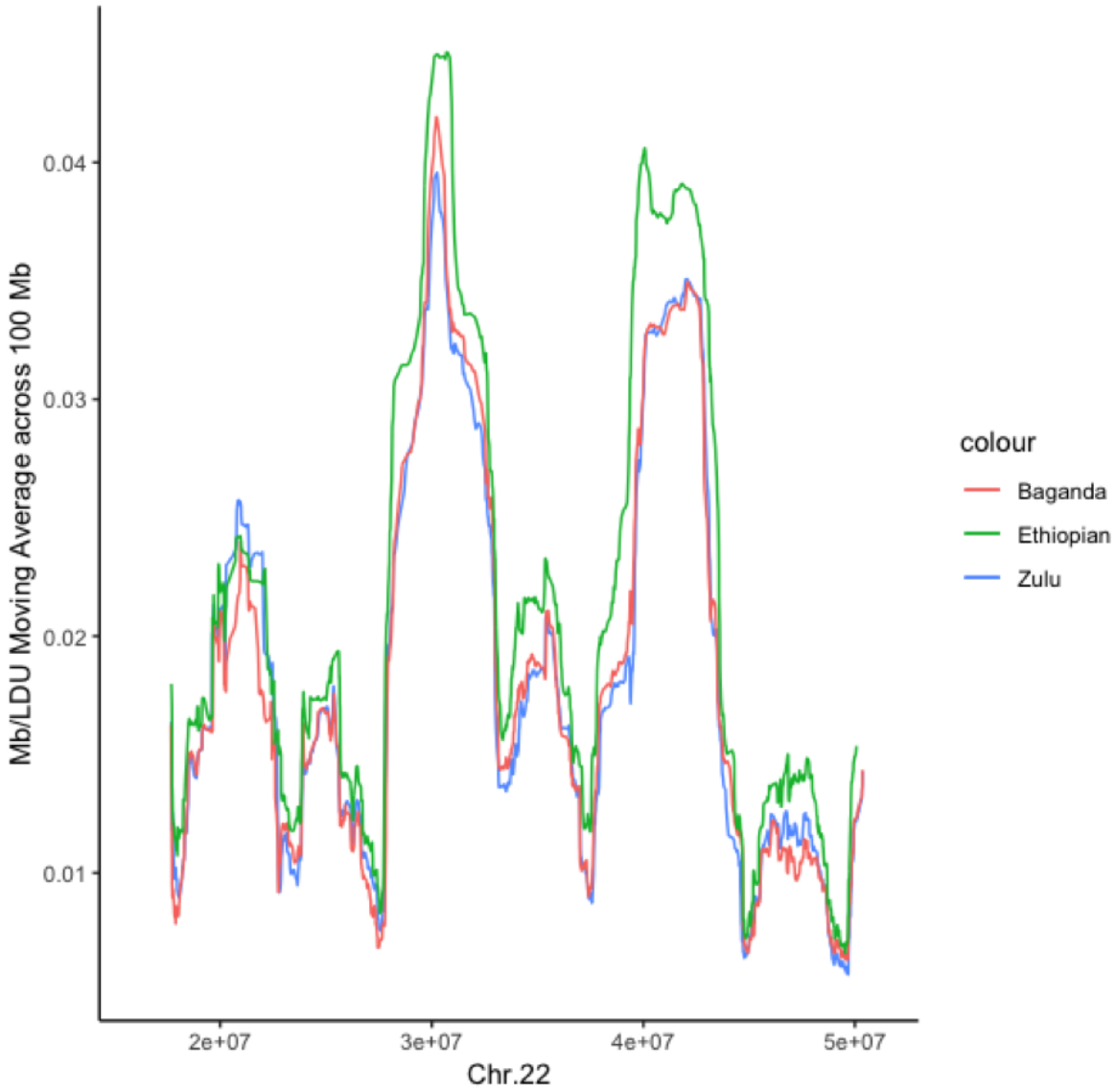


Figure 8.18: Comparison of 100 Mb moving average of Mb/LDU ratio across chr.22 in the three SSA populations. Consistent with the expectation from the overall map lengths for chromosome 22 (Table 6.3), the Ethiopian population represents more extensive LD blocks across the length of the chromosome.

Table 8.22: The LDU extent (KB/LDU) across different genomic regions.

Chr.	**Kb/LDU**	**Kb/LDU (Genic)**	**Coding Genes**		**ncRNA genes**		**Intergenic**
			Exonic	*Intronic*	*Exonic*	*Intronic*	
1	24.36	25.2	49.1	26.29	31.6	24.09	19.94
2	23.05	25.37	41	26.65	20.88	20.79	20.59
3	23.02	24.25	39.87	24.6	23.98	22.76	21.17
4	21.73	22.79	36.75	23.54	26.41	19.75	20.48
5	23.1	25.14	41.01	25.74	28.4	23.36	21.03
6	22.14	23.64	35.78	24.06	21.76	22.77	20.39
7	21.84	23.1	33.56	23.32	24.94	26.17	19.78
8	20.91	21.22	33.22	21.47	18.33	19.9	19.64
9	24.1	22.27	31.9	22.33	35.54	22.88	24.45
10	21.13	22.92	36.55	23.13	21.55	22.8	18.57
11	21.79	22.63	33.78	22.84	18.58	21.6	20.26
12	21.56	22.61	32.14	24.01	20.86	17.45	19.67
13	20.76	21.8	37.49	22.29	18.37	21.37	20.17
14	20.18	23.43	30.93	24.38	20.09	20.69	18.18
15	19.16	22.19	38.71	22.73	24.68	17.74	16.31
16	19.56	18.07	32.95	17.84	20.68	21.74	19.29
17	19.9	21.59	31.76	22.44	23.86	20.37	16.68
18	19.01	20.34	26.52	20.43	17.11	21.53	16.99
19	16.53	16.98	25.33	17.48	20.85	17.57	14.39
20	18.14	19.79	28.32	20.44	16.44	17.65	15.34
21	18.11	17.23	27.29	17.53	17.67	17.03	15.9
22	15.52	18.89	26.85	19.04	19.99	18.77	11.72
Average	20.71	21.89	34.13	22.39	22.39	20.85	18.68

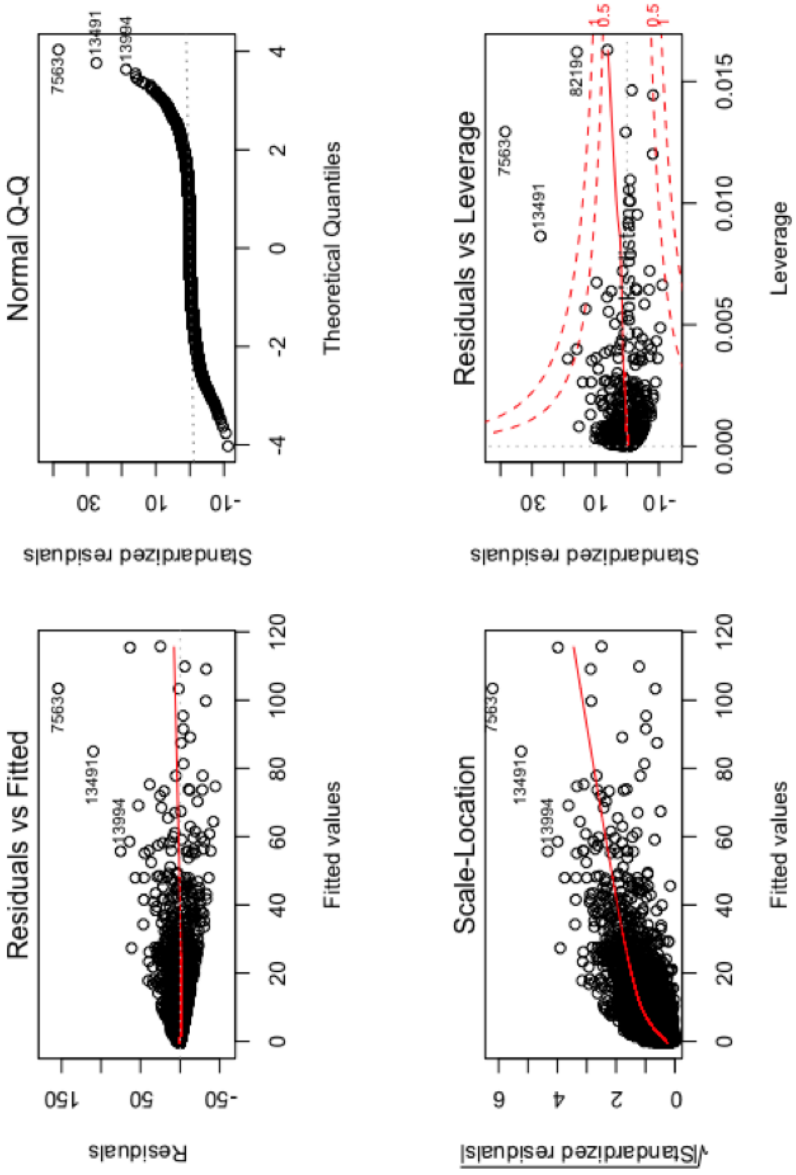
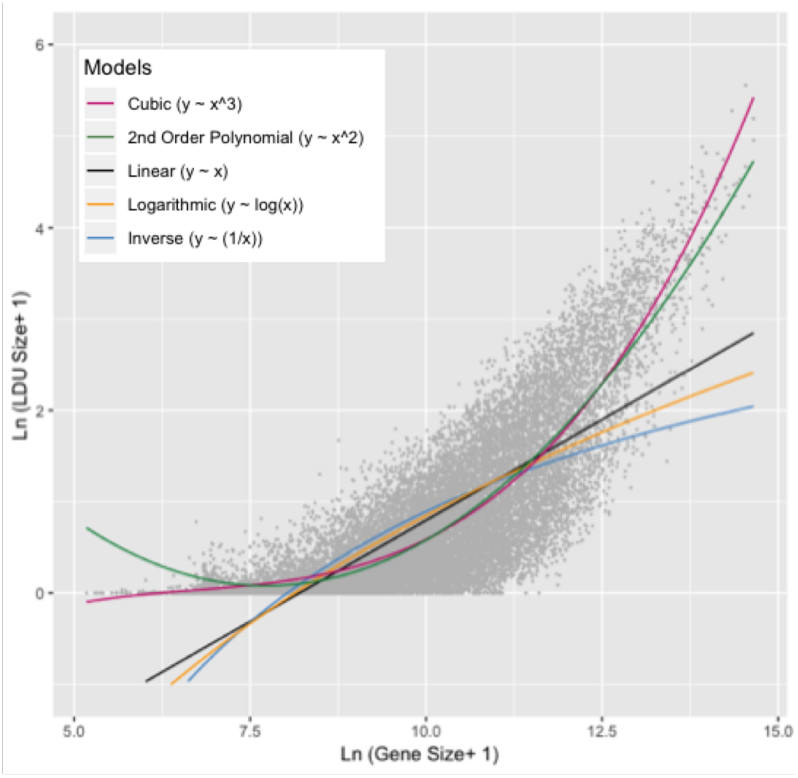
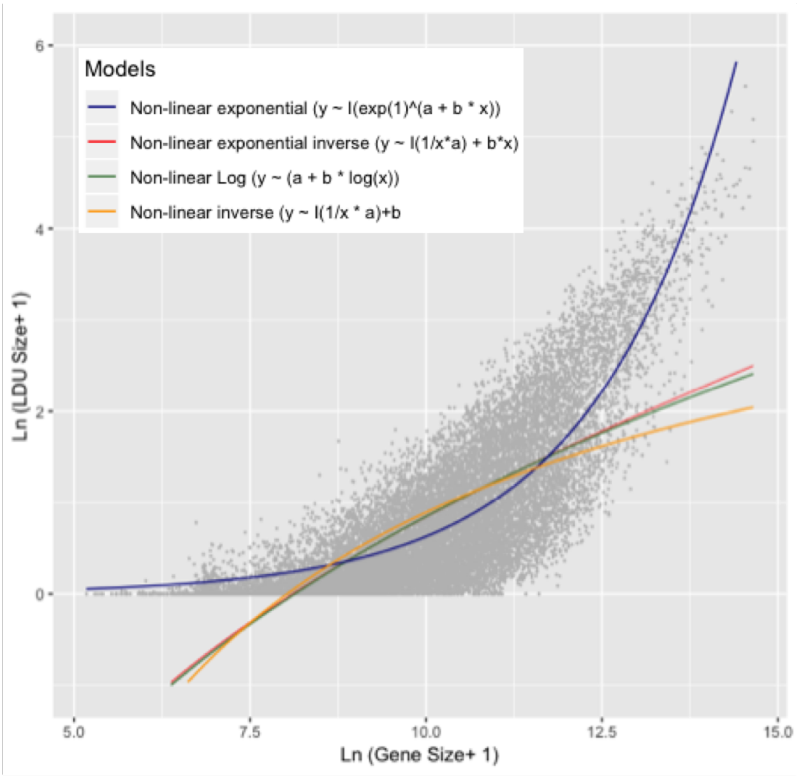


Figure 8.19: Residuals diagnostic plots for the linear regression model for LDU adjustment (using non-transformed data),



(a) Fitted linear regression curves



(b) Fitted non-linear regression curves

Figure 8.20: Fitted regression curves to the transformed data($LDU_T \sim L_T$)

Table 8.23: Wilcoxon comparison of ranked sum $e_L DU$ size across the five categories of genes.

Pairwise test	**Test statistic**	**Standard error**	**Standard test statistic**	**Adjusted P-value**
END- CNM	1,373.89	168.08	8.17	6.30E-13
END- CM	2,913.00	385.97	7.54	4.80E-12
END- MNC	-1,519.89	237.05	-6.41	1.40E-08
NDNE- CM	2,154.68	366.03	5.89	3.90E-07
END- NDNE	-758.31	138.15	-5.49	4.00E-06
NDNE- CNM	615.58	115.17	5.34	9.00E-06
CNM- CM	1,539.10	378.35	4.06	0.0046
NDNE- MNC	761.58	202.99	3.75	0.0165
MNC- CM	1,393.10	413.63	3.37	0.0679
CNM- MNC	-146	224.44	-0.65	0.96659

Table 8.24: Multinomial regression covariates for the fitted model investigating the relationship between the essentiality of the gene (according to OGEE framework [502]) and its respective $e_L DU$ quartile rank.

Parameter Estimates									
Essentiality	**eLDU**	**B**	**Std. Error**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95% CI**	
								Lower Bound	**Upper Bound**
E	Intercept	3.615	0.176	419.949	1	0			
	Q1	0.675	0.257	6.905	1	9.00E-03	1.965	1.187	3.252
	Q2	0.255	0.238	1.145	1	0.285	1.29	0.809	2.057
	Q3	0.323	0.257	1.573	1	0.21	1.381	0.834	2.287
	Q4	0b	.	.	0	.	.	.	.
NE	Intercept	4.582	0.175	685.672	1	0			
	Q1	-0.193	0.256	0.569	1	0.451	0.824	0.499	1.362
	Q2	-0.436	0.237	3.401	1	0.065	0.646	0.406	1.028
	Q3	0.046	0.256	0.032	1	0.858	1.047	0.634	1.728
	Q4	0b	.	.	0	.	.	.	.

†This parameter is set to zero because it is redundant;
E: Essential
NE: Non-Essential

References

[1] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM), an Online catalog of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D789–D798, 2015.

[2] J. McClellan and M.-C. King, "Genetic heterogeneity in human disease," *Cell*, vol. 141, no. 2, pp. 210–217, 2010.

[3] L. E. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, B. W. van Bon, a. Hoischen, B. B. de Vries, H. G. Brunner, and J. a. Veltman, "A de novo paradigm for mental retardation," *Nat Genet*, vol. 42, no. 12, pp. 1109–1112, 2010.

[4] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," 2015.

[5] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and E. A. Consortium, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, pp. 285–291, aug 2016.

[6] D. M. Ruderfer, T. Hamamsy, M. Lek, K. J. Karczewski, D. Kavanagh, K. E. Samocha, M. J. Daly, D. G. Macarthur, M. Fromer, and S. M. Purcell, "Patterns of genic intolerance of rare copy number variation in 59,898 human exomes," *Nat. Genet.*, 2016.

[7] R. Walsh, K. L. Thomson, J. S. Ware, B. H. Funke, J. Woodley, K. J. McGuire, F. Mazzarotto, E. Blair, A. Seller, J. C. Taylor, E. V. Minikel, D. G. MacArthur, M. Farrall, S. A. Cook, and H. Watkins, "Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples," *Genet. Med.*, 2017.

[8] J. Lehmann and A. Libchaber, "Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon.," *RNA*, vol. 14, no. 7, pp. 1264–9, 2008.

[9] D. Perera, R. C. Poulos, A. Shah, D. Beck, J. E. Pimanda, and J. W. H. Wong, "Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes," *Nature*, vol. 532, no. 7598, pp. 259–263, 2016.

[10] L. Duret, "Mutation Patterns in the Human Genome: More Variable Than Expected," *PLoS Biol.*, vol. 7, no. 2, p. e28, 2009.

[11] S. Duchêne, S. Y. W. Ho, and E. C. Holmes, "Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models," *BMC Evol. Biol.*, vol. 15, no. 1, p. 36, 2015.

[12] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *J. Mol. Evol.*, vol. 16, no. 2, pp. 111–120, 1980.

[13] I. Keller, D. Bensasson, and R. A. Nichols, "Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes," *PLoS Genet.*, vol. 3, no. 2, pp. 0185–0191, 2007.

[14] Y. Guo, J. Long, J. He, C.-I. Li, Q. Cai, X.-O. Shu, W. Zheng, and C. Li, "Exome sequencing generates high quality data in non-target regions.," *BMC Genomics*, vol. 13, no. 1, p. 194, 2012.

[15] G. T. Wang, B. Peng, and S. M. Leal, "Variant association tools for quality control and analysis of large-scale sequence and genotyping array data," *Am. J. Hum. Genet.*, vol. 94, no. 5, pp. 770–783, 2014.

[16] M. N. Bainbridge, M. Wang, Y. Wu, I. Newsham, D. M. Muzny, J. L. Jefferies, T. J. Albert, D. L. Burgess, and R. a. Gibbs, "Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities," *Genome Biol.*, vol. 12, no. 7, p. R68, 2011.

[17] L. Cartegni, S. L. Chew, and A. R. Krainer, "Listening to silence and understanding nonsense: exonic mutations that affect splicing.," *Nat. Rev. Genet.*, vol. 3, no. April, pp. 285–298, 2002.

[18] W. F. Mueller, L. S. Larsen, A. Garibaldi, G. W. Hatfield, and K. J. Hertel, "The silent sway of splicing by synonymous substitutions," *J. Biol. Chem.*, vol. 290, no. 46, pp. 27700–27711, 2015.

[19] P. C. Ng and S. Henikoff, "Predicting the Effects of Amino Acid Substitutions on Protein Function," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, no. 1, pp. 61–80, 2006.

[20] S. Teng, A. K. Srivastava, C. E. Schwartz, E. Alexov, and L. Wang, "Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein protein interaction," *Int. J. Comput. Biol. Drug Des.*, vol. 3, no. 4, p. 334, 2010.

[21] D. E. Wildman, M. Uddin, G. Liu, L. I. Grossman, and M. Goodman, "Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 12, pp. 7181–8, 2003.

[22] J. T. Mendell, N. A. Sharifi, J. L. Meyers, F. Martinez-Murillo, and H. C. Dietz, "Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise," *Nat Genet*, vol. 36, no. 10, pp. 1073–8, 2004.

[23] R. A. Veitia, "Dominant negative factors in health and disease.," *J. Pathol.*, vol. 218, no. 4, pp. 409–18, 2009.

[24] Z. E. Sauna and C. Kimchi-Sarfaty, "Understanding the contribution of synonymous mutations to human disease.," *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 683–91, 2011.

[25] L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, and A. R. Krainer, "ESEfinder: A web resource to identify exonic splicing enhancers," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3568–3571, 2003.

[26] Y. Kapustin, E. Chan, R. Sarkar, F. Wong, I. Vorechovsky, R. M. Winston, T. Tatusova, and N. J. Dibb, "Cryptic splice sites and split genes," *Nucleic Acids Res.*, vol. 39, no. 14, pp. 5837–5844, 2011.

[27] D. F. Conrad, J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, and P. Awadalla, "Variation in genome-wide mutation rates within and between human families.," *Nat. Genet.*, vol. 43, no. 7, pp. 712–4, 2011.

[28] G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome variation from population-scale sequencing.," *Nature*, vol. 467, no. 7319, pp. 1061–73, 2010.

[29] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, and J. O. Korbel, "Mapping copy number variation by population-scale genome sequencing.," *Nature*, vol. 470, no. 7332, pp. 59–65, 2011.

