Developing pedigree-based strategies to analyse whole genome sequencing data for complex disorders: learning from schizophrenia

A thesis submitted to
The University of Dublin
for the degree of
Doctorate in Philosophy

by

Cathal Ormond, BA (Mod), MSc

11th January 2023

Discipline of Psychiatry
School of Medicine
Trinity College Dublin
University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Except where otherwise stated, the work has been carried out by the author alone.

Signature of author: .....

Cathal Ormond
January, 2023

Summary

Complex genetic disorders are impacted by a range of DNA variation. Next generation sequencing (NGS) allows for the direct examination of this variation, but large sample numbers are required to identify rare variants in unrelated cohorts. Pedigree-based cohorts can partly resolve this, as densely affected pedigrees are likely to be influenced by the same collection of rare variants. In this thesis, we examined approaches for disease-gene prioritisation from pedigree-based NGS data for complex disorders. As a model phenotype, we consider schizophrenia. A spectrum of rare and common variants is known to increase individual risk for schizophrenia, although identifying such variants remains challenging.

To begin, we examined some issues with variants derived from NGS data. Converting base pair positions between builds of a reference genome can result in instabilities that can impact downstream analysis. We characterised all such unstable positions between two builds of the human reference genome. We replicated these instabilities in whole genome sequencing (WGS) data and showed that removing variants at unstable positions results in variants stable to the conversion process. Next, we developed a novel pipeline for calling copy number variants (CNVs) that takes a consensus of four calling methods. By incorporating relatedness information, we can reclaim lower confidence CNV calls in our consensus approach. We benchmarked this pipeline against a curated "Gold Standard" set of CNV calls and showed that our method outperforms all other comparison pipelines selected.

We examined WGS data from a collection of identical twins discordant for schizophrenia and related disorders. We identified seven rare, deleterious, missense variants present in an affected sample but absent from their co-twin. One impacted gene (*POLG*) has previously been implicated in mood disorders and psychosis. We also identified a rare duplication at chromosome 3q29 private to one affected sample. Duplications in this region have previously been observed in autism and developmental delay.

Next, we investigated WGS data from a cohort of seven Utah pedigrees multiply affected with schizophrenia. We considered an identity-by-state (IBS) filtering approach and prioritised ultra-rare, protein-coding variants in constrained genes. We identified three such variants with a reduced co-segregation pattern in three separate pedigrees. One such gene (*ATP2B2*) has been implicated in common variants associated with schizophrenia

and was found to be nominally associated with schizophrenia in a recent rare-variant case-control analysis.

We evaluated two tools (`pVAAST` and `PERCH`) which aim to prioritise variants from pedigree-based NGS data in a more unified framework compared to the IBS filtering. We found that `pVAAST` correctly identified deleterious variants that followed a Mendelian inheritance pattern using a synthetic phenotype but was unable to identify the three variants prioritised in the IBS analysis of the Utah pedigrees. `PERCH` did not identify several of the deleterious Mendelian variants, and so both tools were removed from future analyses.

To address some of the limitations of previous methodologies, we developed a novel Bayesian model to measure pedigree-based causality from NGS data. We found that our method performed well at identifying the correct variants from the synthetic phenotype and the Utah pedigrees. Additionally, our method identified a rare frameshift variant in *KDM2B* perfectly co-segregating with schizophrenia that was discounted by the IBS analysis. A variant in gene has been recently implicated in schizophrenia from co-segregation in a Japanese pedigree.

Our work has wider implications, making substantial contributions in aiding researchers to elucidate the genetic architecture of pedigree-based NGS data for complex genetic disorders in psychiatry and beyond.

Acknowledgements

First and foremost, I would like to thank my brilliant supervisors, Dr Eleisa Heron and Prof. Aiden Corvin, without whom none of this work would have been possible. Both have given constant encouragement and mentorship, and have taught me how to ask the right research questions - even when I try to run down rabbit holes! I will always be grateful for the many opportunities (research and otherwise) that I have been afforded while working with you both.

I am extremely thankful to all of the study participants who donated DNA that was used in this thesis. I would also like to thank our collaborators who coordinated these samples for use in our research project, particularly Prof. Pat Sullivan, Prof. William Byerley and their respective teams. This research project was funded by Science Foundation Ireland and the National Institutes of Health as part of the Psychiatric Genomics Consortium. Also, I would like to thank Prof. Michael Gill, Ms Geraldine Quinn, Dr Elaine Kenny, Dr Niamh Ryan, Dr Amy Cole, and Dr Siobhán Connolly for their extensive help and advice with this project. This work was supported by TCHPC (Research IT, Trinity College Dublin), and I thank the sysadmins for their help.

I'm grateful to all the other brilliant people in the Neuropsychiatric Genetics Research group and in TTMI who've been amazing to work with over the last few years. Shout out in particular to all the tenants of Office 1.12 past and present (especially Niamh, Elaine, Ciara, and Carlos) for putting up with my endless chatting and distractions! Also, a big thanks to all the other PhD students in the group (in particular, Sarah-Marie, Tom, Fiana, Niall, Albert, and Stephen) for helping to unwind, whether it be cocktails, karaoke or Kenny's - sometimes, all three!

It's not easy doing a PhD even outside a pandemic, but my friends have been a spectacular help these past few years. I would have given up long ago if I didn't have good people around me to keep me sane and centred. I also need to mention my writing companion *el Gato* BOB, and the babies Piper, Riz and MJ, who provided plenty of hijinks. And finally, I'd like to thank my family for their support over the last several years, and a big welcome to the world to my new niece Beth!

“Déanaigí bur nóiceall de,
Níl ann ach fear bur gcuarta.”

Contents

Declaration iii

Summary v

Acknowledgements vii

List of Figures 2

List of Tables 4

Acronyms 5

1 Introduction and Background 9
- 1.1 Genetic Models . . . 9
  - 1.1.1 Overview of Complex Disorders . . . 9
  - 1.1.2 Pedigree-Based Studies . . . 10
- 1.2 Genomic Technologies . . . 11
  - 1.2.1 Next-Generation Sequencing . . . 11
  - 1.2.2 Structural Variants . . . 12
  - 1.2.3 NGS Pedigree Analysis . . . 13
- 1.3 Schizophrenia . . . 14
  - 1.3.1 Phenotype . . . 14
  - 1.3.2 Genomics . . . 14
- 1.4 Aims of the Thesis . . . 17

2 General Methods 19
- 2.1 File Formats . . . 19
  - 2.1.1 FASTQ . . . 19
  - 2.1.2 SAM/BAM/CRAM . . . 20
  - 2.1.3 VCF . . . 20
  - 2.1.4 FAM . . . 21
- 2.2 SNV and Indel Calling . . . 21
  - 2.2.1 Read Alignment and Post-Processing . . . 21
  - 2.2.2 Variant Calling and Joint Genotyping . . . 24
- 2.3 Variant Quality Control . . . 24
  - 2.3.1 Variant Quality Score Recalibration . . . 24
  - 2.3.2 Genotype-Level Metrics . . . 25
- 2.4 Pedigree Consistency . . . 25
- 2.5 Public Databases and Resources . . . 25
  - 2.5.1 Allele Frequency Databases . . . 25
  - 2.5.2 Variant Deleteriousness Metrics . . . 26
  - 2.5.3 Gene Constraint Scores . . . 28

2.6 Variant Annotation . . . 28
2.6.1 vep . . . 28
2.6.2 SnpSift . . . 28

3 Converting Single Nucleotide Variants between Genome Builds 29
3.1 Introduction . . . 29
3.2 Identification of Unstable Positions . . . 32
3.2.1 Chain File . . . 32
3.2.2 Full-Genome Data . . . 32
3.2.3 Algorithm to Identify Novel CUPs . . . 33
3.2.4 Comparison with Assembly Annotation Sets . . . 37
3.3 Application to WGS Data . . . 41
3.3.1 Evaluation Data . . . 41
3.3.2 CUPs in WGS Data . . . 42
3.3.3 Discordance Rates Between Aligned and Converted Data . . . 44
3.4 Conclusions . . . 46

4 Copy Number Variant Calling for Family-Based Sequencing Studies 47
4.1 Introduction . . . 47
4.2 Calling Pipeline . . . 50
4.2.1 Per Individual . . . 50
4.2.2 Per Pedigree . . . 55
4.3 Alternative Strategies . . . 56
4.3.1 Individual Callers . . . 56
4.3.2 Khan et al. . . . 56
4.4 Benchmarking . . . 57
4.4.1 Curated Gold Standard CNV Calls . . . 57
4.4.2 Validation Metrics . . . 59
4.4.3 Results . . . 60
4.5 Conclusions . . . 60

5 Rare Variant Analysis of Discordant Monozygotic Twins 61
5.1 Introduction . . . 61
5.2 Cohort Description . . . 63
5.2.1 Sample Procurement . . . 63
5.2.2 Sample Processing . . . 64
5.2.3 Zygosity Check . . . 66
5.2.4 Discordant Variants . . . 70
5.3 Protein-Coding Variants . . . 70
5.4 Regulatory Variants . . . 74
5.4.1 Rare Deleterious Variants . . . 74
5.4.2 ENCODE Regulatory Features . . . 75
5.5 Germline CNVs . . . 78
5.5.1 Known SCZ CNVs . . . 78
5.5.2 Rare and Pathogenic CNVs . . . 80
5.6 Somatic CNVs . . . 83
5.7 Repeat Expansions . . . 85

5.8 Conclusions . . . 87

6 Rare Variant Analysis of Utah Pedigrees 89
6.1 Introduction . . . 89
6.2 Cohort Description . . . 90
6.2.1 Sample Procurement and Assessment . . . 90
6.2.2 Description of the Pedigrees . . . 91
6.3 WGS Data and Sample QC . . . 94
6.3.1 Batch 1 . . . 94
6.3.2 Batch 2 . . . 95
6.3.3 Quality Control Measures . . . 95
6.3.4 Cross-Platform Biases . . . 97
6.4 Protein-Coding Variants . . . 102
6.5 Copy Number Variants . . . 106
6.6 Conclusions . . . 109

7 Evaluation of Two Software Tools for Disease-Gene Prioritisation 111
7.1 Introduction . . . 111
7.2 Software Tools . . . 112
7.2.1 Description . . . 112
7.2.2 Implementation . . . 113
7.3 Mendelian Phenotype: CEPH 1463 Pedigree . . . 115
7.3.1 Data . . . 115
7.3.2 Pseudo-Causal Variant Selection . . . 116
7.3.3 Results . . . 120
7.3.4 Empirical Significance Calculation . . . 123
7.4 Complex Phenotype: Utah Pedigrees . . . 125
7.5 Conclusions . . . 126

8 A Bayesian Framework for Pedigree-Based Causality 127
8.1 Introduction . . . 127
8.2 Overview . . . 129
8.2.1 Summary of Equations . . . 129
8.2.2 Model Parameters and Assumptions . . . 130
8.3 Bayes Factor . . . 132
8.3.1 Causal Model: Likelihood . . . 132
8.3.2 Neutral Model: Likelihood . . . 134
8.3.3 Prior Sensitivity . . . 135
8.4 General Prior Probability for Causality . . . 137
8.5 Mendelian Phenotype: CEPH 1463 Pedigree . . . 147
8.6 Complex Phenotype: Utah Pedigrees . . . 151
8.7 Conclusions . . . 159

9 Conclusions and Future Work 161
9.1 Chapter 3 Conclusions . . . 161
9.2 Chapter 4 Conclusions . . . 162
9.3 Chapter 5 Conclusions . . . 163

9.4 Chapter 6 Conclusions . . . 164
9.5 Chapter 7 Conclusions . . . 165
9.6 Chapter 8 Conclusions . . . 166
9.7 Final Remarks . . . 167

A Supplementary Information **169**
A.1 WGS Details . . . 169
A.2 Web Resources . . . 171
A.2.1 Chapter 2 . . . 171
A.2.2 Chapter 3 . . . 171
A.2.3 Chapter 4 . . . 172
A.2.4 Chapter 5 . . . 172
A.2.5 Chapter 6 . . . 173
A.2.6 Chapter 7 . . . 173
A.2.7 Chapter 8 . . . 173
A.3 Software Versions . . . 174
A.4 Novel Javascript Code . . . 175

B Mathematical Details for the Bayesian Inference Model **177**
B.1 Definitions and Theorems . . . 177
B.2 Overview . . . 178
B.2.1 Data and Parameters . . . 178
B.2.2 Models . . . 178
B.2.3 Assumptions . . . 179
B.2.4 Variables . . . 179
B.3 Causal Model . . . 180
B.3.1 Uniform Prior . . . 183
B.3.2 Beta Prior . . . 185
B.3.3 Linear Prior . . . 187
B.4 Neutral Model . . . 189
B.4.1 Uniform Priors . . . 191
B.5 Summary of Formulae . . . 192
B.5.1 Causal Model . . . 192
B.5.2 Neutral Model . . . 192
B.6 Algorithms . . . 193

C Published Material **195**

Bibliography **203**

List of Figures

Figure 1.1 An overview of variants with a known association with schizophrenia taken from Singh et al. 16

Figure 2.1 Flowchart of the variant calling, genotyping, and annotation pipelines. 23

Figure 3.1 Visual depiction of chain files. 32
Figure 3.2 Flow chart of the algorithm to identify novel conversion-unstable positions. 35
Figure 3.3 Venn diagrams for the overlap between the three selected assembly annotation sets. 38
Figure 3.4 The proportion of conversion-unstable positions that overlap with the assembly annotation sets. 40

Figure 4.1 An example of a CNV identified from our WGS data. 49
Figure 4.2 A visualisation of the collapsing and merging strategy for a single individual. 52
Figure 4.3 Workflow of the CNV calling pipeline per individual. 53
Figure 4.4 Upset plots showing the intersection between the four CNV calling methods in sample NA12878 54
Figure 4.5 Upset plots for the three samples in the CEPH 1463 trio 56
Figure 4.6 Upset plot for the overlap of CNV regions between the five "Gold Standard" NA12878 call sets. 59

Figure 5.1 Post-zygotic variants before and after the twinning event. 63
Figure 5.2 A heatmap of the pairwise relatedness scores for the 18 pairs of samples in the MZ twin cohort. 67
Figure 5.3 A PCA plot of the MZ twins and a background population. . . . 68
Figure 5.4 `IGV` plot of the bamout output of `HaplotypeCaller` showing three SNVs arising from re-constructed reads. 73
Figure 5.5 Boxplots of the counts of discordant variants within each of the eight regulatory annotation sets. 76
Figure 5.6 Read depth plots of two putative discordant somatic CNV calls which were subsequently rejected as false positives. 85

Figure 6.1 Pedigree diagrams for the seven pedigrees selected for analysis. . 93
Figure 6.2 A PCA plot of the WGS samples and a background population from the 1000 Genomes Project. 96
Figure 6.3 Plot of the first two principal components from the internal PCA step of XPAT. 99
Figure 6.4 Plot of pairs of the first 10 principal components from the internal PCA step of XPAT. 101
Figure 6.5 Pedigree images of the pedigrees that harbour an ultra-rare SNV with reduced co-segregation. 105

Figure 7.1 The CEPH 1463 pedigree with the paternal grandparents removed. 117
Figure 7.2 The pVAAST CLRT scores for all genes harbouring a PCV. 121
Figure 7.3 The BayesSeg scores from PERCH for all genes harbouring a PCV. 122
Figure 7.4 The empirical p-values of the scores from pVAAST and PERCH. . . 124

Figure 8.1 Hypothetical breast cancer pedigree taken from Mohammadi et al. 133
Figure 8.2 The density plots of the parameters under various prior distributions. 134
Figure 8.3 The CEPH 1463 pedigree with a specific phenotype pattern selected. 135
Figure 8.4 The Bayes factor change for both prior distributions of the parameters. 136
Figure 8.5 The performance metrics of the five regression models on the ClinVar training and testing datasets. 142
Figure 8.6 Violin plots for Benign and Pathogenic variants split by whether they were misclassified or not by the regression model. 145
Figure 8.7 Boxplots of the prior probabilities for VUS. 146
Figure 8.8 The performance metrics of the regression model underlying the SCHEMA prior compared to the regression model underlying the General prior. 152

Figure B.1 A simulated breast cancer pedigree taken from Mohammadi et al. 181
Figure B.2 Density functions for the parameter terms 188

List of Tables

1.1 Details of 12 rare CNVs from 11 unique regions with a statistically significant association with schizophrenia. . . . 15
1.2 Details of 11 CNVs from eight unique regions that are nominally associated with schizophrenia. . . . 15

3.1 Details of the stable positions and conversion-unstable positions for the full-genome data. . . . 36
3.2 Counts and proportions of all SNVs present in WGS data. . . . 43
3.3 Discordance rates between converted data and aligned data 45

4.1 Counts of the number of CNVs called for each of the three individuals in the CEPH 1463 trio 53
4.2 A comparison of the CNV calling pipelines. . . . 60

5.1 Phenotypic data for the 19 pairs of MZ twins. . . . 65
5.2 Within-pair concordance metrics for a zygosity check. . . . 69
5.3 Discordant protein-coding variants with a predicted deleterious effect. . 72
5.4 Results from the two-sided t-tests to evaluate the enrichment of discordant variants in various regulatory features. . . . 77
5.5 Putative SCZ-associated CNVs identified in the cohort. . . . 79
5.6 A list of rare discordant CNVs. . . . 81
5.7 A list of CNVs with a predicted pathogenic effect in ClinGen. . . . 82
5.8 A list of 16 selected disorders associated with a multi-nucleotide repeat expansion. . . . 86
5.9 For each of the multi-nucleotide repeat disorders, details across all 34 samples. . . . 87

6.1 Counts of the number of samples sequenced from each of the ten pedigrees. . . . 92
6.2 Counts of the number of samples sequenced from each of the ten pedigrees. . . . 97
6.3 The number of variants remaining after each stage of the prioritisation process across the seven pedigrees. . . . 103
6.4 Details of the three prioritized variants with reduced co-segregation. . . 104

6.5 Gene-level constraint information for the prioritised variants. 104
6.6 Schizophrenia risk CNVs putatively identified in the cohort. 108

7.1 The pseudo-causal variants identified by the filtering process. 119
7.2 The pVAAST scores for the three genes prioritised from Chapter 6. . . . 126

8.1 The selected ClinVar variants broken down by variant type and functional consequence. 139
8.2 Description of the logistic regression models and the predictors for each. 140
8.3 List of the coefficients of regression Model 5. 143
8.4 The proportion of correctly and incorrectly classified variants for each functional consequence (General prior). 144
8.5 The results for the pseudo-causal variants on applying the fully Bayesian inference model to the CEPH 1463 pedigree. 150
8.6 List of the coefficients of the regression models underlying the SCHEMA prior. 152
8.7 The proportion of correctly and incorrectly classified variants for each functional consequence (SCHEMA prior). 153
8.8 The top 10 variants ranked by their posterior probability of causality for three Utah pedigrees. 158

A.1 Versions of the software used in this thesis. 174

Acronyms

`GATK` genome analysis toolkit. 19

`MoChA` Mosaic Chromosomal Alterations. 81

`vep` variant effect predictor. 26

AFR African. 24

ALS amyotrophic lateral sclerosis. 83

AMR admixed American. 24

BAM binary alignment map. 18

BED browsed extensible data. 30

bp base pair. 9

BQSR base quality score recalibration. 20

CADD Combined Annotation Dependent Depletion. 25

CGH comparative genomic hybridization. 11

CNV copy number variant. 11

CRAM compressed alignment map. 18

CUP conversion-unstable position. 31

dbNSFP database of non-synonymous functional prediction. 25

DECIPHER Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. 24

DGV Database of Genomic Variants. 24

DP depth of coverage. 19

DZ dizygotic. 59

EAS East Asian. 24

EGCG Edinburgh Genomics, Clinical Genomics. 19

ENA European Nucleotide Archive. 48

ENCODE Encyclopedia of DNA Elements. 73

EUR European. 24

ExAC exome aggregation consortium. 160

FISH fluorescent *in situ* hybridization. 11

FTD frontotemporal dementia. 83

gnomAD genome aggregation database. 24

GQ genotype quality score. 19

GRC Genome Reference Consortium. 11

GT genotype. 19

GWAS genome-wide association study. 9

IBS identity by state. 12

indel insertion or deletion. 10

IUPAC International Union of Pure and Applied Chemistry. 39

LD linkage disequilibrium. 9

loeuf loss-of-function observed/expected upper-bound fraction. 26

MPC Missense badness, PolyPhen2 and missense Constraint. 25

MQ mapping quality across all reads. 62

MZ monozygotic. 59

NGS next-generation sequencing. 9, 10

PCA principal component analysis. 23

PCR polymerase chain reaction. 10

PGC Psychiatric Genomics Consortium. 13

pLI probability of loss-of-function intolerant. 26

PolyPhen2 Polymorphism Phenotyping v2. 25

QD phred-scaled quality score normalised to read depth. 63

QUAL phred-scaled quality score. 62

SAM sequence alignment map. 18

SAS South Asian. 24

SCHEMA Schizophrenia Exome Meta-Analysis. 14

SIFT Sorting Intolerant From Tolerant. 24

SNP single nucleotide polymorphism. 9

SNV single nucleotide variant. 10

SV structural variant. 11

T2T Telomere-to-Telomere. 157

UCSC University of California, Santa Cruz. 11

URV ultra-rare variant. 14

VCF variant call format. 18

VQSLOD variant quality score, logarithm of odds. 22

VQSR variant quality score recalibration. 22

WES whole-exome sequencing. 10

WGS whole-genome sequencing. 10

Chapter 1

Introduction and Background

Identifying the genetic basis of diseases is important if we are to understand disease pathogenesis and improve patient outcomes. Mendelian disorders, with a simple genetic architecture, are rare, with the bulk of disease morbidity involving a complex interplay of genetic and environmental risk. This complexity necessitates multiple methodological approaches to DNA variant discovery. Studying families, which share genetic variants, can be useful where the emphasis is on variants that are rare in the wider population. The focus of this thesis is learning about and developing methods for analysing rare variants from human sequence data (in Chapter 4, Chapter 7, and Chapter 8), and applying this to schizophrenia, a heritable brain disorder (in Chapter 5, Chapter 6, and Chapter 8). Here we discuss some of the technologies and techniques used to investigate complex genetic disorders before providing a brief overview of the current understanding of schizophrenia genetics.

1.1 Genetic Models

1.1.1 Overview of Complex Disorders

In comparison to Mendelian disorders, complex genetic disorders are influenced by multiple genetic factors and different models for the genetic architecture involved have been proposed (Mitchell, 2012). The common-variant hypothesis states that there are many common DNA variants, each with a small effect on the phenotype. This can range from a modest number of variants (oligogenic) to hundreds, or thousands in a polygenic model. At the other extreme, the rare-variant hypothesis proposes involvement of many rare variants, each with a moderate to large effect on the phenotype. Both hypotheses have merits and should be seen as complementary rather than competitive (Gibson, 2012; Schork et al., 2009). The liability threshold model states that complex disorders have some underlying distribution of risk, either genetic, environmental, or both (Falconer, 1965; Pearson, 1901). An accumulation of risk factors pushing an individual above some threshold results in that individual having the phenotype. This model provides some harmonisation between the common-variant and rare-variant hypotheses.

DNA microarrays have been used for many decades to genotype single nucleotide polymorphisms (SNPs) which are individual DNA base pair (bp) changes to the genome. To be classified as a SNP, a variant must be found in at least 1% of the population. Their frequency, and the development of arrays that can genotype a million SNPs or more (LaFramboise, 2009), made possible the comprehensive genetic analysis of common variants predicted by Risch and Merikangas (Risch & Merikangas, 1996). Over the last two decades, genome-wide association study (GWAS) have used SNP arrays to evaluate the common-variant hypothesis by examining differences in SNP frequencies across a given phenotype in cohorts of unrelated individuals (Bush & Moore, 2012). This work has mapped out a significant contribution of common genetic risk across a wide range of conditions. However, each SNP identified typically represents a linkage disequilibrium (LD) block which may contain many DNA variants, so the causal mechanism of a GWAS association peak can be hard to establish. Significant loci from GWAS can be summarised at an individual level as a polygenic (risk) score, which is the sum of the number of alleles carried by that individual, weighted by the effect sizes of the loci.

1.1.2 Pedigree-Based Studies

The heritability explained by SNPs alone may fall short of the known family-based heritability estimates, a phenomenon known as the "missing heritability" problem (Manolio et al., 2009). One explanation for this is that rare variants, whose signal may not be readily detectable from SNP genotype arrays, account for a substantial proportion of the remaining heritability (Zuk et al., 2014). The rationale is that pathogenic variants with strong effects on a phenotype are likely to be rare in the general population due to purifying selection. Identifying such variants may be particularly important in understanding biological mechanisms that underpin phenotypes. Rare variants that affect the amino-acid chain of protein-coding regions are more readily interpretable in a biological context than tagging SNPs under association peaks, the majority of which are non-coding (Cano-Gamez & Trynka, 2020). Additionally, such rare variants are often amenable to follow-up molecular analyses to provide biological validation to statistical identification. These variants that are not typically captured by SNP arrays could only be systematically investigated with the advent of next-generation sequencing (NGS), described below. However, large sample sizes are required to perform a gene-based burden analysis of rare variants in an unrelated cohort, and greater sizes again are needed to discover specific risk variants (Sanders et al., 2017).

Pedigree-based analyses offer a solution to this issue (DeLisi, 2016; Glahn et al., 2019). A

rare variant present in a founder of a pedigree is more likely to be present in the founder's immediate descendants than in the general population. The assumption is that densely affected pedigrees will be enriched for highly penetrant, rare variants, reducing the need for the extremely large sample sizes that are required in unrelated cohorts (Sullivan et al., 2012). LD-structure differences and population stratification are less likely to occur in closely related individuals within a pedigree, although care should be taken with marry-in individuals who may have different genomic ancestry. Another potential advantage is that there tends to be less variation in environmental factors between individuals in a pedigree compared to unrelated cohorts (Morris et al., 2015), and unaffected individuals within a pedigree can often serve as controls.

1.2 Genomic Technologies

1.2.1 Next-Generation Sequencing

Traditional Sanger sequencing allows for the examination of contiguous DNA sequences but is typically limited to segments of less than 1kbp in length (Crossley et al., 2020). Where longer DNA sequences are required, shotgun sequencing of overlapping segments can be performed (Heather & Chain, 2016). However, this process is not always feasible for several genomic loci at once due to its cost, so genomic regions need to be prioritised in advance by some other method. This issue found a powerful resolution with the widespread use of short-read NGS technologies. In the 2000s, the cost of NGS fell rapidly, and its scalability made it an attractive alternative to Sanger sequencing (Goodwin et al., 2016). While DNA sequencing is typically orders of magnitude more expensive than SNP genotyping, it facilitates the direct evaluation of DNA with no requirement for imputation panels or careful probe design.

NGS involves splitting the DNA into short, contiguous fragments which may be amplified by polymerase chain reaction (PCR) (McCombie et al., 2019). These fragments (typically 100-300 bp in length) are sequenced to generate a read, which contains the ordered DNA nucleotides and their sequencing quality score. Reads can be assembled to re-construct the original genome of the sample (Reinert et al., 2015). Sequencing may be restricted to protein-coding regions, known as whole-exome sequencing (WES), or cover the entire genome, known as whole-genome sequencing (WGS). By analysing the assembled reads, various classes of DNA variants may be called, whose genotypes can be inferred by probabilistic modelling (Van der Auwera et al., 2013). While SNP probes on genotype arrays are chosen to be reasonably common in a given population, NGS has

the ability to examine single nucleotide variants (SNVs) or short insertions or deletions (indels) of any population frequency.

A key component to read alignment and variant calling with NGS data is a reference genome, which is a standardised, representative genome for a given species. The most frequently used human reference genomes are those constructed by the Genome Reference Consortium (GRC) (Church et al., 2011) who to date have released thirty-eight iterative reference builds; the two most recent being GRCh37 (released in 2009) and GRCh38 (released in 2013). The University of California, Santa Cruz (UCSC) Genomics Institute have also released analogous versions of these builds, referred to as hg19 and hg38 respectively (Haeussler et al., 2019). Both GRCh37 and GRCh38 were generated by sequencing DNA from a collection of human donors, predominantly using Sanger sequencing (Genome Reference Consortium, 2010; Consortium, 2013). DNA sequences were combined to form high-confidence contiguous segments known as contigs, which were joined to form a *de novo* assembly of the reference genome. One of the major updates in GRCh38 was the closing of numerous gaps where sequencing had previously not been possible (Schneider et al., 2017).

Updates to the base pair coordinates in the reference genome mean that not all positions are comparable between builds. While the most accurate solution would be the realignment of reads to a common reference genome, this is a computationally expensive task (Guo et al., 2017). This means that resources curated relative to different builds of the reference genome must be pre-processed to make them comparable. Tools exist to convert the coordinates between builds (Haeussler et al., 2019; M. Zhao et al., 2013), but the process is known to have instabilities (Liu et al., 2016).

1.2.2 Structural Variants

A structural variant (SV) is a large-scale change to a chromosome. Microscopic events such as an abnormal number of chromosomes have historically been detected using karyotyping. On a sub-microscopic level, SVs typically take the form of deletions, duplications, insertions, inversion or translocations (Feuk et al., 2006). Copy number variants (CNVs) are simply deletion or duplication events, which are estimated to make up 4.8-9.5% of the human genome (Zarrei et al., 2015). SVs were historically detected by cytogenic techniques, such as fluorescent *in situ* hybridization (FISH) and comparative genomic hybridization (CGH) (K. Wang & Bucan, 2008). Both methods have limitations, the most important of which is a low level of resolution. Also, FISH can only detect SVs in regions targeted by the fluorescent probes, and CGH cannot detect bal-

anced structural changes such as certain translocations or inversions (Weiss et al., 1999).

SNP-arrays allow a higher resolution of SV detection, given the density of SNP probes available on most modern arrays (Coughlin et al., 2012). SV calling algorithms for SNP-arrays are sensitive to probe designs, and so different arrays may not be empowered to detect all types of SVs (Haraksingh et al., 2017). One recommended strategy is to combine the results from multiple calling algorithms to improve the power to detect all variants (Kim et al., 2012), but the results will always be limited by the distribution and density of the probes in the array. Additionally, it is not always possible to determine the breakpoints of SVs with accuracy. NGS technologies have the potential to provide a solution to this issue, given that they examine all base pairs in the genome/exome, and are not dependent on tagging SNPs (M. Zhao et al., 2013). However, there are several different computational approaches to calling CNVs from NGS data, and there is a wide variability in performance of calling software tools (Kosugi et al., 2019). Unlike SNVs and indels, there are no "Best Practices" for calling CNVs from NGS data, so studies are not able to benchmark the ability of their pipeline to detect CNVs.

1.2.3 NGS Pedigree Analysis

Linkage analysis is the *de facto* standard used to identify candidate causal genes or regions in pedigrees. Typically, multiple generations and a minimum number of samples are required for linkage analysis to achieve statistical significance, which is not always feasible, especially for NGS data (Ott et al., 2015). An alternative approach is co-segregation analysis which we refer to as identity by state (IBS) filtering. One strategy is to examine the subset of variants present in affected individuals and absent from unaffected individuals. As with linkage analysis, characteristics common to complex disorders such as reduced penetrance and the presence of phenocopies may also be incorporated into such filtering. This method is non-statistical but has the advantage of simplicity and is a reasonable alternative when linkage analysis is not possible. Variants may be further prioritised by filtering on population-derived metrics such as conservation, deleteriousness, or allele frequency.

While IBS filtering is often implemented, it has its limitations. Firstly, there is no measure of co-segregation, so there is no way to compare results from different pedigrees. For example, we cannot know whether there is more evidence from a large sibship or from a smaller but multi-generational family. Secondly, there is no obvious approach to relaxing the requirement that all affected individuals carry a risk variant consistently across different family structures. Finally, the population-based filtering methods used

to prioritise variants, even if guided by empirical work, are arbitrary and may vary from one research group to another. Some tools have been designed to provide a statistical framework for pedigree-based NGS data analysis (Feng, 2017; Hu et al., 2014), but their novel metrics make them difficult to interpret or to compare to more traditional methods. Additionally, as with most frequentist statistical approaches, they make no statement about how likely any of the tested hypotheses are *a priori*.

1.3 Schizophrenia

1.3.1 Phenotype

Schizophrenia is a debilitating psychiatric disorder with an estimated lifetime prevalence of 1% and a reduction in life expectancy of up to 25 years (Tiihonen et al., 2009). The core features of schizophrenia are: hallucinations, delusions, disorganized speech or behaviour, and "negative symptoms" such as diminished emotional expression and avolition (Association, 2013). Environmental effects are known to have an impact on schizophrenia (Stilo & Murray, 2019), but genetic heritability has been estimated from twin studies at ~ 0.81 (Sullivan & Geschwind, 2019). Furthermore, there is an increased rate of other psychiatric conditions (e.g. bipolar disorder) in first-degree relatives of people with schizophrenia (Lichtenstein et al., 2009) and more recent work indicates a likely shared heritability across many psychiatric disorders (Anttila et al., 2018). As such, understanding the genetic aetiology of schizophrenia may provide wider insight into the genetics of mental disorders.

1.3.2 Genomics

The first large-scale studies of common variants in schizophrenia in 2009 found a handful of significantly associated loci including the major histocompatibility complex (Purcell et al., 2009; Stefansson et al., 2009). Over a decade later, the Psychiatric Genomics Consortium (PGC) wave 3 GWAS identified 287 loci associated with schizophrenia across multiple genomic ancestry groups (Trubetskoy et al., 2022). Parallel analysis of rare CNVs from SNP array data enabled the detection of twelve CNVs with a statistically significant association with schizophrenia (Marshall et al., 2017; Rees et al., 2014). Details of these loci are shown in Table 1.1. Some of these CNVs are associated with related disorders such as bipolar disorder (Green et al., 2016), major depressive disorder (Kendall et al., 2019), intellectual disability (Coe et al., 2014), and autism (Malhotra & Sebat, 2012; Sanders, 2015). Additionally, 11 CNVs were found to be nominally associated with schizophrenia, some of which had a protective effect (see Table 1.2).

Cytoband	Type	OR	Notes	Other
1q21.1	DUP	3.45		ID, ASD, MDD
	DEL	8.35		ID, ASD
2p16.3	DEL	9.01	*NRXN1*	ID, ASD
3q29	DEL	57.65		ID, ASD
7q11.23	DUP	11.35	Williams-Beuren syndrome	ID. ASD
15q11.2	DEL	2.15		ID
15q11-q13	DUP	13.20	Prader-Willi syndrome	MDD, ID, ASD
15q13.3	DEL	7.52		ID, ASD
16p13.11	DUP	2.30		ID
16p11.2, dist.	DEL	20.60		ID, ASD
16p11.2, prox.	DUP	11.52		MDD, BD, ID
22q11.2	DEL	67.70	Velocardiofacial syndrome	ID, ASD

Table 1.1: Details of 12 rare CNVs from 11 unique regions with a statistically significant association with schizophrenia, including the odds ratio (Rees et al., 2014) and other phenotypes also associated with the CNV. The odds ratios for the 16p11.2 distal deletion and the 22q11.2 deletion were taken from Marshall et al. (Marshall et al., 2017). DEL: deletion; DUP: duplication; ID: intellectual disability; ASD: autism spectrum disorder; MDD: major depressive disorder; BD: bipolar disorder; dist.: distal; prox.: proximal.

Cytoband	Type	OR	Notes	Other
7q11.21	DEL/DUP	0.66	*ZNF92*	
7p36.3	DEL/DUP	3.50	*VIPR2*, *WDR60*	
8q22.2	DEL	14.50	*VPS13B*	
9p24.3	DEL/DUP	12.40	*DMRT1*	
13q12.11	DUP	0.36	*ZMYM5*	
22q11.2	DUP	0.15		ID, ASD
Xq28	DUP	0.35	*MAGEA11*	
Xq28, dist.	DUP	8.90		

Table 1.2: Details of 11 CNVs from eight unique regions that are nominally associated with schizophrenia (Marshall et al., 2017). DEL: deletion; DUP: duplication; ID: intellectual disability, ASD: autism spectrum disorder; dist: distal.

Recently, the Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium collated WES data on 24,248 schizophrenia cases and 97,322 controls from across five genomic ancestry super-populations (T. Singh et al., 2022). The analysis focused on ultra-rare variants (URVs) affecting genes that were predicted to be intolerant to loss-of-function variants. The SCHEMA consortium reported 10 genes in which the burden of URVs was significantly higher in cases than controls and suggested that many more genes in which URVs contribute to schizophrenia risk are yet to be discovered. A summary of known rare and common variants implicated in schizophrenia as described by Singh et al. is shown in Figure 1.1.

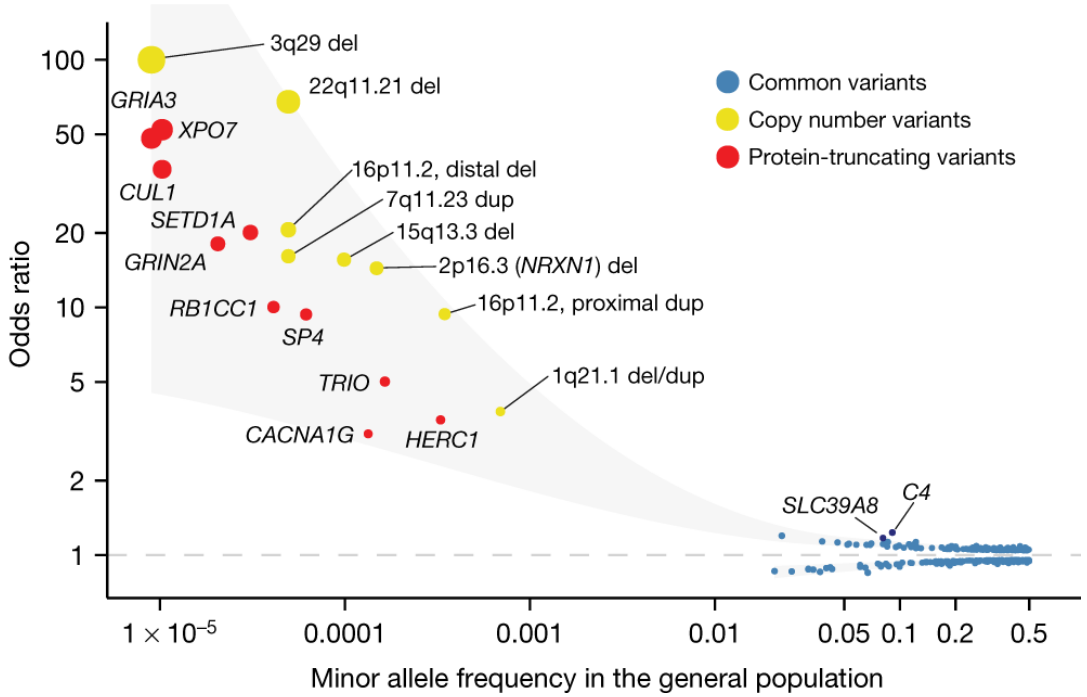


Figure 1.1: An overview of variants with a known association with schizophrenia taken from Singh et al., showing the broadly inverse relationship between the variant's allele frequency (x-axis), and the odds ratio of that variant (y-axis) (T. Singh et al., 2022). Common SNPs from GWAS are shown in blue, rare CNVs are shown in yellow, and genes harbouring ultra-rare, protein-truncating SNVs and indels are shown in red.

1.4 Aims of the Thesis

- To compare two tools (`liftOver` (Haeussler et al., 2019) and `CrossMap` (M. Zhao et al., 2013)) for converting SNVs between genome builds, and to characterize instabilities in the conversion process.
- To develop and benchmark a consensus pipeline for calling CNVs from WGS data in a pedigree-based cohort.
- To investigate the presence of rare, post-zygotic variants from WGS data in a cohort of identical twins discordant for psychiatric diagnosis.
- To analyse rare variants in a cohort of multiplex schizophrenia pedigrees from Utah using an IBS filtering approach.
- To evaluate the strengths and weaknesses of two software tools (`pVAAST` (Hu et al., 2014) and `PERCH` (Feng, 2017)) for disease-gene prioritisation from pedigree-based NGS data.
- To develop a Bayesian framework for measuring pedigree-based causality of rare variants.

Chapter 2

General Methods

In this chapter, we detail some of the general bioinformatics processes applied to prepare our WGS data for the primary analyses. We describe the read alignment pipeline applicable to all NGS data that is based on well-established "Best Practices". For some of our data, this was performed by a sequencing facility, whereas for data obtained from collaborators or from online resources, this was performed locally. Next, we call and genotype SNVs and indels across all samples in a cohort and re-calibrate the variant-level quality control scores to remove lower-confidence calls. Finally, we select various metrics from publicly available databases with which we annotate our variant call sets.

2.1 File Formats

2.1.1 FASTQ

FASTQ files are the standard format for storing raw sequencing reads from Illumina platforms and are considered the *de facto* standard for most other sequencing platforms (Cock et al., 2010). A FASTQ file is organised into four lines per read: the sequence identifier and optional meta-data (line begins with "@"), the raw nucleotide sequence, optional repeat of the title (line begins with "+"), and the per-base quality scores. The quality scores are phred-scaled and are stored as ASCII characters so that one character represents the score of each nucleotide in the read. Paired-end sequencing results in two files: the forward reads and the reverse reads. The files are matched so that the order of the reads is the same for both files, and typically tools will fail to process FASTQ files where the read order is out of sync. For WGS data, these files can be large, so they are typically compressed with `bgzip` and indexed with `tabix` (H. Li, 2011) for quicker access.

For the FASTQ data in this thesis, the tool `FastQC` (see "Web Resources", Subsection A.2.1) was applied to all files identify any potential quality control issues. `FastQC` generates figures on several metrics such as: base pair quality score, GC content (proportion of G or C nucleotides), N content (proportion of no-call bases), sequence duplication levels, etc. Samples which showed low base pair quality scores across their reads were

flagged as not having sufficiently high-quality data to continue analysis. The GC content across all chromosomes in *homo sapiens* is known to be approximately normally distributed, centred around 41% (Lander et al., 2001). Any deviation from this is usually an indicator of the presence of DNA of a different organism. Thus, samples were excluded if the distribution of GC content was multi-modal, as this is likely due to bacterial contamination of the DNA sample, or possibly the presence of tumour samples.

2.1.2 SAM/BAM/CRAM

After read alignment (discussed below), reads are stored in a sequence alignment map (SAM) format (H. Li et al., 2009). At the top of the file is a header section (lines beginning with "@") which contains meta-information about the alignment, such as the chromosomes present, the read group, the commands used to generate the file, etc. Next comes the alignment section with information from one sequencing read per line. The alignment section has 11 mandatory columns which describe the mapping of the read to the reference genome, as well as all information from the FASTQ file. Since SAM files are typically large for WGS data, they can be converted to binary alignment map (BAM) files, which are smaller in size. An efficient alternative to BAM format is compressed alignment map (CRAM) format (Cochrane et al., 2013), which can offer significant storage improvements over BAM format. However, while most tools which process aligned data accept BAM files as an input, many tools are not capable of processing CRAM files so decompression is often required. Typically, we use BAM files when calling variants and compress to CRAM for long-term storage. SAM, BAM and CRAM files were created and manipulated with `samtools` (H. Li et al., 2009).

2.1.3 VCF

The variant call format (VCF) is generated from an alignment file and stores information about genetic variants (Danecek et al., 2011). This file format can be used to describe SNVs, indels, or SVs. As with the alignment files, VCF files begin with a header (lines beginning with a "##") which contains meta-data on the main body of the file, including: the chromosomes present, annotation information about the variants, how the data were generated, variant filters, etc. Next comes a line beginning with a "#" which describes the fields, and following this is the data section, with one variant per line. The mandatory eight fields are: the chromosome, the base pair position of the start of the variant, an ID string, the reference allele, the alternate allele(s), a phred-scaled quality score, any filtering info, and any user-defined information about the variant.

If samples are provided, the ninth column describes the format of the sample-level information, and then data for each sample are displayed as additional columns, one column per sample. Typically, the sample columns contain the genotype (GT) of the individual for the variant, the depth of coverage (DP) at the variant site and the genotype quality score (GQ) which represents the confidence of the genotype call. For diploid chromosomes, the genotypes may be homozygous for the reference allele (represented by "0/0"), heterozygous ("0/1" or "1/0"), homozygous for the alternate allele ("1/1"), or missing ("./."). A variant may have multiples alleles, which is reflected in the genotype. For example, a genotype of 0/2 means an individual is heterozygous for the second allele in the allele column. As with FASTQ files, VCF are typically compressed with `bgzip` and indexed with `tabix`.

2.1.4 FAM

A FAM file describes information about individuals in a cohort, and any family information that may be present. Each line represents one individual, and the file contains six columns: the family ID, the individual ID, the ID of the father, the ID of the mother, the sex, and the phenotype code. Unknown values for the ID of the parents or the sex are coded as a "0", and unknown phenotype values are usually coded as "-9". Using this information, a pedigree diagram for a family may be constructed, and the pairwise relatedness between any two individuals in the file may be estimated. Note that multiple families may be present in the one file.

2.2 SNV and Indel Calling

2.2.1 Read Alignment and Post-Processing

The read alignment and post-processing of WGS data were mostly performed at Edinburgh Genomics, Clinical Genomics (EGCG). This pipeline was broadly based on the well-known genome analysis toolkit (`GATK`) "Best Practices" v3 (Van der Auwera et al., 2013), with some modifications for speed and optimisation. However, where raw data was obtained directly (either from collaborators or downloaded from public resources), this pipeline was applied on local servers. To ensure compatibility, the BAM and FASTQ files obtained from EGCG were examined, and identical parameters were used for the local instance of the pipeline. An overview of this pipeline is shown in Figure 2.1. Source code for the alignment process is available online (see "Web Resources", Subsection A.2.1). Either the GRCh38 reference genome (including decoy, HLA and alternative contigs, GenBank accession: GCA_000001405.15) or the GRCh37 reference genome (GenBank

accession: GCA_000001405.14) was selected, depending on the analysis.

The alignment and post processing consisted of the following steps:

1. pairs of FASTQ files were aligned to the chosen reference genome using the `BWA-MEM` algorithm (H. Li, 2013).

2. the chromosomes were re-ordered and the reads within the chromosomes were sorted using the `ReorderSam` and `SortSam` modules from `picard` respectively (see "Web Resources", Subsection A.2.1).

3. the BAM file was validated with `ValidateSamFile` module from `picard` to ensure that there were no errors with the read alignment or file formatting.

4. PCR duplicates from the sequencing process were marked. BAM files processed by EGCG had duplicates marked with `samblaster` (Faust & Hall, 2014), whereas BAM files processed locally had duplicates marked using the `MarkDuplicates` module from `picard`. Both tools perform comparably, but `samblaster` is optimised for speed.

5. local re-alignment around indels was performed using `GATK` v3.4. This step is unnecessary when later versions of `GATK` are used but was retained for compatibility with data from EGCG.

6. base quality score recalibration (BQSR) was performed to correct for potential errors in the sequencing chemistry and platform using the `BaseRecalibrator` module from `GATK`. The error rates before and after adjustment are plotted by the `AnalyzeCovariates` module of `GATK`, and the adjustments are applied by the `PrintReads` module.

7. the BAM file was validated once more using the `ValidateSamFile` module from `picard`, since this is the final stage of read post-processing.

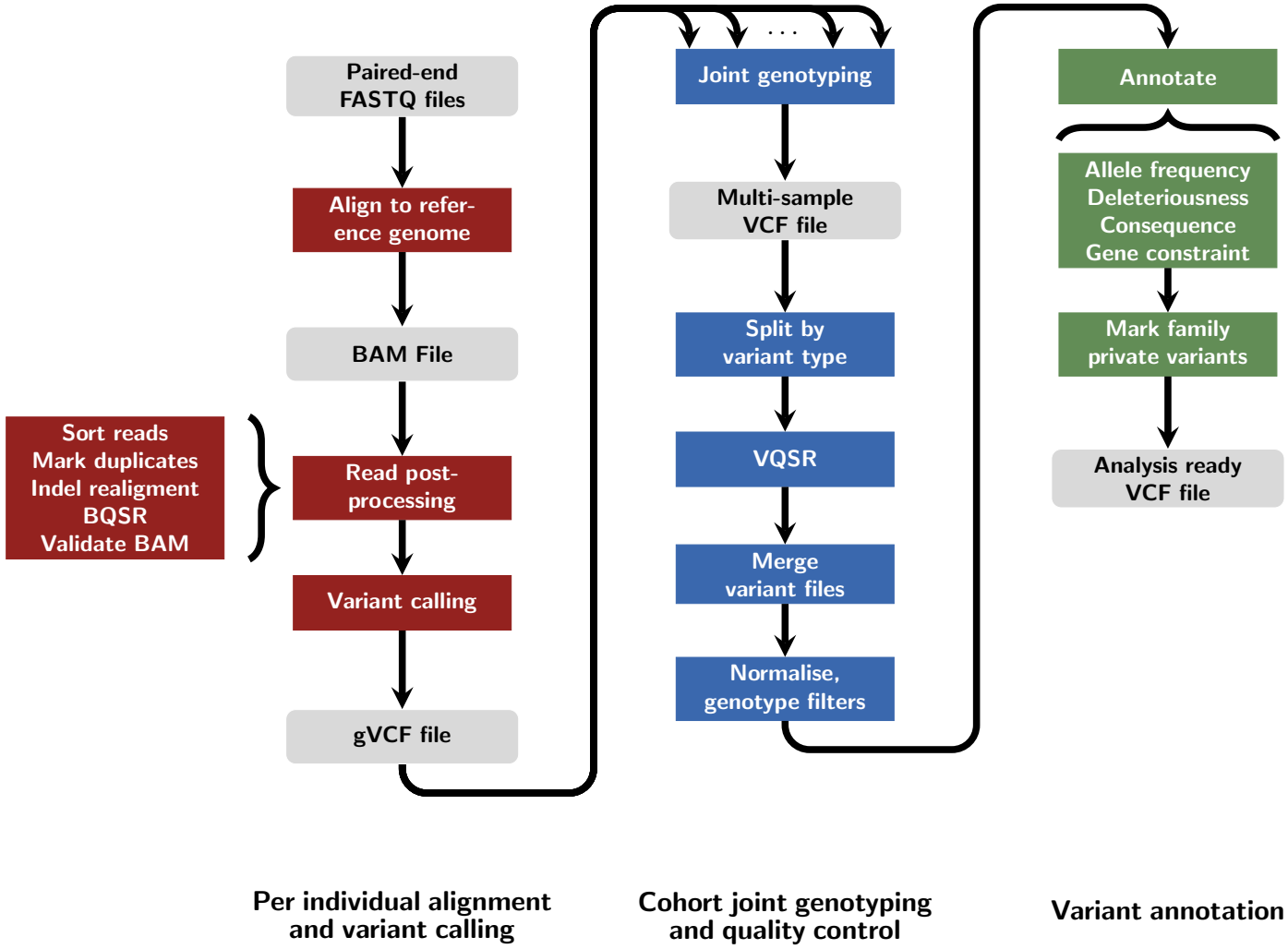


Figure 2.1: Flowchart of the variant calling, genotyping, and annotation pipelines. BQSR: base quality score recalibration; BAM: binary alignment map; gVCF: genomic variant call format; VQSR: variant quality score recalibration.

2.2.2 Variant Calling and Joint Genotyping

Once read alignment and post-processing was performed, variants were called using the `HaplotypeCaller` module from `GATK`. GVCF mode was selected, whereby all positions in the genome are evaluated for having a variant or not, so likelihoods for a site being homozygous reference can be calculated. Instead of a standard VCF file, a genomic VCF (gVCF) file is produced with records for non-variant sites as well as candidate variant sites. For convenience, neighbouring non-variant records in the gVCF file are combined so intervals may be represented as a single record (or block) in which the genotype likelihoods are binned. After variants were called with `HaplotypeCaller` in GVCF mode, genotypes for all samples were assigned jointly using the `GenotypeGVCFs` module from `GATK` v3.8. A newer version of `GATK` was used for this and subsequent steps, since v3.4 had a known bug and was not able to process spanning deletions (i.e. sites where one sample had a deletion, but another sample had an SNV within the deleted region).

2.3 Variant Quality Control

2.3.1 Variant Quality Score Recalibration

After genotyping, variants on the standard 23 pairs of chromosomes were retained. As recommended from the `GATK` "Best Practices", variants whose depth of coverage was greater than five standard deviations above the average coverage across all sites were removed (see "Web Resources", Subsection A.2.1). To remove low-quality variants, variant quality score recalibration (VQSR) is applied to calculate a new metric, the variant quality score, logarithm of odds (VQSLOD) (Van der Auwera et al., 2013). Variants were split by type using the `SelectVariants` module from `GATK`. VQSLOD scores were calculated by the `VariantRecalibrator` module and annotated by the `ApplyRecalibration` module, both from `GATK`. The recommended VQSLOD tranche thresholds are 99.9% for SNVs and 99.0% for indels. For variants which are neither SNVs nor indels (spanning deletions, multi-nucleotide variants, etc.), the following hard filters (recommended for indels from the `GATK` "Best Practices") were applied:

```
QD < 2 || FS > 200 || SOR > 10 || ReadPosRankSum < -20
```

SNVs, indels and other variants were then merged with `CombineVariants` and the VCF file was validated with `ValidateVariants`, both from `GATK`. At this stage, the sample IDs within the VCF files were assessed for consistency and where necessary samples were renamed with `bcftools reheader` (Danecek et al., 2021). Source code

for the joint genotyping process and VQSR is available on GitHub (see "Web Resources", Subsection A.2.1)

2.3.2 Genotype-Level Metrics

During joint genotyping of multiple samples, two different variants may be present at the same site, e.g. two alternate alleles of an SNV, or an SNV in one sample and an indel in another. It is often useful to separate these variants so that they may be processed independently, especially if allele-specific information is being used. To do this, we used the `norm` module from `bcftools`, which separates out the genotype-specific metrics for each alternate allele. Following the split of multi-allelic sites, we required that each sample had a minimum of evidence for a genotype to be correctly called. We required that there were at least 10 reads supporting the genotype, or `DP` $\geqslant 10$. Also, we required that the Phred-scaled GQ score was at least 20. This means that the likelihood for the genotype was at least 100 times greater than the likelihood for the next most likely genotype. These filters were applied using the `VariantFiltration` module from `GATK`. Genotypes which failed these filters were set to missing using the `SelectVariants` module from `GATK`.

2.4 Pedigree Consistency

The software `peddy` was used to check the consistency of the pedigree information with the genetic data (Pedersen & Quinlan, 2017). The following quality control measures are examined: expected versus observed relatedness (by the `KING` algorithm (Manichaikul et al., 2010)), predicted sex concordance, median depth of coverage, and genomic ancestry clustering prediction following principal component analysis (PCA). `peddy` has known bugs when calculating the expected relatedness from complex pedigree structures, such as when half-siblings or consanguinity are present (see "Web Resources", Subsection A.2.1). In this instance, the observed relatedness scores are calculated using `vcftools` (Danecek et al., 2011) which also implements the `KING` algorithm, and the expected relatedness scores are calculated with the `kinship2` package from `R` (Sinnwell et al., 2014).

2.5 Public Databases and Resources

2.5.1 Allele Frequency Databases

The following allele frequency databases are used throughout the thesis:

- The 1000 Genomes Project (Phase III) examines NGS and SNP genotype data on 2,504 individuals from five genomic ancestry groups: African (AFR), admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) (Auton et al., 2015).