[30] S. Clancy and K. M. Shaw, "DNA Deletion and Duplication and the Associated Genetic Disorders," *Nature*, 2008.

[31] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping.," *Nat. Rev. Genet.*, vol. 12, no. 5, pp. 363–76, 2011.

[32] A. W. Pang, J. R. MacDonald, D. Pinto, J. Wei, M. a. Rafiq, D. F. Conrad, H. Park, M. E. Hurles, C. Lee, J. C. Venter, E. F. Kirkness, S. Levy, L. Feuk, and S. W. Scherer, "Towards a comprehensive structural variation map of an individual human genome.," *Genome Biol.*, vol. 11, no. 5, p. R52, 2010.

[33] D. J. Turner, M. Miretti, D. Rajan, H. Fiegler, N. P. Carter, M. L. Blayney, S. Beck, and M. E. Hurles, "Germline rates of de novo meiotic deletions and duplications causing several genomic disorders," *Nat. Genet.*, vol. 40, no. 1, pp. 90–95, 2008.

[34] K. Sander, "Wilhelm Roux and the rest: Developmental theories 1885–1895 BT - Landmarks in Developmental Biology 1883–1924: Historical Essays from Roux's Archives," pp. 16–18, Berlin, Heidelberg: Springer Berlin Heidelberg, 1997.

[35] D. L. Rimoin and K. Hirschhorn, "A history of medical genetics in pediatrics," 2004.

[36] G. W. Beadle and E. L. Tatum, "Genetic Control of Biochemical Reactions in Neurospora.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 27, no. 11, pp. 499–506, 1941.

[37] O. T. Avery and C. M. Macleod, "Studies on the Chemical Inducing Nature Types of the Substance Transformation," *J. Exp. Med.*, vol. 79, no. 2, pp. 137–158, 1994.

[38] S. A. Carter, S. D. Bryce, C. S. Munro, E. Healy, R. Bashir, J. Weissenbach, J. Leblanc-Straceski, R. Kucherlapati, A. Stephenson, J. L. Rees, and T. Strachan, "Linkage Analyses in British Pedigrees Suggest a Single Locus for Darier Disease and Narrow the Location to the Interval between D12S105 and D12S129," 1994.

[39] J. Haldane, "Methods fo detection of autosomal linkage in man," *Ann. Hum. Genet.*, vol. 6, no. 1, pp. 26–65, 1934.

[40] R. Fisher, "The detection of linkage with dominant abnormalities," *Ann. Hum. Genet.*, vol. 6, no. 2, pp. 187–201, 1935.

[41] N. E. Morton, "Sequential tests for the detection of linkage.," *Am. J. Hum. Genet.*, vol. 7, no. 3, pp. 277–318, 1955.

[42] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.," *Nat. Genet.*, vol. 33 Suppl, no. march, pp. 228–37, 2003.

[43] W. Xu, S. B. Bull, L. Mirea, and C. M. T. Greenwood, "Model-Free Linkage Analysis of a Binary Trait BT - Statistical Human Genetics: Methods and Protocols," pp. 317–345, Totowa, NJ: Humana Press, 2012.

[44] D. Botstein, R. White, M. Skolnick, and R. Davis, "Construction of a genetic linkage map in man using restriction fragment length polymorphisms.," *Am. J. Hum. Genet.*, vol. 32, no. 3, pp. 314–331, 1980.

[45] L. Tsui, M. Buchwald, D. Barker, J. Braman, R. Knowlton, J. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic, and A. Et, "Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker," *Science (80-.).*, vol. 230, no. 4729, pp. 1054–1057, 1985.

[46] L. A. Farrer, R. H. Myers, L. A. Cupples, and P. M. Conneally, "Considerations in using linkage analysis as a presymptomatic test for Huntington's disease.," *J. Med. Genet.*, vol. 25, no. 9, pp. 577–88, 1988.

[47] G. Bell, S. Horita, and J. Karam, "A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus.," *Diabetes*, vol. 33, no. 2, pp. 176–183, 1984.

[48] P. H. St George-Hyslop, R. E. Tanzi, R. J. Polinsky, J. L. Haines, L. Nee, P. C. Watkins, R. H. Myers, R. G. Feldman, D. Pollen, D. Drachman, and et Al., "The genetic defect causing familial Alzheimer's disease maps on chromosome 21.," *Science*, vol. 235, no. 4791, pp. 885–90, 1987.

[49] L. G. Lee, C. R. Connell, and W. Bloch, "Allelic discrimination by nick-translation PCR with fluorgenic probes," *Nucleic Acids Res.*, vol. 21, no. 16, pp. 3761–3766, 1993.

[50] N. Risch and K. Merikangas, "The Future of Genetic Studies of Complex Human Diseases," *Science (80-.).*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[51] N. J. Dovichi, "DNA sequencing by capillary electrophoresis," *Electrophoresis*, vol. 41, pp. 2393–2399, 1999.

[52] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[53] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, F. Di V, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, and M. Nodell, "The sequence of the human genome," *Science (80-.).*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[54] S. T. Sherry, M. Ward, and K. Sirotkin, "dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation," *Genome Res.*, vol. 9, no. 8, pp. 677–679, 1999.

[55] S. Wang, N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzotti, K. Hill, A. M. Hurtado, B. Camrena, H. Nicolini, W. Klitz, R. Barrantes, J. A. Molina, N. B. Freimer, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, J. E. Dipierri, E. L. Alfaro, G. Bailliet, N. O. Bianchi, E. Llop, F. Rothhammer, L. Excoffier, and A. Ruiz-Linares, "Geographic patterns of genome admixture in latin American mestizos," *PLoS Genet.*, vol. 4, no. 3, 2008.

[56] N. Rosenberg, L. Huang, E. Jewett, Z. Szpiech, I. Jankovic, and M. Boehnke, "Genome-wide association studies in diverse populations," *Nat. Rev. Genet.*, vol. 11, no. 5, pp. 356–366, 2010.

[57] M. C. Campbell and S. A. Tishkoff, "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 403–33, 2008.

[58] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. a. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, M. T. Pato, T. L. Petryshen, L. N. Kolonel, E. S. Lander, P. Sklar, B. Henderson, J. N. Hirschhorn, and D. Altshuler, "Assessing the impact of population stratification on genetic association studies.," *Nat. Genet.*, vol. 36, no. 4, pp. 388–93, 2004.

[59] M. Slatkin, "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future.," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–85, 2008.

[60] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander, "High-resolution haplotype structure in the human genome.," *Nat. Genet.*, vol. 29, no. 2, pp. 229–232, 2001.

[61] A. R. Collins, "Linkage disequilibrium and association mapping: an introduction.," *Methods Mol. Biol.*, vol. 376, pp. 1–15, 2007.

[62] J. Hey, "What's so hot about recombination hotspots?," 2004.

[63] T. International and H. Consortium, "The International HapMap Project.," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[64] T. International and H. Consortium, "The International HapMap Project.," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[65] A.-C. Syvanen, "Toward genome-wide SNP genotyping," *Nat. Genet.*, vol. 37, no. June, pp. S5–S10, 2005.

[66] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh, "Complement factor H polymorphism in age-related macular degeneration.," *Science*, vol. 308, no. 5720, pp. 385–9, 2005.

[67] J. Fadista, A. K. Manning, J. C. Florez, and L. Groop, "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants," *Eur. J. Hum. Genet.*, no. November 2015, pp. 1–4, 2016.

[68] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?," *Am. J. Hum. Genet.*, vol. 69, no. 1, pp. 124–37, 2001.

[69] T. Langaee and M. Ronaghi, "Genetic variation analyses by Pyrosequencing," 2005.

[70] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," 2008.

[71] T. J. Albert, M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. a. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock, and R. a. Gibbs, "Direct selection of human genomic loci by microarray hybridization.," *Nat. Methods*, vol. 4, no. 11, pp. 903–905, 2007.

[72] L. G. Biesecker, "Exome sequencing makes medical genomics a reality.," *Nat. Genet.*, vol. 42, no. 1, pp. 13–14, 2010.

[73] N. Siva, "1000 Genomes project," *Nat. Biotechnol.*, vol. 26, no. 3, pp. 256–256, 2008.

[74] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, E. E. Eichler, D. A. Nickerson, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, M. Bamshad, and J. Shendure, "Targeted capture and massively parallel sequencing of 12 human exomes.," *Nature*, vol. 461, no. 7261, pp. 272–6, 2009.

[75] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333–351, 2016.

[76] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases.," *Nat. Genet.*, vol. 40, no. 6, pp. 695–701, 2008.

[77] E. P. Hong and J. W. Park, "Sample size and statistical power calculation in genetic association studies.," *Genomics and Informatics*, vol. 10, no. 2, pp. 117–22, 2012.

[78] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nat. Rev. Genet.*, vol. 15, no. 5, pp. 335–346, 2014.

[79] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. a. Todd, "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.," *Science (80-.).*, vol. 324, no. 5925, pp. 387–389, 2009.

[80] M. A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagace, B. Neale, K. S. Lo, P. Schumm, L. Torkvist, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. B. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, and M. J. Daly, "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease," *Nat Genet*, vol. 43, pp. 1066–1073, nov 2011.

[81] J. J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, W. Wu, R. Corominas, Á. Peoples, A. Koren, A. Gore, S. Kang, G. N. Lin, J. Estabillo, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L. M. Iakoucheva, Y. Li, J. Wang, and J. Sebat, "Whole-genome sequencing in autism identifies hot spots for de novo germline mutation," *Cell*, vol. 151, no. 7, pp. 1431–1442, 2012.

[82] R. J. Pengelly, W. Tapper, J. Gibson, M. Knut, R. Tearle, A. Collins, and S. Ennis, "Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations," *BMC Genomics*, vol. 16, no. 1, p. 666, 2015.

[83] Y. Li, C. Sidore, H. M. M. Kang, M. Boehnke, and G. R. Abecasis, "Low-coverage sequencing: implications for design of complex trait association studies.," *Genome Res.*, vol. 21, no. 6, pp. 940–951, 2011.

[84] G. Gibson, "Rare and common variants: twenty arguments," *Nat. Rev. Genet.*, vol. 13, no. 2, pp. 135–145, 2012.

[85] J. T. Dudley, Y. Kim, L. Liu, G. J. Markov, K. Gerold, R. Chen, A. J. Butte, and S. Kumar, "Human genomic disease variants: A neutral evolutionary explanation," 2012.

[86] S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P.-P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti, "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak," *Science (80-.).*, vol. 345, no. 6202, pp. 1369–72, 2014.

[87] D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, and K. Hatzikotoulas, "The African Genome Variation Project shapes medical genetics in Africa," *Nature*, vol. 517, 2015.

[88] H. L. Rehm, "Disease-targeted sequencing: a cornerstone in the clinic," apr 2013.

[89] A. C. Need, V. Shashi, Y. Hitomi, K. Schoch, K. V. Shianna, M. T. McDonald, M. H. Meisler, and D. B. Goldstein, "Clinical application of exome sequencing in undiagnosed genetic conditions," *J. Med. Genet.*, 2012.

[90] Y. Yaping, M. D. M., R. J. G., B. M. N., W. Alecia, W. P. A., B. Alicia, B. Joke, X. Fan, N. Zhiyv, H. Matthew, P. Richard, B. M. Reza, L. M. S., K. Amelia, P. Peter, S. Jennifer, W. Min, D. Yan, P. S. E., L. J. R., B. A. L., G. R. A., and E. C. M., "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders," *N. Engl. J. Med.*, vol. 369, no. 16, pp. 1502–1511, 2013.

[91] Y. Yang, M. DM, F. Xia, and et Al, "Molecular findings among patients referred for clinical whole-exome sequencing," *JAMA*, vol. 312, pp. 1870–1879, nov 2014.

[92] J. Meienberg, R. Bruggmann, K. Oexle, and G. Matyas, "Clinical sequencing: is WGS the better WES?," *Hum. Genet.*, vol. 135, no. 3, pp. 359–362, 2016.

[93] J. C. Taylor, H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier, A. Rimmer, A. Kanapin, G. Lunter, S. Fiddy, C. Allan, A. R. Aricescu, M. Attar, C. Babbs, J. Becq, D. Beeson, C. Bento, P. Bignell, E. Blair, V. J. Buckle, K. Bull, O. Cais, H. Cario, H. Chapel, R. R. Copley, R. Cornall, J. Craft, K. Dahan, E. E. Davenport, C. Dendrou, O. Devuyst, A. L. Fenwick, J. Flint, L. Fugger, R. D. Gilbert, A. Goriely, A. Green, I. H. Greger, R. Grocock, A. V. Gruszczyk, R. Hastings,

E. Hatton, D. Higgs, A. Hill, C. Holmes, M. Howard, L. Hughes, P. Humburg, D. Johnson, F. Karpe, Z. Kingsbury, U. Kini, J. C. Knight, J. Krohn, S. Lamble, C. Langman, L. Lonie, J. Luck, D. McCarthy, S. J. McGowan, M. F. McMullin, K. A. Miller, L. Murray, A. H. Nemeth, M. A. Nesbit, D. Nutt, E. Ormondroyd, A. B. Oturai, A. Pagnamenta, S. Y. Patel, M. Percy, N. Petousi, P. Piazza, S. E. Piret, G. Polanco-Echeverry, N. Popitsch, F. Powrie, C. Pugh, L. Quek, P. A. Robbins, K. Robson, A. Russo, N. Sahgal, P. A. van Schouwenburg, A. Schuh, E. Silverman, A. Simmons, P. S. Sorensen, E. Sweeney, J. Taylor, R. V. Thakker, I. Tomlinson, A. Trebes, S. R. F. Twigg, H. H. Uhlig, P. Vyas, T. Vyse, S. A. Wall, H. Watkins, M. P. Whyte, L. Witty, B. Wright, C. Yau, D. Buck, S. Humphray, P. J. Ratcliffe, J. I. Bell, A. O. M. Wilkie, D. Bentley, P. Donnelly, and G. McVean, "Factors influencing success of clinical genome sequencing across a broad spectrum of disorders," *Nat Genet*, vol. 47, pp. 717–726, jul 2015.

[94] J. A. O'Rawe, S. Ferson, and G. J. Lyon, "Accounting for uncertainty in DNA sequencing data," 2015.

[95] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat. Genet.*, vol. 43, no. 5, pp. 491–498, 2011.

[96] J. D. Wall, L. F. Tang, B. Zerbe, M. N. Kvale, P. Y. Kwok, C. Schaefer, and N. Risch, "Estimating genotype error rates from high-coverage next-generation sequence data," *Genome Res.*, vol. 24, no. 11, pp. 1734–1739, 2014.

[97] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, and M. A. DePristo, "A universal SNP and small-indel variant caller using deep neural networks," *Nat. Biotechnol.*, sep 2018.

[98] H. Li and J. Wren, "Toward better understanding of artifacts in variant calling from high-coverage samples," 2014.

[99] M. G. Dozmorov, I. Adrianto, C. B. Giles, E. Glass, S. B. Glenn, C. Montgomery, K. L. Sivils, L. E. Olson, T. Iwayama, W. M. Freeman, C. J. Lessard, and J. D. Wren, "Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data.," *BMC Bioinformatics*, vol. 16 Suppl 1, no. Suppl 13, p. S10, 2015.

[100] K. V. Fuentes Fajardo, D. Adams, C. E. Mason, M. Sincan, C. Tifft, C. Toro, C. F. Boerkoel, W. Gahl, and T. Markello, "Detecting false-positive signals in exome sequencing," *Hum. Mutat.*, vol. 33, no. 4, pp. 609–613, 2012.

[101] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," 2015.

[102] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, "Characterizing and measuring bias in sequence data.," *Genome Biol.*, vol. 14, no. 5, p. R51, 2013.

[103] M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler, "Genetic variation and the de novo assembly of human genomes," *Nat. Rev. Genet.*, vol. 16, no. 11, pp. 627–640, 2015.

[104] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly.," *Proc. Natl. Acad. Sci.*, vol. 98, no. 17, pp. 9748–53, 2001.

[105] C. Alkan, S. Sajjadian, and E. E. Eichler, "Limitations of next-generation genome sequence assembly," *Nat. Methods*, vol. 8, no. 1, pp. 61–65, 2010.

[106] N. Nagarajan and M. Pop, "Sequence assembly demystified.," *Nat. Rev. Genet.*, vol. 14, no. 3, pp. 157–67, 2013.

[107] C.-Y. Lee, Y.-C. Chiu, L.-B. Wang, Y.-L. Kuo, E. Y. Chuang, L.-C. Lai, and M.-H. Tsai, "Common applications of next-generation sequencing technologies in genomic research," *Transl. Cancer Res.*, vol. 2, no. 1, pp. 33–45, 2013.

[108] C. Bleidorn, "Third generation sequencing: technology and its potential impact on evolutionary biodiversity research," *Syst. Biodivers.*, vol. 2000, no. January, pp. 1–8, 2015.

[109] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-time DNA sequencing from single polymerase molecules.," *Science*, vol. 323, no. 5910, pp. 133–8, 2009.

[110] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss, "The potential and challenges of nanopore sequencing.," *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1146–53, 2008.

[111] H. Lee, J. Gurtowski, and S. Yoo, "Error correction and assembly complexity of single molecule sequencing reads," *bioRxiv*, pp. 1–17, 2014.

[112] Y. Feng, Y. Zhang, C. Ying, D. Wang, and C. Du, "Nanopore-based fourth-generation DNA sequencing technology," 2015.

[113] J. Clarke, H.-c. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing.," *Nat. Nanotechnol.*, vol. 4, no. 4, pp. 265–270, 2009.

[114] T. Laver, J. Harrison, P. A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme, "Assessing the performance of the Oxford Nanopore Technologies MinION," *Biomol. Detect. Quantif.*, vol. 3, pp. 1–8, 2015.

[115] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the MinION nanopore sequencer," *Nat. Methods*, vol. 12, no. 4, pp. 351–356, 2015.

[116] V. Boža, B. Brejová, and T. Vinař, "DeepNano: Deep recurrent neural networks for base calling in MinION Nanopore reads," *PLoS One*, vol. 12, no. 6, 2017.

[117] P. M. Ashton, S. Nair, T. Dallman, S. Rubino, W. Rabsch, S. Mwaigwisya, J. Wain, and J. O'Grady, "MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island," *Nat Biotechnol*, vol. 33, no. 3, pp. 296–300, 2015.

[118] J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. J. Baum, B. Becker-Ziaja, J. P. Boettcher, M. Cabeza-Cabrerizo, Á. Camino-Sánchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. G. Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallasch, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, T. Racine, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. J. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O. Shea, A. M. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Wölfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keïta, A. Rambaut, P. Formenty, S. Günther, and M. W. Carroll, "Real-time, portable genome sequencing for Ebola surveillance," *Nature*, vol. 530, no. 7589, pp. 228–232, 2016.

[119] X. Ma, E. Stachler, and K. Bibby, "Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization," *bioRxiv*, p. 99960, 2017.

[120] C. F. Minervini, C. Cumbo, P. Orsini, L. Anelli, A. Zagaria, L. Impera, N. Coccaro, C. Brunetti, A. Minervini, P. Casieri, G. Tota, A. Russo Rossi, G. Specchia, and F. Albano, "Mutational analysis in BCR-ABL1 positive leukemia by deep sequencing based on nanopore MinION technology," *Exp. Mol. Pathol.*, vol. 103, no. 1, pp. 33–37, 2017.

[121] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biol.*, vol. 11, no. 5, p. 207, 2010.

[122] A. Celesti, F. Celesti, M. Fazio, P. Bramanti, and M. Villari, "Are Next-Generation Sequencing Tools Ready for the Cloud?," 2017.

[123] J. G. Reid, A. Carroll, N. Veeraraghavan, M. Dahdouli, A. Sundquist, A. English, M. Bainbridge, S. White, W. Salerno, C. Buhay, F. Yu, D. Muzny, R. Daly, G. Duyk, R. A. Gibbs, and E. Boerwinkle, "Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline.," *BMC Bioinformatics*, vol. 15, no. 1, p. 30, 2014.

[124] R. C. Green, J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, R. L. Nussbaum, J. M. O'Daniel, K. E. Ormond, H. L. Rehm, M. S. Watson, M. S. Williams, L. G. Biesecker, and American College of Medical Genetics and Genomics, "ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.," *Genet. Med.*, vol. 15, no. 7, pp. 565–74, 2013.

[125] Z. Lin, A. B. Owen, and R. B. Altman, "Genomic Research and Human Subject Privacy," *Science (80-.).*, vol. 305, no. 5681, pp. 183–183, 2004.

[126] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genet.*, vol. 4, no. 8, 2008.