- The genome aggregation database (gnomAD) includes population-level allele frequencies for a range of super- and sub-populations of genetic ancestry (Karczewski et al., 2020). Version 2.1.1 of gnomAD is considered by the authors to be the preferred version for analysis of protein-coding regions due to the large sample numbers (125,748 exomes). Version 3.1 of gnomAD is composed of 76,156 WGS samples and so is preferred for examining non-coding variants. Additionally, a collection of structural variants is compiled for 14,891 individuals (Collins et al., 2020).

- The Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) collates a list of CNVs and their allele frequencies in the general population (Firth et al., 2009).

- The Database of Genomic Variants (DGV), a similar project which aims to catalogue structural variants from healthy controls (MacDonald et al., 2014).

2.5.2 Variant Deleteriousness Metrics

The functional interpretation of genetic variants can be challenging, even for variants in protein-coding genes. If we have not observed a variant in an individual (or many individuals) with a phenotype, we have no evidence to implicate that variant with the phenotype. Various scores (known as deleteriousness metrics) aim to predict how damaging a variant is or how likely a variant is to be implicated in diseases or disorders in general. These scores can be used to remove variants that are unlikely to be disease-causing candidates. Some examples of commonly used metrics are described below and are used in aspects of the work in this thesis.

Sorting Intolerant From Tolerant (SIFT) scores can be calculated for all non-synonymous variants and are based on the prediction of whether the amino acid substitution will affect protein function or not (Ng & Henikoff, 2003). Given a query protein, related proteins are examined to identify if similar amino acid substitutions are observed in protein sequence databases. Substitutions not observed are assumed to be selected against, and so variants resulting in these amino acid substitutions are predicted to be deleterious.

SIFT generates a score from 0 to 1, with smaller scores representing deleterious variants.

Polymorphism Phenotyping v2 (PolyPhen2) scores can also be calculated for all non-synonymous variants and are based on amino acid substitutions (Adzhubei et al., 2013). A naïve Bayes classifier is used to predict the functional importance of a variant from a set of sequence-based as well as protein-structure-based features. Two datasets are used to train the model (HumDiv and HumVar), which give varying degrees of performance accuracy. PolyPhen2 generates a score from 0 to 1, but the scale is in the opposite direction to SIFT; variants with high scores are predicted to be damaging.

Combined Annotation Dependent Depletion (CADD) is a machine learning model built using over 60 genomic features, including deleteriousness, conservation, genetic functional consequence, epigenomic modification, etc. (Rentzsch et al., 2019). It is trained on a set of evolutionarily simulated variants rather than curated benign/pathogenic sets, which allows all positions in the genome to be scored for a given alternate allele. This is especially useful for considering non-coding variants. CADD reports raw scores as the output of the penalised logistic regression model, but more frequently used are the ranked, Phred-transformed C-scores. A CADD C-score of 20, for example, indicates that a variant is in the top 1% of all variants when ranked by the raw CADD score. The C-scores range from 0.001 to 99, with larger scores being more deleterious.

The Missense badness, PolyPhen2 and missense Constraint (MPC) score combines three measures of deleteriousness for missense variants (Samocha et al., 2017). Instead of assuming a uniform distribution of observed/expected missense variants, a transcript of a gene may be split into segments where missense variants are enriched/depleted. This identifies regions of the transcript that are constrained for missense variants. Additionally, all potential amino acid substitutions across overlapping transcripts for a given variant may be evaluated, and the "badness" score represents the fold enrichment of amino acid substitutions in constrained versus unconstrained regions. Both missense constraint and missense badness were combined with the PolyPhen2 score as a composite predictor of deleteriousness. The MPC ranges from 0 to 5, with higher scores being more deleterious.

A useful resource for the above metrics is the database of non-synonymous functional prediction (dbNSFP) which collates transcript-specific information on all potential non-synonymous SNVs, over 84 million variants (Liu et al., 2020). Included is a wide range of variant-level information, including many deleteriousness predictors. Position information is given for both the GRCh37 and GRCh38 reference genomes.

2.5.3 Gene Constraint Scores

Due to negative selection, deleterious variants are expected to occur less frequently in the genome than benign variants (Lek et al., 2016). The difference between the observed and the expected number of variants in a gene can indicate how tolerant that gene is to different variant categories. Lek et al. quantified this deviation from the expected number of protein-truncating variants which results in a probability of loss-of-function intolerant (pLI) score. This gene-based constraint score ranges from 0 to 1, and the authors classify genes with pLI $>$ 0.9 as highly intolerant to loss-of-function variants. Karczewski et al. calculated a modified version of this from gnomAD called the oe score (Karczewski et al., 2020). The loss-of-function observed/expected upper-bound fraction (loeuf) is recommended by the authors as a measure of gene constraint. This loeuf score should be used as a continuous metric between 0 and 1, but the authors suggest that genes with loeuf $<$ 0.35 may be considered constrained. The gene-constraint scores described here are also included in dbNSFP.

2.6 Variant Annotation

2.6.1 `vep`

Variants were annotated with external databases and resources using the variant effect predictor (`vep`) (McLaren et al., 2016). Since all gene-based information queried by `vep` is specific to individual transcripts, care must be taken when variants overlap multiple transcripts, or even multiple genes. The "`--per_gene`" flag selects one transcript per gene (determined by a pre-defined hierarchy, typically the canonical transcript) and reports one annotation report ("consequence") per overlapping gene. In addition to the default resources, other databases can be supplied to `vep` for annotation. In particular the following information was manually supplied: allele frequencies from gnomAD v2.1.1 or v3, CADD v1.6 scores and functional prediction metrics from dbNSFP v4.1.

2.6.2 `SnpSift`

We are often interested in examining variants that may only be present in a particular pedigree out of a jointly genotyped cohort. We annotated these variants with the `private` module from `SnpSift` (Cingolani et al., 2012). `SnpSift` takes a VCF file and a FAM file as input and outputs a VCF file with a "Private" tag annotated for each record that is only found in one pedigree.

Chapter 3

Converting Single Nucleotide Variants between Genome Builds

NGS studies require a high-quality reference genome for SNV calling. Although the two most recent builds of the human genome are widely used, position information is typically not directly comparable between them. Re-aligning positions to a particular genome build is computationally expensive, and so tools are used to convert data from one build to another. However, the positions of converted SNVs do not always match SNVs derived from aligned data and in some instances, SNVs are known to change chromosome when converted. In this chapter, we describe a novel algorithm to identify positions that are unstable when converting between human genome reference builds. These positions are detected independent of the conversion tools and are determined by the chain files. Pre-excluding SNVs at these positions, prior to conversion, results in SNVs that are stable to conversion. This work has been published (Ormond et al., 2021) and is included in Appendix C for reference.

3.1 Introduction

The human reference genome is fundamental to genome assembly and variant calling for NGS studies (Church et al., 2011; Guo et al., 2017). Without a reference, *de novo* assembly of each sequenced genome would need to take place, which is computationally intensive and in certain scenarios may result in a poor quality assembly (Treangen & Salzberg, 2011). The current builds of the human reference genome (GRCh37 and GRCh38) are the most widely adopted builds for genomic analysis. However, further iterations are inevitable as GRCh38 also contains a much larger collection of unlocalized (known sequence and chromosome but position unknown) and unplaced (known sequence, but chromosome and position unknown) contigs, as well as including alternate contigs (known alternate representations of specific regions of the genome to account for population differences) (Schneider et al., 2017). Different builds result in different genome assemblies which will impact downstream analysis of genomic variants (Guo et al., 2017).

Each genome build brings improvements, but updates to the base pair coordinates mean that not all positions are comparable between builds. Because a wealth of annotation information is available for GRCh37 and many pipelines and tools are still based on this older version (Guo et al., 2017), researchers are sometimes hesitant to switch to the newer version. Where a newer build is adopted, a similar problem arises when trying to compare new sequences to data aligned to an older build: both data sets must be aligned to the same build to be comparable. Obviously, re-aligning sequence data to the newest build will typically provide the most accurate base pair position information, but this can be quite computationally expensive (Guo et al., 2017). Also, the raw sequence data required for alignment, if available, can be large, so long-term storage may not be feasible.

An alternative approach to re-alignment is to convert between genome builds using tools such as `liftOver` (provided as part of the Genome Browser tool (Haeussler et al., 2019) hosted by the UCSC Genomics Institute), `CrossMap` (H. Zhao et al., 2014) or `Remap` (hosted by the National Centre for Biotechnology Information (NCBI) (Agarwala et al., 2018)). This process is aided by a chain file, which provides a mapping of contiguous positions from one build to another. The ability to convert between builds using these tools has proved vital, allowing the integration of a wide range of SNV annotation databases and sequence data, regardless of how they were originally aligned, for example gnomAD (Karczewski et al., 2019), CADD (Rentzsch et al., 2019) and dbNSFP (Liu et al., 2016).

For those who do choose to convert between GRCh37 and GRCh38, there are known problems with this conversion process, particularly for SNVs. In the online user guide for the UCSC Genome Browser, the authors note that "*occasionally, a chunk of sequence may be moved to an entirely different chromosome*" (see "Web Resources", Subsection A.2.2). This is echoed in Liu et al., where the authors note that after converting the dbNSFP database to other builds using `liftOver`, "*there are a few SNVs whose coordinates in hg38 and hg19 . . . have inconsistent chromosome numbers*" (Liu et al., 2016). This phenomenon can inevitably prove problematic for downstream analyses. Taking a real world example, suppose we wish to examine variants in protein-coding regions of the genome, prioritised using CADD scores. Consider the T>A missense substitution at position 15690247 on chromosome 22 of GRCh38 (chr22:c.15690247T>A), contained in the first exon of *POTEH*. CADD v1.6 gives the variant a C-score of 20.8, indicating that it is in the top percentile of all ranked deleterious variants. If we convert the position to GRCh37 (using either `liftOver` or `CrossMap`), this variant maps to position 19553586

on chromosome 14, where the reference allele is still T (chr14:c.19553586T>A) but the variant is now in the first exon of *POTEG*. CADD v1.6 for GRCh37 gives this variant a C-score of 0.009, indicating that it is now in the bottom percentile of all ranked deleterious variants in the genome.

Pan et al. (2019) examined SNVs from data aligned under a range of bioinformatics pipelines to data converted between GRCh37 and GRCh38 using both `liftOver` and `CrossMap` (Pan et al., 2019). The authors noted that on average 1% of SNVs did not convert from GRCh37 to GRCh38, and an average of 5% of SNVs did not convert from GRCh38 to GRCh37. Furthermore, on average 1.5% of SNVs which were successfully converted were not found in the corresponding aligned data, a trend that was more pronounced when converting from GRCh38 to GRCh37. Such discordant sites were noted to be low-confidence calls, have lower average read depth, and have a higher than average GC content. The authors urged caution when converting SNVs between builds.

Recently, Luu et al. (2020) benchmarked six tools (including `liftOver`, `CrossMap` and `Remap`) for converting multi-base pair regions derived from epigenetic data from GRCh37 to GRCh38 (Luu et al., 2020). The authors found a high degree of correlation between the six tools but noted that gapped regions in both chain files can result in conversion failure, or even regions mapping to incorrect locations. A guideline to improve conversion is offered, which involves removing input data which overlap with the gapped regions, as well as removing input data which map to multiple regions or alternate contigs. However, if this strategy were applied to SNV data, some variants may not necessarily be removed, such as those in un-gapped regions which also change chromosome under conversion.

Here we present a novel algorithm to identify base pair positions in the human genome which exhibit unstable behaviour when converting between genome reference builds. In addition, we are providing the list of these unstable positions for the two most recent builds (GRCh37 and GRCh38) on GitHub (see "Web Resources", Subsection A.2.2). This list can be used to pre-exclude SNVs prior to conversion to remove potentially problematic variants, resulting in stable SNVs and improving the quality of sequencing data post-conversion.

3.2 Identification of Unstable Positions

3.2.1 Chain File

A chain file provides a mapping of the analogous positions from one genome build to another. Given a sequence of DNA on both builds, it details the count (and hence position) of contiguous bases where the sequences match and allows for gaps to be present in either build. A visual depiction is given in Figure 3.1 below. Chain files may only be used in one direction i.e. from the source build to the target build. When creating a chain file, tools such as `BLAST` (Altschul et al., 1990) or `BLAT` (Kent, 2002) are used to ensure that the overall matching regions have a sufficiently high proportion of matching base pairs, known as sequence identity (e.g. at least 98%, see "Web Resources", Subsection A.2.2). This allows for a small number of differences in the sequence between matched regions, which may arise due to errors in the genome build being corrected in minor patches. Chain files mapping between GRCh37 and GRCh38 (one for each direction) were obtained from the `liftOver` website hosted by the UCSC Genomics Institute (see "Web Resources", Subsection A.2.2), since these files were recommended by the selected conversion tools.

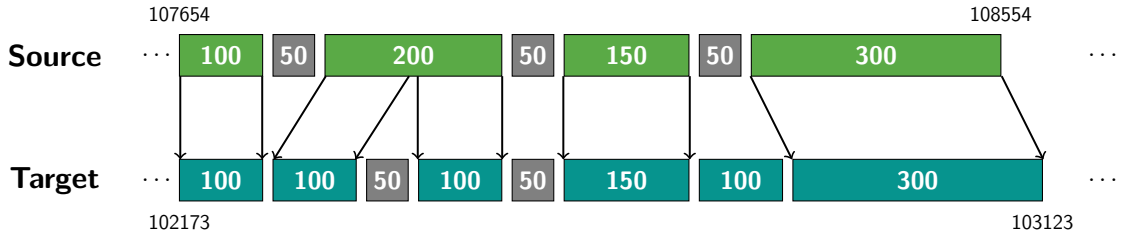


Figure 3.1: A visual depiction of a chain file showing the pairwise mapping between the source and target, allowing for gapped regions in either build (grey blocks). Identifying the contiguous bases and gaps allows for positions in one build to be mapped to another.

3.2.2 Full-Genome Data

Genome build conversion tools use base pair position information only, so it is possible to examine the stability of all base pair positions in the genome. This allows the behaviour of all potential SNVs to be examined when converting between builds, rather than just a subset that might be found on an individual sample's genome. To this effect, browsed extensible data (BED) files were created containing an entry for each base pair position in both the GRCh37 and GRCh38 reference genomes, which we refer to as the full-genome

data. This includes positions that are not typically amenable to short-read WGS, such as known gaps in the genome assembly. Positions on the unplaced, unlocalized and alternate contigs were not included in the input data, and so only the standard 23 pairs of chromosomes were considered. The mitochondrial chromosome was excluded since variant calling on the mitochondrial chromosome often uses a separate reference genome (Bandelt et al., 2014). Each entry in the input BED file was given a label containing the original chromosome and start position for unique identification, and the file was split by chromosome for parallelisation (Tange, 2011). This generated 3,095,677,412 positions for GRCh37 and 3,088,269,832 positions for GRCh38.

3.2.3 Algorithm to Identify Novel Conversion-Unstable Positions

To identify base pair positions that are unstable in the conversion process (defined below), each input file was converted from the source build to the target build and then back to the source build again (see Figure 3.2 below). Entries in the output files were extracted if they satisfied one of the following conditions:

- positions which failed on the first conversion ("Reject_1");
- positions which mapped to a different chromosome on the first conversion ("CHR_Jump_1");
- positions which failed on the second conversion ("Reject_2");
- positions which did not map back to the original chromosome on the second conversion ("CHR_Jump_2"); and
- positions which did not map back to the original position on the second conversion ("POS_Jump")

We refer to these collectively as conversion-unstable positions (CUPs), and all other positions are referred to as stable. Note that entries in the Reject_1 category are typically identified by the conversion tool, so the latter four entries are what we refer to collectively as novel CUPs. Reject_1 and CHR_Jump_1 positions were removed prior to the second conversion (from the target build back to the source build). Despite not being included in the input data, entries that mapped to the unplaced, unlocalized, and alternate contigs were retained in the CHR_Jump_1 and CHR_Jump_2 categories to ensure each base pair position in the source build had an accurate category designation. Both `liftOver` and `CrossMap` were used for the conversion (see "Web Resources", Subsection A.2.2). `Remap` was not considered as its input file is limited to 250,000 entries, which is much

smaller than the lengths of the input chromosomes. The same chain files were used by both `liftOver` and `CrossMap`, allowing us to also check the robustness of CUP identification, as a consensus between tools would give higher confidence in the output. This algorithm was run twice, once for the GRCh37 build as the source and once for the GRCh38 build as the source.

Both `liftOver` and `CrossMap` gave identical output for the same input data (see Table 3.1). On GRCh37, approximately 11.3Mbp of novel CUPs were identified (representing 0.37% of the build) and on GRCh38 20Mbp of novel CUPs were identified (0.65% of the build). For both builds, a successive application of the algorithm on the stable positions using either tool did not identify any additional base pair positions for any of the CUP categories, as expected.

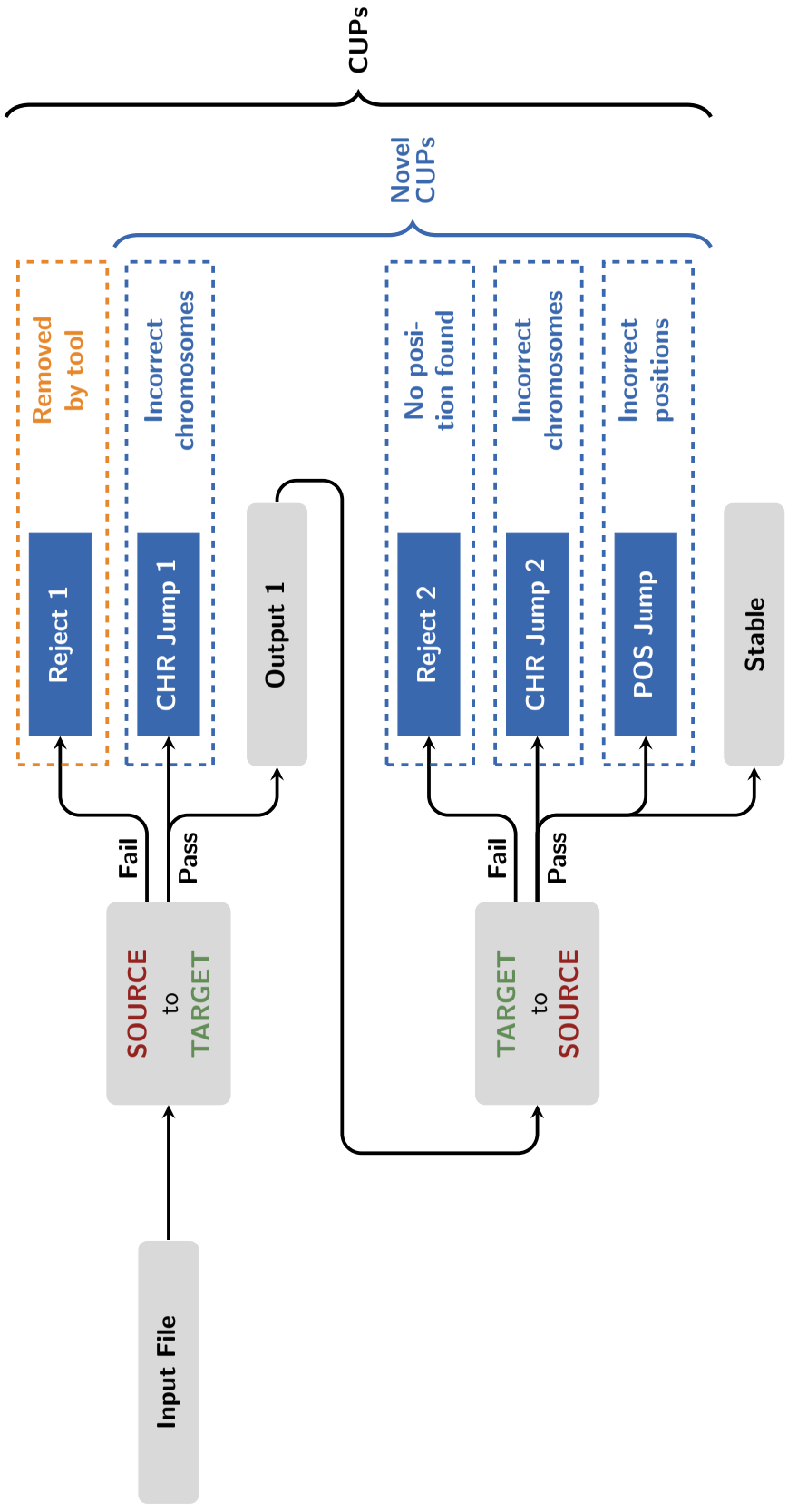


Figure 3.2: Flow chart of the algorithm to identify novel conversion-unstable positions.

Category	GRCh37 to GRCh38 (bp)	% of Source	GRCh38 to GRCh37 (bp)	% of Source
All	3,095,677,412	100.000	2,859,470,792	92.370
Reject_1	234,712,067	7.582	-	-
CHR_Jump_1	1,494,553	0.048	-	-
Reject_2	-	-	100,180	0.003
CHR_Jump_2	-	-	799,922	0.026
POS_Jump	-	-	8,907,439	0.288
Stable	2,859,470,792	92.370	2,849,663,251	92.053
Novel CUPs	-	-	11,302,094	0.365

(a)

Category	GRCh38 to GRCh37 (bp)	% of Source	GRCh37 to GRCh38 (bp)	% of Source
All	3,088,269,832	100.000	2,862,067,878	92.675
Reject_1	218,510,733	7.076	-	-
CHR_Jump_1	7,691,221	0.249	-	-
Reject_2	-	-	73,770	0.002
CHR_Jump_2	-	-	292,083	0.009
POS_Jump	-	-	12,038,774	0.390
Stable	2,862,067,878	92.675	2,849,663,251	92.274
Novel CUPs	-	-	20,095,848	0.651

(b)

Table 3.1: Details of the stable positions and conversion-unstable positions (CUPs) for the full-genome data for **(a)** GRCh37 as the source and **(b)** GRCh38 as the source, including the number of base pairs (bp) for each category, and the proportion of the source genome build covered (%). Novel CUP category names are highlighted in green.

3.2.4 Comparison with Assembly Annotation Sets

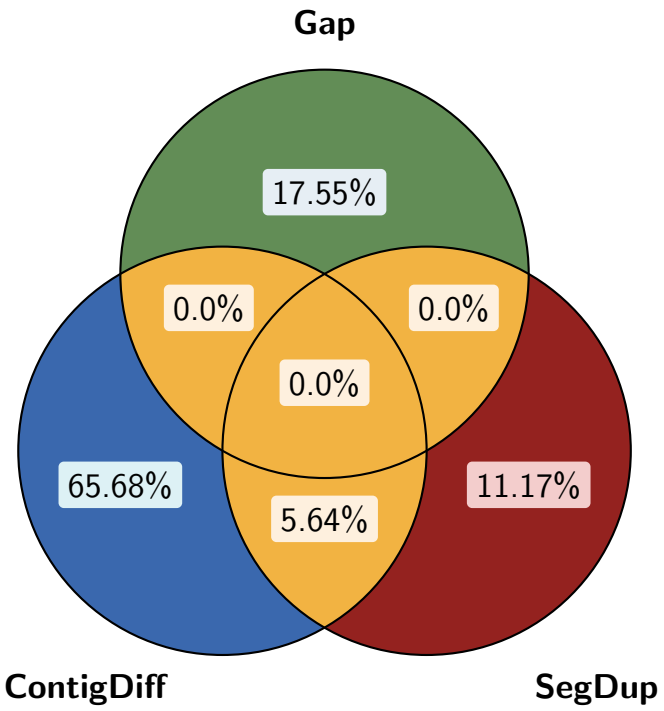
To better understand the possible reasons for CUPs occurring, we also identified where these positions originated in the genome. Given the reconstruction of some contigs in the development of GRCh38 (Schneider et al., 2017), one explanation for base pair positions being rejected during a conversion is that the position is not in the target build. Additionally, in the online support forum for the UCSC Genome Browser, it is noted that variants may change chromosomes between builds because they lie in repetitive regions or segmental duplications (see "Web Resources", Subsection A.2.2). In an attempt to isolate the source of each CUP, the following assembly annotation sets were obtained from the UCSC Table browser (Haeussler et al., 2019) for both genome builds (table ID given in brackets):

- gaps in the build (gap): regions that are not present in the build, including telomeres, the short arms of specific chromosomes and gaps between known contigs. The centromeres are present in the GRCh37 gap set (as they did not form part of the assembly) but not in the GRCh38 gap set and so were removed from the GRCh37 gap set prior to comparison.

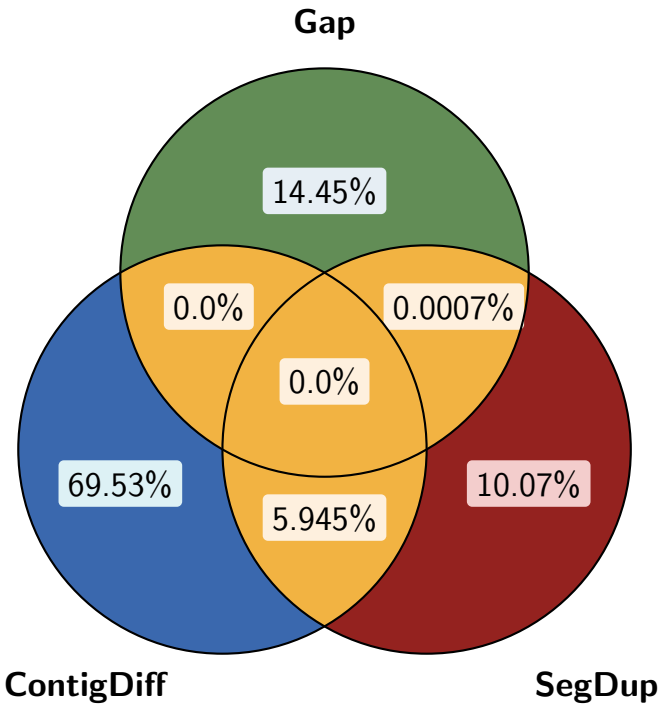
- differences between contigs (hg38ContigDiff): regions that are different in the GRCh38 and GRCh37 builds due to updates in individual contigs.

- segmental duplications (genomicSuperDups): regions longer than 1kb that have a high degree of sequence identity with other regions.

Given the overlap between these sets, positions unique to each of the three sets, as well as positions which were present in more than one set (multiple), or no set (other) were considered (see Figure 3.3). It is worth noting that the "multiple" set on GRCh37 was composed entirely of the intersection between the contig differences and segmental duplications. The same was virtually true for GRCh38, with a very small proportion (0.0007%) arising from the intersection between the gaps and segmental duplications. In both cases, the "multiple" set accounted for less than 6% of all positions in the selected assembly annotation sets. For the CUPs identified above, contiguous entries were collapsed into multi-base pair regions using `bedtools` (Quinlan & Hall, 2010), to allow for quicker comparison with the assembly annotation sets. The proportion of overlap in CUP category A of assembly annotation set B is defined as $\frac{|A \cap B|}{|A|}$, and was computed using `bedtools`.



(a)



(b)

Figure 3.3: Venn diagrams for the overlap between the three selected assembly annotation sets (green: gaps; blue: contig differences; red: segmental duplications; yellow: multiple sets) for **(a)** GRCh37 and **(b)** GRCh38.

For both builds, the proportion of overlap for each CUP category across all the assembly annotation sets was at least 97.5% for all except the Reject_1 category on GRCh37, where the proportion was 69.2% (see Figure 3.4). However, the centromeres that were removed from the gap set (which do not overlap with the other assembly annotation sets) account for an additional 29.4% of the Reject_1 category, giving a total overlap proportion explained of 98.6%. The CUPs in the "Other" set for all categories were examined using the UCSC Genome Browser (Haeussler et al., 2019), but there was no consistent overlap between these positions and any other assembly annotation track.

For both builds, the Reject_1 category is dominated by the gap and contig differences sets. This is a highly plausible explanation for these base pair positions as the conversion tools will fail when regions of the genome are not present (or have been updated) in the target build. For example, the centromeres were broadly reconstructed during the assembly of GRCh38, so it is not surprising that they feature in the Reject_1 category on GRCh37. The novel CUPs are largely composed of the intersection between the contig differences and segmental duplications. If a region is contained in both a segmental duplication and a contig difference, this may indicate that the region is better placed in another part of the genome, which would explain the conversion instability.

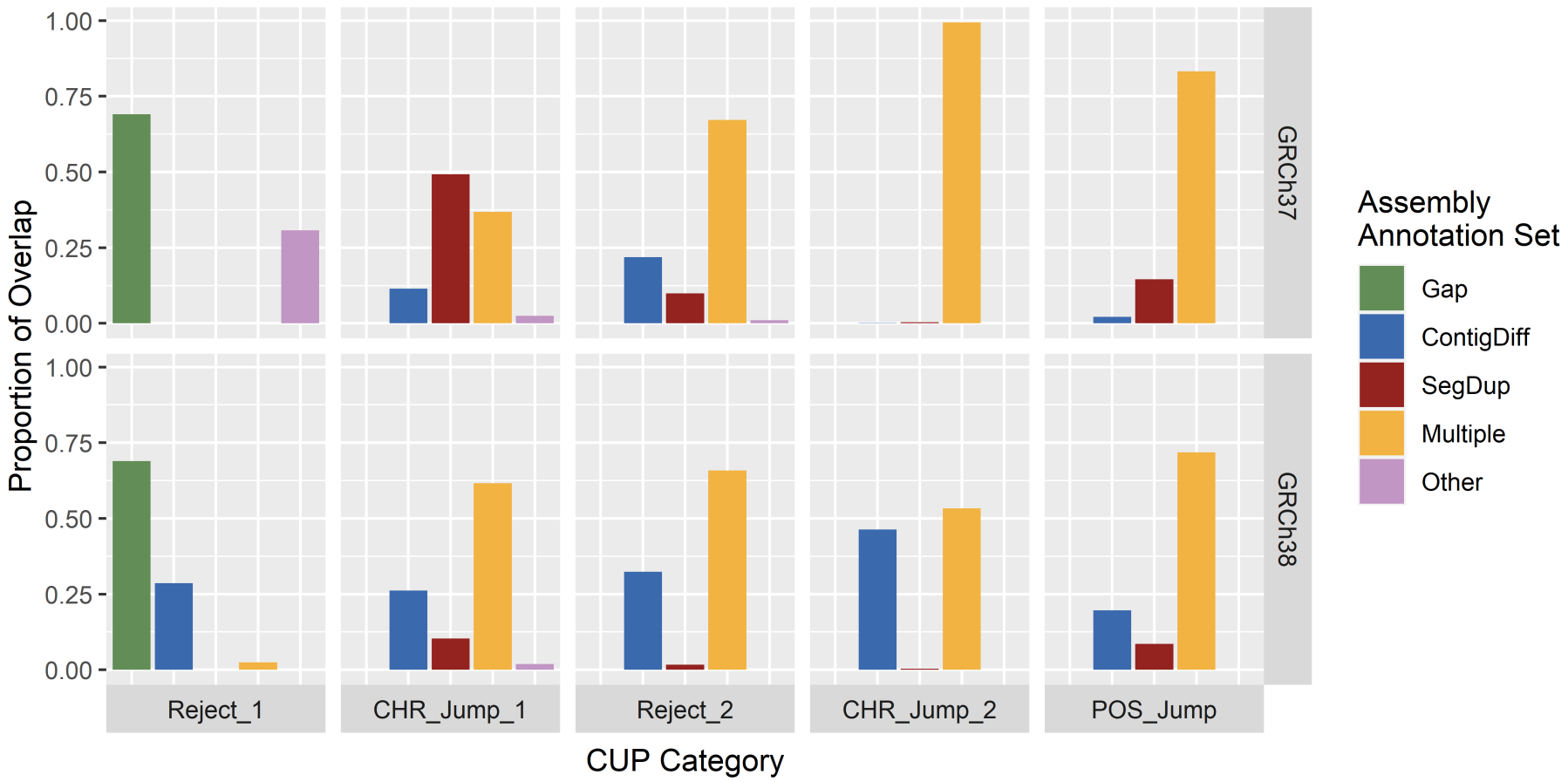


Figure 3.4: The proportion of conversion-unstable positions (CUPs) that overlap with the assembly annotation sets, for the GRCh37 (top panel) and GRCh38 (bottom panel) builds. Here, "Multiple" represents positions present in one or more of the assembly annotation sets and "Other" represents positions present in none of the assembly annotation sets (this includes the centromeres for GRCh37). Gap: gaps in the assembly; ContigDiff: differences in contigs between builds; SegDup: segmental duplications.

3.3 Application to WGS Data

3.3.1 Evaluation Data

As a proof of principle, the well-characterized NA12877 and NA12878 samples for the CEPH 1463 family were used to examine the behaviour of SNVs from WGS data when converting between builds. High confidence variant calls for both samples were obtained from the publicly available Illumina Platinum Genomics project in VCF format on both GRCh37 and GRCh38 (see "Web Resources", Subsection A.2.2). These variants were identified using several variant calling algorithms, and were validated using the genotypes in the samples' parents and children to remove Mendelian errors (Eberle et al., 2017). As we are only considering the behaviour of SNVs and aim to compare the WGS data with the full-genome data, only biallelic SNVs were extracted for both samples. Each variant was given a unique ID containing the original aligned position, reference allele, alternate allele, and source build for ease of identification.

A slightly modified version of the above algorithm was implemented using the `LiftoverVcf` module from `picard` rather than `liftOver`, as `liftOver` does not handle VCF file format. The `LiftoverVcf` module is based on `liftOver` but additionally checks the reference allele of each variant with the target reference genome, removing any sites where there is a mismatch. `CrossMap` can accommodate VCF file format, and updates the reference allele to that of the target build where there is a discrepancy and returns a failure if the alternate allele is the same as the updated reference allele on the target build. If a reference allele was updated to an ambiguous base (denoted by International Union of Pure and Applied Chemistry (IUPAC) codes), these were manually removed and considered a mismatch. For the WGS data, two additional output categories were included for variants which failed due to reference-allele mismatches on the first conversion (Mismatch_1) or on the second conversion (Mismatch_2).

Since individual base pair positions are converted independently of one another, variants which are present in any of the novel CUPs can also be excluded prior to conversion to ensure all variants are stable and data are of high quality. These filtered data were compared with the output from the algorithm on the original data to confirm that both methods are equivalent. In addition to the VCF data files, BED files were generated using position information extracted from the VCF data. This allowed us to apply our original position-based algorithm (that used the `liftOver` and `CrossMap` tools) as a sanity check to ensure that both versions of the algorithm behaved the same.

3.3.2 CUPs in WGS Data

NA12877 had 3,518,008 SNVs on GRCh37 and 3,576,396 SNVs on GRCh38. NA12878 had 3,523,638 SNVs on GRCh37 and 3,594,064 SNVs on GRCh38. Each of these represent approximately 0.1% of the full genome data for their respective build. For both samples, the CUPs identified from the VCF data were contained within the CUPs identified from the corresponding BED data, as expected. The only positions from the VCF data that were not contained in the BED data were the mismatch categories. Furthermore, the CUPs identified from the BED positions from the WGS data were contained within the respective full-genome CUPs. `liftOver` and `CrossMap` broadly agreed on the CUPs derived from the VCF data, with differences arising purely due to how each tool treats the reference allele in the target build, including ambiguous bases (Mismatch_1, Mismatch_2).

The same stable SNVs were identified from the filtered data (variants at novel CUPs excluded) as for the original, unfiltered WGS data when the algorithm was applied to both (see Table 3.2). Also, the only variants removed by the algorithm from the filtered data were those in the Reject_1 and mismatch categories. As expected, no additional variants in the CUP categories were identified on a successive application of the algorithm to either the original data or to the filtered data. The SNVs at novel CUPs represented approximately 0.13% of SNVs on either build.

Pan et al. reported conversion failure rates for WGS data of on average 1% from GRCh37 to GRCh38 and 5% from GRCh38 to GRCh37, noting that the SNVs that failed tended to have much lower depth of coverage, and may represent false-positive variant calls (Pan et al., 2019). Here, we observe much lower tool conversion failure rates of 0.14% from GRCh37 to GRCh38 and 0.72% from GRCh38 to GRCh37 for the WGS data. We note that this dataset is a particularly clean and accurate set of SNVs (Eberle et al., 2017), which may account for the decrease in conversion failure rates compared to the previous study. However, the trend in performance is in the same direction; converting from GRCh37 to GRCh38 results in fewer conversion failures than GRCh38 to GRCh37. While Pan et al. showed that read depth and variant quality may have an impact on discordance rates, the variants examined here did not have this information available, thus we were unfortunately not able to assess these aspects of the novel CUPs.

Source	Category	liftOver				CrossMap			
		Original		Filtered		Original		Filtered	
		Count	%	Count	%	Count	%	Count	%
GRCh37	All	3,523,638	100.000	3,518,229	100.000	3,523,638	100.000	3,518,229	100.000
	Reject_1	4,947	0.140	4,947	0.141	4,947	0.140	4,947	0.141
	Mismatch_1	20,533	0.583	19,976	0.568	20,510	0.582	19,959	0.567
	Mismatch_2	128	0.004	0	0.000	123	0.003	0	0.000
	Novel CUPs	4,724	0.134	0	0.000	4,735	0.134	0	0.000
	Stable	3,493,306	99.139	3,493,306	99.292	3,493,323	99.140	3,493,323	99.292
GRCh38	All	3,594,064	100.000	3,588,396	100.000	3,594,064	100.000	3,588,396	100.000
	Reject_1	25,852	0.719	25,852	0.720	25,852	0.719	25,852	0.720
	Mismatch_1	16,772	0.467	15,741	0.439	16,740	0.466	15,726	0.438
	Mismatch_2	85	0.002	0	0.000	81	0.002	0	0.000
	Novel CUPs	4,552	0.127	0	0.000	4,573	0.127	0	0.000
	Stable	3,546,803	98.685	3,546,803	98.841	3,546,818	98.685	3,546,818	98.841

Table 3.2: Counts and proportions (%) of all SNVs present in WGS data for sample NA12878 broken down by genome build (GRCh37, GRCh38), conversion tool (`liftOver` or `CrossMap`) and whether the original or filtered data was considered. All novel conversion-unstable positions (CUPs) have been combined into one entry in the table (novel CUPs, highlighted in grey).

3.3.3 Discordance Rates Between Aligned and Converted Data

Given that re-aligned data is considered the most accurate method to derive variants, we compared SNVs from converted data to aligned data on both builds to evaluate the error rates of the conversion process. The position discordance rate (computed using `bedtools`) was defined as the proportion of SNVs in the converted data whose position did not match that of a variant in the aligned data. The genotype discordance rate (computed using `GenotypeConcordance` from `picard`) was defined as the proportion of SNVs in the converted data whose position matched a variant in the aligned data, but whose genotype did not match.

The combined position and genotype discordance rates were on average 3.07% when converting from GRCh38 to GRCh37, and 1.68% when converting from GRCh37 to GRCh38 (see Table 3.3). When variants in the novel CUPs were pre-excluded, these rates reduced to 2.97% and 1.61% respectively. This is higher than the average discordance rate observed by Pan et al. of 1.5%, however these rates are not directly comparable. The average discordance rate from Pan et al. is taken across all bioinformatics pipelines, across both builds and across both tools. Although Pan et al. do not provide the exact rates to compare, our discordance rates are broadly in line with those observed in their Figure 6A (Pan et al., 2019). As with the conversion failure rates, both this study and Pan et al. found converting from GRCh38 to GRCh37 yields higher discordance rates. We note that the genotype discordance rates are quite low at an average of 0.0011% for both builds (see Table 3.3). This indicates that when the position of a variant has been correctly converted, the genotype is also highly likely to be correct.

Finally, we examined the position and genotype discordance rates for SNVs at the novel CUP categories with the aligned data (see Table 3.3). The position discordance rates overall are much higher for variants in the novel CUP categories compared with the filtered data, with an average of 83.2% on GRCh37 and 61.2% on GRCh38. Similarly, the genotype discordance rates for variants at novel CUPs is higher than the filtered data, with an average of 1.4% on GRCh37 and 0.4% on GRCh38. These rates indicate that variants at CUPs are less likely to be identified in the aligned target build and give support to our recommended strategy of removing them prior to conversion.

Sample	Tool	Category	GRCh38 to GRCh37		GRCh37 to GRCh38	
			Pos Disc (%)	Geno Disc (%)	Pos Disc (%)	Geno Disc (%)
NA12877	liftOver	original	3.0694	0.0012	1.8119	0.0012
NA12877	liftOver	filtered	2.9754	0.0010	1.7369	0.0010
NA12877	liftOver	novel CUPs	83.3893	1.5965	60.5490	0.3661
NA12877	CrossMap	original	3.0703	0.0012	1.8123	0.0012
NA12877	CrossMap	filtered	2.9759	0.0010	1.7371	0.0010
NA12877	CrossMap	novel CUPs	83.4450	1.5965	60.6054	0.3661
NA12878	liftOver	original	3.0691	0.0012	1.5544	0.0013
NA12878	liftOver	filtered	2.9654	0.0011	1.4794	0.0011
NA12878	liftOver	novel CUPs	82.9226	1.1436	61.8269	0.4978
NA12878	CrossMap	original	3.0700	0.0012	1.5550	0.0013
NA12878	CrossMap	filtered	2.9658	0.0011	1.4798	0.0011
NA12878	CrossMap	novel CUPs	82.9851	1.1436	61.8749	0.4978

Table 3.3: Discordance rates between converted data and aligned data for position (Pos Disc) and genotype (Geno Disc), for both WGS samples, both conversion tools, and comparing original data, filtered data or variants at conversion-unstable positions (CUPs). The entries containing the novel CUPs are shaded in grey.

3.4 Conclusions

Here, we have replicated the previously observed phenomenon whereby a small proportion of SNVs change chromosome when they are converted to another genome build (Liu et al., 2016). Additionally, we have identified all novel sites where base pair position information does not behave as expected, or where a one-to-one mapping between positions on both builds is not present. The novel CUPs represent 0.37% of the GRCh37 build and 0.65% of the GRCh38 build. We have clearly highlighted the care that must be taken when converting between genome builds to ensure high quality data. Unless the user is familiar with the instabilities we have described, we recommend the simple strategy devised here of removing variants at novel CUPs to ensure high confidence data when converting SNVs between builds of the human genome.

Chapter 4

Copy Number Variant Calling for Family-Based Sequencing Studies

Calling CNVs from short-read WGS data remains an ongoing challenge, and no "Best Practice" guidelines exist. A commonly implemented approach is to take a consensus of multiple calling methods to derive high-quality calls. However, this limits the CNV discovery to the collective strengths of the chosen methods, which may have been selected arbitrarily. Here we discuss a novel consensus CNV calling pipeline for family-based WGS data. By taking relatedness information into account, we were able to recover CNV calls that would have been removed due to lack of a consensus. We benchmarked our pipeline using a curated "Gold Standard" call set and showed that our method performs well overall, and out-performs a comparable calling pipeline designed for family-based data. This work was formulated jointly with Dr Niamh Ryan unless otherwise specified.

4.1 Introduction

CNVs are a form of SV defined as a deletion or duplication of a region in a genome that spans at least 50 bp in size. An estimated 4.9-9.5% of the human genome contains a CNV (Zarrei et al., 2015) and much work has been done to examine the contribution of CNVs to both Mendelian and complex genetic disorders (Girirajan et al., 2011; Stankiewicz & Lupski, 2010; Weischenfeldt et al., 2013). Hybridization-based techniques such as array CGH and SNP microarrays have been used historically to detect and genotype CNVs (Alkan et al., 2011). However, such methods are highly dependent on the design of the hybridization probes and so are limited in the size of the variants that they can detect, as well as lacking the resolution to accurately detect their breakpoints.

NGS technologies can provide greater accuracy at CNV calling compared to previous methods (Zhou et al., 2018). Many computational approaches leverage discrepancies in read alignments to identify putative regions that exhibit copy number changes (M. Zhao et al., 2013). Paired end read (PR) or split read (SR) tools detect CNVs by examining where the paired-end reads are significantly different from the expected insert size for a

collection of reads, or if one read in a pair doesn't map properly with its mate (Pirooznia et al., 2015). An example of a duplication is shown in Figure 4.1, where we can see that the insert size of the reads at the breakpoints is much larger than that of reads away from the CNV breakpoints. Read depth (RD) tools examine the number of reads within a region, under the assumption that this is correlated with the copy number of that segment of DNA (Pirooznia et al., 2015). Significant increases or decreases in read depth can indicate the breakpoint of a duplication or deletion event respectively. Returning to Figure 4.1, there is a noticeable increase in the read depth between the breakpoints compared to the flanking regions.

Despite the wealth of software tools and methods, there are no "Best Practices" for CNV calling from NGS data. A commonly used strategy is to use multiple tools to call CNVs and accept a consensus of the tools, which can provide a level of validation (Friedrich et al., 2020; Zarate et al., 2020). Despite the advantages of the consensus approach, individual calling methods are often chosen for arbitrary reasons and the performance of some of these ensembles may not have been formally examined. A comprehensive evaluation of SV calling methods showed that while there is no single method that can detect all variants, some tools are optimised for specific classes of variants (Kosugi et al., 2019). Also, some specific calling methods (as well as categories of calling methods) perform better together when a consensus approach is required. While this is useful for any analysis making use of pairs of methods, it is insufficient for more complex ensembles.

Another drawbacks of a consensus approach is that if a specific tool is able to detect certain variants that others cannot, these CNV calls will be lost due to lack of support. We refer to calls identified by one tool only in a consensus as singleton calls. Khan et al. attempted to partially resolve this issue using family call data which can provide further evidence for CNV calls with low levels of support from a consensus (Khan et al., 2018). Singleton CNVs in an individual were retained if another member of the pedigree had the same CNV call identified by a consensus of tools. In studies of individuals who are closely related, we might expect that the breakpoints for the same CNV are more comparable than in an unrelated cohort. This strategy can enable consensus approaches to reclaim the utility of the individual tools while maintaining some level of control on false positives with in-family validation. In this chapter, we describe a novel family-based consensus approach for CNV calling using four different calling methods. We evaluated the performance of our consensus calls on a set of curated "Gold Standard" CNV calls and compared this to previously published CNV calling methods for pedigree data.

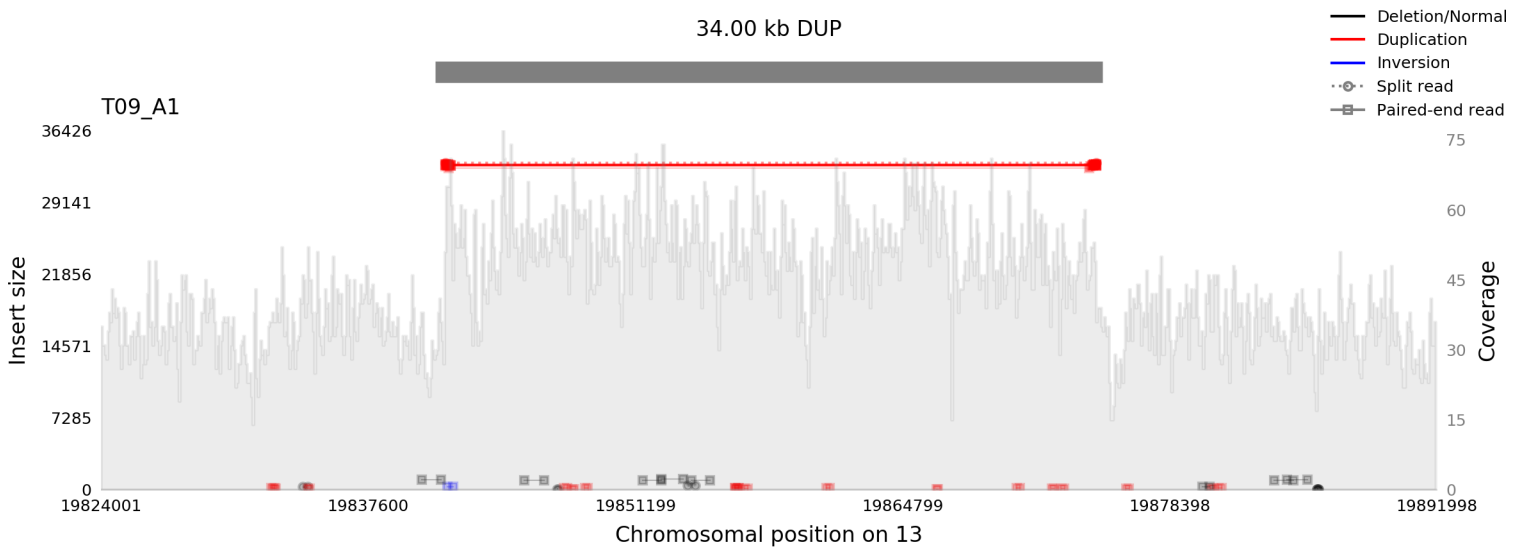


Figure 4.1: An example of a 34 kbp duplication on chromosome 13 in an individual identified from our WGS data (information described in Chapter 5), visualised using `samplot` (Belyeu et al., 2021). The left-side y-axis shows the insert size between the paired-end reads, represented by the black, blue, and red blocks connected with horizontal lines. The right-side y-axis shows the depth of coverage, represented as the grey histogram. At the breakpoints of the CNV region (marked with a black horizontal bar at the top), the insert size of some reads is much higher than other reads, and the depth of coverage changes compared to flanking regions.

4.2 Calling Pipeline

We selected a trio of individuals from the CEPH 1463 pedigree (proband: NA12878; father: NA12891; mother NA12892) on which to call CNVs. Sample NA12878 has been studied extensively, and several sequencing technologies have been used to characterize CNVs in this genome (Haraksingh et al., 2017; Rao et al., 2020; Zook et al., 2016). WGS had been performed to approximately 50× coverage on all samples as part of the Illumina Platinum Genomes project (Eberle et al., 2017), and publicly available FASTQ files were obtained from the European Nucleotide Archive (ENA), project number PRJEB3381. All paired-end FASTQ files were examined using `FastQC` and `samtools` (H. Li et al., 2009) to screen for DNA contamination or degradation, and none was observed. Reads were aligned to the GRCh38 reference genome and standard pre-processing was applied as described in Subsection 2.2.1.

4.2.1 Per Individual

Our consensus approach combines two PR/SR tools and two RD tools. For deletions and duplications larger than 1kbp, taking a consensus of callers both within and across these calling classes has shown a reasonable improvement in CNV detection compared to using the tools on their own (Kosugi et al., 2019). The tools we selected which implement PR/SR calling were `Manta` (Chen et al., 2016) and `LUMPY` (Layer et al., 2014). These two tools have been used in several CNV consensus calling approaches such as `bcbio-nextgen` (Chapman et al., 2021), `sv-callers` (Kuzniar et al., 2020) and `Parliament2` (Zarate et al., 2020), and have been shown to perform well individually and as a pair (Gong et al., 2020). The RD tools selected were `ERDS` (Zhu et al., 2012) and `CNVnator` (Abyzov et al., 2011). This pair of tools together has been shown to outperform several other RD-based callers for NGS data (Trost et al., 2018). All four calling methods were run using the default settings recommended by the authors. Based on recommendations from the online documentation of `CNVnator`, we chose a bin size of 50bp since our data has an average depth of coverage of 50×. `ERDS` requires SNV and indel calls, which were derived from the input BAM files using the `HaplotypeCaller` module from `GATK` following Subsection 2.2.2 above.

On a preliminary evaluation of the calling methods, we observed that they sometimes generate several largely overlapping CNV calls which appear to represent one single copy number event. This was more prevalent in the output of the two PR/SR callers than the RD callers, likely because multiple read pairs close to one another may behave

similarly for the same copy number event but may be called separately. To eliminate such repetition, we implemented a collapsing strategy on the raw calls from each separate tool to identify sets of equivalent CNVs, comparable to that described in Trost et al. (Trost et al., 2018). These overlapping regions were reduced to single regions as follows:

- If two CNVs of the same type (either deletion or duplication) overlap reciprocally by at least 25%, then they are added to the same set.
- If only one of the two CNVs is already in a set, then the other is added to that set. If both CNVs are already in sets, then the two sets are combined.
- Once all sets have been created, each set is collapsed down to one region by taking the union of all CNVs using `bedtools` (Quinlan & Hall, 2010).

A depiction of this collapsing method is shown in Figure Figure 4.2. A consensus CNV call across all tools is generated by merging calls of the same type that overlap reciprocally by 50%, first considering calls within calling method types (`CNVnator` vs `ERDS`, and `LUMPY` vs `Manta`), and then across the resulting calling method types (PR/SR vs RD). A depiction of this merging of CNV calls across call sets is also shown in Figure 4.2. CNV calls were annotated with which calling method(s) identified the region.

CNV calling is known to be confounded by repeat and low complexity regions (RLCR), which Trost et al. defined as:

1. assembly gaps, (UCSC "gap" table);
2. segmental duplications (UCSC "genomicSuperDups" table);
3. the pseudo-autosomal regions of the sex chromosomes.

It is worth noting that Trost et al. originally included repeat regions identified by `RepeatMasker` (Smit et al., 2015), but noted that this reduced sensitivity to detect rare, genic CNVs and so we excluded this from the RLCR definition (Trost et al., 2018). In our analysis, CNV calls for which over 75% of their length comprise of RLCR were removed. Finally, variants of length less than 1kbp were removed, since the RD callers are known to be limited to this resolution (Trost et al., 2018). A workflow diagram for the calling pipeline is shown in Figure 4.3.

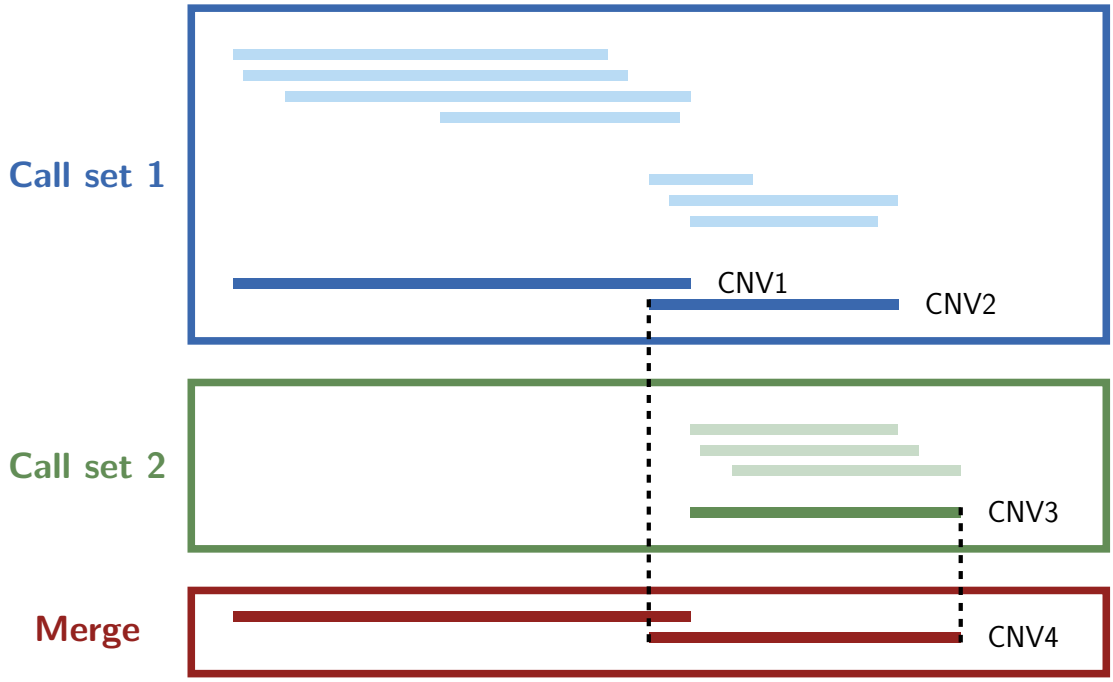


Figure 4.2: A visualisation of the collapsing and merging strategy for a single individual, showing call sets from two CNV callers. The paler horizontal blocks represent CNV calls from the same tool that are collapsed down to represent the one site within an individual (darker blocks). Note that in Call set 1, two distinct regions (CNV1 and CNV2) are formed of calls that satisfy the overlap criteria. Then across multiple call sets, we merge two CNV calls that overlap reciprocally by 50% (CNV2 and CNV3) by taking the union, indicated by the dashed vertical lines (CNV4).