[127] P. Makrythanasis and S. E. Antonarakis, "High-throughput sequencing and rare genetic diseases," in *Mol. Syndromol.*, vol. 3, pp. 197–203, 2012.

[128] Y. Li, B. Pawlik, N. Elcioglu, M. Aglan, H. Kayserili, G. Yigit, F. Percin, F. Goodman, G. Nürnberg, A. Cenani, J. Urquhart, B. D. Chung, S. Ismail, K. Amr, A. D. Aslanger, C. Becker, C. Netzer, P. Scambler, W. Eyaid, H. Hamamy, J. Clayton-Smith, R. Hennekam, P. Nürnberg, J. Herz, S. A. Temtamy, and B. Wollnik, "LRP4 Mutations Alter Wnt/β-Catenin Signaling and Cause Limb and Kidney Malformations in Cenani-Lenz Syndrome," *Am. J. Hum. Genet.*, vol. 86, no. 5, pp. 696–706, 2010.

[129] P. Kumar, J. Radhakrishnan, M. a. Chowdhary, and P. F. Giampietro, "Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department.," *Mayo Clin. Proc.*, vol. 76, no. 8, pp. 777–83, 2001.

[130] D. A. Stevenson and J. C. Carey, "Contribution of malformations and genetic disorders to mortality in a children's hospital.," *Am. J. Med. Genet. A*, vol. 126A, no. 4, pp. 393–7, 2004.

[131] I. Barroso, M. Gurnell, V. E. F. Crowley, M. Agostini, J. W. Schwabe, M. A. Soos, G. L. Maslen, T. D. M. Williams, H. Lewis, A. J. Schafer, V. K. K. Chatterjee, and S. O'Rahilly, "Dominant negative mutations in human PPAR gamma associated with severe insulin resistance, diabetes mellitus and hypertension," *Nature*, vol. 402, pp. 880–883 ST – Dominant negative mutations in human, 1999.

[132] C. F. Wright, D. R. FitzPatrick, and H. V. Firth, "Paediatric genomics: Diagnosing rare disease in children," 2018.

[133] P. W. Yoon, R. S. Olney, M. J. Khoury, W. M. Sappenfield, G. F. Chavez, and D. Taylor, "Contribution of birth defects and genetic diseases to pediatric hospitalizations. A population-based study," *Arch. Pediatr. Adolesc. Med.*, vol. 151, no. 11, pp. 1096–1103, 1997.

[134] R. Shaheen, K. Szymanska, B. Basu, N. Patel, N. Ewida, E. Faqeih, A. Al Hashem, N. Derar, H. Alsharif, M. A. Aldahmesh, A. M. Alazami, M. Hashem, N. Ibrahim, F. M. Abdulwahab, R. Sonbul, H. Alkuraya, M. Alnemer, S. Al Tala, M. Al-Husain, H. Morsy, M. Z. Seidahmed, N. Meriki, M. Al-Owain, S. AlShahwan, B. Tabarki, M. A. Salih, T. Faquih, M. El-Kalioby, M. Ueffing, K. Boldt, C. V. Logan, D. A. Parry, N. Al Tassan, D. Monies, A. Megarbane, M. Abouelhoda, A. Halees, C. A. Johnson, and F. S. Alkuraya, "Characterizing the morbid genome of ciliopathies," *Genome Biol.*, vol. 17, no. 1, p. 242, 2016.

[135] A. Telenti, L. C. T. Pierce, W. H. Biggs, J. di Iulio, E. H. M. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, and J. C. Venter, "Deep sequencing of 10,000 human genomes," *Proc. Natl. Acad. Sci.*, vol. 113, no. 42, pp. 11901–11906, 2016.

[136] OmicX, "OMICtools."

[137] A. Nekrutenko and J. Taylor, "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility," *Nat. Rev. Genet.*, vol. 13, no. 9, pp. 667–672, 2012.

[138] J. A. Glasel, "Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios," *Biotechniques*, vol. 18, no. 1, pp. 62–63, 1995.

[139] S. R. Gallagher, "Quantitation of DNA and RNA with absorption and fluorescence spectroscopy.," *Curr. Protoc. Mol. Biol.*, vol. Appendix 3, no. January, p. 3D, 2011.

[140] L. Zhu, Y. Zhang, W. Zhang, S. Yang, J.-Q. Chen, and D. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," *BMC Genomics*, vol. 10, no. 1, p. 47, 2009.

[141] S. R. Head, H. Kiyomi Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, and P. Ordoukhanian, "Library construction for next-generation sequencing: Overviews and challenges," *Biotechniques*, vol. 56, no. 2, pp. 61–77, 2014.

[142] Illumina, "MultiplexPE_SamplePrep.book," no. December 2008, pp. 1–23, 2008.

[143] R. J. Pengelly, J. Gibson, G. Andreoletti, A. Collins, C. J. Mattocks, and S. Ennis, "A SNP profiling panel for sample tracking in whole-exome sequencing studies.," *Genome Med.*, vol. 5, no. 9, p. 89, 2013.

[144] N. Aziz, Q. Zhao, L. Bry, D. K. Driscoll, B. Funke, J. S. Gibson, W. W. Grody, M. R. Hegde, G. A. Hoeltge, D. G. B. Leonard, J. D. Merker, R. Nagarajan, L. A. Palicki, R. S. Robetorye, I. Schrijver, K. E. Weck, and K. V. Voelkerding, "College of American Pathologists' Laboratory Standards for Next-Generation Sequencing Clinical Tests," *Arch. Pathol. Lab. Med.*, vol. 139, no. 4, pp. 481–493, 2015.

[145] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Res.*, vol. 8, no. 3, pp. 186–194, 1998.

[146] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, p. 10, 2011.

[147] S. Andrews, "FastQC: A quality control tool for high throughput sequence data.," *Babraham Bioinforma.*, 2010.

[148] R. Lindner and C. C. Friedel, "A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq," *PLoS One*, vol. 7, no. 12, 2012.

[149] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, 1970.

[150] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[151] M. T. W. Ebbert, M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, J. Duce, J. S. K. Kauwe, and P. G. Ridge, "Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches," *BMC Bioinformatics*, vol. 17, no. S7, p. 239, 2016.

[152] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[153] R. Bao, L. Huang, J. Andrade, W. Tan, W. a. Kibbe, H. Jiang, and G. Feng, "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing," *Lib. Acad.*, vol. 13, pp. 67–82, 2014.

[154] W. Zhou, T. Chen, H. Zhao, A. K. Eterovic, F. Meric-Bernstam, G. B. Mills, and K. Chen, "Bias from removing read duplication in ultra-deep sequencing experiments," *Bioinformatics*, vol. 30, no. 8, pp. 1073–1080, 2014.

[155] S. Tian, H. Yan, M. Kalmbach, and S. L. Slager, "Impact of post-alignment processing in variant discovery from whole exome data," *BMC Bioinformatics*, vol. 17, no. 1, p. 403, 2016.

[156] N. Homer and S. F. Nelson, "Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA," *Genome Biol.*, vol. 11, no. 10, p. R99, 2010.

[157] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, "Dindel: Accurate indel calls from short-read data," *Genome Res.*, vol. 21, no. 6, pp. 961–973, 2011.

[158] H. Li, "Exploring single-sample snp and indel calling with whole-genome de novo assembly," *Bioinformatics*, vol. 28, no. 14, pp. 1838–1844, 2012.

[159] P. Carnevali, J. Baccash, A. L. Halpern, I. Nazarenko, G. B. Nilsen, K. P. Pant, J. C. Ebert, A. Brownley, M. Morenzoni, V. Karpinchyk, B. Martin, D. G. Ballinger, and R. Drmanac, "Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads," *J. Comput. Biol.*, vol. 19, no. 3, pp. 279–292, 2012.

[160] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," 2010.

[161] Y. Xue, Y. Chen, Q. Ayub, N. Huang, E. V. Ball, M. Mort, A. D. Phillips, K. Shaw, P. D. Stenson, D. N. Cooper, and C. Tyler-Smith, "Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing," *Am. J. Hum. Genet.*, vol. 91, no. 6, pp. 1022–1032, 2012.

[162] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline," *Curr. Protoc. Bioinforma.*, no. SUPL.43, 2013.

[163] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data," *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 443–451, 2011.

[164] G. Jun, M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, and H. M. Kang, "Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data," *Am. J. Hum. Genet.*, vol. 91, no. 5, pp. 839–848, 2012.

[165] S. organisation, "The Variant Call Format (VCF) Version 4 . 2 Specification," *Online Resour.*, pp. 1–28, 2015.

[166] R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks, "Scaling accurate genetic variant discovery to tens of thousands of samples," *bioRxiv*, 2017.

[167] C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman, "Disease gene identification strategies for exome sequencing.," *Eur. J. Hum. Genet.*, 2012.

[168] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, 2010.

[169] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins.," *Nucleic Acids Res.*, vol. 40, no. Web Server issue, pp. W452–7, 2012.

[170] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Curr. Protoc. Hum. Genet.*, no. SUPPL.76, 2013.

[171] S. Chun and J. C. Fay, "Identification of deleterious mutations within three human genomes," *Genome Res.*, vol. 19, no. 9, pp. 1553–1561, 2009.

[172] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nat. Methods*, vol. 7, no. 8, pp. 575–576, 2010.

[173] H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. A. Barker, K. J. Edwards, I. N. M. Day, and T. R. Gaunt, "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models," *Hum. Mutat.*, vol. 34, no. 1, pp. 57–65, 2013.

[174] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, B. J. O. Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants.," *Nat. Genet.*, vol. 46, no. 3, pp. 310–315, 2014.

[175] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, 2010.

[176] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. D. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res.*, vol. 15, no. 8, pp. 1034–1050, 2005.

[177] M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie, "Identifying novel constrained elements by exploiting biased substitution patterns," in *Bioinformatics*, vol. 25, 2009.

[178] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe, A. K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, A. L. Martins, T. Massingham, I. Moltke, B. J. Raney, M. D. Rasmussen, J. Robinson, A. Stark, A. J. Vilella, J. Wen, X. Xie, M. C. Zody, J. Baldwin, T. Bloom, C. Whye Chin, D. Heiman, R. Nicol, C. Nusbaum, S. Young, J. Wilkinson, K. C. Worley, C. L. Kovar, D. M. Muzny, R. A. Gibbs, A. Cree, H. H. Dihn, G. Fowler, S. Jhangiani, V. Joshi, S. Lee, L. R. Lewis, L. V. Nazareth, G. Okwuonu, J. Santibanez, W. C. Warren, E. R. Mardis, G. M. Weinstock, R. K. Wilson, K. Delehaunty, D. Dooling, C. Fronik, L. Fulton, B. Fulton, T. Graves, P. Minx, E. Sodergren, E. Birney, E. H. Margulies, J. Herrero, E. D. Green, D. Haussler, A. Siepel, N. Goldman, K. S. Pollard, J. S. Pedersen, E. S. Lander, and M. Kellis, "A high-resolution map of human evolutionary constraint using 29 mammals," *Nature*, vol. 478, no. 7370, pp. 476–482, 2011.

[179] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs," *Hum. Mutat.*, vol. 37, no. 3, pp. 235–241, 2016.

[180] G. Yeo and C. B. Burge, "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals," *J. Comput. Biol.*, vol. 11, pp. 377–394, mar 2004.

[181] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, "Human Splicing Finder: an online bioinformatics tool to predict splicing signals," *Nucleic Acids Res.*, vol. 37, pp. e67–e67, may 2009.

[182] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff," *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, 2012.

[183] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.

[184] R. Sun, Q. Deng, I. Hu, B. C.-Y. Zee, and M. H. Wang, "A clustering approach to identify rare variants associated with hypertension," *BMC Proc.*, vol. 10, pp. 153–157, oct 2016.

[185] P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. D. Phillips, and D. N. Cooper, "The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine," 2014.

[186] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott, "ClinVar: Public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862–D868, 2016.

[187] D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J.-B. Cazier, and P. Donnelly, "Choice of transcripts and software has a large effect on variant annotation," *Genome Med.*, vol. 6, no. 3, p. 26, 2014.

[188] Y. Fu, Z. Liu, S. Lou, J. Bedford, X. J. Mu, K. Y. Yip, E. Khurana, and M. Gerstein, "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer," *Genome Biol.*, vol. 15, no. 10, p. 480, 2014.

[189] K. Eilbeck, A. Quinlan, and M. Yandell, "Settling the score: variant prioritization and Mendelian disease," *Nat. Rev. Genet.*, 2017.

[190] S. De Summa, G. Malerba, R. Pinto, A. Mori, V. Mijatovic, and S. Tommasi, "GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data," *BMC Bioinformatics*, vol. 18, no. S5, p. 119, 2017.

[191] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls," *Nat. Biotechnol.*, 2014.

[192] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Stromberg, and G. T. Marth, "Bamtools: A C++ API and toolkit for analyzing and managing BAM files," *Bioinformatics*, vol. 27, no. 12, pp. 1691–1692, 2011.

[193] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, "LUMPY: a probabilistic framework for structural variant discovery," *Genome Biol.*, vol. 15, no. 6, p. R84, 2014.

[194] B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, S. Loeillet, P. Legoix-né, A. Nicolas, O. Delattre, and E. Barillot, "SVDetect: A tool to identify genomic structural variations from paired-end and mate-pair sequencing data," *Bioinformatics*, vol. 26, no. 15, pp. 1895–1896, 2010.

[195] D. Earl, K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W. K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. A. Fonseca, I. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, S. Koren, S. P. Yang, W. Wu, W. C. Chou, A. Srivastava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. A. Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. E. Green, D. Haussler, I. Korf, and B. Paten, "Assemblathon 1: A competitive assessment of de novo short read assembly methods," 2011.

[196] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs, and E. E. Eichler, "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nat. Genet.*, vol. 41, no. 10, pp. 1061–1067, 2009.

[197] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis, "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation," *Nat. Methods*, vol. 6, no. 9, pp. 677–681, 2009.

[198] F. Hormozdiari, I. Hajirasouliha, P. Dao, F. Hach, D. Yorukoglu, C. Alkan, E. E. Eichler, and S. C. Sahinalp, "Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery," *Bioinformatics*, vol. 26, no. 12, 2010.

[199] J. O. Korbel, A. Abyzov, X. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein, "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome Biol.*, vol. 10, no. 2, p. R23, 2009.

[200] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael, "A geometric approach for classification and comparison of structural variants.," *Bioinformatics*, vol. 25, no. 12, pp. i222–30, 2009.

[201] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, "Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm," *Nucleic Acids Res.*, vol. 39, no. 10, 2011.

[202] V. Bansal, C. Dorn, M. Grunert, S. Klaassen, R. Hetzer, F. Berger, and S. R. Sperling, "Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with tetralogy of fallot," *PLoS One*, vol. 9, no. 1, 2014.

[203] P. Guan and W.-K. Sung, "Structural variation detection using next-generation sequencing data: A comparative technical review," *Methods*, vol. 102, pp. –, 2016.

[204] E. Karakoc, C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson, and E. E. Eichler, "Detection of structural variants and indels within exome data," *Nat. Methods*, vol. 9, no. 2, pp. 176–178, 2011.

[205] L. Tattini, R. D'Aurizio, and A. Magi, "Detection of Genomic Structural Variants from Next-Generation Sequencing Data," *Front. Bioeng. Biotechnol.*, vol. 3, 2015.

[206] G. Bihl and A. Meyers, "Recurrent renal stone disease - Advances in pathogenesis and clinical management," *Lancet*, vol. 358, no. United Kingdom LG - English PT - Journal: Conference Paper EM - 200100 DD - 20010912, pp. 651–656, 2001.

[207] J. D. Sibonga and R. Pietrzyk, "Evidence Report : Risk of Renal Stone Formation.," *Evid. Rep. NASA Hum. Res. Progr.*, 2017.

[208] E. N. Taylor, M. J. Stampfer, and G. C. Curhan, "Diabetes mellitus and the risk of nephrolithiasis," *Kidney Int.*, vol. 68, no. 3, pp. 1230–1235, 2005.

[209] L. Borghi, T. Meschi, A. Guerra, A. Briganti, T. Schianchi, F. Allegri, and A. Novarini, "Essential arterial hypertension and stone disease," *Kidney Int.*, vol. 55, no. 6, pp. 2397–2406, 1999.

[210] H. M. Kramer and G. Curhan, "The association between gout and nephrolithiasis: The National Health and Nutrition Examination Survey III, 1988-1994," *Am. J. Kidney Dis.*, vol. 40, no. 1, pp. 37–42, 2002.

[211] Y. K. Tan, D. Y. Cha, and M. Gupta, "Management of Stones in Abnormal Situations," 2013.

[212] P. M. Ferraro, A. D'Addessi, and G. Gambaro, "When to suspect a genetic disorder in a patient with renal stones, and why.," *Nephrol. Dial. Transplant*, vol. 28, no. 4, pp. 811–20, 2013.

[213] N. Mohebbi, P. M. Ferraro, G. Gambaro, and R. Unwin, "Tubular and genetic disorders associated with kidney stones," *Urolithiasis*, pp. 1–11, 2016.

[214] O. A. Raheem, Y. S. Khandwala, R. L. Sur, K. R. Ghani, and J. D. Denstedt, "Burden of Urolithiasis: Trends in Prevalence, Treatments, and Costs," 2017.

[215] V. O. Edvardsson, D. S. Goldfarb, J. C. Lieske, L. Beara-Lasic, F. Anglani, D. S. Milliner, and R. Palsson, "Hereditary causes of kidney stones and chronic kidney disease," *Pediatr. Nephrol.*, vol. 28, no. 10, pp. 1923–1942, 2013.

[216] G. Rumsby, "Genetic defects underlying renal stone disease," 2016.

[217] G. Bollée, C. Dollinger, L. Boutaud, D. Guillemot, A. Bensman, J. Harambat, P. Deteix, M. Daudon, B. Knebelmann, and I. Ceballos-Picot, "Phenotype and genotype characterization of adenine phosphoribosyltransferase deficiency.," *J. Am. Soc. Nephrol.*, vol. 21, no. 4, pp. 679–688, 2010.

[218] P. Cochat and G. Rumsby, "Primary Hyperoxaluria.," *N. Engl. J. Med.*, vol. 369, no. 7, pp. 649–658, 2013.

[219] E. L. Williams, D. Bockenhauer, W. G. Van'T Hoff, N. Johri, C. Laing, M. D. Sinha, R. Unwin, A. Viljoen, and G. Rumsby, "The enzyme 4-hydroxy-2-oxoglutarate aldolase is deficient in primary hyperoxaluria type 3," *Nephrol. Dial. Transplant.*, vol. 27, no. 8, pp. 3191–3195, 2012.

[220] N. Issler, S. Dufek, R. Kleta, D. Bockenhauer, N. Smeulders, and W. van't Hoff, "Epidemiology of paediatric renal stone disease: a 22-year single centre experience in the UK," *BMC Nephrol.*, vol. 18, no. 1, p. 136, 2017.

[221] J. Halbritter, M. Baum, A. M. Hynes, S. J. Rice, D. T. Thwaites, Z. S. Gucev, B. Fisher, L. Spaneas, J. D. Porath, D. a. Braun, A. J. Wassner, C. P. Nelson, V. Tasic, J. a. Sayer, and F. Hildebrandt, "Fourteen monogenic genes account for 15% of nephrolithiasis/nephrocalcinosis.," *J. Am. Soc. Nephrol.*, vol. 26, no. 3, pp. 543–51, 2015.

[222] M. J. Stechman, N. Y. Loh, and R. V. Thakker, "Genetics of hypercalciuric nephrolithiasis: Renal stone disease," in *Ann. N. Y. Acad. Sci.*, vol. 1116, pp. 461–484, 2007.

[223] K. K. Frick and D. A. Bushinsky, "Molecular Mechanisms of Primary Hypercalciuria," *J. Am. Soc. Nephrol.*, vol. 14, pp. 1082–1095, apr 2003.

[224] M. Audran and E. Legrand, "Hypercalciuria.," *Joint. Bone. Spine*, vol. 67, no. 6, pp. 509–15, 2000.

[225] O. W. Moe and O. Bonny, "Genetic hypercalciuria.," *J. Am. Soc. Nephrol.*, vol. 16, no. 3, pp. 729–745, 2005.

[226] D. J. Hunter, M. D. Lange, H. Snieder, A. J. MacGregor, R. Swaminathan, R. V. Thakker, and T. D. Spector, "Genetic contribution to renal function and electrolyte balance: a twin study.," *Clin. Sci. (Lond).*, vol. 103, no. 3, pp. 259–65, 2002.

[227] C. Polito, A. La Manna, B. Nappi, J. Villani, and R. Di Toro, "Idiopathic hypercalciuria and hyperuricosuria: Family prevalence of nephrolithiasis," *Pediatr. Nephrol.*, vol. 14, no. 12, pp. 1102–1104, 2000.