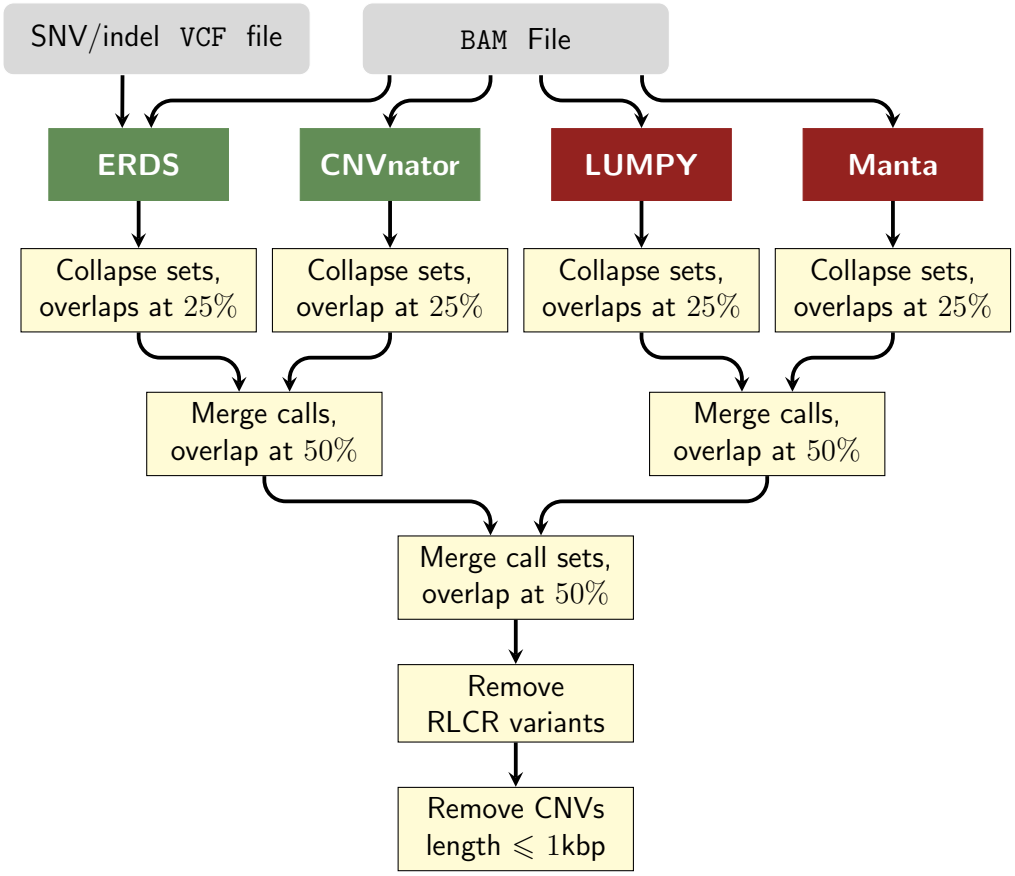
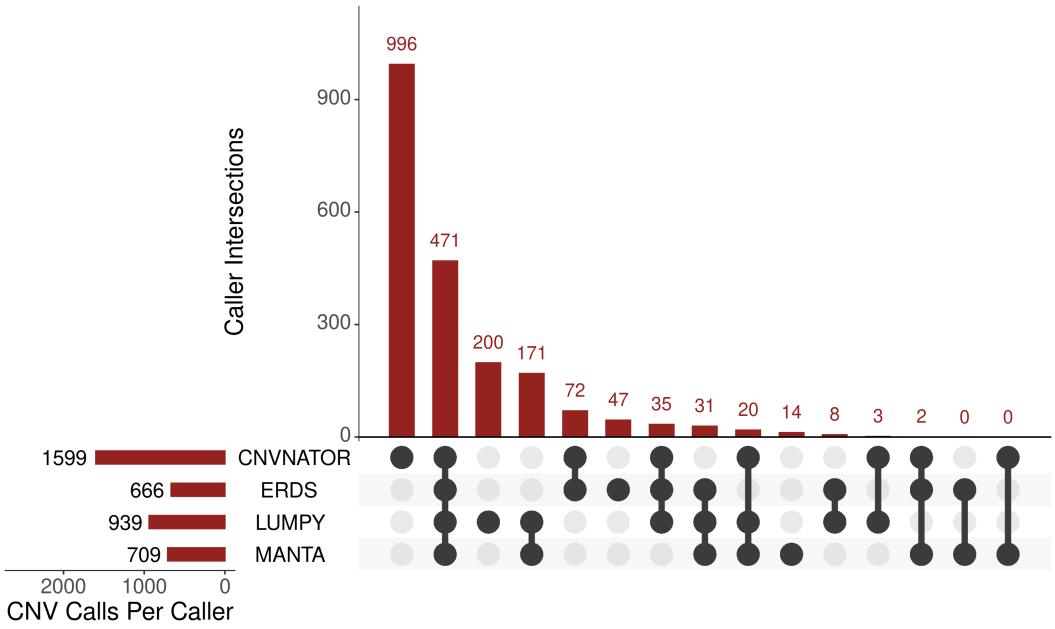


Figure 4.3: Workflow of the CNV calling pipeline per individual. PR/SR callers are shown in red, and RD callers are shown in green. RLCR: repeat/low-complexity region.

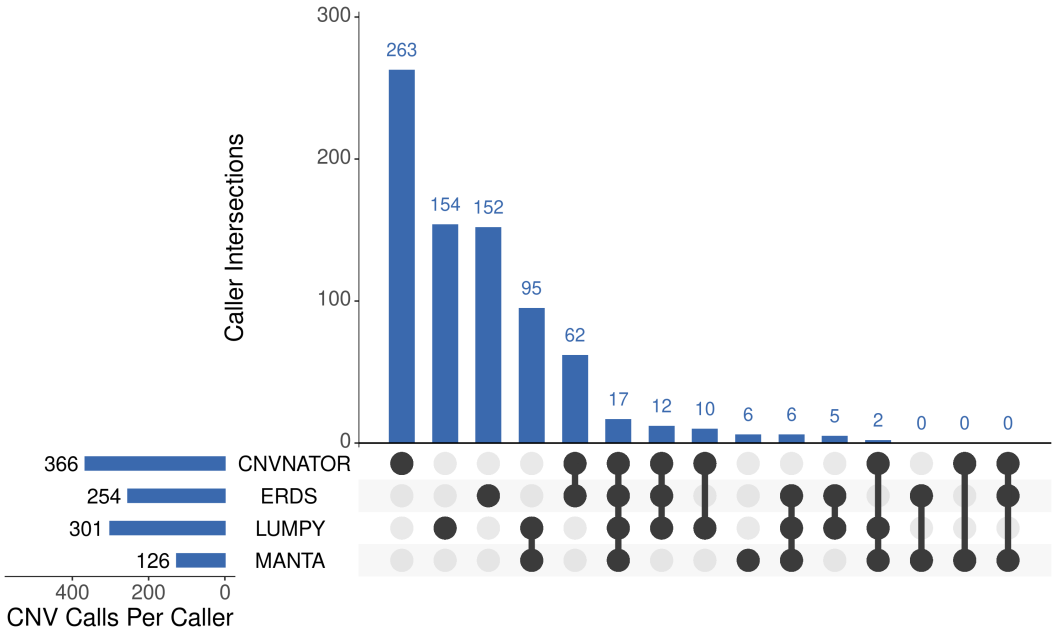
We applied the above method to the CEPH 1463 trio, and the number of CNVs identified in each sample is shown in Table 4.1. Upset plots for the intersection between each caller for sample NA12878 are shown in Figure 4.4. We can see for NA12878 that there are 1,257 singleton deletions (60.7%) and 575 singleton duplications (73.3%), representing 64.2% of all CNVs called for that sample. `CNVnator` had the highest number of singletons and `Manta` had the lowest number, across both deletions and duplications.

Sample	DEL	DUP	Total
NA12878	2,070	784	2,854
NA12891	2,867	1,115	3,982
NA12892	2,535	798	3,333

Table 4.1: Counts of the number of CNVs called for each of the three individuals in the CEPH 1463 trio. DEL: deletion; DUP: duplication.



(a)

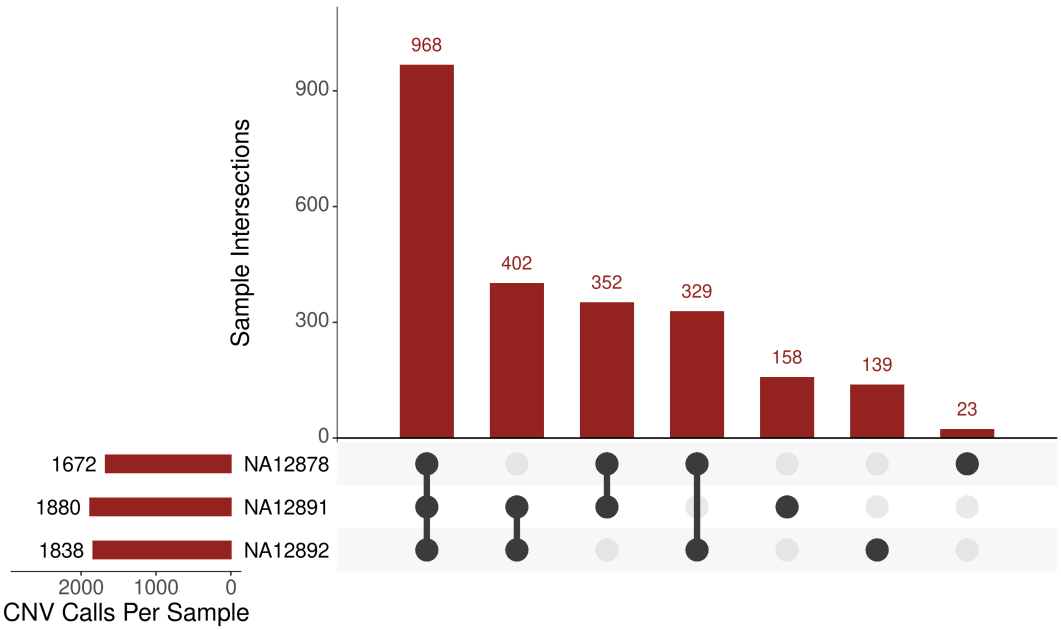


(b)

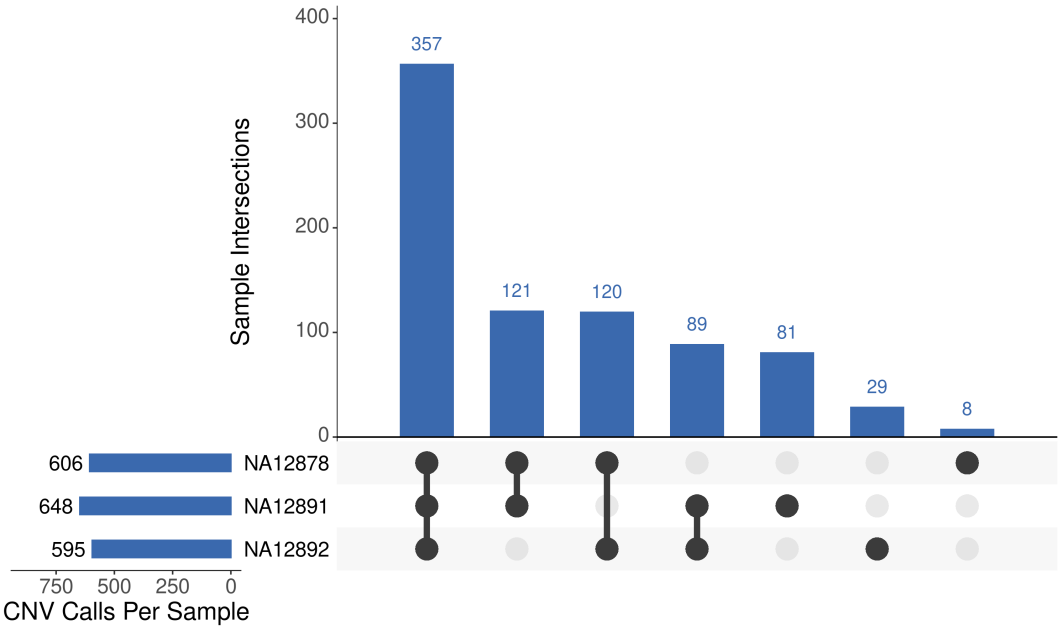
Figure 4.4: Upset plots showing the intersection and relative complements between the four CNV calling methods in sample NA12878 for: **(a)** deletions; and **(b)** duplications. The black dots indicate which caller is included in the intersection or complement, and the bar chart above indicates the number of CNV calls per section. The horizontal bars at the bottom left indicate the total number of CNVs called by each caller.

4.2.2 Per Pedigree

All calls within a pedigree were combined, again taking the union of calls of the same type with 50% reciprocal overlap. Following Khan et al., singleton calls that were not detected by at least two callers in any of the individual's direct relatives were removed (Khan et al., 2018). This ensured that the final list of CNVs for any individual in the pedigree either had support from at least two calling methods or was also present with confidence in a relative. We applied this strategy to the CEPH 1463 trio, resulting in 2,371 deletions and 805 duplications. Upset plots for the overlap between the three samples is shown in Figure 4.5.



(a)



(b)

Figure 4.5: Upset plots for the three samples in the CEPH 1463 trio, split by: **(a)** deletions; and **(b)** duplications.

4.3 Alternative Strategies

4.3.1 Individual Callers

We compared the final list of CNV calls for NA12878 to the calls that were identified by each of the four callers alone. As before, we implemented the collapsing of overlapping calls, removed calls RLCRs, and retained those at least 1kbp in length. Since we no longer have a consensus, each call will be a singleton, so we relaxed the requirement for a direct relative to carry the CNV. Additionally, we extracted the CNV calls that were identified by the PR/SR callers and the RD callers separately. Given that we have two callers each, we kept the check in direct relatives to retain singletons and applied the pipeline as before.

4.3.2 Khan et al.

We also benchmarked the pipeline described in Khan et al. against our own (Khan et al., 2018) to assess their relative performances. The authors applied this pipeline to samples sequenced at an average of 16× coverage on build GRCh37, and used two different calling methods to those we implement here: `cn.mops` (Klambauer et al., 2012) instead

of ERDS, and DELLY (Rausch et al., 2012) instead of Manta. Given that our evaluation samples were sequenced to an average of 50× coverage with 101bp read length, we adjusted some of the input parameters of the calling methods to account for this. The documentation for cn.mops recommends that a window size be selected such that on average 100 reads are present in each window. Given our sequencing data, there will be approximately 100 reads fully or partially contained in a window length of 200 bp. Khan et al. removed CNVs with fewer than four supporting reads from LUMPY and those with fewer than three supporting reads from DELLY. For both of these tools, CNVs with genotype quality less than 20.0 were also removed. Since only three samples were evaluated in our analysis, we did not remove samples based on outlier CNV counts for any of the calling methods, as was done in Khan et al., where over 300 samples were sequenced.

4.4 Benchmarking

4.4.1 Curated Gold Standard CNV Calls

The generation of a CNV call set for NA12878 was done by a member of our research group (Dr Niamh Ryan) and is described in this Subsection for reference. Despite the extensive study of sample NA12878, it is difficult to fully characterise all detectable CNVs in their genome, since no single technology can detect all variants. With this in mind, Dr Ryan examined the following studies which attempted to curate a list of CNV calls detected using a variety of technologies:

- **DGV** - the Database of Genomic Variants is a catalogue of curated SVs observed in the general population taken from multiple studies and resources (MacDonald et al., 2014). Included are a set of CNV calls for NA12878 predominantly taken from various Phases of The 1000 Genomes Project.

- **LUMPY** - in the companion paper to LUMPY (Layer et al., 2014), the authors generated a list of CNV calls for NA12878 that had been validated by PacBio and/or Illumina Moleculo long-read sequencing.

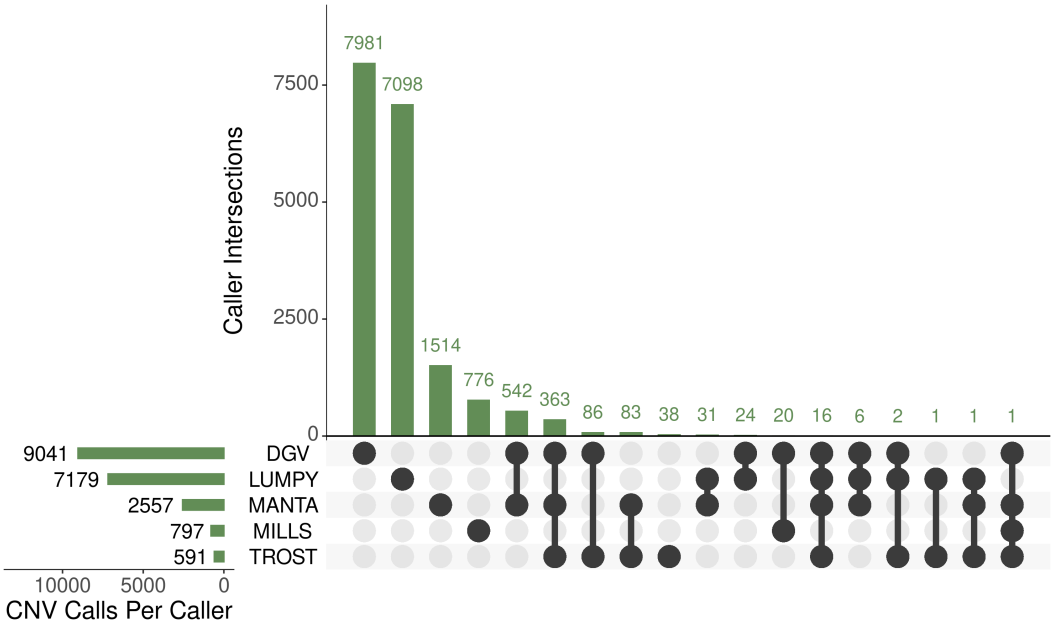
- **Manta** - in the companion paper to Manta (Chen et al., 2016), the authors considered pedigree consistent CNV calls from all 17 members of the CEPH 1463 pedigree, generated using pindel (Ye et al., 2009) and DELLY (Rausch et al., 2012).

- **Mills** - Mills et al. constructed a map of CNVs from 185 individuals (including

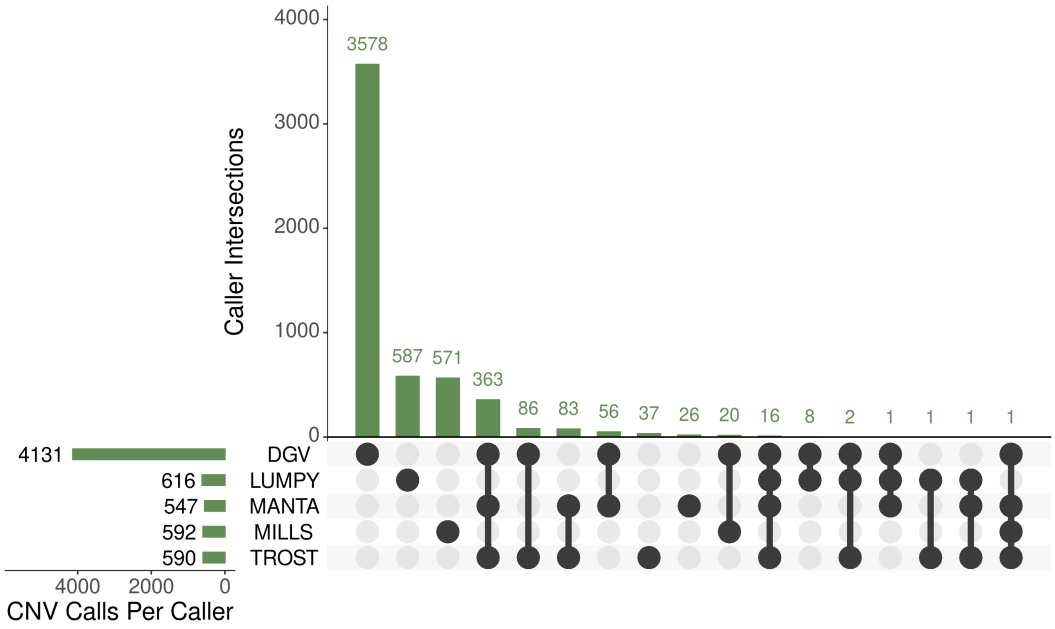
NA12878) using a wide variety of sequencing strategies and variant callers (Mills et al., 2011).

- **Trost** - in an evaluation of RD tools (Trost et al., 2018), the authors used a list of previously generated CNV calls for NA12878 using PacBio long-read sequencing as well as Illumina short- and long-read sequencing.

The website of the DGV notes that since not all CNV calls were generated from the same source, what one study calls a deletion, another may call a duplication (see "Web Resources", Subsection A.2.3). Therefore, for the purposes of the benchmarking, Dr Ryan considered the CNV regions only, and did not match for CNV type. There was a relatively low overlap across the five call sets; an upset plot is shown in Figure 4.6. One explanation for this is that CNVs were called with different technologies across the five call sets (SNP genotype arrays, aCGH, cytogenic techniques, short-/long-read sequencing, etc.), so it is reasonable that some methods will detect CNVs that others cannot. There were 1,176 CNV regions (6.3%) on the autosomal chromosomes present in at least two call sets, out of a total of 18,583 unique CNV regions. Of these, 638 were greater than 1kbp in length, which is the recommended length for the RD callers (Trost et al., 2018). This final list of CNV regions was used as the curated "Gold Standard" list.



(a)



(b)

Figure 4.6: Upset plot for the overlap of CNV regions between the five "Gold Standard" NA12878 call sets for **(a)** all CNV calls; and **(b)** CNV calls with length greater than 1kbp. For readability, empty intersections are not displayed.

4.4.2 Validation Metrics

To evaluate the accuracy of the pipelines considered, we calculated the precision of the CNV call sets for NA12878 relative to the curated "Gold Standard" CNV calls. Recall (or sensitivity) is defined as the proportion of curated "Gold Standard" CNV calls that each query call set identified and is a measure of the pipeline's ability to detect true positives. Another often used metric is the precision (or positive predictive value), defined as the proportion of the input call set that is present in the curated "Gold Standard". However, since the overlap of the five individual "Gold Standard" sets is relatively small, it is possible that true positive variant calls were only identified by one of the five studies, given the variety of sequencing/calling technologies used. Since the curated "Gold Standard" is simply a set of high-confidence CNV calls for NA12878, the absence of a CNV in this set does not necessarily indicate that it is a false positive. Therefore, we will not use precision as a benchmarking metric for this curated "Gold Standard" set.

4.4.3 Results

The curated "Gold Standard" set was generated on build GRCh37, so we converted the final list of CNV calls of length at least 1kbp from our pipeline and alternative pipelines to this build using `liftOver` (Haeussler et al., 2019). The results of the benchmarking are shown in Table 4.2. This analysis showed that our method outperformed the alternate calling pipeline by Khan et al., while also performing better than any of the constituent tools individually.

Pipeline	Total		In GS	Recall
	GRCh38	GRCh37		
Our Method	2,255	1,964	557	87.30%
PR/SR	1,161	1,045	539	84.48%
`LUMPY`	1,201	1,072	538	84.33%
`Manta`	808	763	519	81.35%
RD	2,030	1,810	465	72.88%
`ERDS`	888	764	445	69.75%
`CNVnator`	1,940	1,749	434	68.03%
Khan et al.	6,778	6,115	382	59.87%

Table 4.2: A comparison of the CNV calling pipelines. Shown are the number of CNV calls on build GRCh38 and GRCh37 (following `liftOver`), the number of CNV calls present in the curated "Gold Standard" (GS) set out of a total of 638, and the recall values.

4.5 Conclusions

Here we have introduced a novel consensus CNV calling pipeline designed for pedigree based NGS data, by selecting calling methods and classes known to support one another. Following previous work, our pipeline is able to reclaim lower-confidence CNV calls by considering calls from close relatives. We have shown that our pipeline performs well at identifying a curated list of "Gold Standard" CNV calls from sample NA12878, and out-performs an alternate pipeline designed for the same data.

Chapter 5

Rare Variant Analysis of Discordant Monozygotic Twins

Monozygotic twins are often thought to have identical genomes, but recent work has shown that early post-zygotic events can result in a spectrum of DNA variants that are different between twins. Such variants may explain phenotypic discordance and contribute to disease aetiology. Here we performed whole genome sequencing in 17 pairs of MZ twins discordant for schizophrenia and related disorders. We identified seven genes harbouring rare, predicted deleterious SNVs that were private to an affected sample in the cohort. Four of the genes implicated had been reported to carry rare deleterious variants in two previous case-control schizophrenia WES studies. A discordant missense variant in *POLG* was observed in an individual with major depressive disorder. Deleterious variants in this gene have been previously implicated in mood disorders and psychosis in both human and mouse studies. Additionally, we identified seven rare genic CNVs private to an affected sample, one of which was predicted to be pathogenic and has been observed in autism and developmental delay cases.

5.1 Introduction

Monozygotic (MZ), or identical twins, occur when a zygote divides into two separate embryos, and dizygotic (DZ) twins, or non-identical twins, occur when two ova are fertilised separately during the same pregnancy. MZ twins are often described as sharing 100% of their genomes, compared to DZ twins or other non-twin siblings who share approximately 50% of their genomes. Significant differences in concordance rates between MZ and DZ twins classically indicate a genetic factor for a given phenotype. If the phenotype could be explained purely by genetic factors, then we would expect that the concordance rates between MZ twins to be close to 100%. Twin studies have several advantages over case-control and other family study types, as twins are the same age and typically have similar or comparable exposure to many environmental factors, (e.g. childhood trauma, urbanicity, etc.) compared to non-twin siblings. Since non-shared factors may contribute to phenotypic differences between siblings, twin studies have been used to provide

insights into the genetic aetiology of many diseases and disorders (Kato et al., 2005), including schizophrenia.

Despite their similarity, MZ twins do not always have identical genomes. As with all humans, post-zygotic DNA variation (i.e. variants that occur after fertilisation) can be present in the twins' genomes, but depending if the variation occurs before or after the single fertilized zygote splits in two (known as the twinning event), they may be private to an MZ twin (Jonsson et al., 2021). Post-zygotic variation can occur spontaneously during DNA replication or can be induced by mutagens, resulting in somatic mosaicism, i.e. the presence of different genomes in different cells within the same individual (Griffiths et al., 2008) However, if these post-zygotic variants occur before the specification of the embryonic cells that will eventually become germ cells (known as primordial germ cells), they may be found in most cells in the body and may present as germline variants. A depiction of this is shown in Figure 5.1 (Jonsson et al., 2021).

In Polymeropoulos et al., the authors observe that: "the [phenotypic] discordance could be explained by the hypothesis that the . . . phenotype will remain silent unless released by environmental and other non-familial stressors" (Polymeropoulos et al., 1993). Under this hypothesis, both twins share a common genetic risk which alone is insufficient to be causal for the phenotype, but rare, post-zygotic variation present in the affected twin increases their disease-risk. Since these post-zygotic variants are typically not examined during twin heritability analyses, they would be mistakenly counted as part of the environmental or even non-additive genetic effects (S. M. Singh et al., 2020). Rare variants of interest are therefore *de novo* events within a twin pair, i.e. where one individual has exactly one more copy of the allele of interest than their co-twin. We refer to these as discordant variants.

Discordant SNVs and indels have been shown to be causal for several Mendelian disorders for which MZ twins are discordant, for example: Darier's Disease (Sakuntabhai et al., 1999), Van der Woude Syndrome (Kondo et al., 2002) and otopalatodigital syndrome spectrum disorders (Robertson et al., 2006). Discordance for trinucleotide repeat expansion length in the *FMR1* gene is thought to be causal for the discordance of fragile X syndrome (Helderman-van den Enden et al., 1999). Large discordant chromosomal abnormalities such as aneuploidy have been observed, resulting in MZ twins discordant for Down's syndrome (Dahoun et al., 2008), Patau syndrome (Taylor et al., 2008) and even sex (Zech et al., 2008).

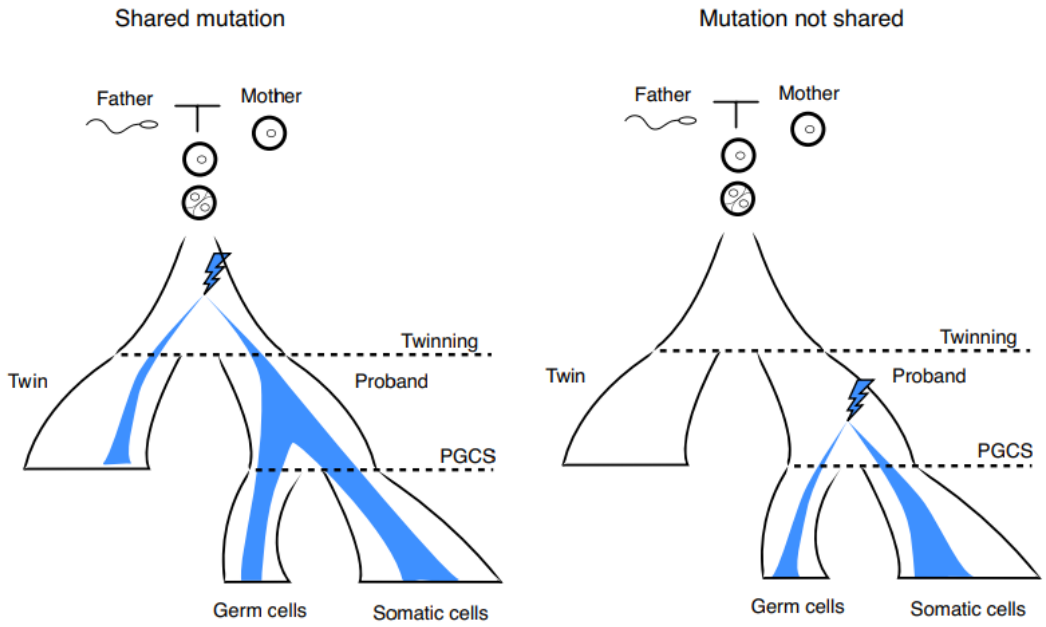


Figure 5.1: Post-zygotic variants may be shared in a twin pair if they occur before the twinning event (left), or private to a twin if they occur after twinning (right). If they occur prior to primordial germ cell specification (PGCS), they may be present in most germ and somatic cells. This figure is adapted from Figure 1 of Jonsson et al. (Jonsson et al., 2021).

For psychiatric disorders, as with other complex traits, the genetic contribution of discordant variation is less clear. Older work focused on discordant CNVs, but while some variants had been observed in schizophrenia samples (Castellani et al., 2014), such discordances were not widely replicated (Bloom et al., 2013; Laplana et al., 2014). More recently, NGS analyses have identified post-zygotic variation in MZ twins discordant for schizophrenia (Castellani et al., 2017; Tang et al., 2017), autism (Huang et al., 2019), and Tourette's syndrome (Vadgama et al., 2019) and so further investigation is warranted. In this Chapter we examine WGS data from a cohort of MZ twins discordant for schizophrenia and related disorders and investigate various classes of variants that may be increasing the affected individuals' risk for their respective phenotype.

5.2 Cohort Description

5.2.1 Sample Procurement

The schizophrenia and bipolar twin study in Sweden (STAR) has collected data on 462 MZ and DZ twin pairs with schizophrenia or bipolar disorder. The clinical assessment and DNA sampling of the cohort are described in Johansson et al., described here briefly (Johansson et al., 2019). The participants in this study were originally identified through the Swedish twin register (STR) (Lichtenstein et al., 2006) and the National patient re-

gister (NPR), which is administered by the social board of health and welfare. Potential participants were invited to the STAR study if one twin in a pair had a registered treatment episode of schizophrenia or bipolar disorder (diagnoses according to International Statistical Classification of Diseases: ICD-8: 295 or 296, ICD-9: 295 or 296 or ICD-10: F20, F30 or F31). Cases were categorized as schizophrenia (SCZ; ICD-10: F20), schizoaffective disorder (SAD; ICD-10: F25), bipolar disorder (BD; ICD-10: F31), major depressive disorder (MDD; ICD-10: F32-F33) or not affected by any of these diagnoses.

From the STAR cohort we selected all available phenotype discordant MZ twin pairs in which one twin was diagnosed with either SCZ, SAD, or BD, and the co-twin was unaffected (see Table 5.1). If the co-twin had been diagnosed with MDD previously, we included the twin pair only if the severity level of the depression was mild and without occurrence of psychotic symptoms. If only one sample in a twin pair had a diagnosis (known as "narrow discordance"), we use the suffix "_A" to refer to the affected sample and the suffix "_U" to refer to the unaffected sample. If the co-twin had a diagnosis of MDD (known as "broad discordance"), we use the suffix "_A1" to refer to the non-MDD sample and the suffix "_A2" to refer to the sample with MDD. DNA samples were sent to us from our collaborators in the Karolinska Institutet, Sweden.

5.2.2 Sample Processing and WGS Data

DNA concentrations were quantified using Qubit, and the quality of DNA was determined by agarose gel electrophoresis by a member of our research team (Dr Amy Cole). All samples were found to have sufficient DNA concentration to be sent for whole genome sequencing at EGCG. The pair T19 failed quality control metrics for sequencing and were excluded from the study. All FASTQ files received from EGCG were examined using `FastQC` and `samtools` to screen for DNA contamination or degradation, but none were flagged at this stage. Reads had been aligned to GRCh38 by EGCG and variants were called as described in Subsection 2.2.2. Genotype calling was performed jointly across all samples, and variant quality score recalibration (VQSR) was performed on SNVs and indels separately (see Subsection 2.3.1).

The software `peddy` (Pedersen & Quinlan, 2017) was used to check for relatedness in all samples jointly as described in Section 2.4. To further examine the pairwise relatedness in the cohort, we selected a subset of high-confidence variants. SNVs were retained if they passed the following filters from the jointly genotyped data:

a) phred-scaled quality score (QUAL) $>$ 1000.0;

Pair ID	Sex	Age at Sampling	Twin 1		Twin 2		Disc
			ID	Pheno	ID	Pheno	
T01	M	35	T01_A1	SCZ	T01_A2	MDD	Broad
T02	M	38	T02_A	SAD	T02_U	None	Narrow
T03	F	34	T03_A	SCZ	T03_U	None	Narrow
T04	F	30	T04_A	BD	T04_U	None	Narrow
T05	M	25	T05_A1	SAD	T05_A2	MDD	Broad
T06	F	65	T06_A	SAD	T06_U	None	Narrow
T07	F	61	T07_A	BD	T07_U	None	Narrow
T08	F	60	T08_A	SAD	T08_U	None	Narrow
T09	F	59	T09_A	BD	T09_U	None	Narrow
T10	M	58	T10_A	SCZ	T10_U	None	Narrow
T11	F	52	T11_A	BD	T11_U	None	Narrow
T12	M	50, 51	T12_A	BD	T12_U	None	Narrow
T13	M	48	T13_A1	SCZ	T13_A2	MDD	Broad
T14	M	50, 51	T14_A	BD	T14_U	None	Narrow
T15	M	43	T15_A1	SAD	T15_A2	MDD	Broad
T16	M	46	T16_A	BD	T16_U	None	Narrow
T17	F	45	T17_A	BD	T17_U	None	Narrow
T18	F	27	T18_A1	SAD	T18_A2	MDD	Broad
T19	F	38	T19_A	SAD	T19_U	None	Narrow

Table 5.1: Phenotypic data for the 19 pairs of MZ twins. For the discordance (Disc), "broad" indicates that both samples have a diagnosis and "narrow" indicates that only one sample has a diagnosis. BD: bipolar disorder; MDD: major depressive disorder; SAD: schizoaffective disorder; SCZ: schizophrenia.

b) DP $>$ 100;

c) mapping quality across all reads (MQ) $>$ 5.0;

d) VQSLOD $>$ 10.0;

e) phred-scaled quality score normalised to read depth (QD) $>$ 5.0.

These thresholds were obtained by manually examining the density plots for each of the respective metrics to remove those at the lower end of the distribution. We then performed LD pruning on the remaining SNVs using the `--indep-pairwise` command from `plink` (Purcell et al., 2007) with parameters "50 5 0.2". Finally, we calculated the pairwise relatedness scores with `vcftools` as described in Section 2.4. A heatmap

describing these relatedness scores is shown in Figure 5.2.

It was expected that each sample in the cohort should be perfectly related to themselves and to their co-twin (red in the heatmap) and unrelated to all other samples (white in the heatmap). However, sample T04_U appears to be related to all other twin samples. Because of this relatedness issue, twin pair T04 was excluded from our cohort. Additionally, from this heatmap it appears that some twin pairs have negative relatedness scores with the other twin samples (displayed as blue in the heatmap). The relatedness scores are estimates of the true proportion of shared DNA, and some amount of variability is expected. However, the negative scores for twin pair T05 are more striking than all other twin pairs. A negative `KING` relatedness score can often indicate the presence of different genomic ancestry groups (Manichaikul et al., 2010). The PCA performed by `peddy` identified that all samples were predicted to have European genomic ancestry with the exception of twin pair T05 who were predicted to have East Asian genomic ancestry (see Figure 5.3), which is consistent with the relatedness scores from the heatmap.

5.2.3 Zygosity Check

The zygosity of both samples within a twin pair was estimated using the above-described set of high-confidence variants. Treating one sample in a pair as the "truth" sample, we evaluated the sensitivity and genotype concordance using the `GenotypeConcordance` module from `picard`. The sensitivity measures the proportion of variants in the call set that are present in the truth set, and the genotype concordance measures the proportion of variants with matching genotypes out of those which match a position in the truth set. The results of this are displayed in Table 5.2 below. We can see that within each pair, the sensitivity was at least 99.8% for all samples, and the genotype concordance rate was at least 99.999% for all samples. This confirms that all samples within each twin pair are monozygotic as expected.

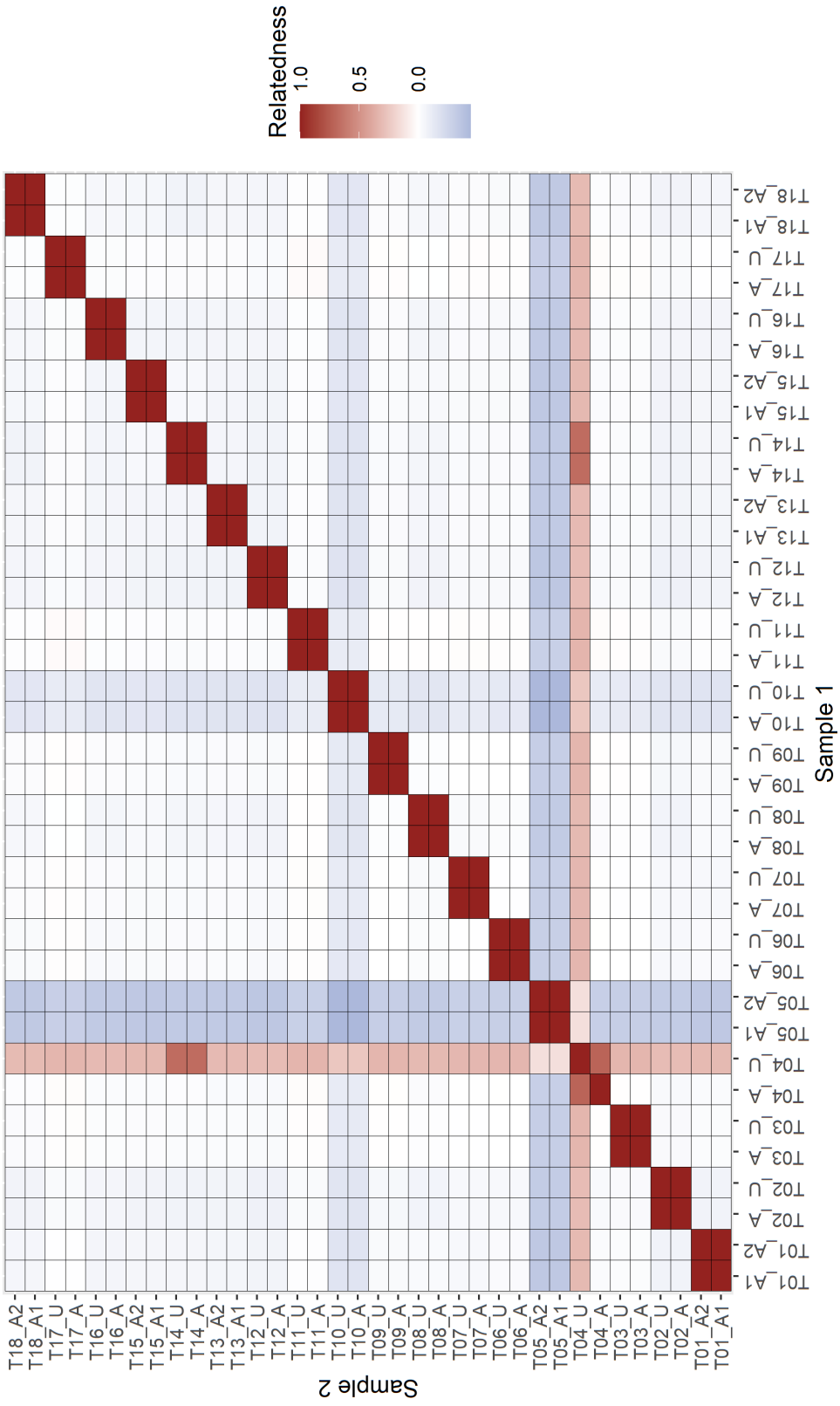


Figure 5.2: A heatmap of the pairwise relatedness scores for the 18 pairs of samples in the MZ twin cohort. Observed relatedness scores were generated via the KING algorithm from `vcftools`.

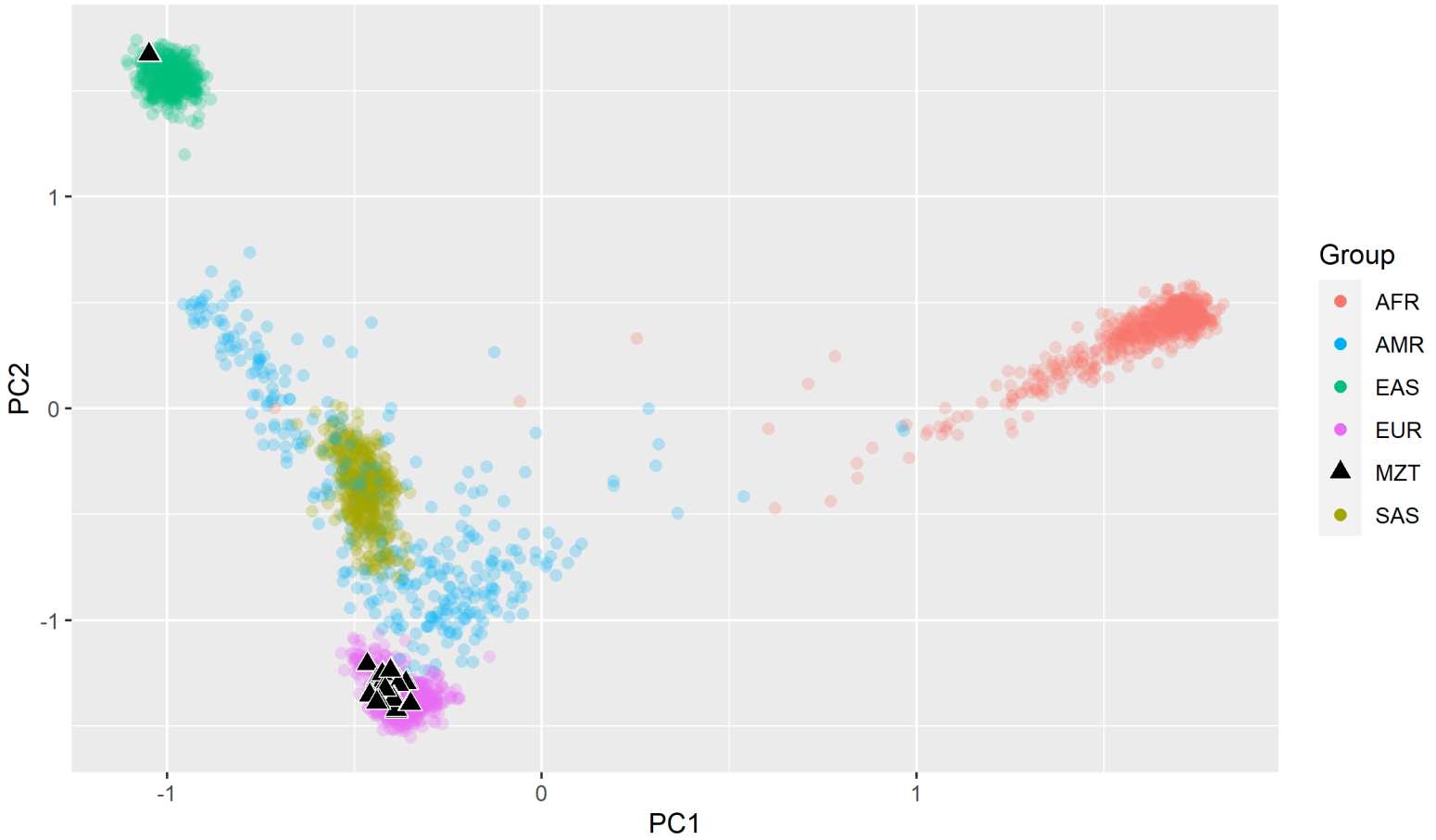


Figure 5.3: A plot of the first two principal components of the MZ twins and a background population from the 1000 Genomes Project, re-generated from the output of peddy using the R statistical package (R Core Team, 2013). AFR: African; AMR: admixed Americas; EAS: East Asian; EUR: European; MZT: monozygotic twins; SAS: South Asian.

Twin Pair	Truth Sample	Call Sample	Sens (%)	Geno Conc (%)
T01	T01_A1	T01_A2	99.8994	99.9998
	T01_A2	T01_A1	99.9058	99.9998
T02	T02_A	T02_U	99.8854	99.9997
	T02_U	T02_A	99.8908	99.9997
T03	T03_A	T03_U	99.8925	99.9997
	T03_U	T03_A	99.8910	99.9997
T05	T05_A1	T05_A2	99.8848	99.9997
	T05_A2	T05_A1	99.8904	99.9997
T06	T06_A	T06_U	99.8949	99.9997
	T06_U	T06_A	99.8955	99.9997
T07	T07_A	T07_U	99.8924	99.9997
	T07_U	T07_A	99.8912	99.9997
T08	T08_A	T08_U	99.8881	99.9997
	T08_U	T08_A	99.8949	99.9997
T09	T09_A	T09_U	99.8931	99.9997
	T09_U	T09_A	99.8906	99.9997
T10	T10_A	T10_U	99.8944	99.9998
	T10_U	T10_A	99.8902	99.9998
T11	T11_A	T11_U	99.8686	99.9997
	T11_U	T11_A	99.8867	99.9997
T12	T12_A	T12_U	99.8920	99.9998
	T12_U	T12_A	99.8981	99.9998
T13	T13_A1	T13_A2	99.8739	99.9997
	T13_A2	T13_A1	99.8952	99.9997
T14	T14_A	T14_U	99.8709	99.9997
	T14_U	T14_A	99.8892	99.9997
T15	T15_A1	T15_A2	99.8871	99.9997
	T15_A2	T15_A1	99.8936	99.9997
T16	T16_A	T16_U	99.8986	99.9997
	T16_U	T16_A	99.8881	99.9997
T17	T17_A	T17_U	99.8953	99.9997
	T17_U	T17_A	99.8950	99.9997
T18	T18_A1	T18_A2	99.9005	99.9998
	T18_A2	T18_A1	99.8876	99.9998

Table 5.2: Within-pair concordance metrics on a set of high-confidence SNVs for the 17 MZ twin pairs. Included are the sensitivity (Sens) and the genotype concordance rate (Geno Conc).

5.2.4 Discordant Variants

After the above quality control measures, 17 pairs of twins were carried forward for analysis. Joint genotyping and VQSR was re-applied to these 34 samples so that the excluded samples would not contribute to the variant metrics prior to quality control measures being applied. Multi-allelic sites were split into bi-allelic sites to further identify the pathogenic allele, as per Subsection 2.3.2. Any variant with QUAL < 100.0 was considered to be low-quality and removed (Castellani et al., 2017). Finally, we applied standard genotype specific filters to all SNVs and indels (see Subsection 2.3.2). After removing lower quality variants, each sample had an average of 44,306 discordant SNVs and indels across the genome. Given that short read WGS detects approximately 4,000,000 variants per genome (Lappalainen et al., 2019), this implies approximately 1.1% of the variants detected in each sample are discordant. As the estimated error rate is 0.1% for short-read WGS on Illumina HiSeq technologies (Fox et al., 2014), this number is unlikely to be attributed solely to sequencing errors.

5.3 Protein-Coding Variants

Our first analysis was to identify rare, damaging, protein-coding variants which may be implicated in the phenotypic discordance. To this end, discordant variants were annotated with `vep` including: functional impact; predicted deleteriousness (SIFT and PolyPhen-2), and allele frequency (1000 Genomes Project and gnomAD v2.1.1), as described in Section 2.6. Only SNVs were considered at this step, as SIFT and PolyPhen do not provide scores for indels. To identify rare, putatively pathogenic variants, the following filters were applied:

i) variant was present in the coding sequence of the canonical transcript of a protein-coding gene as determined by RefSeq (O'Leary et al., 2016);

ii) SIFT was "deleterious" or PolyPhen was "damaging";

iii) the allele frequency was $<1\%$ or absent in the appropriate population groups in the 1000 Genomes Project and gnomAD databases; and

iv) variant was not observed in any other samples within the cohort.

Thirteen rare, predicted-deleterious discordant SNVs were identified across nine unique genes, described in Table 5.3 below. Ten of the prioritised variants were present in an affected twin across seven genes, and three variants were present in unaffected individuals across two genes. All SNVs were missense variants, with `vep` IMPACT classification

of "MODERATE". At three of these genes (*CSMD3*, *FOXN1*, and *TIMP1*), several discordant variants were found to be in close proximity (<15bp) in the same individual. This behaviour is somewhat unexpected given the strict filtering requirements, and one explanation for such a phenomenon is that these variants are in fact a sequencing artefact.

To investigate this, we re-called variants from the original BAM files at a 300bp window up and downstream from the variant sites with the same parameters used previously. The window size of 300bp was chosen since the read length is 150bp, so this window size would be expected to show the behaviour of the majority of the reads affecting the variant sites. We also included the `--bamout` option from `HaplotypeCaller` to extract the locally re-aligned reads and assembled haplotypes. Then we visualised the region from the bamout files using the Integrative Genomics Viewer (`IGV`), (Robinson et al., 2011). An example of such is given in Figure 5.4 below. The variant base pairs that were in close proximity only appeared on the same re-constructed reads, which likely occurred due to local re-alignment of the reads around indels by `HaplotypeCaller` during variant calling. Hence these variants in close proximity are likely due to the same indel event.

Of particular note, an individual with major depressive disorder carried a missense SNV in *POLG* (DNA subunit polymerase-γ), which plays a role in mitochondrial DNA replication. This gene was found to be expressed in multiple brain tissue types according to the Genotype Tissue Expression (GTEx) database (Keen & Moore, 2015). In a mouse model study, samples which carried a specific missense variant in *POLG* exhibited symptoms consistent with mood disorders (Kasahara et al., 2016). In humans, deleterious non-synonymous variants in this gene were found to be significantly enriched in bipolar cases compared to controls (Kasahara et al., 2017). Case reports have noted psychiatric symptoms in *POLG* variant carriers, such as recurrent major depression (Verhoeven et al., 2011) and psychosis (Hakonen et al., 2005).

None of the 10 missense SNVs appeared in the SCHEMA database (T. Singh et al., 2022), but rare SNVs at four of these genes (*FAM90A1*, *FOXN1*, *KRTAP10-6* and *POLG*) had been reported in two previous schizophrenia WES studies (Genovese et al., 2016; Howrigan et al., 2020).

Chr	Pos	rsID	Ref	Alt	Gene	HGVSp	SIFT	PolyPhen	Carrier	Phenotype
chr6	31356433	rs12721827	G	A	*HLA-B*	T118I	D	B	T03_U	None
chr8	112244420	rs1300679966	C	A	*CSMD3*	W3459L	D	D	T06_U	None
chr8	112244426	rs1204369873	T	A	*CSMD3*	N3457I	D	D	T06_U	None
chr9	96932219	rs112610837	C	T	*NUTM2G*	P172S	T	D	T13_A1	SCZ
chr12	8224750	rs201155866	A	C	*FAM90A1*	V28G	T	D	T18_A2	MDD
chr15	89320970	rs752971760	C	G	*POLG*	S926T	T	D	T01_A2	MDD
chr17	28530789	rs1385768054	A	C	*FOXN1*	S291R	D	D	T07_A	BD
chr17	28530791	rs371766542	C	A	*FOXN1*	S291R	D	D	T07_A	BD
chr17	28530802	rs1220808552	G	C	*FOXN1*	S295T	D	D	T07_A	BD
chr17	28881251	-	C	T	*FLOT2*	A347T	D	D	T09_A	BD
chr22	44592351	rs367621282	G	C	*KRTAP10-6*	P45R	D	D	T10_A	SCZ
chrX	47585615	rs1478486447	C	A	*TIMP1*	A134D	D	D	T18_A2	MDD
chrX	47585618	rs1417127009	A	G	*TIMP1*	Q135R	D	D	T18_A2	MDD

Table 5.3: Discordant protein-coding variants with a predicted deleterious effect. Each variant is annotated with: genomic positions (GRCh38), rs identification numbers, the reference and alternative alleles, the gene harbouring the variant, the amino acid substitution (HGVSp), pathogenicity scores from SIFT (D: deleterious; T: tolerated) and PolyPhen (D: damaging; B: benign), the sample carrying the variant, and their phenotype. All prioritized variants are missense SNVs. BD: bipolar disorder; MDD: major depressive disorder; SCZ: schizophrenia.