[228] M. J. Stechman, N. Y. Loh, and R. V. Thakker, "Genetic causes of hypercalciuric nephrolithiasis," 2009.

[229] D. Prié, V. Huart, N. Bakouh, G. Planelles, O. Dellis, B. Gérard, P. Hulin, F. Benqué-Blanchet, C. Silve, B. Grandchamp, and G. Friedlander, "Nephrolithiasis and Osteoporosis Associated with Hypophosphatemia Caused by Mutations in the Type 2a Sodium–Phosphate Cotransporter," *N. Engl. J. Med.*, 2002.

[230] D. Rendina, G. Mossetti, G. De Filippo, M. Cioffi, and P. Strazzullo, "Fibroblast growth factor 23 is increased in calcium nephrolithiasis with hypophosphatemia and renal phosphate leak," *J. Clin. Endocrinol. Metab.*, 2006.

[231] V. Walker, E. M. Stansbridge, and D. G. Griffin, "Demography and biochemistry of 2800 patients from a renal stones clinic," *Ann. Clin. Biochem.*, 2013.

[232] C. A. Wagner, I. Rubio-Aliaga, J. Biber, and N. Hernando, "Genetic diseases of renal phosphate handling," 2014.

[233] D. Riccardi and E. M. Brown, "Physiology and pathophysiology of the calcium-sensing receptor in the kidney," *AJP Ren. Physiol.*, 2010.

[234] A. Martin, V. David, and L. D. Quarles, "Regulation and Function of the FGF23/Klotho Endocrine Pathways," *Physiol. Rev.*, 2012.

[235] V. Walker and D. G. Griffin, *Renal calcium and urate handling and diet in idiopathic calcium stone formers.* 2015.

[236] D. S. Goldfarb, M. E. Fischer, Y. Keich, and J. Goldberg, "A twin study of genetic and dietary influences on nephrolithiasis: A report from the Vietnam Era Twin (VET) Registry," *Kidney Int.*, vol. 67, no. 3, pp. 1053–1061, 2005.

[237] J. C. Loredo-Osti, N. M. Roslin, J. Tessier, T. M. Fujiwara, K. Morgan, and A. Bonnardeaux, "Segregation of urine calcium excretion in families ascertained for nephrolithiasis: Evidence for a major gene," *Kidney Int.*, 2005.

[238] M. Monga, B. Macias, E. Groppo, and A. Hargens, "Genetic Heritability of Urinary Stone Risk in Identical Twins," *J. Urol.*, vol. 175, pp. 2125–2128, jun 2006.

[239] J. C. Lieske, S. T. Turner, S. N. Edeh, J. A. Smith, and S. L. Kardia, "Heritability of urinary traits that contribute to nephrolithiasis," *Clin. J. Am. Soc. Nephrol.*, 2014.

[240] D. G. Griffin, *A family-based study investigating the genetic basis of calcium-containing kiney stones.* PhD thesis, University of Southampton, 2001.

[241] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: Text mining and data integration of disease-gene associations," *Methods*, vol. 74, pp. 83–89, 2015.

[242] K. Semagn, R. Babu, S. Hearne, and M. Olsen, "Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement," 2014.

[243] X. S. Wang, K. Diener, D. Jannuzzi, D. Trollinger, T. H. Tan, H. Lichenstein, M. Zukowski, and Z. Yao, "Molecular cloning and characterization of a novel protein kinase with a catalytic domain homologous to mitogen-activated protein kinase kinase kinase.," *J. Biol. Chem.*, vol. 271, no. 49, pp. 31607–31611, 1996.

[244] H. Ichijo, "Induction of Apoptosis by ASK1, a Mammalian MAPKKK That Activates SAPK/JNK and p38 Signaling Pathways," *Science (80-.).*, vol. 275, no. 5296, pp. 90–94, 1997.

[245] F. Y. Ma, G. H. Tesch, and D. J. Nikolic-Paterson, "ASK1/p38 signaling in renal tubular epithelial cells promotes renal fibrosis in the mouse obstructed kidney.," *Am. J. Physiol. Renal Physiol.*, vol. 307, no. 11, pp. F1263–73, 2014.

[246] C. Stambe, "p38 Mitogen-Activated Protein Kinase Activation and Cell Localization in Human Glomerulonephritis: Correlation with Renal Injury," *J. Am. Soc. Nephrol.*, vol. 15, no. 2, pp. 326–336, 2004.

[247] N. I. Dmitrieva, D. V. Bulavin, A. J. Fornace, and M. B. Burg, "Rapid activation of G2/M checkpoint after hypertonic stress in renal inner medullary epithelial (IME) cells is protective and requires p38 kinase.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 8, pp. 184–189, 2002.

[248] A. Del Arco and J. Satrústegui, "Identification of a novel human subfamily of mitochondrial carriers with calcium-binding domains," *J. Biol. Chem.*, vol. 279, no. 23, pp. 24701–24713, 2004.

[249] G. Fiermonte, E. Paradies, S. Todisco, C. M. T. Marobbio, and F. Palmieri, "A Novel Member of Solute Carrier Family 25 (SLC25A42) Is a Transporter of Coenzyme A and Adenosine 3,5-Diphosphate in Human Mitochondria," *J. Biol. Chem.*, vol. 284, no. 27, pp. 18152–18159, 2009.

[250] A. Okada, T. Yasui, S. Hamamoto, M. Hirose, Y. Kubota, Y. Itoh, K. Tozawa, Y. Hayashi, and K. Kohri, "Genome-wide analysis of genes related to kidney stone formation and elimination in the calcium oxalate nephrolithiasis model mouse: detection of stone-preventive factors and involvement of macrophage activity.," *J. Bone Miner. Res.*, vol. 24, no. 5, pp. 908–924, 2009.

[251] E. Purevjav, J. Varela, M. Morgado, D. L. Kearney, H. Li, M. D. Taylor, T. Arimura, C. L. Moncman, W. McKenna, R. T. Murphy, S. Labeit, M. Vatta, N. E. Bowles, A. Kimura, A. M. Boriek, and J. A. Towbin, "Nebulette mutations are associated with dilated cardiomyopathy and endocardial fibroelastosis," *J. Am. Coll. Cardiol.*, vol. 56, no. 18, pp. 1493–1502, 2010.

[252] K. Sintiprungrat, N. Singhto, and V. Thongboonkerd, "Characterization of calcium oxalate crystal-induced changes in the secretome of U937 human monocytes," *Mol. BioSyst.*, vol. 12, no. 3, pp. 879–889, 2016.

[253] C. Yin and N. Wang, "Kidney injury molecule-1 in kidney disease," *Ren. Fail.*, vol. 38, no. 10, pp. 1567–1573, 2016.

[254] D. Banville, S. Ahmad, R. Stocco, and S. H. Shen, "A novel protein-tyrosine phosphatase with homology to both the cytoskeletal proteins of the band 4.1 family and junction-associated guanylate kinases," *J. Biol. Chem.*, vol. 269, no. 35, pp. 22320–22327, 1994.

[255] J. Miyazaki, K. Ito, T. Fujita, Y. Matsuzaki, T. Asano, M. Hayakawa, T. Asano, and Y. Kawakami, "Progression of Human Renal Cell Carcinoma via Inhibition of RhoA-ROCK Axis by PARG1.," *Transl. Oncol.*, vol. 10, pp. 142–152, apr 2017.

[256] A. Munck, C. Böhm, N. M. Seibel, Z. H. Hosseini, and W. Hampe, "Hu-K4 is a ubiquitously expressed type 2 transmembrane protein associated with the endoplasmic reticulum," *FEBS J.*, vol. 272, no. 7, pp. 1718–1726, 2005.

[257] C. Cruchaga, C. M. Karch, S. C. Jin, B. A. Benitez, Y. Cai, R. Guerreiro, O. Harari, J. Norton, J. Budde, S. Bertelsen, A. T. Jeng, B. Cooper, T. Skorupa, D. Carrell, D. Levitch, S. Hsu, J. Choi, M. Ryten, J. Hardy, M. Ryten, D. Trabzuni, M. E. Weale, A. Ramasamy, C. Smith, C. Sassi, J. Bras, J. R. Gibbs, D. G. Hernandez, M. K. Lupton, J. Powell, P. Forabosco, P. G. Ridge, C. D. Corcoran, J. T. Tschanz, M. C. Norton, R. G. Munger, C. Schmutz, M. Leary, F. Y. Demirci, M. N. Bamne, X. Wang, O. L. Lopez, M. Ganguli, C. Medway, J. Turton, J. Lord, A. Braae, I. Barber, K. Brown, P. Passmore, D. Craig, J. Johnston, B. McGuinness, S. Todd, R. Heun, H. Kölsch, P. G. Kehoe, N. M. Hooper, E. R. Vardy, D. M. Mann, S. Pickering-Brown, K. Brown, N. Kalsheker, J. Lowe, K. Morgan, A. David Smith, G. Wilcock, D. Warden, C. Holmes, P. Pastor, O. Lorenzo-Betancor, Z. Brkanac, E. Scott, E. Topol, K. Morgan, E. Rogaeva, A. B. Singleton, J. Hardy, M. I. Kamboh, P. St George-Hyslop, N. Cairns, J. C. Morris, J. S. K. Kauwe, and A. M. Goate, "Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease," *Nature*, vol. 505, no. 7484, pp. 550–554, 2013.

[258] M. J. Barron, S. J. Brookes, C. E. Draper, D. Garrod, J. Kirkham, R. C. Shore, and M. J. Dixon, "The cell adhesion molecule nectin-1 is critical for normal enamel formation in mice," *Hum. Mol. Genet.*, vol. 17, no. 22, pp. 3509–3520, 2008.

[259] D. A. Braun, J. A. Lawson, H. Y. Gee, J. Halbritter, S. Shril, W. Tan, D. Stein, A. J. Wassner, M. A. Ferguson, Z. Gucev, B. Fisher, L. Spaneas, J. Varner, J. A. Sayer, D. Milosevic, M. Baum, V. Tasic, and F. Hildebrandt, "Prevalence of Monogenic Causes in Pediatric Patients with Nephrolithiasis or Nephrocalcinosis," *Clin. J. Am. Soc. Nephrol.*, vol. 11, no. 4, pp. 664–672, 2016.

[260] G. Meroni, B. Franco, N. Archidiacono, S. Messali, G. Andolfi, M. Rocchi, and A. Ballabio, "Characterization of a cluster of sulfatase genes on Xp22.3 suggests gene duplications in an ancestral pseudoautosomal region," *Hum. Mol. Genet.*, 1996.

[261] Weizmann Institute of Science, "GeneCards- DNAJA4 Gene."

[262] S. H. Obligado and D. S. Goldfarb, "The association of nephrolithiasis with hypertension and obesity: A review," 2008.

[263] M. Thiri, K. Honda, K. Kashiwase, A. Mabuchi, H. Suzuki, K. Watanabe, M. Nakayama, T. Watanabe, K. Doi, K. Tokunaga, and E. Noiri, "High-density Association Mapping and Interaction Analysis of PLA2R1 and HLA Regions with Idiopathic Membranous Nephropathy in Japanese," vol. 6, p. 38189, dec 2016.

[264] Y. Lin, Q. Zhang, Z. A. Zhong, Z. Xu, S. He, F. Rao, Y. Liu, J. Tang, F. Wang, H. Liu, J. Xie, H. Wu, S. Wang, X. Li, Z. Shan, C. Deng, Z. Liao, H. Deng, H. Liao, Y. Xue, W. Chen, X. Zhan, B. Zhang, and S. Wu, "Whole Genome Sequence Identified a Rare Homozygous Pathogenic Mutation of the DSG2 Gene in a Familial Arrhythmogenic Cardiomyopathy Involving Both Ventricles," *Cardiology*, vol. 138, no. 1, pp. 41–54, 2017.

[265] A. Perrot, P. Tomasov, E. Villard, R. Faludi, P. Melacini, J. Lossie, N. Lohmann, P. Richard, M. De Bortoli, A. Angelini, A. Varga-Szemes, S. R. Sperling, T. Simor, J. Veselka, C. Özcelik, and P. Charron, "Mutations in NEBL encoding the cardiac

Z-disk protein nebulette are associated with various cardiomyopathies," *Arch. Med. Sci.*, vol. 12, no. 2, pp. 263–278, 2016.

[266] J. V. Bonventre, "Kidney injury molecule-1 (KIM-1): a urinary biomarker and much more," *Nephrol. Dial. Transplant.*, vol. 24, no. 11, p. 3265, 2009.

[267] J. Satrústegui, B. Pardo, and A. del Arco, "Mitochondrial Transporters as Novel Targets for Intracellular Calcium Signaling," *Physiol. Rev.*, 2007.

[268] E. Féraille and a. Doucet, "Sodium-potassium-adenosinetriphosphatase-dependent sodium transport in the kidney: hormonal control.," *Physiol. Rev.*, 2001.

[269] V. R. Price, C. A. Reed, W. Lieberthal, and J. H. Schwartz, "ATP depletion of tubular cells causes dissociation of the zonula adherens and nuclear translocation of beta-catenin and LEF-1.," *J. Am. Soc. Nephrol.*, 2002.

[270] A. Zuk, J. V. Bonventre, D. Brown, and K. S. Matlin, "Polarity, integrin, and extracellular matrix dynamics in the postischemic rat kidney.," *Am. J. Physiol.*, 1998.

[271] B. Y. Reed, W. L. Gitomer, H. J. Heller, C. H. Ming, M. Lemke, P. Padalino, and C. Y. C. Pak, "Identification and characterization of a gene with base substitutions associated with the absorptive hypercalciuria phenotype and low spinal bone density," *J. Clin. Endocrinol. Metab.*, vol. 87, no. 4, pp. 1476–1485, 2002.

[272] F. R. Spivacow, E. E. del Valle, A. L. Negri, E. Fradinger, A. Abib, and P. Rey, "Biochemical diagnosis in 3040 kidney stone formers in Argentina," *Urolithiasis*, vol. 43, no. 4, pp. 323–330, 2015.

[273] E. M. Worcester and F. L. Coe, "New Insights Into the Pathogenesis of Idiopathic Hypercalciuria," *Semin. Nephrol.*, 2008.

[274] T. Long, M. Hicks, H.-C. Yu, W. H. Biggs, E. F. Kirkness, C. Menni, J. Zierer, K. S. Small, M. Mangino, H. Messier, S. Brewerton, Y. Turpaz, B. A. Perkins, A. M. Evans, L. A. D. Miller, L. Guo, C. T. Caskey, N. J. Schork, C. Garner, T. D. Spector, J. C. Venter, and A. Telenti, "Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites," *Nat. Genet.*, 2017.

[275] L. Guo, M. V. Milburn, J. A. Ryals, S. C. Lonergan, M. W. Mitchell, J. E. Wulff, D. C. Alexander, A. M. Evans, B. Bridgewater, L. Miller, M. L. Gonzalez-Garay, and C. T. Caskey, "Plasma metabolomic profiles enhance precision medicine for volunteers of normal health," *Proc. Natl. Acad. Sci.*, 2015.

[276] S.-Y. Shin, E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, T. M. T. H. E. R. M. Consortium, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector, and N. Soranzo, "An atlas of genetic influences on human blood metabolites," *Nat. Genet.*, vol. 46, p. 543, may 2014.

[277] H. H. Draisma, R. Pool, M. Kobl, R. Jansen, A. K. Petersen, A. A. Vaarhorst, I. Yet, T. Haller, A. Demirkan, T. Esko, G. Zhu, S. Böhringer, M. Beekman, J. B. Van Klinken, W. Römisch-Margl, C. Prehn, J. Adamski, A. J. De Craen, E. M. Van Leeuwen, N. Amin, H. Dharuri, H. J. Westra, L. Franke, E. J. De Geus, J. J. Hottenga, G. Willemsen, A. K. Henders, G. W. Montgomery, D. R. Nyholt, J. B. Whitfield, B. W. Penninx, T. D. Spector, A. Metspalu, P. Eline Slagboom, K. W. Van Dijk, P. A. 'T Hoen, K. Strauch, N. G. Martin, G. J. B. Van Ommen, T. Illig, J. T. Bell, M. Mangino, K. Suhre, M. I. McCarthy, C. Gieger, A. Isaacs, C. M. Van Duijn, and D. I. Boomsma, "Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels," *Nat. Commun.*, 2015.

[278] D. J. Weiner, E. M. Wigdor, S. Ripke, R. K. Walters, J. A. Kosmicki, J. Grove, K. E. Samocha, J. I. Goldstein, A. Okbay, J. Bybjerg-Grauholm, T. Werge, D. M. Hougaard, J. Taylor, D. Skuse, B. Devlin, R. Anney, S. J. Sanders, S. Bishop, P. B. Mortensen, A. D. Børglum, G. D. Smith, M. J. Daly, E. B. Robinson, M. Bækvad-Hansen, A. Dumont, C. Hansen, T. F. Hansen, D. Howrigan, M. Mattheisen, J. Moran, O. Mors, M. Nordentoft, B. Nørgaard-Pedersen, T. Poterba, J. Poulsen, C. Stevens, V. Anttila, P. Holmans, H. Huang, L. Klei, P. H. Lee, S. E. Medland, B. Neale, L. A. Weiss, L. Zwaigenbaum, T. W. Yu, K. Wittemeyer, A. J. Willsey, E. M. Wijsman, T. H. Wassink, R. Waltes, C. A. Walsh, S. Wallace, J. A. Vorstman, V. J. Vieland, A. M. Vicente, H. Van Engeland, K. Tsang, A. P. Thompson, P. Szatmari, O. Svantesson, S. Steinberg, K. Stefansson, H. Stefansson, M. W. State, L. Soorya, T. Silagadze, S. W. Scherer, G. D. Schellenberg, S. Sandin, E. Saemundsen, G. A. Rouleau, B. Rogé, K. Roeder, W. Roberts, J. Reichert, A. Reichenberg, K. Rehnström, R. Regan, F. Poustka, C. S. Poultney, J. Piven, D. Pinto, M. A. Pericak-Vance, M. Pejovic-Milovancevic, M. G. Pedersen, C. B. Pedersen, A. D. Paterson, J. R. Parr, A. T. Pagnamenta, G. Oliveira, J. I. Nurnberger, M. T. Murtha, S. Mouga, E. M. Morrow, D. M. DeLuca, A. P. Monaco, N. Minshew, A. Merikangas, W. M. McMahon, S. G. McGrew, I. Martsenkovsky, D. M. Martin, S. M. Mane, P. Magnusson, T. Magalhaes, E. Maestrini, J. K. Lowe, C. Lord, P. Levitt, C. L. Martin, D. H. Ledbetter, M. Leboyer, A. S. LeCouteur, C. Ladd-Acosta, A. Kolevzon, S. M. Klauck, S. Jacob, B. Iliadou, C. M. Hultman, I. Hertz-Picciotto, R. Hendren, C. S. Hansen, J. L. Haines, S. J. Guter, D. E. Grice, J. M. Green, A. Green, A. P. Goldberg, C. Gillberg, J. Gilbert, L. Gallagher, C. M. Freitag, E. Fombonne, S. E. Folstein, B. Fernandez, M. D. Fallin, A. G. Ercan-Sencicek, S. Ennis, F. Duque, E. Duketis, R. Delorme, S. DeRubeis, M. V. DeJonge, G. Dawson, M. L. Cuccaro, C. T. Correia, J. Conroy, I. C. Conceição, A. G. Chiocchetti, P. B. Celestino-Soper, J. Casey, R. M. Cantor, C. Cafe, S. Brennan, T. Bourgeron, P. F. Bolton, S. Bölte, N. Bolshakova, C. Betancur, R. Bernier, A. L. Beaudet, A. Battaglia, V. H. Bal, G. Baird, A. J. Bailey, J. S. Bader, E. Bacchelli, E. Anagnostou, D. Amaral, J. Almeida, J. D. Buxbaum, A. Chakravarti, E. H. Cook, H. Coon, D. H. Geschwind, M. Gill, H. Hakonarson, J. Hallmayer, A. Palotie, S. Santangelo, J. S. Sutcliffe, and D. E. Arking, "Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders," *Nat. Genet.*, 2017.

[279] D. J. Stavropoulos, D. Merico, R. Jobling, S. Bowdin, N. Monfared, B. Thiruvahindrapuram, T. Nalpathamkalam, G. Pellecchia, R. K. C. Yuen, M. J. Szego, R. Z. Hayeems, R. Z. Shaul, M. Brudno, M. Girdea, B. Frey, B. Alipanahi, S. Ahmed, R. Babul-Hirji, R. B. Porras, M. T. Carter, L. Chad, A. Chaudhry, D. Chitayat, S. J. Doust, C. Cytrynbaum, L. Dupuis, R. Ejaz, L. Fishman, A. Guerin, B. Hashemi, M. Helal, S. Hewson, M. Inbar-Feigenberg, P. Kannu, N. Karp, R. H. Kim, J. Kro-

nick, E. Liston, H. MacDonald, S. Mercimek-Mahmutoglu, R. Mendoza-Londono, E. Nasr, G. Nimmo, N. Parkinson, N. Quercia, J. Raiman, M. Roifman, A. Schulze, A. Shugar, C. Shuman, P. Sinajon, K. Siriwardena, R. Weksberg, G. Yoon, C. Carew, R. Erickson, R. A. Leach, R. Klein, P. N. Ray, M. S. Meyn, S. W. Scherer, R. D. Cohn, and C. R. Marshall, "Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine," *npj Genomic Med.*, vol. 1, no. 1, p. 15012, 2016.

[280] L. Liu, S. Oza, D. Hogan, J. Perin, I. Rudan, J. E. Lawn, S. Cousens, C. Mathers, and R. E. Black, "Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: An updated systematic analysis," *Lancet*, vol. 385, no. 9966, pp. 430–440, 2015.

[281] S. E. McCandless, J. W. Brunger, and S. B. Cassidy, "The burden of genetic disease on inpatient care in a children's hospital.," *Am. J. Hum. Genet.*, vol. 74, no. 1, pp. 121–7, 2004.

[282] D. T. Miller, M. P. Adam, S. Aradhya, L. G. Biesecker, A. R. Brothman, N. P. Carter, D. M. Church, J. A. Crolla, E. E. Eichler, C. J. Epstein, W. A. Faucett, L. Feuk, J. M. Friedman, A. Hamosh, L. Jackson, E. B. Kaminsky, K. Kok, I. D. Krantz, R. M. Kuhn, C. Lee, J. M. Ostell, C. Rosenberg, S. W. Scherer, N. B. Spinner, D. J. Stavropoulos, J. H. Tepperberg, E. C. Thorland, J. R. Vermeesch, D. J. Waggoner, M. S. Watson, C. L. Martin, and D. H. Ledbetter, "Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies," *Am. J. Hum. Genet.*, vol. 86, no. 5, pp. 749–764, 2010.

[283] Y. Xue, A. Ankala, W. R. Wilcox, and M. R. Hegde, "Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing," *Genet. Med.*, vol. 17, no. 6, pp. 444–451, 2015.

[284] H. Lee, J. L. Deignan, N. Dorrani, S. P. Strom, S. Kantarci, F. Quintero-Rivera, K. Das, T. Toy, B. Harry, M. Yourshaw, M. Fox, B. L. Fogel, J. A. Martinez-Agosto, D. A. Wong, V. Y. Chang, P. B. Shieh, C. G. S. Palmer, K. M. Dipple, W. W. Grody, E. Vilain, and S. F. Nelson, "Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders," *JAMA*, vol. 312, no. 18, p. 1880, 2014.

[285] Z. Stark, D. Schofield, K. Alam, W. Wilson, N. Mupfeki, I. Macciocca, R. Shrestha, S. M. White, and C. Gaff, "Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement," *Genet. Med.*, vol. 19, no. 8, pp. 867–874, 2017.

[286] S. E. Soden, C. J. Saunders, L. K. Willig, E. G. Farrow, L. D. Smith, J. E. Petrikin, J.-B. LePichon, N. A. Miller, I. Thiffault, D. L. Dinwiddie, G. Twist, A. Noll, B. A. Heese, L. Zellmer, A. M. Atherton, A. T. Abdelmoity, N. Safina, S. S. Nyp, B. Zuccarelli, I. A. Larson, A. Modrcin, S. Herd, M. Creed, Z. Ye, X. Yuan, R. A. Brodsky, and S. F. Kingsmore, "Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders," *Sci. Transl. Med.*, vol. 6, no. 265, pp. 265ra168–265ra168, 2014.

[287] Y.-h. Jiang, R. K. C. Yuen, X. Jin, M. Wang, N. Chen, X. Wu, J. Ju, J. Mei, Y. Shi, M. He, G. Wang, J. Liang, Z. Wang, D. Cao, M. T. Carter, C. Chrysler, I. E. Drmic, J. L. Howe, L. Lau, C. R. Marshall, D. Merico, T. Nalpathamkalam,

B. Thiruvahindrapuram, A. Thompson, M. Uddin, S. Walker, J. Luo, E. Anagnostou, L. Zwaigenbaum, R. H. Ring, J. Wang, C. Lajonchere, J. Wang, A. Shih, P. Szatmari, H. Yang, G. Dawson, Y. Li, and S. W. Scherer, "Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing.," *Am. J. Hum. Genet.*, vol. 93, no. 2, pp. 249–63, 2013.

[288] C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. M. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, R. Leach, R. Klein, R. Tearle, T. Bo, R. Pfundt, H. G. Yntema, B. B. A. de Vries, T. Kleefstra, H. G. Brunner, L. E. L. M. Vissers, and J. A. Veltman, "Genome sequencing identifies major causes of severe intellectual disability," *Nature*, vol. 511, no. 7509, pp. 344–347, 2014.

[289] H. C. Martin, G. E. Kim, A. T. Pagnamenta, Y. Murakami, G. L. Carvill, E. Meyer, R. R. Copley, A. Rimmer, G. Barcia, M. R. Fleming, J. Kronengold, M. R. Brown, K. A. Hudspith, J. Broxholme, A. Kanapin, J. B. Cazier, T. Kinoshita, R. Nabbout, D. Bentley, G. McVean, S. Heavin, Z. Zaiwalla, T. McShane, H. C. Mefford, D. Shears, H. Stewart, M. A. Kurian, I. E. Scheffer, E. Blair, P. Donnelly, L. K. Kaczmarek, and J. C. Taylor, "Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis," *Hum. Mol. Genet.*, vol. 23, no. 12, pp. 3200–3211, 2014.

[290] C. J. Saunders, N. A. Miller, S. E. Soden, D. L. Dinwiddie, A. Noll, N. A. Alnadi, N. Andraws, M. L. Patterson, L. A. Krivohlavek, J. Fellis, S. Humphray, P. Saffrey, Z. Kingsbury, J. C. Weir, J. Betley, R. J. Grocock, E. H. Margulies, E. G. Farrow, M. Artman, N. P. Safina, J. E. Petrikin, K. P. Hall, S. F. Kingsmore, R. James, E. H. Margulies, E. G. Farrow, M. Artman, P. Safina, J. E. Petrikin, K. P. Hall, and S. F. Kingsmore, "Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units," *Sci. Transl. Med.*, vol. 4, no. 154, pp. 154ra135–154ra135, 2012.

[291] J. G. Gall and M. L. Pardue, "Formation and Detection of Rna-Dna Hybrid Molecules in Cytological Preparations*," *Proc Natl Acad Sci U S A.*, vol. 63, no. 1, pp. 378–383, 1969.

[292] J. G. Bauman, J. Wiegant, P. Borst, and P. van Duijn, "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA," *Exp. Cell Res.*, vol. 128, no. 2, pp. 485–490, 1980.

[293] a. Kuwano, S. a. Ledbetter, W. B. Dobyns, B. S. Emanuel, and D. H. Ledbetter, "Detection of deletions and cryptic translocations in Miller-Dieker syndrome by in situ hybridization.," *Am. J. Hum. Genet.*, vol. 49, no. 4, pp. 707–714, 1991.

[294] J. R. Lupski, R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, D. F. Barker, J. M. Killian, C. A. Garcia, A. Chakravarti, and P. I. Patel, "DNA duplication associated with Charcot-Marie-Tooth disease type 1A," *Cell*, vol. 66, no. 2, pp. 219–232, 1991.

[295] M. R. Speicher and N. P. Carter, "The new cytogenetics: blurring the boundaries with molecular biology," *Nat. Rev. Genet.*, vol. 6, no. 10, pp. 782–792, 2005.

[296] B. J. Trask, "Human cytogenetics: 46 chromosomes, 46 years and counting," *Nat. Rev. Genet.*, vol. 3, no. 10, pp. 769–778, 2002.

[297] a. Gozzetti and M. M. Le Beau, "Fluorescence in situ hybridization: uses and limitations.," *Semin. Hematol.*, vol. 37, no. 4, pp. 320–333, 2000.

[298] M. M. Weiss, M. A. Hermsen, G. A. Meijer, N. C. van Grieken, J. P. Baak, E. J. Kuipers, and P. J. van Diest, "Comparative genomic hybridisation.," *Mol. Pathol.*, vol. 52, no. 5, pp. 243–51, 1999.

[299] a. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.," *Science*, vol. 258, no. 5083, pp. 818–821, 1992.

[300] R. Bishop, "Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance," 2010.

[301] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy Number Variation in Human Health, Disease, and Evolution," *Annu. Rev. Genomics Hum. Genet.*, vol. 10, no. 1, pp. 451–481, 2009.

[302] H. V. Firth, S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. V. Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter, "DECIPHER v9.16- Release July 2017," *DECIPHER*, no. v9.16, 2017.

[303] H. V. Firth, S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. V. Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter, "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources," *Am. J. Hum. Genet.*, vol. 84, no. 4, pp. 524–533, 2009.

[304] I. Lappalainen, J. Lopez, L. Skipper, T. Hefferon, J. D. Spalding, J. Garner, C. Chen, M. Maguire, M. Corbett, G. Zhou, J. Paschall, V. Ananiev, P. Flicek, and D. M. Church, "DbVar and DGVa: Public archives for genomic structural variation," *Nucleic Acids Res.*, vol. 41, no. D1, 2013.

[305] K. Lin, G. Bonnema, G. Sanchez-Perez, and D. De Ridder, "Making the difference: Integrating structural variation detection tools," *Brief. Bioinform.*, vol. 16, no. 5, pp. 852–864, 2014.

[306] K. H. Utami, A. M. Hillmer, I. Aksoy, E. G. Y. Chew, A. S. M. Teo, Z. Zhang, C. W. H. Lee, P. J. Chen, C. C. Seng, P. N. Ariyaratne, S. L. Rouam, L. S. Soo, S. Yousoof, I. Prokudin, G. Peters, F. Collins, M. Wilson, A. Kakakios, G. Haddad, A. Menuet, O. Perche, S. K. H. Tay, K. W. K. Sung, X. Ruan, Y. Ruan, E. T. Liu, S. Briault, R. V. Jamieson, S. Davila, and V. Cacheux, "Detection of Chromosomal Breakpoints in Patients with Developmental Delay and Speech Disorders," *PLoS One*, vol. 9, no. 3, pp. 1–10, 2014.

[307] G. Vandeweyer and R. F. Kooy, "Balanced translocations in mental retardation," *Hum. Genet.*, vol. 126, no. 1, pp. 133–147, 2009.

[308] J. A. Fantes, E. Boland, J. Ramsay, D. Donnai, M. Splitt, J. A. Goodship, H. Stewart, M. Whiteford, P. Gautier, L. Harewood, S. Holloway, F. Sharkey, E. Maher, V. van Heyningen, J. Clayton-Smith, D. R. Fitzpatrick, and G. C. M. Black, "{FISH} Mapping of De Novo Apparently Balanced Chromosome Rearrangements Identifies Characteristics Associated with Phenotypic Abnormality," *Am. J. Hum. Genet.*, vol. 82, no. 4, pp. 916–926, 2008.

[309] J. S. Bae, N. K. Kim, C. Lee, S. C. Kim, H. R. Lee, H. R. Song, K. B. Park, H. W. Kim, S. H. Lee, H. Y. Kim, S. C. Lee, C. Jeong, M. S. Park, W. J. Yoo, C. Y. Chung, I. H. Choi, O. H. Kim, W. Y. Park, and T. J. Cho, "Comprehensive genetic exploration of skeletal dysplasia using targeted exome sequencing," *Genet. Med.*, 2016.

[310] K. A. Geister and S. A. Camper, "Advances in Skeletal Dysplasia Genetics," *Annu. Rev. Genomics Hum. Genet.*, 2015.

[311] L. Bonafe, V. Cormier-Daire, C. Hall, R. Lachman, G. Mortier, S. Mundlos, G. Nishimura, L. Sangiorgi, R. Savarirayan, D. Sillence, J. Spranger, A. Superti-Furga, M. Warman, and S. Unger, "Nosology and classification of genetic skeletal disorders: 2015 revision," *Am. J. Med. Genet. Part A*, 2015.

[312] C. M. Hall, "International Nosology and Classification of Constitutional Disorders of Bone (2001)," *Am. J. Med. Genet.*, 2002.

[313] S. Maddirevula, S. Alsahli, L. Alhabeeb, N. Patel, F. Alzahrani, H. E. Shamseldin, S. Anazi, N. Ewida, H. S. Alsaif, J. Y. Mohamed, A. M. Alazami, N. Ibrahim, F. Abdulwahab, M. Hashem, M. Abouelhoda, D. Monies, N. Al Tassan, M. Alshammari, A. Alsagheir, M. Z. Seidahmed, S. Sogati, M. S. Aglan, M. H. Hamad, M. A. Salih, A. A. Hamed, N. Alhashmi, A. Nabil, F. Alfadli, G. M. H. Abdel-Salam, H. Alkuraya, W. O. Peitee, W. T. Keng, A. Qasem, A. M. Mushiba, M. S. Zaki, M. R. Fassad, M. Alfadhel, S. Alexander, Y. Sabr, S. Temtamy, A. V. Ekbote, S. Ismail, G. A. Hosny, G. A. Otaify, K. Amr, S. Al Tala, A. O. Khan, T. Rizk, A. Alaqeel, A. Alsiddiky, A. Singh, S. Kapoor, A. Alhashem, E. Faqeih, R. Shaheen, and F. S. Alkuraya, "Expanding the phenome and variome of skeletal dysplasia," *Genet. Med.*, 2018.

[314] S. Lazarus, A. Zankl, and E. L. Duncan, "Next-generation sequencing: A frameshift in skeletal dysplasia gene discovery," 2014.

[315] C. Aristidou, C. Koufaris, A. Theodosiou, M. Bak, M. M. Mehrjouy, F. Behjati, G. Tanteles, V. Christophidou-Anastasiadou, N. Tommerup, and C. Sismani, "Accurate breakpoint mapping in apparently balanced translocation families with discordant phenotypes using whole genome mate-pair sequencing," *PLoS One*, vol. 12, no. 1, 2017.

[316] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The Database of Genomic Variants: A curated collection of structural variation in the human genome," *Nucleic Acids Res.*, vol. 42, no. D1, 2014.

[317] A. Sirmaci, M. Spiliopoulos, F. Brancati, E. Powell, D. Duman, A. Abrams, G. Bademci, E. Agolini, S. Guo, B. Konuk, A. Kavaz, S. Blanton, M. C. Digilio, B. Dallapiccola, J. Young, S. Zuchner, and M. Tekin, "Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia," 2011.

[318] M. J. McMillin, J. E. Below, K. M. Shively, A. E. Beck, H. I. Gildersleeve, J. Pinner, G. R. Gogola, J. T. Hecht, D. K. Grange, D. J. Harris, D. L. Earl, S. Jagadeesh, S. G. Mehta, S. P. Robertson, J. M. Swanson, E. M. Faustman, H. C. Mefford, J. Shendure, D. A. Nickerson, and M. J. Bamshad, "Mutations in ECEL1 cause distal arthrogryposis type 5D," *Am. J. Hum. Genet.*, vol. 92, no. 1, pp. 150–156, 2013.

[319] K. Dieterich, S. Quijano-Roy, N. Monnier, J. Zhou, J. Fauré, D. A. Smirnow, R. Carlier, C. Laroche, P. Marcorelles, S. Mercier, A. Mégarbané, S. Odent, N. Romero, D. Sternberg, I. Marty, B. Estournet, P. S. Jouk, J. Melki, and J. Lunardi, "The neuronal endopeptidase ECEL1 is associated with a distinct form of recessive distal arthrogryposis," *Hum. Mol. Genet.*, vol. 22, no. 8, pp. 1483–1492, 2013.

[320] R. Shaheen, M. Al-Owain, A. O. Khan, M. S. Zaki, H. A. A. Hossni, R. Al-Tassan, W. Eyaid, and F. S. Alkuraya, "Identification of three novel ECEL1 mutations in three families with distal arthrogryposis type 5D," *Clin. Genet.*, 2014.

[321] S. J. Patil, G. K. Rai, V. Bhat, V. A. Ramesh, H. A. Nagarajaram, J. Matalia, and S. R. Phadke, "Distal arthrogryposis type 5D with a novel ECEL1 gene mutation," *Am. J. Med. Genet. Part A*, 2014.

[322] C. P. Barnett, E. J. Todd, R. Ong, M. R. Davis, V. Atkinson, R. Allcock, N. Laing, and G. Ravenscroft, "Distal arthrogryposis type 5D with novel clinical features and compound heterozygous mutations in ECEL1," *Am. J. Med. Genet. Part A*, 2014.

[323] S. Shaaban, F. Duzcan, C. Yildirim, W. M. Chan, C. Andrews, N. A. Akarsu, and E. C. Engle, "Expanding the phenotypic spectrum of ECEL1-related congenital contracture syndromes," *Clin. Genet.*, 2014.

[324] A. R. Hamzeh, P. Nair, M. Mohamed, F. Saif, N. Tawfiq, M. Khalifa, M. T. Al-Ali, and F. Bastaki, "A Novel Variant in the Endothelin-Converting Enzyme-Like 1 (ECEL1) Gene in an Emirati Child," *Med. Princ. Pract.*, 2017.

[325] E. L. Stattin, J. Johansson, S. Gudmundsson, A. Ameur, S. Lundberg, M. L. Bondeson, and M. Wilbe, "A novel ECEL1 mutation expands the phenotype of distal arthrogryposis multiplex congenita type 5D to include pretibial vertical skin creases," *Am. J. Med. Genet. Part A*, 2018.

[326] A. Rai, R. D. Puri, and S. R. Phadke, "Extending the phenotype and an ECEL1 gene mutation in distal arthrogryposis type 5D," *Clin. Dysmorphol.*, vol. 27, no. 4, 2018.

[327] U. Ullmann, L. D'Argenzio, S. Mathur, T. Whyte, R. Quinlivan, C. Longman, M. E. Farrugia, A. Manzur, T. Willis, H. Jungbluth, M. Pitt, S. Cirak, L. Feng, W. Stewart, R. Mein, R. Phadke, C. Sewry, A. Sarkozy, and F. Muntoni, "ECEL1 gene related contractural syndrome: Long-term follow-up and update on clinical and pathological aspects," sep 2018.

[328] C. W. Ockeloen, M. H. Willemsen, S. De Munnik, B. W. Van Bon, N. De Leeuw, A. Verrips, S. G. Kant, E. A. Jones, H. G. Brunner, R. L. Van Loon, E. E. Smeets, M. M. Van Haelst, G. Van Haaften, A. Nordgren, H. Malmgren, G. Grigelioniene, S. Vermeer, P. Louro, L. Ramos, T. J. Maal, C. C. Van Heumen, H. G. Yntema, C. E. Carels, and T. Kleefstra, "Further delineation of the KBG syndrome phenotype caused by ANKRD11 aberrations," *Eur. J. Hum. Genet.*, 2015.

[329] K. Walz, D. Cohen, P. M. Neilsen, J. Foster, F. Brancati, K. Demir, R. Fisher, M. Moffat, N. E. Verbeek, K. Bjørgo, A. Lo Castro, P. Curatolo, G. Novelli, C. Abad, C. Lei, L. Zhang, O. Diaz-Horta, J. I. Young, D. F. Callen, and M. Tekin, "Characterization of ANKRD11 mutations in humans and mice related to KBG syndrome," *Hum. Genet.*, 2015.

[330] K. Low, T. Ashraf, N. Canham, J. Clayton-Smith, C. Deshpande, A. Donaldson, R. Fisher, F. Flinter, N. Foulds, A. Fryer, K. Gibson, I. Hayes, A. Hills, S. Holder, M. Irving, S. Joss, E. Kivuva, K. Lachlan, A. Magee, V. McConnell, M. McEntagart, K. Metcalfe, T. Montgomery, R. Newbury-Ecob, F. Stewart, P. Turnpenny, J. Vogt, D. Fitzpatrick, M. Williams, and S. Smithson, "Clinical and genetic aspects of KBG syndrome," *Am. J. Med. Genet. Part A*, 2016.