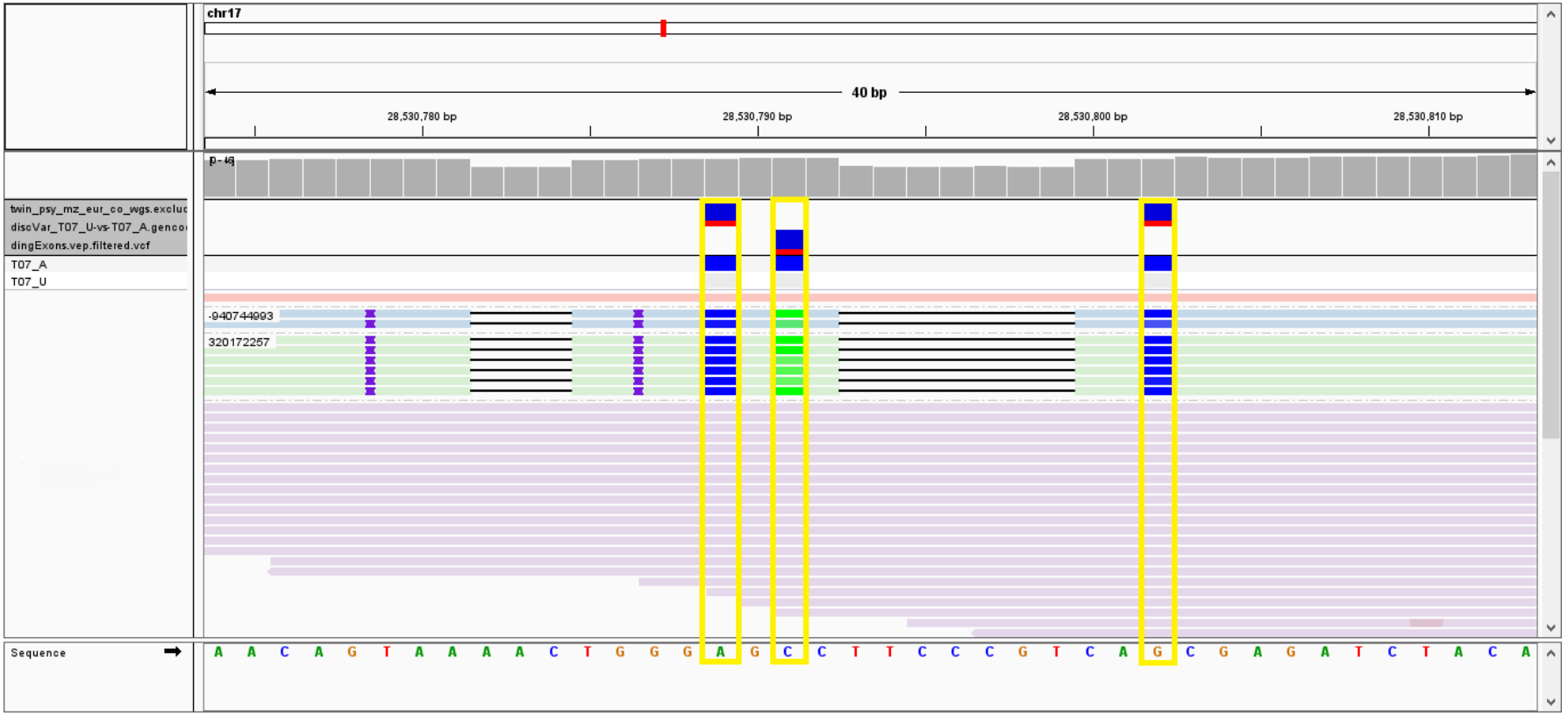


Figure 5.4: `IGV` plot of the bamout output of `HaplotypeCaller` for sample T07_A, showing three SNVs (highlighted in yellow) in *FOXN1* arising from re-constructed reads due to local re-alignment around indels. Reconstructed reads are displayed in colour, regular reads are shown in grey.

5.4 Regulatory Variants

5.4.1 Rare Deleterious Variants

To investigate whether discordant variants across the genome had a predicted regulatory effect, sites were annotated using RegulomeDB (Boyle et al., 2012). RegulomeDB curates a collection of known and predicted regulatory elements across the genome and scores each variant according to the accumulation of evidence that the site has a regulatory effect (see "Web Resources", Subsection A.2.4). A RegulomeDB rank of 2 represents evidence of transcription factor binding (ChIP-seq data) and a transcription factor motif, as well as evidence that the variant lies under a DNase hypersensitive peak. A variant with a rank of 1 requires evidence suggesting it is within a known expression quantitative trait locus (eQTL), as well as some of the same evidence as a rank of 2. Discordant variants with predicted regulatory effect from RegulomeDB have been previously observed in MZ twins discordant for schizophrenia (Tang et al., 2017). The online version of RegulomeDB is aligned to GRCh37, so we downloaded the database to incorporate it into our analysis. Variants at unstable positions were removed as described in Chapter 3, and the positions for each variant were converted to GRCh38 using `liftOver`.

Variants were annotated with `vep` including: RegulomeDB scores on GRCh38, CADD v1.6 scores, and allele frequency (1000 Genomes Project and gnomAD v3.1). Given that variants with a regulatory effect can occur in non-coding regions, CADD was used to estimate deleteriousness since it is defined for all positions in the human genome. Additionally, gnomAD v3.1 was used since it has a higher collection of WGS samples than v2.1.1. The following filters were applied to all discordant variants across the genome:

i) RegulomeDB rank of the variant was 1 or 2;

ii) CADD Phred-like scores greater than 20.0;

iii) the allele frequency was $<1\%$ or absent in the appropriate population groups in the 1000 Genomes Project and gnomAD databases; and

iv) variants were not observed in any other samples within the cohort.

After applying the above filters, no variants were retained.

5.4.2 ENCODE Regulatory Features

We also evaluated whether there was an accumulation of discordant variants in regulatory regions in affected compared to unaffected individuals. We compiled a list of features with a known regulatory effect from the Encyclopedia of DNA Elements (ENCODE) (Dunham et al., 2012). Specifically, we selected:

i) proximal and distal enhancers;

ii) canonical promoter-like signals;

iii) DNase hypersensitivity sites;

iv) anchors from chromatin loops (ChIA-PET); and

v) transcription factor binding sites with footprints.

To this list we added:

vi) brain specific open chromatin regions (Bryois et al., 2018; de la Torre-Ubieta et al., 2018; Fullard et al., 2017, 2019);

vii) brain specific enhancers (D. Wang et al., 2018); and

viii) proximal promoters of protein-coding transcripts from GENCODE (Frankish et al., 2019).

All discordant variants passing QC metrics were subset to these eight regulatory annotation regions. Since the variants were assumed to be post-zygotic, they could be treated as independent events within a twin pair. However, it is possible that some variants may be part of the same linkage disequilibrium (LD) block. Due to a low overlap with reference data taken from the 1000 Genomes Project, we were unable to evaluate whether LD structure was present within the discordant variants, which may have an effect on Type I and Type II error rates for hypothesis testing. A two-tailed t-test at 95% significance was performed using the R software package between samples with or without a diagnosis to evaluate whether there was a significant difference in mean count of discordant variants overlapping a given regulatory annotation. However, no significant difference in mean counts was observed for any regulatory feature (see Figure 5.5 and Table 5.4 below).

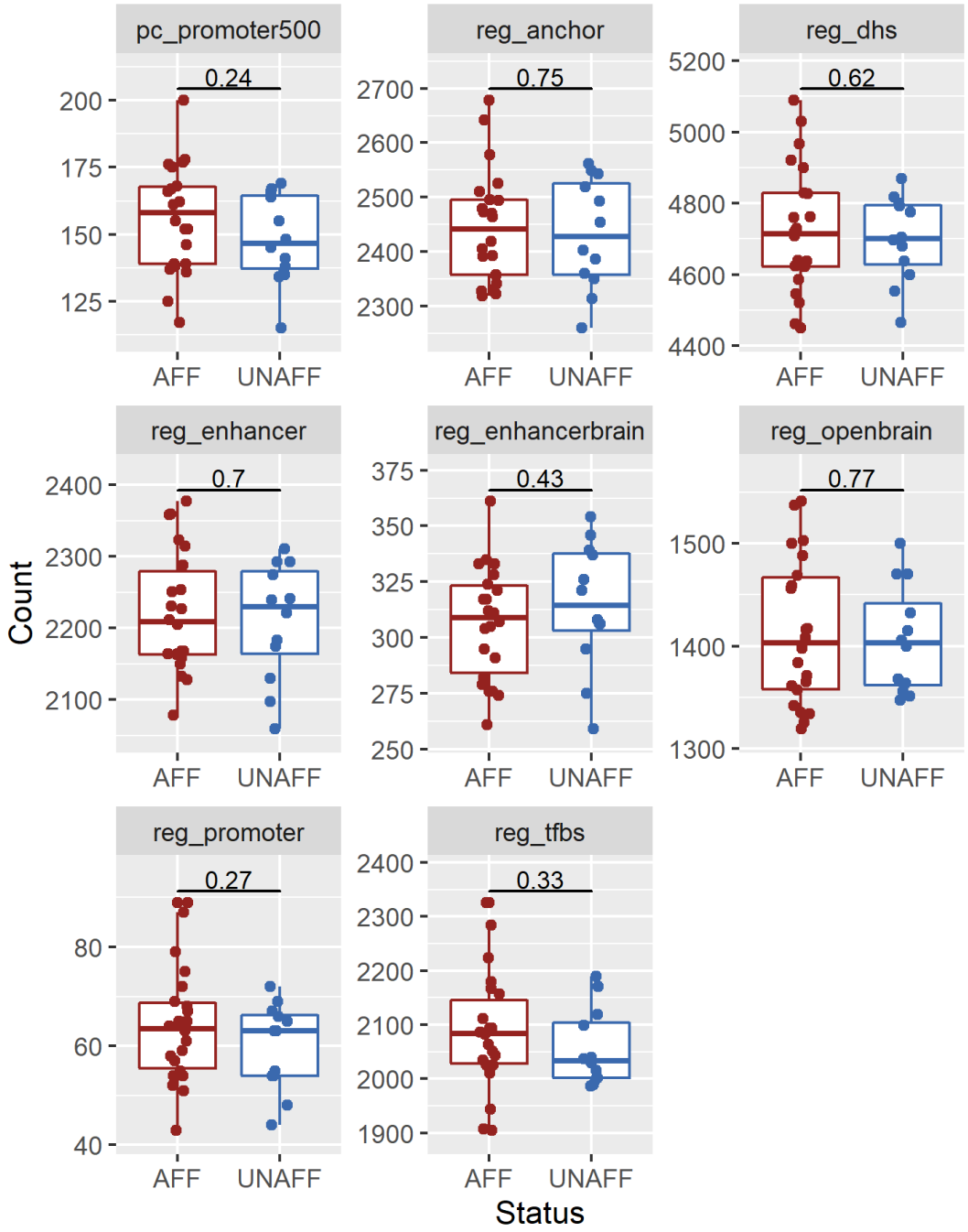


Figure 5.5: Boxplots of the counts of discordant variants within each of the eight regulatory annotation sets. The long name of the regulatory feature is shown in Table 5.4 below The 22 affected samples are displayed in red, and the 12 unaffected samples are displayed in blue. The p-value from a two-tailed t-test is displayed above each pair of boxplots. AFF: affected; UNAFF: unaffected.

Regulatory Feature	Description	Affected		Unaffected		d	t	p
		Mean	SD	Mean	SD			
pc_promoter500	Promoters of protein-coding transcripts	155.82	19.95	148.08	16.63	0.421	-1.206	0.239
reg_anchor	Anchors from chromatin loops	2,444.41	102.50	2,432.75	101.38	0.114	-0.319	0.752
reg_dhs	DNase hypersensitivity sites	4,725.36	177.42	4,699.42	119.91	0.171	-0.506	0.616
reg_enhancer	Proximal/distal enhancers	2,221.55	84.42	2,209.75	82.05	0.142	-0.397	0.695
reg_enhancerbrain	Brain-specific enhancers	306.45	25.03	314.33	28.67	0.293	0.800	0.433
reg_openbrain	Brain-specific open chromatin regions	1,412.96	70.75	1,406.58	52.25	0.102	-0.299	0.767
reg_promoter	Promoter-like signals	63.95	11.49	60.00	8.79	0.387	-1.121	0.272
reg_tfbs	Transcription factor binding sites	2,086.23	107.08	2,056.08	71.01	0.332	-0.982	0.334

Table 5.4: Results from the two-sided t-tests to evaluate the enrichment of discordant variants in various regulatory features between affected samples ($n = 22$) and unaffected samples ($n = 12$). Included is the mean and standard deviation (SD) of the counts, Cohen's d, the t-test statistic, and the p-value.

5.5 Germline CNVs

Another likely source of post-zygotic variation that may be driving phenotypic discordance is CNVs. Germline CNVs were called following the approach detailed in Chapter 4. For each sample, CNVs were combined with those of their co-twin to identify concordant and discordant CNVs. As described in Chapter 4, two CNVs were said to be the same if they had a 50% reciprocal overlap.

5.5.1 Known SCZ-Associated CNVs

First, we screened all CNVs (concordant and discordant) against a list of 23 rare CNVs previously implicated in schizophrenia (see Table 1.1 and Table 1.2). The breakpoints for these CNVs were originally given for GRCh37 but were converted to GRCh38 using `liftOver` for this analysis. Initially, it appeared that four SCZ-associated CNVs were present in both samples of the six twin pairs, all of whom had narrow discordance (see Table 5.5a). However, five of the six CNV calls were identified by `LUMPY` only in both twin pairs. For these CNVs, the proportion of reads supporting a CNV at either breakpoint was generally low in one or both twin pairs (see Table Table 5.5b). Therefore, we excluded these five CNV calls from further analyses.

The remaining CNV (a duplication on chromosome 13q12.11) was identified by all four CNV calling algorithms in both samples of twin pair T09. In a discovery association analysis, this CNV was noted to have a protective effect but was only nominally significantly associated with schizophrenia (Marshall et al., 2017). Interestingly, the affected individual in this twin pair T09 also has a rare, deleterious, discordant protein-coding variant in the *FLOT2* gene (see Table 5.3). *FLOT2* (Flotillin-2) has been shown to be involved in neuronal differentiation (Hanafusa & Hayashi, 2019) and flotillins are known to interact with the NR2A and NR2B subunits of N-methyl-D-aspartate receptors (Swanwick et al., 2009).

Chr	Start	End	ID	Twin Pair	Affected				Unaffected			
					C	E	L	M	C	E	L	M
chr3	195945160	197641345	3q29_DEL	T06			×				×	
chr16	21776296	22592576	16p12.1_DEL	T08			×				×	
chr13	19841000	19874999	13q12.11_DUP	T09	×	×	×	×	×	×	×	×
chr15	22776711	28851112	15q11.2_DEL	T11			×				×	
chr15	22775347	28851109	15q11.2_DEL	T12			×				×	
chr15	22776709	28851099	15q11.2_DEL	T17			×				×	

(a)

ID	Twin Pair	Affected			Unaffected		
		PE	DP_S (%)	DP_E (%)	PE	DP_S (%)	DP_E (%)
3q29_DEL	T06	4	89 (4%)	73 (5%)	4	131 (3%)	61 (7%)
16p12.1_DEL	T08	7	51 (14%)	59 (12%)	4	57 (7%)	68 (6%)
15q11.2_DEL	T11	10	53 (19%)	21 (48%)	11	40 (28%)	33 (33%)
15q11.2_DEL	T12	4	34 (12%)	35 (11%)	9	34 (26%)	30 (30%)
15q11.2_DEL	T17	4	25 (16%)	28 (14%)	4	41 (10%)	32 (4%)

(b)

Table 5.5: Putative SCZ-associated CNVs identified in the cohort. The start and end points are given for the GRCh38 reference genome. **(a)** A breakdown of which of the four callers identified the CNV (C - `CNVnator`; E - `ERDS`; L - `LUMPY`; M - `Manta`). **(b)** For the three CNV regions identified by `LUMPY` alone, the number of paired end (PE) reads that support the event, the read depth at the start of the CNV (DP_S) and the read depth at the end of the CNV (DP_E). Also shown beside the DP is the proportion of PE reads at the start or end of the CNV.

5.5.2 Rare and Pathogenic Discordant CNVs

Next, we examined discordant CNVs in the cohort. As described in Chapter 4, any CNV that was found in one sample of the pair and identified by only one calling algorithm was removed. Therefore, discordant CNVs have the support of at least two calling algorithms. Variants were then removed if they had at least a 50% reciprocal overlap with any common CNVs (i.e. frequency at least 1% in the appropriate population group) in the following public databases: gnomAD (Karczewski et al., 2019), the DECIPHER study (Firth et al., 2009), and the DGV (MacDonald et al., 2014), as described in Section 2.5. As the DECIPHER and gnomAD databases were curated relative to the GRCh37 genome build, the CNV files were converted to this build using `liftOver`.

After applying the above filters, seven rare CNVs that overlapped gene regions were identified in affected individuals in the cohort (see Table 5.6). However, the same number of discordant CNVs were identified in the unaffected twins. Of note, a duplication on chromosome 3q29 was observed in the affected sample of twin pair T17. While 3q29 deletions are associated with schizophrenia, 3q29 duplications have been implicated in autism spectrum disorders and developmental delay (Rehm et al., 2015).

The presence of this 3q29 duplication prompted us to examine a more extensive list of CNVs annotated by the NIH Clinical Genomics (ClinGen) CNV database as implicated in psychiatric or neurodevelopmental disorders (Rehm et al., 2015). Any discordant CNV that had a 50% reciprocal overlap with a variant labelled as "Pathogenic" in ClinGen (UCSC "iscaPathogenic" table) was retained, regardless of population frequency. Since ClinGen collates CNV calls from a wide collection of sources, each of which may use different reference material for CNV calling, it is not possible to know if the type of pathogenic CNV matches that of the CNV call in our data. Hence, CNV calls were not matched for type at this stage. Pathogenic CNVs were retained if the associated phenotype was psychiatric or neurodevelopmental in nature. Fourteen CNVs with a clinical impact were identified across the samples (see Table 5.7), but only the 3q29 duplication was present solely in affected individuals.

Chr	Start	End	Length	Locus	Type	Sample	Pheno	Path
chr1	32500722	32539739	39,017	1p35.1	DEL	T15_A2	MDD	
chr1	221964250	227275254	5,311,004	1q41-42.13	DUP	T05_A2	MDD	
chr3	38053588	48061842	10,008,254	3p22.2-21.31	DUP	T06_U	None	
chr3	195940567	197638156	1,697,589	3q29	DUP	T17_A	BD	×
chr5	180634984	180636040	1,056	5q35.3	DEL	T14_U	None	
chr7	98392886	98394241	1,355	7q21.3	DEL	T15_A2	MDD	
chr10	92847856	92849207	1,351	10q23.33	DEL	T02_A	SAD	
chr11	19858200	19861999	3,799	11p15.1	DUP	T17_U	None	
chr12	676066	1623296	947,230	12p13.33	DEL	T06_U	None	×
chr12	120201498	120204299	2,801	12q24.23	DEL	T16_A	BD	
chr13	50013200	50015499	2,299	13q14.2	DEL	T08_U	None	
chr19	2390200	2391236	1,036	19p13.3	DEL	T14_U	None	
chr19	11261852	11262999	1,147	19p13.2	DEL	T01_A1	SCZ	
chr22	39243400	39245099	1,699	22q13.1	DEL	T07_U	None	

Table 5.6: A list of rare discordant CNVs, including: the positions (GRCh38); length; type; the carrier sample ID; their phenotype (Pheno); and if they are annotated as pathogenic (Path). DEL: deletion; DUP: duplication; BD: bipolar disorder; MDD: major depressive disorder; SAD: schizoaffective disorder; SCZ: schizophrenia

Chr	Start	End	Length	Locus	Phenotype	Type	Disc		All
							A	U	
chr1	143707655	148368205	4,660,550	1q21.1-21.2	ASD; DD; GDD; ADHD; Seizures; ID	DEL	0	1	3
chr2	95974322	97579728	1,605,406	2q11.1-11.2	DD; ASD; ID; Seizures	DEL	4	1	7
chr3	1880400	6681218	4,800,818	3p26.3-26.1	DD; GDD	DEL	0	1	9
						DUP	1	2	5
chr3	195940567	197638156	1,697,589	3q29	ASD; DD; GDD; Seizures	DUP	1	0	1
chr5	141631941	142199762	567,821	5q31.3	DD; GDD	DUP	1	0	21
						DEL	1	0	27
chr7	66392490	76087266	9,694,776	7q11.21-11.23	DD; ID	DEL	1	0	7
						DUP	1	1	20
chr12	676066	1623296	947,230	12p13.33	DD; GDD; Seizures	DEL	0	1	1
chr12	120442095	121624256	1,182,161	12q24.31	GDD, Seizures	DUP	2	1	25
chr16	21935407	29107962	7,172,555	16p12.2-11.2	ID; DD; ASD	DUP	1	1	12
chr16	28470489	29397846	927,357	16p12.1-11.2	DD; Seizures; ASD; GDD; ID	DUP	1	0	11
chr17	36271243	37995798	1,724,555	17q12	DD; Seizures; GDD; ID; ASD; Anorexia	DEL	2	1	11
chrX	66699870	84875177	18,175,307	Xq12-21.1	DD	DUP	2	1	5
chrX	69751300	78489674	8,738,374	Xq13.1-21.1	DD	DUP	1	0	9
						DEL	1	0	7
chrX	94704302	98524458	3,820,156	Xq21.33	Seizures	DUP	1	0	3

Table 5.7: A list of CNVs with a predicted pathogenic effect in ClinGen, including: chromosome (Chr), start and end positions (GRCh38), length, the associated phenotypes, CNV type, the number of affected (A) and unaffected (U) samples who carried a discordant variant (Disc), and the overall number of samples who carried the variant (All). DEL: deletion; DUP: duplication; ADHD: Attention Deficit/Hyperactivity Disorder; ASD: Autism Spectrum Disorder; DD: Developmental Delay; GDD: Global Developmental Delay; ID: Intellectual Disability.

5.6 Somatic CNVs

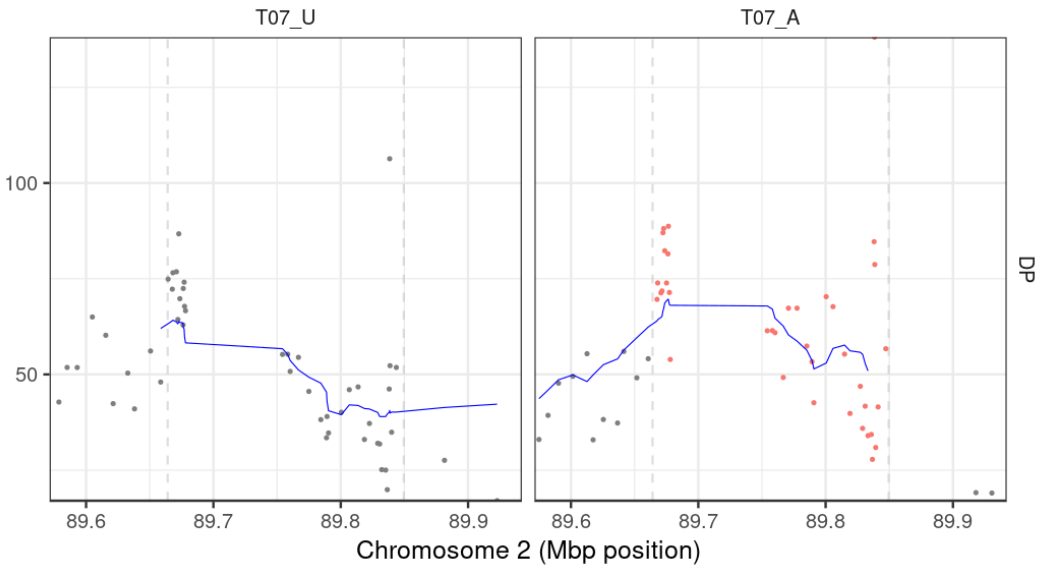
As noted above, we assumed that the post-zygotic variation occurred sufficiently early during embryogenesis, so they present as germline variants. However, somatic mosaicism can also have an effect on phenotypic discordance between twins (S. M. Singh et al., 2020). The typical average depth of coverage for WGS data is often not sufficient to detect somatic mosaicism present in a low proportion of cells for SNVs and indels. However, the tool Mosaic Chromosomal Alterations (`MoChA`) was designed to investigate the presence of somatic CNVs from genotype array or from NGS data (Loh et al., 2018). `MoChA` operates similarly to the read-depth based callers from Chapter 4, but instead of taking the BAM files as an input, it takes read-depth information from phased SNVs and indels of the samples.

`MoChA` was applied to the jointly genotyped short variant VCF file with default settings, including a list of regions to exclude for WGS data that was provided with the tool. As part of the process, multi-allelic SNVs and indels are normalised and genotypes with DP$<$10 or GQ $<$ 20 are set to missing, as described in Subsection 2.3.2. A workflow was recommended by the authors of the tool, which was provided on the tool's GitHub page (see "Web Resources", Subsection A.2.4). As with germline CNVs in Chapter 4, deletions or duplications for which at least 50% of their length comprised of RLCR were removed (see Subsection 4.2.1). Any somatic CNV call that had also been identified in an individual's germline call set in Section 5.5 above was removed. Finally, CNV calls present in both samples of a twin pair were removed to focus solely on discordant somatic CNVs.

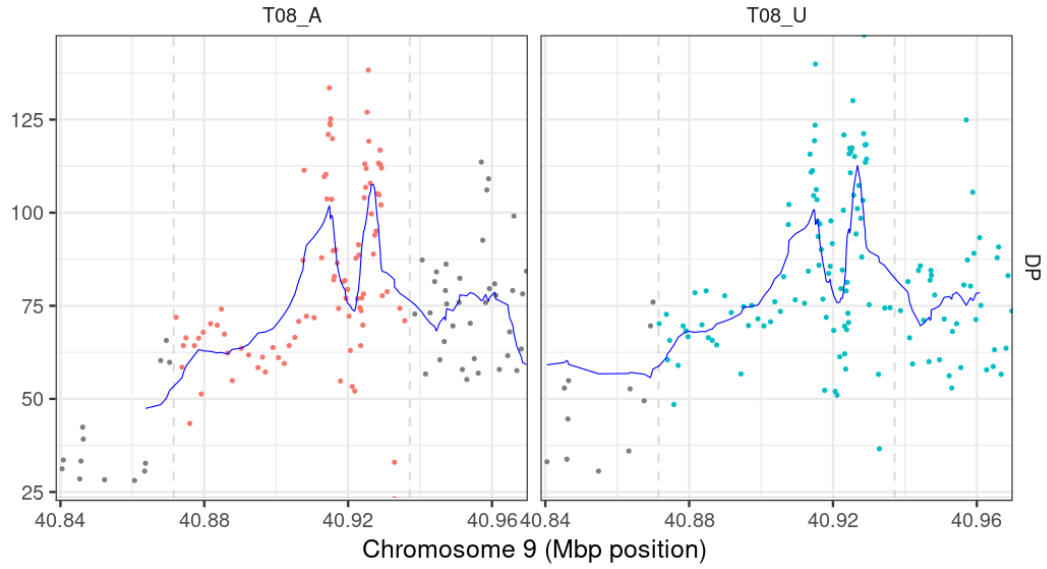
When `MoChA` was applied to the jointly genotyped data, 940 somatic CNV calls were identified across all samples (449 deletions and 491 duplications), with an average of 27.6 calls per sample. The subsequent filtering resulted in 25 putative discordant somatic CNVs (five deletions and 20 duplications), ranging in length from 9.5kbp to 404.1kbp. To confirm that the calls were truly discordant, the read-depth profiles for the regions of interest were plotted for carriers and their co-twin using the `mocha_plot.R` script provided with the tool. This script plots the read depth of the SNVs and indels in the CNV region for all samples provided, allowing us to compare the discordant calls within a twin pair. If the read depth plots appeared similar between both twin pairs despite only being called by `MoChA` in one sample, then the CNV was rejected as a false positive discordant call.

Three of the prioritised CNVs did not generate a read depth plot, so they were rejected as being low confidence calls. For each of the remaining 22 CNV calls, the read depth profile was almost identical between the two samples. For example, a putative discordant CNV call present in sample T07_A is shown in Figure 5.6a below. We can see that although the CNV was only identified in one sample the read-depth profile is virtually the same in the supposed non-carrier. Additionally, in Figure 5.6b, there appears to be CNV calls in both samples, but in T08_A the end breakpoint of the CNV is different to that of the co-twin despite their similar read-depth profiles. We queried these behaviours with the author of `MoChA` who informed us that: "*Phasing does not work very well with WGS data and there are many artifacts that tend to skew the variant allele fraction causing to call many false positives*" (G. Genovese, personal communication, 15/01/20).