[331] A. Goldenberg, F. Riccardi, A. Tessier, R. Pfundt, T. Busa, P. Cacciagli, Y. Capri, C. Coutton, A. Delahaye-Duriez, T. Frebourg, V. Gatinois, A. M. Guerrot, D. Genevieve, F. Lecoquierre, A. Jacquette, P. Khau Van Kien, B. Leheup, S. Marlin, A. Verloes, V. Michaud, G. Nadeau, C. Mignot, P. Parent, M. Rossi, A. Toutain, E. Schaefer, C. Thauvin-Robinet, L. Van Maldergem, J. Thevenon, V. Satre, L. Perrin, C. Vincent-Delorme, A. Sorlin, C. Missirian, L. Villard, J. Mancini, P. Saugier-Veber, and N. Philip, "Clinical and molecular findings in 39 patients with KBG syndrome caused by deletion or mutation of ANKRD11," *Am. J. Med. Genet. Part A*, 2016.

[332] M. L. De Bernardi, I. Ivanovski, S. G. Caraffi, I. Maini, M. E. Street, A. Bayat, M. Zollino, F. R. Lepri, M. Gnazzo, E. Errichiello, A. Superti-Furga, and L. Garavelli, "Prominent and elongated coccyx, a new manifestation of KBG syndrome associated with novel mutation in ANKRD11," *Am. J. Med. Genet. Part A*, vol. 176, pp. 1991–1995, aug 2018.

[333] S. Miyatake, N. Okamoto, Z. Stark, M. Nabetani, Y. Tsurusaki, M. Nakashima, N. Miyake, T. Mizuguchi, A. Ohtake, H. Saitsu, and N. Matsumoto, "ANKRD11 variants cause variable clinical features associated with KBG syndrome and Coffin-Siris-like syndrome," *J. Hum. Genet.*, 2017.

[334] C. Li, Q. Liu, N. Li, W. Chen, L. Wang, Y. Wang, Y. Yu, and X. Cao, "EAPF/Phafin-2, a novel endoplasmic reticulum-associated protein, facilitates TNF-α-triggered cellular apoptosis through endoplasmic reticulum-mitochondrial apoptotic pathway," *J. Mol. Med.*, vol. 86, no. 4, pp. 471–484, 2008.

[335] M. Esteller, J. Garcia-Foncillas, E. Andion, S. N. Goodman, O. F. Hidalgo, V. Vanaclocha, S. B. Baylin, and J. G. Herman, "Inactivation of the DNA-Repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents," *N. Engl. J. Med.*, vol. 343, pp. 1350–1354, nov 2000.

[336] S. L. Gerson, "MGMT: its role in cancer aetiology and cancer therapeutics," *Nat. Rev. Cancer*, vol. 4, no. 4, pp. 296–307, 2004.

[337] M. Reza Jabalameli, I. Briceno, J. Martinez, I. Briceno, R. J. Pengelly, S. Ennis, and A. Collins, "Aarskog-Scott syndrome: phenotypic and genetic heterogeneity," *AIMS Genet.*, vol. 3, no. 1, pp. 49–59, 2016.

[338] A. C. Lidral, L. M. Moreno, and S. A. Bullard, "Genetic Factors and Orofacial Clefting," *Semin. Orthod.*, 2008.

[339] M. J. Dixon, M. L. Marazita, T. H. Beaty, and J. C. Murray, "Cleft lip and palate: Understanding genetic and environmental influences," 2011.

[340] L. Arias Uruena, I. Briceno Balcazar, J. Martinez Lozano, A. Collins, and D. A. Uricoechea Patino, "Clinical Aspects associated with Syndromic forms of Orofacial Clefts in a Colombian population.," *Colomb. medica (Cali, Colomb.*, 2015.

[341] D. Aarskog, "A familial syndrome of short stature associated with facial dysplasia and genital anomalies," *J. Pediatr.*, vol. 77, pp. 856–861, jun 1970.

[342] C. I. Scott, "Unusual facies, joint hypermobility, genital anomaly and short stature: a new dysmorphic syndrome.," *Birth Defects Orig. Artic. Ser.*, vol. 7, no. 6, pp. 240–246, 1971.

[343] A. Orrico, L. Galli, M. L. Cavaliere, L. Garavelli, J.-P. Fryns, E. Crushell, M. M. Rinaldi, A. Medeira, and V. Sorrentino, "Phenotypic and molecular characterisation of the Aarskog-Scott syndrome: a survey of the clinical variability in light of FGD1 mutation analysis in 46 patients.," *Eur. J. Hum. Genet.*, vol. 12, no. 1, pp. 16–23, 2004.

[344] L. Estrada, E. Caron, and J. L. Gorski, "Fgd1, the Cdc42 guanine nucleotide exchange factor responsible for faciogenital dysplasia, is localized to the subcortical actin cytoskeleton and Golgi membrane.," *Hum. Mol. Genet.*, vol. 10, no. 5, pp. 485–95, 2001.

[345] A. Orrico, L. Galli, J. Clayton-Smith, and J.-P. Fryns, "Clinical utility gene card for: Aarskog–Scott Syndrome (faciogenital dysplasia) – update 2015," *Eur. J. Hum. Genet.*, vol. 23, apr 2015.

[346] A. Orrico, L. Galli, J. Clayton-Smith, and J.-P. Fryns, "Clinical utility gene card for: Aarskog–Scott syndrome (faciogenital dysplasia)," *Eur. J. Hum. Genet.*, vol. 19, nov 2011.

[347] A. Orrico, L. Galli, L. Faivre, J. Clayton-Smith, S. M. Azzarello-Burri, J. M. Hertz, S. Jacquemont, R. Taurisano, I. Arroyo Carrera, E. Tarantino, K. Devriendt, D. Melis, T. Thelle, U. Meinhardt, and V. Sorrentino, "Aarskog-Scott syndrome: Clinical update and report of nine novel mutations of the FGD1 gene," *Am. J. Med. Genet. Part A*, vol. 152, no. 2, pp. 313–318, 2010.

[348] A. S. Teebi, J. K. Rucquoi, and M. S. Meyn, "Aarskog syndrome: report of a family with review and discussion of nosology.," *Am. J. Med. Genet.*, vol. 46, no. 5, pp. 501–509, 1993.

[349] M. Pérez-Coria, J. J. Lugo-Trampe, M. Zamudio-Osuna, I. P. Rodríguez-Sánchez, A. Lugo-Trampe, B. de la Fuente-Cortez, L. D. Campos-Acevedo, and L. E. Martínez-de Villarreal, "Identification of novel mutations in Mexican patients with Aarskog–Scott syndrome," *Mol. Genet. Genomic Med.*, vol. 3, no. 3, pp. 197–202, 2015.

[350] L. Logie and M. Porteous, "Intelligence and development in Aarskog syndrome," *Arch. Dis. Child.*, vol. 79, pp. 359–360, oct 1998.

[351] A. Orrico, L. Galli, S. Buoni, G. Hayek, A. Luchetti, S. Lorenzini, M. Zappella, M. G. Pomponi, and V. Sorrentino, "Attention-deficit/hyperactivity disorder (ADHD) and variable clinical expression of Aarskog–Scott syndrome due to a novel FGD1 gene mutation (R408Q)," *Am. J. Med. Genet. Part A*, vol. 135A, no. 1, pp. 99–102, 2005.

[352] R. B. Nayak, L. Ambika, G. S. Bhogale, and a. Pandurangi, "Mania with Aarskog-Scott syndrome.," *Indian Pediatr.*, vol. 49, no. 4, pp. 327–8, 2012.

[353] A. S. Teebi, K. K. Naguib, S. Al-Awadi, and Q. A. Al-Saleh, "New autosomal recessive faciodigitogenital syndrome.," *J. Med. Genet.*, vol. 25, pp. 400–406, jun 1988.

[354] A. E. Roberts, J. E. Allanson, M. Tartaglia, and B. D. Gelb, "Noonan syndrome," *Lancet*, vol. 381, no. 9863, pp. 333–342, 2013.

[355] A. Al Kaissi, T. Bieganski, D. Baranska, F. B. Chehida, H. Gharbi, M. B. Ghachem, L. Hendaoui, H. Safi, and K. Kozlowski, "Robinow syndrome: Report of two cases and review of the literature," 2007.

[356] P. Smpokou, D. J. Zand, K. N. Rosenbaum, and M. L. Summar, "Malignancy in Noonan syndrome and related disorders," 2015.

[357] J. P. Fryns, "Aarskog syndrome: The changing phenotype with age," *Am. J. Med. Genet.*, vol. 43, no. 1-2, pp. 420–427, 1992.

[358] R. Galupa and E. Heard, "X-chromosome inactivation: New insights into cis and trans regulation," 2015.

[359] A. Jogiya and C. Sandy, "Mild optic nerve hypoplasia with retinal venous tortuosity in aarskog (facial-digital-genital) syndrome," *Ophthalmic Genet*, vol. 26, no. 3, pp. 139–141, 2005.

[360] R. J. Andrassy, S. Murthy, and M. M. Woolley, "Aarskog syndrome: Significance for the surgeon," *J. Pediatr. Surg.*, vol. 14, no. 4, pp. 462–464, 1979.

[361] R. V. Mikelsaar and I. W. Lurie, "Atypical case of Aarskog syndrome.," *J. Med. Genet.*, vol. 29, pp. 349–350, may 1992.

[362] J. P. Fryns and H. Van den Berghe, "On the occurrence of macroorchidism and mental handicap in the Aarskog syndrome.," 1989.

[363] a. S. Teebi, K. K. Naguib, S. Al-Awadi, and Q. a. Al-Saleh, "New autosomal recessive faciodigitogenital syndrome.," *J. Med. Genet.*, vol. 25, no. 6, pp. 400–406, 1988.

[364] A. S. Teebi and S. A. al Awadi, "Kuwait type faciodigitogenital syndrome.," *J. Med. Genet.*, vol. 28, p. 805, nov 1991.

[365] E. Bawle, M. Tyrkus, S. Lipman, D. Bozimowski, and J. M. Opitz, "Aarskog syndrome: Full male and female expression associated with an X-autosome translocation," *Am. J. Med. Genet.*, vol. 17, no. 3, pp. 595–602, 1984.

[366] C. Redin, S. Le Gras, O. Mhamdi, V. Geoffroy, C. Stoetzel, M.-C. Vincent, P. Chiurazzi, D. Lacombe, I. Ouertani, F. Petit, M. Till, A. Verloes, B. Jost, H. B. Chaabouni, H. Dollfus, J.-L. Mandel, and J. Muller, "Targeted high-throughput sequencing for diagnosis of genetically heterogeneous diseases: efficient mutation detection in Bardet-Biedl and Alström Syndromes," *J. Med. Genet.*, vol. 49, no. 8, pp. 502–512, 2012.

[367] A. Jugessur, F. Rahimov, R. T. Lie, A. J. Wilcox, H. K. Gjessing, R. M. Nilsen, T. T. Nguyen, and J. C. Murray, "Genetic variants in IRF6 and the risk of facial clefts: Single-marker and haplotype-based analyses in a population-based case-control study of facial clefts in Norway," *Genet. Epidemiol.*, 2008.

[368] T. M. Zucchero, M. E. Cooper, B. S. Maher, S. Daack-Hirsch, B. Nepomuceno, L. Ribeiro, D. Caprau, K. Christensen, Y. Suzuki, J. Machida, N. Natsume, K.-I. Yoshiura, A. R. Vieira, I. M. Orioli, E. E. Castilla, L. Moreno, M. Arcos-Burgos, A. C. Lidral, L. L. Field, Y.-e. Liu, A. Ray, T. H. Goldstein, R. E. Schultz, M. Shi,

M. K. Johnson, S. Kondo, B. C. Schutte, M. L. Marazita, and J. C. Murray, "Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate.," *N. Engl. J. Med.*, 2004.

[369] A. D. Person, S. Beiraghi, C. M. Sieben, S. Hermanson, A. N. Neumann, M. E. Robu, J. R. Schleiffarth, C. J. Billington, H. Van Bokhoven, J. M. Hoogeboom, J. F. Mazzeu, A. Petryk, L. A. Schimmenti, H. G. Brunner, S. C. Ekker, and J. L. Lohr, "WNT5A mutations in patients with autosomal dominant Robinow syndrome," *Dev. Dyn.*, vol. 239, no. 1, pp. 327–337, 2010.

[370] R. A. Pagon and P. R.A., "GeneTests: An online genetic information resource for health care providers," 2006.

[371] Illumina, "TruSight^{TM} One Sequencing Panel," 2015.

[372] R. K. Dale, B. S. Pedersen, and A. R. Quinlan, "Pybedtools: A flexible Python library for manipulating genomic datasets and annotations," *Bioinformatics*, vol. 27, no. 24, pp. 3423–3424, 2011.

[373] M. W. . O. B. . P. H. . J. B. C. . Y. H. . S. H. . A. M. . T. A. . T. Y. . T. M. . L. P. C. . D. W. . cynddl ; Erik Ziegler ; diego0020 ; Yury V. Zaytsev ; T and Michael Waskom ; Olga Botvinnik ; Paul Hobson ; John B. Cole ; Yaroslav Halchenko ; Stephan Hoyer ; Alistair Miles ; Tom Augspurger ; Tal Yarkoni ; Tobias Megies ; Luis Pedro Coelho ; Daniel Wehner ; cynddl ; Erik Ziegler ; diego0020 ; Yury V. Zaytsev ; T, "seaborn: v0.5.0," *zenodo*, 2014.

[374] F. Hahne and R. Ivanek, "Visualizing genomic data using Gviz and bioconductor," in *Methods Mol. Biol.*, vol. 1418, pp. 335–351, 2016.

[375] K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano, "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity.," *Nat. Genet.*, vol. 48, no. 12, pp. 1581–1586, 2016.

[376] C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, and X. Liu, "Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies," *Hum. Mol. Genet.*, vol. 24, no. 8, pp. 2125–2137, 2015.

[377] A. J. Masino, E. T. Dechene, M. C. Dulik, A. Wilkens, N. B. Spinner, I. D. Krantz, J. W. Pennington, P. N. Robinson, and P. S. White, "Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology," *BMC Bioinformatics*, vol. 15, no. 1, p. 248, 2014.

[378] W. McKinney and P. D. Team, "Pandas - Powerful Python Data Analysis Toolkit," *Pandas - Powerful Python Data Anal. Toolkit*, 2015.

[379] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma'ayan, "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins.," *Database (Oxford).*, vol. 2016, p. baw100, 2016.

[380] INSERM, "Orphanet: an online database of rare diseases and orphan drugs."

[381] C. F. Wright, T. W. Fitzgerald, W. D. Jones, S. Clayton, J. F. McRae, M. Van Kogelenberg, D. A. King, K. Ambridge, D. M. Barrett, T. Bayzetinova, A. P. Bevan, E. Bragin, E. A. Chatzimichali, S. Gribble, P. Jones, N. Krishnappa, L. E. Mason, R. Miller, K. I. Morley, V. Parthiban, E. Prigmore, D. Rajan, A. Sifrim, G. J. Swaminathan, A. R. Tivey, A. Middleton, M. Parker, N. P. Carter, J. C. Barrett, M. E. Hurles, D. R. Fitzpatrick, and H. V. Firth, "Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data," *Lancet*, vol. 385, no. 9975, pp. 1305–1314, 2015.

[382] M. Simbolo, M. Gottardi, V. Corbo, M. Fassan, A. Mafficini, G. Malpeli, R. T. Lawlor, and A. Scarpa, "DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples," *PLoS One*, vol. 8, no. 6, 2013.

[383] R. L. Hood, M. A. Lines, S. M. Nikkel, J. Schwartzentruber, C. Beaulieu, M. J. Nowaczyk, J. Allanson, C. A. Kim, D. Wieczorek, J. S. Moilanen, D. Lacombe, G. Gillessen-Kaesbach, M. L. Whiteford, C. R. D. Quaio, I. Gomy, D. R. Bertola, B. Albrecht, K. Platzer, G. McGillivray, R. Zou, D. R. McLeod, A. E. Chudley, B. N. Chodirker, J. Marcadier, J. Majewski, D. E. Bulman, S. M. White, and K. M. Boycott, "Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome," *Am. J. Hum. Genet.*, vol. 90, no. 2, pp. 308–313, 2012.

[384] M. A. Monroy, D. D. Ruhl, X. Xu, D. K. Granner, P. Yaciuk, and J. C. Chrivia, "Regulation of cAMP-responsive element-binding protein-mediated transcription by the SNF2/SWI-related protein, SRCAP.," *J. Biol. Chem.*, vol. 276, pp. 40721–6, nov 2001.

[385] K. Nagasaki, T. Asami, H. Sato, Y. Ogawa, T. Kikuchi, A. Saitoh, T. Ogata, and M. Fukami, "Long-term follow-up study for a patient with Floating-Harbor syndrome due to a hotspot SRCAP mutation," *Am. J. Med. Genet. Part A*, vol. 164, no. 3, pp. 731–735, 2014.

[386] M. P. Adam and L. Hudgins, "Kabuki syndrome: A review," 2005.

[387] D. Well, S. Blanchard, J. Kaplan, P. Guilford, F. Gibson, J. Walsh, P. Mburu, A. Varela, J. Levilliers, M. Weston, D., P. Kelley, M., W. Kimberling, J., M. Wagenaar, F. Levi-Acobas, D. Larget-Piet, A. Munnich, K. Steel, P., S. D. Brown, M., and C. Petit, "Defective myosin VIIA gene responsible for Usher syndrome type IB," *Nature*, vol. 374, no. 6517, pp. 60–61, 1995.

[388] Y. Sun, J. Chen, H. Sun, J. Cheng, J. Li, Y. Lu, Y. Lu, Z. Jin, Y. Zhu, X. Ouyang, D. Yan, P. Dai, D. Han, W. Yang, R. Wang, X. Liu, and H. Yuan, "Novel missense mutations in MYO7A underlying postlingual high- or low-frequency non-syndromic hearing impairment in two large families from China.," *J. Hum. Genet.*, vol. 56, no. 1, pp. 64–70, 2011.

[389] J. W. Witherspoon and K. G. Meilleur, "Review of RyR1 pathway and associated pathomechanisms.," *Acta Neuropathol. Commun.*, vol. 4, no. 1, p. 121, 2016.

[390] D. X. Bharucha-Goebel, M. Santi, L. Medne, K. Zukosky, J. Dastgir, P. B. Shieh, T. Winder, G. Tennekoon, R. S. Finkel, J. J. Dowling, N. Monnier, and C. G. Bonnemann, "Severe congenital RYRI-associated myopathy: The expanding clinicopathologic and genetic spectrum," *Neurology*, vol. 80, no. 17, pp. 1584–1589, 2013.

[391] H. Isaacs and M. E. Badenhorst, "Dominantly inherited malignant hyperthermia (MH) in the King-Denborough Syndrome," *Muscle and Nerve*, vol. 15, no. 6, pp. 740–742, 1992.

[392] R. Robinson, D. Carpenter, M. A. Shaw, J. Halsall, and P. Hopkins, "Mutations in RYR1 in malignant hypertheraiia and central core disease," 2006.

[393] L. C. Cantley, "The phosphoinositide 3-kinase pathway.," *Science*, vol. 296, no. 5573, pp. 1655–7, 2002.

[394] C. Bárcena, V. Quesada, A. De Sandre-Giovannoli, D. a. Puente, J. Fernández-Toral, S. Sigaudy, A. Baban, N. Lévy, G. Velasco, and C. López-Otín, "Exome sequencing identifies a novel mutation in PIK3R1 as the cause of SHORT syndrome.," *BMC Med. Genet.*, vol. 15, no. 1, p. 51, 2014.

[395] D. A. Dyment, A. C. Smith, D. Alcantara, J. A. Schwartzentruber, L. Basel-Vanagaite, C. J. Curry, I. K. Temple, W. Reardon, S. Mansour, M. R. Haq, R. Gilbert, O. J. Lehmann, M. R. Vanstone, C. L. Beaulieu, J. Majewski, D. E. Bulman, M. O'Driscoll, K. M. Boycott, and A. M. Innes, "Mutations in PIK3R1 cause SHORT syndrome," *Am. J. Hum. Genet.*, vol. 93, no. 1, pp. 158–166, 2013.

[396] X. Liu, L. Wang, K. Zhao, P. R. Thompson, Y. Hwang, R. Marmorstein, and P. a. Cole, "The structural basis of protein acetylation by the p300/CBP transcriptional coactivator.," *Nature*, vol. 451, no. 7180, pp. 846–850, 2008.

[397] A. S. Turnell, G. S. Stewart, R. J. a. Grand, S. M. Rookes, A. Martin, H. Yamano, S. J. Elledge, and P. H. Gallimore, "The APC/C and CBP/p300 cooperate to regulate transcription and cell-cycle progression.," *Nature*, vol. 438, no. 7068, pp. 690–695, 2005.

[398] R. C. M. Hennekam, "Rubinstein-Taybi syndrome.," *Eur. J. Hum. Genet.*, vol. 14, no. 9, pp. 981–985, 2006.

[399] R. C. Hennekam and J. M. Van Doorne, "Oral aspects of Rubinstein-Taybi syndrome.," *Am. J. Med. Genet. Suppl.*, vol. 6, pp. 42–47, 1990.

[400] M. I. Ferrante, A. Zullo, A. Barra, S. Bimonte, N. Messaddeq, M. Studer, P. Dollé, and B. Franco, "Oral-facial-digital type I protein is required for primary cilia formation and left-right axis specification.," *Nat. Genet.*, vol. 38, no. 1, pp. 112–7, 2006.

[401] M. L. de Luna, M. L. Raspa, and J. Ibargoyen, "Oral-Facial-Digital Type 1 Syndrome of Papillon-Léage and Psaume," *Pediatr. Dermatol.*, vol. 9, pp. 52–56, mar 1992.

[402] M. I. Ferrante, G. Giorgio, S. A. Feather, A. Bulfone, V. Wright, M. Ghiani, A. Selicorni, L. Gammaro, F. Scolari, A. S. Woolf, O. Sylvie, L. Bernard, S. Malcolm, R. Winter, A. Ballabio, and B. Franco, "Identification of the gene for oral-facial-digital type I syndrome," *Am J Hum Genet*, vol. 68, no. 3, pp. 569–576, 2001.

[403] Y. Tsurusaki, T. Kosho, K. Hatasaki, Y. Narumi, K. Wakui, Y. Fukushima, H. Doi, H. Saitsu, N. Miyake, and N. Matsumoto, "Exome sequencing in a family with an X-linked lethal malformation syndrome: Clinical consequences of hemizygous truncating OFD1 mutations in male patients," *Clin. Genet.*, vol. 83, no. 2, pp. 135–144, 2013.

[404] M. Tétreault, M. Gonzalez, M. J. Dicaire, P. Allard, K. Gehring, D. Leblanc, N. Leclerc, R. Schondorf, J. Mathieu, S. Zuchner, and B. Brais, "Adult-onset painful axonal polyneuropathy caused by a dominant NAGLU mutation," *Brain*, vol. 138, no. 6, pp. 1477–1483, 2015.

[405] C. Basso, D. Corrado, B. Bauce, and G. Thiene, "Arrhythmogenic right ventricular cardiomyopathy," *Circ. Arrhythmia Electrophysiol.*, vol. 5, no. 6, pp. 1233–1246, 2012.

[406] A. M. Bertoli-Avella, E. Gillis, H. Morisaki, J. M. A. Verhagen, B. M. de Graaf, G. van de Beek, E. Gallo, B. P. T. Kruithof, H. Venselaar, L. A. Myers, S. Laga, A. J. Doyle, G. Oswald, G. W. A. van Cappellen, I. Yamanaka, R. M. van der Helm, B. Beverloo, A. de Klein, L. Pardo, M. Lammens, C. Evers, K. Devriendt, M. Dumoulein, J. Timmermans, H. T. Bruggenwirth, F. Verheijen, I. Rodrigus, G. Baynam, M. Kempers, J. Saenen, E. M. Van Craenenbroeck, K. Minatoya, R. Matsukawa, T. Tsukube, N. Kubo, R. Hofstra, M. J. ose Goumans, J. A. Bekkers, J. W. Roos-Hesselink, I. M. B. H. van de Laar, H. C. Dietz, L. Van Laer, T. Morisaki, M. W. Wessels, and B. L. Loeys, "Mutations in a TGF-β ligand, TGFB3, cause syndromic aortic aneurysms and dissections," *J. Am. Coll. Cardiol.*, vol. 65, no. 13, pp. 1324–1336, 2015.

[407] E. T. Tonkin, T.-J. Wang, S. Lisgo, M. J. Bamshad, and T. Strachan, "NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome.," *Nat. Genet.*, vol. 36, no. 6, pp. 636–641, 2004.

[408] M. I. Boyle, C. Jespersgaard, K. Brøndum-Nielsen, A.-M. Bisgaard, and Z. Tümer, "Cornelia de Lange Syndrome.," *Clin. Genet.*, no. March, pp. 38–41, 2014.

[409] S. Rohatgi, D. Clark, A. D. Kline, L. G. Jackson, J. Pie, V. Siu, F. J. Ramos, I. D. Krantz, and M. A. Deardorff, "Facial diagnosis of mild and variant CdLS: Insights from a dysmorphologist survey," *Am. J. Med. Genet. Part A*, vol. 152A, pp. 1641–1653, jul 2010.

[410] A. D. Kline, M. Grados, P. Sponseller, H. P. Levy, N. Blagowidow, C. Schoedel, J. Rampolla, D. K. Clemens, I. Krantz, A. Kimball, C. Pichard, and D. Tuchman, "Natural history of aging in Cornelia de Lange syndrome," *Am. J. Med. Genet. Part C Semin. Med. Genet.*, vol. 145, no. 3, pp. 248–260, 2007.

[411] L. Jackson, A. D. Kline, M. A. Barr, and S. Koch, "de Lange syndrome: A clinical review of 310 individuals," in *Am. J. Med. Genet.*, vol. 47, pp. 940–946, 1993.

[412] A. D. Kline, I. D. Krantz, A. Sommer, M. Kliewer, L. G. Jackson, D. R. FitzPatrick, A. V. Levin, and A. Selicorni, "Cornelia de Lange syndrome: Clinical review, diagnostic and scoring systems, and anticipatory guidance," 2007.

[413] F. Lalatta, S. Russo, B. Gentilin, L. Spaccini, C. Boschetto, F. Cavalleri, M. Masciadri, C. Gervasini, A. Bentivegna, P. Castronovo, and L. Larizza, "Prenatal/neonatal pathology in two cases of Cornelia de Lange syndrome harboring novel mutations of NIPBL," *Genet. Med.*, vol. 9, no. 3, pp. 188–194, 2007.

[414] J. E. Allanson, R. C. Hennekam, and M. Ireland, "De Lange syndrome: subjective and objective comparison of the classical and mild phenotypes.," *J. Med. Genet.*, vol. 34, no. 8, pp. 645–50, 1997.

[415] L. A. Gillis, J. McCallum, M. Kaur, C. DeScipio, D. Yaeger, A. Mariani, A. D. Kline, H.-h. Li, M. Devoto, L. G. Jackson, and I. D. Krantz, "NIPBL mutational analysis in 120 individuals with Cornelia de Lange syndrome and evaluation of genotype-phenotype correlations.," *Am. J. Hum. Genet.*, vol. 75, no. 4, pp. 610–623, 2004.

[416] J. Oliveira, C. Dias, E. Redeker, E. Costa, J. Silva, M. Reis Lima, J. T. Den Dunnen, and R. Santos, "Development of NIPBL Locus-Specific Database Using LOVD: From Novel Mutations to Further Genotype-Phenotype Correlations in Cornelia de Lange Syndrome," *Hum. Mutat.*, vol. 31, no. 11, pp. 1216–1222, 2010.

[417] Z. a. Bhuiyan, M. Klein, P. Hammond, a. van Haeringen, M. M. a. M. Mannens, I. Van Berckelaer-Onnes, and R. C. M. Hennekam, "Genotype-phenotype correlations of 39 patients with Cornelia De Lange syndrome: the Dutch experience.," *J. Med. Genet.*, vol. 43, no. 7, pp. 568–575, 2006.

[418] M. E. Teresa-Rodrigo, J. Eckhold, B. Puisac, A. Dalski, M. C. Gil-Rodríguez, D. Braunholz, C. Baquero, M. Hernández-Marcos, J. C. de Karam, M. Ciero, F. Santos-Simarro, P. Lapunzina, J. Wierzba, C. H. Casale, F. J. Ramos, G. Gillessen-Kaesbach, F. J. Kaiser, and J. Pié, "Functional characterization of NIPBL physiological splice variants and eight splicing mutations in patients with cornelia de lange syndrome," *Int. J. Mol. Sci.*, vol. 15, no. 6, pp. 10350–10364, 2014.

[419] K. Bhalla, Y. Luo, T. Buchan, M. A. Beachem, G. F. Guzauskas, S. Ladd, S. J. Bratcher, R. J. Schroer, J. Balsamo, B. R. DuPont, J. Lilien, and A. K. Srivastava, "Alterations in CDH15 and KIRREL3 in Patients with Mild to Severe Intellectual Disability," *Am. J. Hum. Genet.*, vol. 83, no. 6, pp. 703–713, 2008.

[420] G. Popescu, A. Robert, J. R. Howe, and A. Auerbach, "Reaction mechanism determines NMDA receptor response to repetitive stimulation.," *Nature*, vol. 430, no. 7001, pp. 790–793, 2004.

[421] S. Endele, G. Rosenberger, K. Geider, B. Popp, C. Tamer, I. Stefanova, M. Milh, F. Kortum, A. Fritsch, F. K. Pientka, Y. Hellenbroich, V. M. Kalscheuer, J. Kohlhase, U. Moog, G. Rappold, A. Rauch, H. H. Ropers, S. von Spiczak, H. Tonnies, N. Villeneuve, L. Villard, B. Zabel, M. Zenker, B. Laube, A. Reis, D. Wieczorek, L. Van Maldergem, and K. Kutsche, "Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes," *Nat Genet*, vol. 42, no. 11, pp. 1021–1026, 2010.

[422] G. Lesca, G. Rudolf, N. Bruneau, N. Lozovaya, A. Labalme, N. Boutry-Kryza, M. Salmi, T. Tsintsadze, L. Addis, J. Motte, S. Wright, V. Tsintsadze, A. Michel, D. Doummar, K. Lascelles, L. Strug, P. Waters, J. de Bellescize, P. Vrielynck, A. de Saint Martin, D. Ville, P. Ryvlin, A. Arzimanoglou, E. Hirsch, A. Vincent, D. Pal, N. Burnashev, D. Sanlaville, and P. Szepetowski, "GRIN2A mutations in acquired epileptic aphasia and related childhood focal epilepsies and encephalopathies with speech and language dysfunction.," *Nat. Genet.*, vol. 45, no. 9, pp. 1061–6, 2013.

[423] J. R. Lemke, D. Lal, E. M. Reinthaler, I. Steiner, M. Nothnagel, M. Alber, K. Geider, B. Laube, M. Schwake, K. Finsterwalder, A. Franke, M. Schilhabel, J. A. Jähn, H. Muhle, R. Boor, W. Van Paesschen, R. Caraballo, N. Fejerman, S. Weckhuysen, P. De Jonghe, J. Larsen, R. S. Møller, H. Hjalgrim, L. Addis, S. Tang, E. Hughes, D. K. Pal, K. Veri, U. Vaher, T. Talvik, P. Dimova, R. Guerrero López, J. M. Serratosa, T. Linnankivi, A.-E. Lehesjoki, S. Ruf, M. Wolff, S. Buerki, G. Wohlrab,

J. Kroell, A. N. Datta, B. Fiedler, G. Kurlemann, G. Kluger, A. Hahn, D. E. Haberlandt, C. Kutzer, J. Sperner, F. Becker, Y. G. Weber, M. Feucht, H. Steinböck, B. Neophythou, G. M. Ronen, U. Gruber-Sedlmayr, J. Geldner, R. J. Harvey, P. Hoffmann, S. Herms, J. Altmüller, M. R. Toliat, H. Thiele, P. Nürnberg, C. Wilhelm, U. Stephani, I. Helbig, H. Lerche, F. Zimprich, B. A. Neubauer, S. Biskup, and S. von Spiczak, "Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes.," *Nat. Genet.*, vol. 45, no. 9, pp. 1067–72, 2013.

[424] G. L. Carvill, B. M. Regan, S. C. Yendle, B. J. O'Roak, N. Lozovaya, N. Bruneau, N. Burnashev, A. Khan, J. Cook, E. Geraghty, L. G. Sadleir, S. J. Turner, M.-H. Tsai, R. Webster, R. Ouvrier, J. A. Damiano, S. F. Berkovic, J. Shendure, M. S. Hildebrand, P. Szepetowski, I. E. Scheffer, and H. C. Mefford, "GRIN2A mutations cause epilepsy-aphasia spectrum disorders.," *Nat. Genet.*, vol. 45, no. 9, pp. 1073–1076, 2013.

[425] J.-P. . Legius, E ; Schrander-Stumpel, C ; Schollen, E ; Pulles-Heintzberger, C ; Gewellig, M ; Fryns, "PTPN11 mutations in LEOPARD syndrome," *J. Med. Genet.*, vol. 39, no. 8, pp. 571–574, 2002.

[426] M. Tartaglia, E. L. Mehler, R. Goldberg, G. Zampino, H. G. Brunner, H. Kremer, I. van der Burgt, a. H. Crosby, A. Ion, S. Jeffery, K. Kalidas, M. a. Patton, R. S. Kucherlapati, and B. D. Gelb, "Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome.," *Nat. Genet.*, vol. 29, no. 4, pp. 465–8, 2001.

[427] M. Tartaglia, C. M. Niemeyer, X. Song, J. Buechner, A. Jung, K. Hählen, H. Hasle, J. D. Licht, and B. D. Gelb, "Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia , myelodysplastic syndromes and acute myeloid leukemia," *Nat. Genet.*, vol. 34, no. 2, pp. 148–151, 2003.

[428] M. Maheshwari, J. Belmont, S. Fernbach, T. Ho, L. Molinari, I. Yakub, F. Yu, A. Combes, J. Towbin, W. J. Craigen, and R. Gibbs, "PTPN11 Mutations in Noonan syndrome type I: detection of recurrent mutations in exons 3 and 13," *Hum. Mutat.*, vol. 20, no. 4, pp. 298–304, 2002.

[429] C. Martin and Y. Zhang, "The diverse functions of histone lysine methylation.," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 11, pp. 838–49, 2005.

[430] E. Toska, H. U. Osmanbeyoglu, P. Castel, C. Chan, R. C. Hendrickson, M. Elkabets, M. N. Dickler, M. Scaltriti, C. S. Leslie, S. A. Armstrong, and J. Baselga, "PI3K pathway regulates ER-dependent transcription in breast cancer through the epigenetic regulator KMT2D," *Science (80-.).*, vol. 355, no. 6331, pp. 1324–1330, 2017.

[431] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K.-I. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, and J. Shendure, "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.," *Nat. Genet.*, vol. 42, no. 9, pp. 790–793, 2010.

[432] M. C. Hannibal, K. J. Buckingham, S. B. Ng, J. E. Ming, A. E. Beck, M. J. Mcmillin, H. I. Gildersleeve, A. W. Bigham, H. K. Tabor, H. C. Mefford, J. Cook, K. I. Yoshiura, T. Matsumoto, N. Matsumoto, N. Miyake, H. Tonoki, K. Naritomi,

T. Kaname, T. Nagai, H. Ohashi, K. Kurosawa, J. W. Hou, T. Ohta, D. Liang, A. Sudo, C. A. Morris, S. Banka, G. C. Black, J. Clayton-Smith, D. A. Nickerson, E. H. Zackai, T. H. Shaikh, D. Donnai, N. Niikawa, J. Shendure, and M. J. Bamshad, "Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome," *Am. J. Med. Genet. Part A*, vol. 155, no. 7, pp. 1511–1516, 2011.

[433] N. Matsumoto and N. Niikawa, "Kabuki make-up syndrome: a review.," *Am. J. Med. Genet. C. Semin. Med. Genet.*, vol. 117C VN -, no. 1, pp. 57–65, 2003.

[434] H. Kawame, M. C. Hannibal, L. Hudgins, and R. a. Pagon, "Phenotypic spectrum and management issues in Kabuki syndrome.," *J. Pediatr.*, vol. 134, no. 4, pp. 480–5, 1999.

[435] M. C. Digilio, B. Marino, A. Toscano, A. Giannotti, and B. Dallapiccola, "Congenital heart defects in Kabuki syndrome," *Am. J. Med. Genet.*, vol. 100, pp. 269–274, may 2001.

[436] S. M. Yuan, "Congenital heart defects in Kabuki syndrome," 2013.

[437] G. S. Kobayashi, L. Alvizi, D. Y. Sunaga, P. Francis-West, A. Kuta, B. V. P. Almada, S. G. Ferreira, L. C. de Andrade-Lima, D. F. Bueno, C. E. Raposo-Amaral, C. F. Menck, and M. R. Passos-Bueno, "Susceptibility to DNA Damage as a Molecular Mechanism for Non-Syndromic Cleft Lip and Palate," *PLoS One*, vol. 8, no. 6, 2013.

[438] S. S. Kalia, K. Adelman, S. J. Bale, W. K. Chung, C. Eng, J. P. Evans, G. E. Herman, S. B. Hufnagel, T. E. Klein, B. R. Korf, K. D. McKelvey, K. E. Ormond, C. S. Richards, C. N. Vlangos, M. Watson, C. L. Martin, and D. T. Miller, "Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics," *Genet. Med.*, vol. 19, p. 249, nov 2016.

[439] M. Ambrose and R. A. Gatti, "Pathogenesis of ataxia-telangiectasia: the next generation of ATM functions.," 2013.

[440] A. Tanaka, S. Weinel, N. Nagy, M. O'Driscoll, J. E. Lai-Cheong, C. L. Kulp-Shorten, A. Knable, G. Carpenter, S. A. Fisher, M. Hiragun, Y. Yanase, M. Hide, J. Callen, and J. A. McGrath, "Germline mutation in ATR in autosomal- dominant oropharyngeal cancer syndrome," 2012.

[441] J. Murthy, K. G. Seshadri, P. V. Ramanan, A. Rajamani, and A. Hussain, "A Case of Cleft Lip and Palate Associated with Seckel Syndrome," *Cleft Palate-Craniofacial J.*, vol. 41, no. 2, pp. 202–205, 2004.

[442] A. Rajamani, V. Kamat, J. Murthy, and S. A. Hussain, "Anesthesia for cleft lip surgery in a child with Seckel syndrome - A case report," *Paediatr. Anaesth.*, vol. 15, no. 4, pp. 338–341, 2005.

[443] Q. M. M. P. P. D. Napolitano R Maruotti GM, "Prenatal diagnosis of Seckel syndrome on 3-dimensional sonography and magnetic resonance imaging," *J. Ultrasound Med. (ISSN 0278-4297, 1550-9613)*, vol. 2, no. 3, pp. 369–374, 2009.

[444] Deciphering Developmental Disorders Study, "Prevalence and architecture of de novo mutations in developmental disorders," *Nature*, vol. 542, no. 7642, pp. 433–438, 2017.

[445] D. Zhi and R. Chen, "Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic Mendelian diseases by exome sequencing," *PLoS One*, vol. 7, no. 2, 2012.

[446] LOVD-team, "LOVD 3.0 FGD1 gene homepage," 2016.

[447] E. Aten, Y. Sun, R. Almomani, G. W. E. Santen, T. Messemaker, S. M. Maas, M. H. Breuning, and J. T. Den Dunnen, "Exome Sequencing Identifies A Branch Point Variant in Aarskog-Scott Syndrome," *Hum. Mutat.*, vol. 34, no. 3, pp. 430–434, 2013.

[448] I. Illumina, "TruSight One Sequencing Panel Kits Training Support," *Illumina Propr.*, no. January, 2016.

[449] C. Pommerenke, R. Geffers, B. Bunk, S. Eberth, and H. G. Drexler, "Enhanced whole exome sequencing by higher DNA insert lengths," *BMC Genomics*, pp. 1–5, 2016.

[450] M. K. Sakharkar, V. T. K. Chow, and P. Kangueane, "Distributions of exons and introns in the human genome.," *In Silico Biol.*, vol. 4, no. 4, pp. 387–93, 2004.

[451] M. E. K. Niemi, H. C. Martin, D. L. Rice, G. Gallone, S. Gordon, M. Kelemen, K. McAloney, J. McRae, E. J. Radford, S. Yu, J. Gecz, N. G. Martin, C. F. Wright, D. R. Fitzpatrick, H. V. Firth, M. E. Hurles, and J. C. Barrett, "Common genetic variants contribute to risk of rare severe neurodevelopmental disorders," *Nature*, 2018.

[452] J. H. Nadeau, "Modifier genes in mice and humans," 2001.

[453] D. Greene, S. Richardson, and E. Turro, "Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases," *Am. J. Hum. Genet.*, 2016.

[454] H. C. Martin, W. D. Jones, J. Stephenson, J. Handsaker, G. Gallone, J. F. McRae, E. Prigmore, P. Short, M. Niemi, J. Kaplanis, E. Radford, N. Akawi, M. Balasubramanian, J. Dean, R. Horton, A. Hulbert, D. S. Johnson, K. Johnson, D. Kumar, S. A. Lynch, S. G. Mehta, J. Morton, M. J. Parker, M. Splitt, P. D. Turnpenny, P. C. Vasudevan, M. Wright, C. F. Wright, D. R. FitzPatrick, H. V. Firth, M. E. Hurles, and J. C. Barrett, "Quantifying the contribution of recessive coding variation to developmental disorders," *bioRxiv*, 2017.

[455] Y. Bokinni, "Kabuki syndrome revisited," 2012.

[456] N. Bögershausen and B. Wollnik, "Unmasking Kabuki syndrome," 2013.

[457] E. Dawson, G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, D. Carter, M. Papaspyridonos, S. Livingstone, R. Ganske, E. Lõhmussaar, J. Zernant, N. Tõnisson, M. Remm, R. MĠgi, T. Puurand, J. Vilo, A. Kurg, K. Rice, P. Deloukas, R. Mott, A. Metspalu, D. R. Bentley, L. R. Cardon, and I. Dunham, "A first-generation linkage disequilibrium map of human chromosome 22," *Nature*, 2002.

[458] N. H. Barton and S. P. Otto, "Evolution of recombination due to random drift," *Genetics*, 2005.

[459] N. H. Barton, "Mutation and the evolution of recombination," 2010.

[460] S. M. Fullerton, A. B. Carvalho, and A. G. Clark, "Local rates of recombination are positively correlated with GC content in the human genome [4]," 2001.

[461] A. Yu, C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber, "Comparison of human genetic and sequence-based physical maps," 2001.

[462] A. V. Smith, D. J. Thomas, H. M. Munro, and G. R. Abecasis, "Sequence features in regions of weak and strong linkage disequilibrium," *Genome Res.*, 2005.

[463] T. Ohta, "Linkage disequilibrium due to random genetic drift in finite subdivided populations," *Proc. Natl. Acad. Sci.*, 1982.

[464] B. Charlesworth, "The effects of deleterious mutations on evolution at linked sites," *Genetics*, 2012.

[465] G. A. T. McVean and B. Charlesworth, "The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation," *Genetics*, 2000.

[466] H. J. Muller, "Some Genetic Aspects of Sex," *Am. Nat.*, 1932.

[467] H. J. Muller, "The relation of recombination to mutational advance," *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, 1964.

[468] S. A. Tishkoff and B. C. Verrelli, "Patterns of human genetic diversity: implications for human evolutionary history and disease," *Annu. Rev. Genomics Hum. Genet.*, 2003.

[469] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler, "The structure of haplotype blocks in the human genome," *Science (80-.).*, 2002.

[470] S. Service, J. DeYoung, M. Karayiorgou, J. L. Roos, H. Pretorious, and G. Bedoya, "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies," *Nat Genet*, vol. 38, 2006.

[471] E. Jakkula, K. Rehnström, T. Varilo, O. P. Pietiläinen, T. Paunio, N. L. Pedersen, U. DeFaire, M. R. Järvelin, J. Saharinen, N. Freimer, S. Ripatti, S. Purcell, A. Collins, M. J. Daly, A. Palotie, and L. Peltonen, "The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population," *Am. J. Hum. Genet.*, 2008.

[472] A. J. Jeffreys, L. Kauppi, and R. Neumann, "Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex," *Nat Genet*, vol. 29, 2001.

[473] S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean, "A common sequence motif associated with recombination hot spots and genome instability in humans," *Nat. Genet.*, 2008.

[474] F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy, "PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.," *Science*, 2010.

[475] F. Pratto, K. Brick, P. Khil, F. Smagulova, G. V. Petukhova, and R. D. Camerini-Otero, "Recombination initiation maps of individual human genomes," *Science*, vol. 346, p. 1256442, nov 2014.

[476] N. E. Morton, W. Zhang, P. Taillon-Miller, S. Ennis, P. Y. Kwok, and A. Collins, "The optimal measure of allelic association," *Proc Natl Acad Sci U S A*, vol. 98, 2001.

[477] N. Maniatis, A. Collins, C. F. Xu, L. C. McCarthy, D. R. Hewett, and W. Tapper, "The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis," *Proc Natl Acad Sci U S A*, vol. 99, 2002.

[478] A. Collins, S. Ennis, P. Taillon-Miller, P. Y. Kwok, and N. E. Morton, "Allelic association with SNPs: Metrics, populations, and the linkage disequilibrium map," *Hum. Mutat.*, 2001.

[479] W. Zhang, A. Collins, N. Maniatis, W. Tapper, and N. E. Morton, "Properties of linkage disequilibrium (LD) maps," *Proc Natl Acad Sci*, vol. 99, 2002.

[480] W. Zhang, A. Collins, J. Gibson, W. J. Tapper, S. Hunt, and P. Deloukas, "Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps," *Proc Natl Acad Sci U S A*, vol. 101, 2004.

[481] G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, "The Fine-Scale Structure of Recombination Rate Variation in the Human Genome," *Science (80-.).*, 2004.

[482] G. S. Jacobs, T. J. Sluckin, and T. Kivisild, "Refining the use of linkage disequilibrium as a robust signature of selective sweeps," *Genetics*, 2016.

[483] J. Gibson, W. Tapper, W. Zhang, N. Morton, and A. Collins, "Cosmopolitan linkage disequilibrium maps," *Hum. Genomics*, vol. 2, no. 1, p. 20, 2005.

[484] C. de Filippo, K. Bostoen, M. Stoneking, and B. Pakendorf, "Bringing together linguistic and genetic evidence to test the Bantu expansion," *Proc. R. Soc. B Biol. Sci.*, 2012.

[485] G. B. Busby, G. Band, Q. Si Le, M. Jallow, E. Bougama, V. D. Mangano, L. N. Amenga-Etego, A. Enimil, T. Apinjoh, C. M. Ndila, A. Manjurano, V. Nyirongo, O. Doumba, K. A. Rockett, D. P. Kwiatkowski, and C. C. Spencer, "Admixture into and within sub-Saharan Africa," *Elife*, 2016.

[486] J. A. Hodgson, C. J. Mulligan, A. Al-Meeri, and R. L. Raaum, "Early Back-to-Africa Migration into the Horn of Africa," *PLoS Genet.*, 2014.

[487] S. Li, C. Schlebusch, and M. Jakobsson, "Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples," *Proc. R. Soc. B Biol. Sci.*, 2014.

[488] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir, "A high-performance computing toolset for relatedness and principal component analysis of SNP data," *Bioinformatics*, 2012.

[489] A. Collins and N. E. Morton, "Mapping a disease locus by allelic association.," *Proc. Natl. Acad. Sci. U. S. A.*, 1998.

[490] A. Collins, C. Lonjou, and N. E. Morton, "Genetic epidemiology of single-nucleotide polymorphisms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, pp. 15173–15177, dec 1999.

[491] N. Spataro, J. A. Rodríguez, A. Navarro, and E. Bosch, "Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology," *Hum. Mol. Genet.*, 2017.

[492] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas, "PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements," *Nucleic Acids Res.*, 2017.

[493] I. Bartha, J. Di Iulio, J. C. Venter, and A. Telenti, "Human gene essentiality," *Nat. Rev. Genet.*, 2018.

[494] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res.*, 2016.

[495] Y. E. Zhang, M. D. Vibranovski, P. Landback, G. A. Marais, and M. Long, "Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome," *PLoS Biol.*, 2010.

[496] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. Von Mering, "STRING v10: Protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, 2015.

[497] A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson, "A high-resolution recombination map of the human genome," *Nat. Genet.*, vol. 31, no. 3, pp. 241–247, 2002.

[498] C. J. Bae, K. Douka, and M. D. Petraglia, "On the origin of modern humans: Asian perspectives," 2017.

[499] E. Patin, M. Lopez, R. Grollemund, P. Verdu, C. Harmant, H. Quach, G. Laval, G. H. Perry, L. B. Barreiro, A. Froment, E. Heyer, A. Massougbodji, C. Fortes-Lima, F. Migot-Nabias, G. Bellis, J. M. Dugoujon, J. B. Pereira, V. Fernandes, L. Pereira, L. Van Der Veen, P. Mouguiama-Daouda, C. D. Bustamante, J. M. Hombert, and L. Quintana-Murci, "Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America," *Science (80-.).*, 2017.

[500] A. G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C. D. Palmer, G. K. Chen, K. Wang, S. G. Buxbaum, E. L. Akylbekova, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L. A. Cupples, S. L. Deming, W. R. Diver, J. Divers, M. Fornage, E. M. Gillanders, J. Glessner, C. C. Harris, J. J. Hu, S. A. Ingles, W. Isaacs, E. M. John, W. H. Kao, B. Keating, R. A. Kittles, L. N. Kolonel, E. Larkin, L. L. Marchand, L. H. McNeill, R. C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G. J. Papanicolaou, M. F. Press, B. M. Psaty, A. P. Reiner, S. S. Rich, J. L. Rodriguez-Gil, J. I. Rotter, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. J. Thun, M. A. Tucker, Z. Wang, J. K.

Wiencke, J. S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, X. Zhu, S. Redline, J. N. Hirschhorn, B. E. Henderson, H. A. Taylor, A. L. Price, H. Hakonarson, S. J. Chanock, C. A. Haiman, J. G. Wilson, D. Reich, and S. R. Myers, "The landscape of recombination in African Americans," *Nature*, 2011.

[501] J. Gibson, W. Tapper, S. Ennis, and A. Collins, "Exome-based linkage disequilibrium maps of individual genes: Functional clustering and relationship to disease," *Hum. Genet.*, 2013.

[502] W. H. Chen, G. Lu, X. Chen, X. M. Zhao, and P. Bork, "OGEE v2: An update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines," *Nucleic Acids Res.*, 2017.

[503] K. Y. Popadin, M. Gutierrez-Arcelus, T. Lappalainen, A. Buil, J. Steinberg, S. I. Nikolaev, S. W. Lukowski, G. A. Bazykin, V. B. Seplyarskiy, P. Ioannidis, E. M. Zdobnov, E. T. Dermitzakis, and S. E. Antonarakis, "Gene age predicts the strength of purifying selection acting on gene expression variation in humans," *Am. J. Hum. Genet.*, 2014.

[504] T. Domazet-Lošo and D. Tautz, "An ancient evolutionary origin of genes associated with human genetic diseases," *Mol. Biol. Evol.*, 2008.

[505] E. K. Maxwell, C. E. Schnitzler, P. Havlak, N. H. Putnam, A. D. Nguyen, R. T. Moreland, and A. D. Baxevanis, "Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: Implications for modeling disease genetics in animals," *BMC Evol. Biol.*, 2014.

[506] J. Seok, H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. López, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H. Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S. F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, and R. G. Tompkins, "Genomic responses in mouse models poorly mimic human inflammatory diseases," *Proc. Natl. Acad. Sci.*, 2013.

[507] K. L. Korunes and M. A. Noor, "Gene conversion and linkage: effects on genome evolution and speciation," *Mol. Ecol.*, 2017.

[508] L. Pagani, T. Kivisild, A. Tarekegn, R. Ekong, C. Plaster, I. Gallego Romero, Q. Ayub, S. Q. Mehdi, M. G. Thomas, D. Luiselli, E. Bekele, N. Bradman, D. J. Balding, and C. Tyler-Smith, "Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool," *Am. J. Hum. Genet.*, 2012.

[509] I. L. Berg, R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler, and A. J. Jeffreys, "Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations," *Proc. Natl. Acad. Sci.*, 2011.

[510] S. Berger, M. Schlather, G. L. De Campos, S. Weigend, R. Preisinger, M. Erbe, and H. Simianer, "A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions," *PLoS One*, 2015.

[511] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow, "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res.*, 2005.

[512] I. L. Berg, R. Neumann, K. W. G. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, and A. J. Jeffreys, "PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans," *Nat. Genet.*, 2010.

[513] J. G. Hussin, A. Hodgkinson, Y. Idaghdour, J. C. Grenier, J. P. Goulet, E. Gbeha, E. Hip-Ki, and P. Awadalla, "Recombination affects accumulation of damaging and disease-associated mutations in human populations," *Nat. Genet.*, 2015.

[514] R. J. Pengelly, A. Vergara-Lope, D. Alyousfi, M. R. Jabalameli, and A. Collins, "Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation," *Brief. Bioinform.*, pp. bbx110–bbx110, aug 2017.

[515] S. Roy, C. Coldren, A. Karunamurthy, N. S. Kip, E. W. Klee, S. E. Lincoln, A. Leon, M. Pullambhatla, R. L. Temple-Smolkin, K. V. Voelkerding, C. Wang, and A. B. Carter, "Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists," 2018.

[516] D. C. Jeffares, C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C. Dessimoz, J. Bähler, and F. J. Sedlazeck, "Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast," *Nat. Commun.*, 2017.

[517] M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, "Long reads: their purpose and place," *Hum. Mol. Genet.*, 2018.

[518] S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand, "Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics," *Nucleic Acids Res.*, 2018.

[519] J. C. Mu, P. Tootoonchi Afshar, M. Mohiyuddin, X. Chen, J. Li, N. Bani Asadi, M. B. Gerstein, W. H. Wong, and H. Y. Lam, "Leveraging long read sequencing from a single individual to provide a comprehensive resource for benchmarking variant calling methods," *Sci. Rep.*, 2015.

[520] E. P. Consortium, I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. a. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B.-K. B.-K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kelllis, P. Kheradpour, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. J. S. L. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. a. L. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. a. Feingold, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. C. Giddings, T. R. Gingeras, R. Guigó, T. J. T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, J. a. Starnatoyannopoulos, S. a. Tennebaum, Z. Weng, K. P.

White, B. Wold, Y. Yu, J. Wrobel, B. a. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, M. L. Eaton, A. Dobin, T. Lassmann, A. Tanzer, J. Lagarde, W. Lin, C. Xue, B. a. Williams, C. Zaleski, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandu, L. Schaeffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. H. Wang, Y. Hayashizaki, A. Reymond, S. E. Antonarakis, G. J. Hannon, Y. Ruan, P. Carninci, C. a. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, L. L. Grasfeder, P. G. Giresi, A. Battenhouse, N. C. Sheffield, K. a. Showers, D. London, A. a. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Z. Zhang, P. a. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, T. Wang, D. Winter, D. Keefe, V. R. Iyer, K. S. Sandhu, M. Zheng, P. Wang, J. Gertz, J. Vielmetter, E. C. Partridge, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, M. a. Muratet, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, J. S. Newberry, S. E. Levy, D. M. Absher, W. H. Wong, M. J. Blow, A. Visel, L. a. Pennachio, L. Elnitski, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, C. Davidson, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. a. Hendrix, T. Hunt, I. Jungreis, M. Kay, E. Khurana, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, E. Tapanari, M. L. Tress, M. J. van Baren, S. Washieti, L. Wilming, A. Zadissa, Z. Zhengdong, M. Brent, D. Haussler, A. Valencia, A. Raymond, N. Addleman, R. P. Alexander, R. K. Auerbach, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyenger, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Larnarre-Vincent, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, D. Raha, L. Ramirez, B. Reed, M. Shi, T. Slifer, H. Witt, L. Wu, X. Xu, K.-K. Yan, X. Yang, K. Struhl, S. M. Weissman, S. a. Tenebaum, L. O. Penalva, S. Karmakar, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, A. Victorsen, T. Auer, L. Centarin, M. Eichenlaub, F. Gruhl, S. Heerman, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, G. Jain, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, A. K. Johnson, E. M. Johnson, T. M. Kutyavin, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, M. E. Sanchez, R. S. Sandstrom, A. O. Shafer, A. B. Stergachis, S. Thomas, B. Vernot, J. Vierstra, S. Vong, M. a. Weaver, Y. Yan, M. Zhang, J. a. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, J. a. Stamatoyannopoulos, K. Beal,

A. Brazma, P. Flicek, N. Johnson, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. a. Schaub, W. Miller, P. J. Bickel, B. Banfai, N. P. Boley, H. Huang, J. J. Li, W. S. Noble, J. a. Bilmes, O. J. Buske, A. O. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, and L. Lochovsky, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, 2012.

[521] I. Ionita-Laza, K. Mccallum, B. Xu, and J. D. Buxbaum, "A spectral approach integrating functional genomic annotations for coding and noncoding variants," *Nat. Genet.*, 2016.

[522] P. J. Short, J. F. McRae, G. Gallone, A. Sifrim, H. Won, D. H. Geschwind, C. F. Wright, H. V. Firth, D. R. Fitzpatrick, J. C. Barrett, and M. E. Hurles, "De novo mutations in regulatory elements in neurodevelopmental disorders," *Nature*, 2018.

[523] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[524] A. Hofherr, C. Seger, F. Fitzpatrick, T. Busch, E. Michel, J. Luan, L. Osterried, F. Linden, A. Kramer-Zucker, B. Wakimoto, C. Schütze, N. Wiedemann, A. Artati, J. Adamski, G. Walz, E. R. S. Kunji, C. Montell, T. Watnick, and M. Köttgen, "The mitochondrial transporter SLC25A25 links ciliary TRPP2 signaling and cellular metabolism," *PLOS Biol.*, 2018.

[525] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. Fitzpatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. A. Hurst, J. Jähn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S. M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. V. Vooren, R. J. Wapner, A. O. Wilkie, C. F. Wright, A. T. Vulto-Van Silfhout, N. D. Leeuw, B. B. De Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis, and P. N. Robinson, "The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data," *Nucleic Acids Res.*, 2014.

[526] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson, "Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies," *Am. J. Hum. Genet.*, 2009.

[527] S. Bauer, S. Köhler, M. H. Schulz, and P. N. Robinson, "Bayesian ontology querying for accurate and noise-tolerant semantic searches," *Bioinformatics*, 2012.

[528] P. N. Robinson, S. Köhler, A. Oellrich, S. M. Genetics, K. Wang, C. J. Mungall, S. E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, C. Gilissen, M. Haendel, and D. Smedley, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Res.*, vol. 24, no. 2, pp. 340–348, 2014.

[529] A. A. Philippakis, D. R. Azzariti, S. Beltran, A. J. Brookes, C. A. Brownstein, M. Brudno, H. G. Brunner, O. J. Buske, K. Carey, C. Doll, S. Dumitriu, S. O. Dyke, J. T. den Dunnen, H. V. Firth, R. A. Gibbs, M. Girdea, M. Gonzalez, M. A. Haendel, A. Hamosh, I. A. Holm, L. Huang, M. E. Hurles, B. Hutton, J. B. Krier, A. Misyura,

C. J. Mungall, J. Paschall, B. Paten, P. N. Robinson, F. Schiettecatte, N. L. Sobreira, G. J. Swaminathan, P. E. Taschner, S. F. Terry, N. L. Washington, S. Züchner, K. M. Boycott, and H. L. Rehm, "The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery," *Hum. Mutat.*, 2015.

[530] N. Akawi, J. McRae, M. Ansari, M. Balasubramanian, M. Blyth, A. F. Brady, S. Clayton, T. Cole, C. Deshpande, T. W. Fitzgerald, N. Foulds, R. Francis, G. Gabriel, S. S. Gerety, J. Goodship, E. Hobson, W. D. Jones, S. Joss, D. King, N. Klena, A. Kumar, M. Lees, C. Lelliott, J. Lord, D. McMullan, M. O'Regan, D. Osio, V. Piombo, E. Prigmore, D. Rajan, E. Rosser, A. Sifrim, A. Smith, G. J. Swaminathan, P. Turnpenny, J. Whitworth, C. F. Wright, H. V. Firth, J. C. Barrett, C. W. Lo, D. R. FitzPatrick, and M. E. Hurles, "Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families," *Nat. Genet.*, 2015.

[531] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith, "A systematic survey of loss-of-function variants in human protein-coding genes," *Science (80-.).*, 2012.

[532] D. Anderson and T. Lassmann, "A phenotype centric benchmark of variant prioritisation tools," *npj Genomic Med.*, 2018.

[533] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genet. Med.*, 2015.

[534] A. Keinan and A. G. Clark, "Recent explosive human population growth has resulted in an excess of rare genetic variants," *Science (80-.).*, 2012.

[535] G. McVean, "The structure of linkage disequilibrium around a selective sweep," *Genetics*, 2007.

[536] G. E. Liu, C. Alkan, L. Jiang, S. Zhao, and E. E. Eichler, "Comparative analysis of Alu repeats in primate genomes," *Genome Res.*, 2009.

[537] J. Jurka, "Evolutionary impact of human Alu repetitive elements," 2004.

[538] Y. Mostovoy, M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie, P. Marks, J. Lee, C. Chu, C. Lin, Z. Dzakula, H. Cao, S. A. Schlebusch, K. Giorda, M. Schnall-Levin, J. D. Wall, and P. Y. Kwok, "A hybrid approach for de novo human genome sequence assembly and phasing," *Nat. Methods*, 2016.

[539] M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H. Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C. S. Chin,

Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P. Y. Kwok, C. E. Mason, E. E. Schadt, and A. Bashir, "Assembly and diploid architecture of an individual human genome via single-molecule technologies," *Nat. Methods*, 2015.