Based on the above, all 25 prioritised CNV calls were rejected, and we conclude that there does not appear to be conclusive evidence for the presence of discordant somatic CNVs in these samples.



(a)



(b)

Figure 5.6: Read depth plots of two putative discordant somatic CNV calls which were subsequently rejected as false positives. The CNV of interest has points coloured in red, and other CNVs are coloured in blue. **(a)** A duplication detected in T07_A, with a similar profile in both twins. **(b)** A duplication detected in T08_A, but with different breakpoints to the CNV of the co-twin, despite similar profiles in both twins.

5.7 Multi-Nucleotide Repeat Expansions

Multi-nucleotide repeat expansions (i.e. where a short segment of DNA is repeated many times) are known to play a role in certain neurological disorders. However, there is mounting evidence that they may also play a role in psychiatric disorders (Xiao et al., 2021). For example, a GGCCCC repeat in *C9ORF72* is known to be causal for fronto-temporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) (DeJesus-Hernandez et al., 2011), but carriers are reported to be likely to experience psychotic symptoms (Devenney et al., 2017). A genome-wide enrichment of repeat expansions (also known as short-tandem repeats) has been observed in autism samples (Mitra et al., 2021; Trost et al., 2020) and recent work has shown an enrichment of rare, exon-disrupting repeats in schizophrenia samples (Mojarad et al., 2022). We examined whether the repeats causal for known disorders were present in the cohort. We selected 15 multi-nucleotide repeat expansions from Orr et al. (Orr & Zoghbi, 2007), and added the *C9ORF72* repeat. The location and pathogenic repeat count for these disorders is described in Table 5.8.

Multi-nucleotide repeat expansions were called from the BAM files of all 34 MZ twin

Disorder	Alias	Locus	Repeat	Count
Fragile X Syndrome	FRAXA	Xq27.3	CGC	200
Fragile X-Associated Tremor Ataxia Syndrome	FXTAS	Xq27.3	CGG	60
Fragile XE Syndrome	FRAXE	Xq28	CCG	200
Friedreich ataxia	FRDA	9q21.11	GAA	200
Myotonic Dystrophy 1	DM1	19q13.32	CTG	50
Myotonic Dystrophy 2	DM2	3q21.3	CCTG	75
Spinobulbar Muscular Atrophy	SBMA	Xq11-12	CAG	38
Huntington Disease	HD	4p16.3	CAG	36
Spinocerebellar Ataxia 1	SCA1	6p22.3	CAG	39
Spinocerebellar Ataxia 2	SCA2	12q24.12	CAG	32
Spinocerebellar Ataxia 3	SCA3	14q32.12	CAG	61
Spinocerebellar Ataxia 6	SCA6	19p13.13	CAG	20
Spinocerebellar Ataxia 7	SCA7	3p14.1	CAG	37
Spinocerebellar Ataxia 17	SCA17	6q27	CAG	47
Dentatorubropallidoluysian atrophy	DRPLA	12p13.31	CAG	49
Amyotrophic Lateral Sclerosis	ALS	9p21.2	GGGGCC	30

Table 5.8: A list of 16 selected disorders associated with a multi-nucleotide repeat expansion, and their pathogenic repeat count threshold.

samples using `ExpansionHunter` (Dolzhenko et al., 2017) with default parameters. Variant regions for the 16 multi-nucleotide repeat disorders were taken from the variant catalogue provided with the tool. `ExpansionHunter` was developed for PCR-free WGS data, whereas our data were generated with PCR-based methods. Since PCR-free data can result in a more even depth of coverage across the genome, we accounted for this difference by examining the coverage for each repeat region to ensure it was not low. The average depth of coverage was at least 14$\times$ for all multi-nucleotide repeat regions across all samples, and so the data were sufficient for repeat expansion calling (see Table 5.9).

While we did observe some differences in the repeat counts within some twin pairs,

none of the samples had repeat counts above the specified pathogenic threshold for any disorder (see Table 5.9).

Disorder	Average DP		Repeat Count		Threshold
	Mean	Range	Mean	Range	
FRAXA	25.3	14.0 - 37.8	13.6	4 - 30	200
FXTAS	33.5	19.1 - 54.2	13.6	4 - 30	60
FRAXE	42.5	33.6 - 54.0	6.3	4 - 17	200
FRDA	49.2	39.8 - 65.4	12.9	8 - 21	200
DM1	42.0	34.3 - 49.4	13.9	10 - 23	50
DM2	29.8	18.0 - 47.1	15.5	15 - 19	75
SBMA	40.2	32.4 - 52.6	23.4	18 - 27	38
HD	43.8	34.6 - 54.7	19.3	16 - 26	36
SCA1	28.0	16.7 - 47.6	31.5	29 - 37	39
SCA2	41.5	34.4 - 49.7	22.2	22 - 23	32
SCA3	37.6	25.6 - 53.0	20.9	18 - 25	61
SCA6	34.4	24.6 - 48.8	12.6	11 - 13	20
SCA7	43.3	37.6 - 51.3	10.6	8 - 13	37
SCA17	42.3	32.6 - 51.8	37.4	36 - 43	47
DRPLA	35.5	29.4 - 43.6	19.8	19 - 21	49
ALS	25.3	14.0 - 37.8	4.2	2 - 8	30

Table 5.9: For each of the multi-nucleotide repeat disorders, details across all 34 samples of the average depth of coverage (DP) and the repeat counts, as well as the pathogenic count threshold.

5.8 Conclusions

Here we report a WGS study where we assessed discordant post-zygotic variation in 17 MZ twins discordant for schizophrenia or a related disorder. We have investigated a broad range of genomic variation, from SNVs (both protein-altering and regulatory), to CNVs and repeat expansions. A rigorous filtering strategy identified 10 rare, deleterious, discordant, protein-coding SNVs across seven genes, each present in an affected member

of the cohort (see Table 5.3). A missense variant in *POLG* was observed in an individual with bipolar disorder, and this gene has been previously implicated in mood disorders. We also identified seven rare, discordant CNVs present in affected samples only. One such variant was a duplication in the 3q29 region in the affected sample of twin pair T17. While only deletions in this region have been shown to be associated with schizophrenia, this region has also been implicated in autism and developmental delay. This study is important as it contributes novel findings to the current body of literature for variants implicated in schizophrenia and related disorders and provides a framework for future studies.

Chapter 6

Rare Variant Analysis of Utah Pedigrees

Family-based studies offer a unique opportunity to map rare genetic risk variants, since risk in multiplex pedigrees is more likely to be influenced by the same collection of variants than in an unrelated cohort. Here we examine WGS data from 41 individuals across seven pedigrees multiply affected by schizophrenia. We applied an IBS filtering pipeline to search for protein-coding variants that co-segregated with disease status and further prioritised these based off results from the recent SCHEMA analysis. We identified deleterious missense variants in three genes (*ATP2B2*, *SLC25A28*, and *GSK3A*) that co-segregated with disease in three of the pedigrees. The most compelling evidence is from *ATP2B2*, which is involved in intracellular calcium homeostasis, is expressed in multiple brain tissue types, and is predicted to be intolerant to loss-of-function and missense variants.

6.1 Introduction

Rare variant analyses are frequently proposed as one response to the missing heritability problem for complex disorders, providing a complementary approach to common variant studies such as GWAS (Gibson, 2012). Indeed, 12 rare CNVs are known to confer substantial risk for schizophrenia (Rees & Kirov, 2021), and mounting evidence supports a contribution from rare protein-coding variants (Purcell et al., 2014; T. Singh et al., 2016). Recently, the Schizophrenia Exome Meta-Analysis (SCHEMA) consortium collated WES data on 24,248 schizophrenia cases and 97,322 controls from across multiple genomic ancestry super-populations (T. Singh et al., 2022). They examined the contribution of three types of variants to schizophrenia:

- protein-truncating variants (PTVs), also known as loss-of-function (LoF) variants, defined as: stop gained, frameshift, splice acceptor or splice donor variants;
- highly deleterious missense variants, with an MPC score $\geqslant 3$; and

- moderately deleterious missense variants, with $2 \leqslant \text{MPC} < 3$.

The first two types are referred to as Class I variants and the last type is referred to as Class II variants. The analysis focused on ultra-rare variants (URVs) that affected genes that were predicted to be intolerant to LoF variants. The SCHEMA consortium reported 10 genes in which the burden of URVs was significantly higher in cases when aggregated across both classes of variants. Highly deleterious missense variants were found to have as strong an effect size on schizophrenia as PTVs. Class II variants were also enriched in cases compared to controls, but the effect size was more modest than for Class I variants. These signals persisted even when these 10 genes were removed, suggesting that many more genes in which URVs contribute to schizophrenia risk are yet to be discovered.

Despite the success of the SCHEMA study, a major limiting factor is the sample sizes required to examine the URVs. Family-based studies offer a unique, alternative opportunity to identify and evaluate URVs, since risk in large multiplex pedigrees are more likely to be influenced by the same subset of variants compared to an unrelated cohort (Glahn et al., 2019). This reduces the need for extremely large sample sizes. Here, we examine WGS data from a collection of pedigrees multiply affected by schizophrenia. Since the SCHEMA analysis focused on ultra-rare variants, we begin by considering family-private variants, which have previously been shown to be enriched in multiplex ASD pedigrees (Wilfert et al., 2021). Given the modest sample numbers, we apply an IBS filtering approach to identify variants with reasonable co-segregation patterns with schizophrenia, as described in Subsection 1.2.3 above. Then, we prioritise the classes of variants examined from the SCHEMA analysis to find candidate causal variants within each pedigree.

6.2 Cohort Description

6.2.1 Sample Procurement and Assessment

Methods used in sample ascertainment and assessment were approved by the University of Utah Institutional Review Board (W. Byerley, personal communication, 10/06/21). Multiplex pedigrees were identified by screening hospitalized patients with diagnoses of schizophrenia. Following written informed consent, subjects were interviewed by a clinician using the Schedule for Affective Disorders and Schizophrenia-Lifetime Version ("SADS-L") (Endicott & Spitzer, 1978). Medical records were obtained for any individual who received psychiatric care. The interview results and any medical records were then presented to a diagnostic panel comprising two clinicians who played no role in as-

certainment or assessment. Consensual diagnoses were made using Research Diagnostic Criteria ("RDC") (Spitzer et al., 1978).

6.2.2 Description of the Pedigrees

Ten pedigrees were identified in which either schizophrenia (SCZ), bipolar disorder (BD) or major depressive disorder (MDD) was present in at least 4 individuals. Other psychiatric phenotypes were also observed such as obsessive-compulsive disorder (OCD) and suicide (SUI). Thirty-eight individuals across these pedigrees had previously undergone WGS (Batch 1) and 23 additional individuals were selected for sequencing at EGCG (Batch 2), giving a total of 61 individuals across the ten pedigrees (see Table 6.1). From this cohort, we selected the seven pedigrees in which schizophrenia was the dominant phenotype for analysis. However, all samples in the cohort were included in the quality control measures, since some of the tools used require a minimum number of samples.

Pedigree ID	In-Family		Marry-In		Phenotypes					Total
	AFF	UN	AFF	UN	SCZ	BD	MDD	OTH	UN	
K1480	4	-	-	1	4	-	-	-	1	5
K1494	4	-	-	1	4	-	-	-	1	5
K1501	5	2	-	1	4	-	-	1	3	8
K1509	2	1	1	1	-	3	-	-	2	5
K1524	5	-	-	-	5	-	-	-	-	5
K1527	5	1	-	1	5	-	-	-	2	7
K1545	5	-	-	1	5	-	-	-	1	6
K1546	5	2	-	1	5	-	-	-	3	8
K1561	6	-	-	1	1	4	1	-	1	7
K2159	4	-	-	1	3	1	-	-	1	5
Total	45	6	1	9	36	8	1	1	15	61

Table 6.1: Counts of the number of samples sequenced from each of the ten pedigrees, broken down by whether they were within the family or married in, and by their phenotype. AFF: affected (any diagnosis); BD: bipolar disorder; MDD: major depressive disorder; OTH: other related phenotype; SCZ: schizophrenia; UN: unaffected.

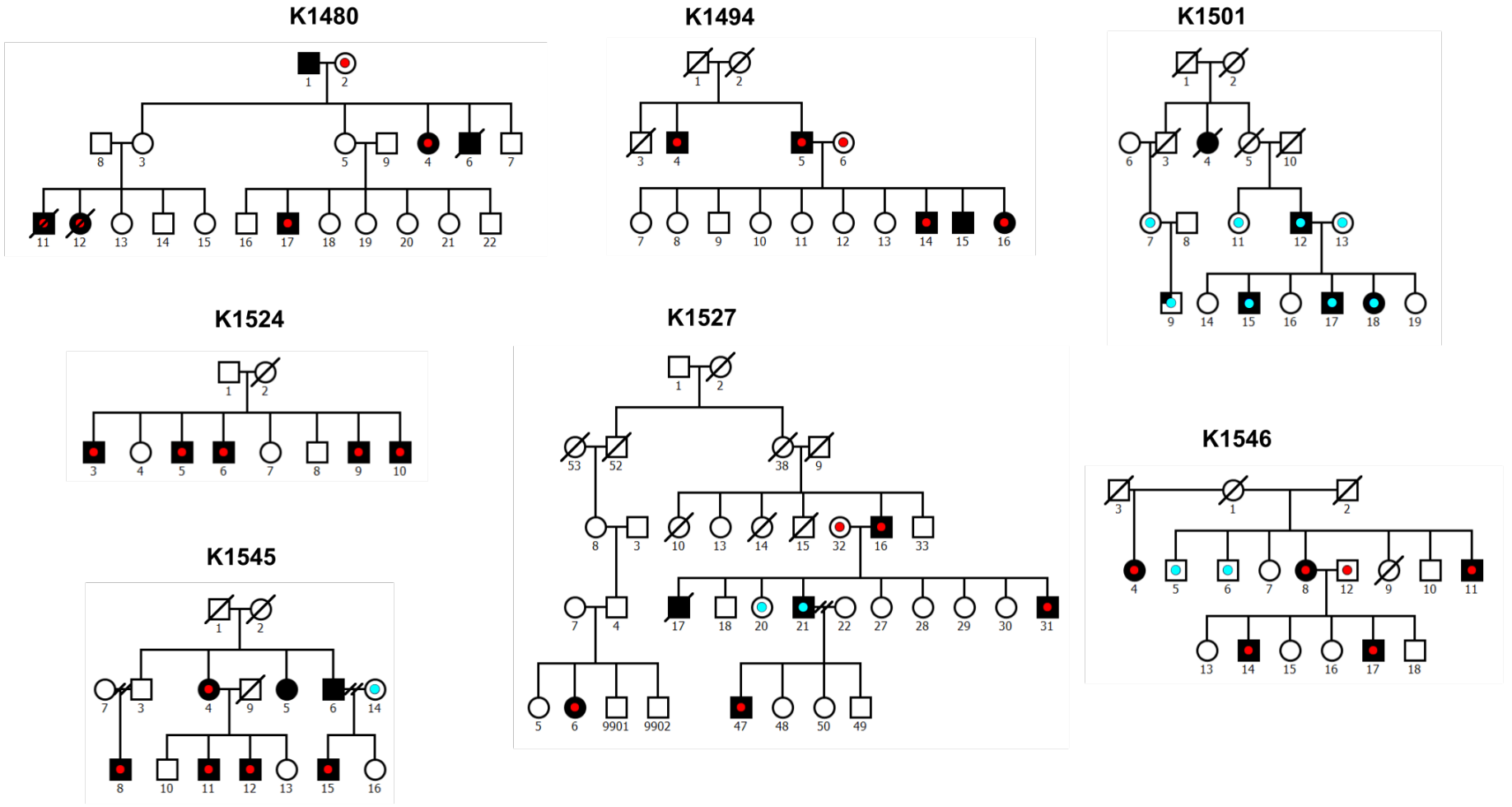


Figure 6.1: Pedigree diagrams for the seven pedigrees selected for analysis. Individuals fully shaded have a diagnosis of schizophrenia, and individuals with the top left quarter shaded have a diagnosis of OCD. Individuals marked with a coloured dot (red: Batch 1; blue: Batch 2) underwent WGS.

6.3 WGS Data and Sample QC

6.3.1 Batch 1

WGS had been performed on the 38 samples in this batch by MedGenome, Inc. on an Illumina HiSeqX to an average depth of coverage of at least $30\times$ per sample (W. Byerley, personal communication, 30/06/21). FASTQ files for 22 samples and BAM files for the remaining 16 samples were transferred to local servers for analysis. These BAM files had been aligned to the GRCh38 reference genome using the Sentieon Genomics proprietary pipeline (Freed et al., 2017). Although this pipeline is modelled on the `GATK` "Best Practices", we decided to convert the BAM files to paired-end FASTQ files to re-run the `GATK` alignment and variant calling pipeline. We converted the BAM files following guidelines from the `GATK` v3 website which recommends stripping all alignment information and shuffling the order of the reads prior to conversion to FASTQ. This is advised since deduplication algorithms sometimes retain the first read observed in a set of duplicates, so shuffling the order should minimize any biases from previous ordering. The steps in converting the BAM to FASTQ were as follows:

- Index and sort the file with `samtools`, so `picard` will not fail due to unsorted data;
- Revert the BAM to strip out any alignment information with `picard RevertSam`;
- Select any reads that have duplicated read names with `samtools` and remove them with `picard FilterSamReads`;
- Verify the information between read pairs, and fix if required using `picard FixMateInformation`;
- Add a default read group containing the sample ID with `picard AddOrReplaceReadGroups`;
- Validate that the SAM file has no errors with `picard ValidateSamFile`;
- Split the BAM file into SAM files with a maximum of 10,000,000 reads per file using the GNU `split` command;
- Shuffle each of the split SAM files using the GNU `shuf` command;
- Merge all the shuffled BAM files with `samtools`; and
- Convert the shuffled BAM file to paired-end FASTQ files with `picard SamToFastq`

Bash code for the above process can be found on GitHub (see "Web Resources", Subsection A.2.5). Following this process, we performed read alignment, data pre-processing, and variant calling on all 38 samples as described in Section 2.2 above.

6.3.2 Batch 2

Based on the availability of DNA and the samples that had previously been sequenced, we selected 23 additional samples for WGS. DNA concentrations were quantified locally by Nanodrop and the quality of DNA was determined by agarose gel electrophoresis by a member of our research group (Dr Amy Cole). All samples were sent to EGCG for WGS on an Illumina HiSeqX to an average depth of coverage of at least $30\times$ per sample. Sample K1501_9 failed quality control metrics at EGCG and was not carried forward for sequencing. All FASTQ files received from EGCG were examined using `FastQC` and `samtools` (H. Li et al., 2009) to screen for DNA contamination or degradation, but no sample from the 7 pedigrees was excluded on this basis.

6.3.3 Quality Control Measures

The software `peddy` (Pedersen & Quinlan, 2017) was used for sex, ancestry and relatedness checking for all samples jointly as described in Section 2.4. Sample K1527_20 was flagged as part of the sex check, as this sample appeared to be female from the pedigree information but was classified as male from the genetic data. We noticed that this sample also had a relatedness of close to 1.0 with sample K1527_21, who is a male. Both the pedigree IDs and DNA tube IDs of these two samples are similar, so it is likely that there was a sample ID mix-up when sending the DNA for sequencing. Therefore, sample K1527_20 was removed from this analysis. Examining the relatedness metrics also revealed that sample K1524_3 was unrelated to their four siblings, as well as all other samples in the cohort, and so was removed from this analysis. All other samples had pairwise observed relatedness scores that were consistent with expected relatedness scores. The PCA revealed that all samples from nine of the pedigrees were predicted to have European genomic ancestry, and all five samples from pedigree K1545 were predicted to have admixed American genomic ancestry (see Figure 6.2). Since the SCHEMA study was a multi-ancestry analysis, we did not exclude any pedigrees based on the results of the PCA.

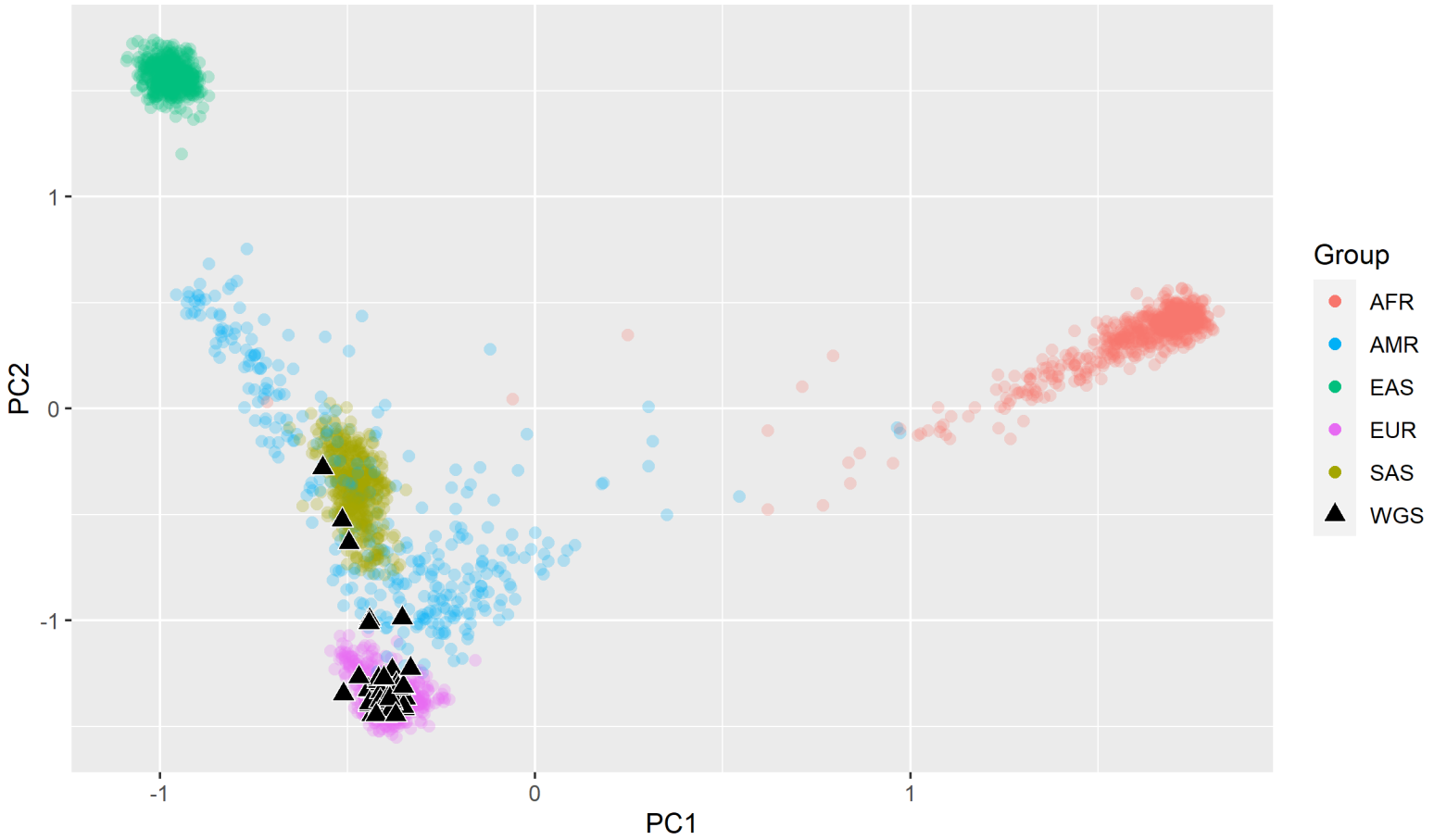


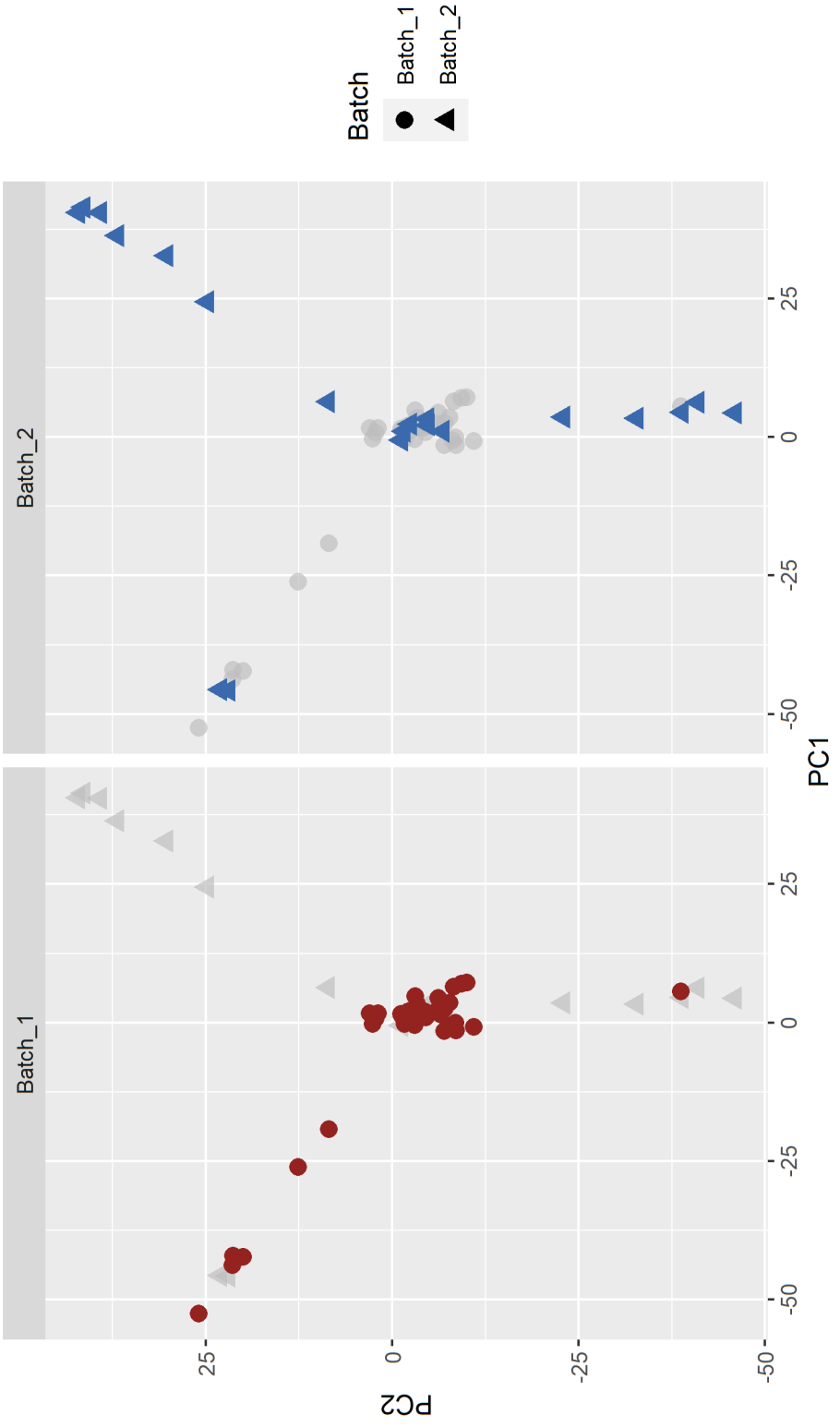
Figure 6.2: A plot of the first two principal components of the WGS samples and a background population from the 1000 Genomes Project, re-generated from the output of peddy using R (R Core Team, 2013). AFR: African; AMR: admixed Americas; EAS: East Asian; EUR: European; SAS: South Asian. WGS: all 57 WGS samples that passed QC criteria.

6.3.4 Cross-Platform Biases

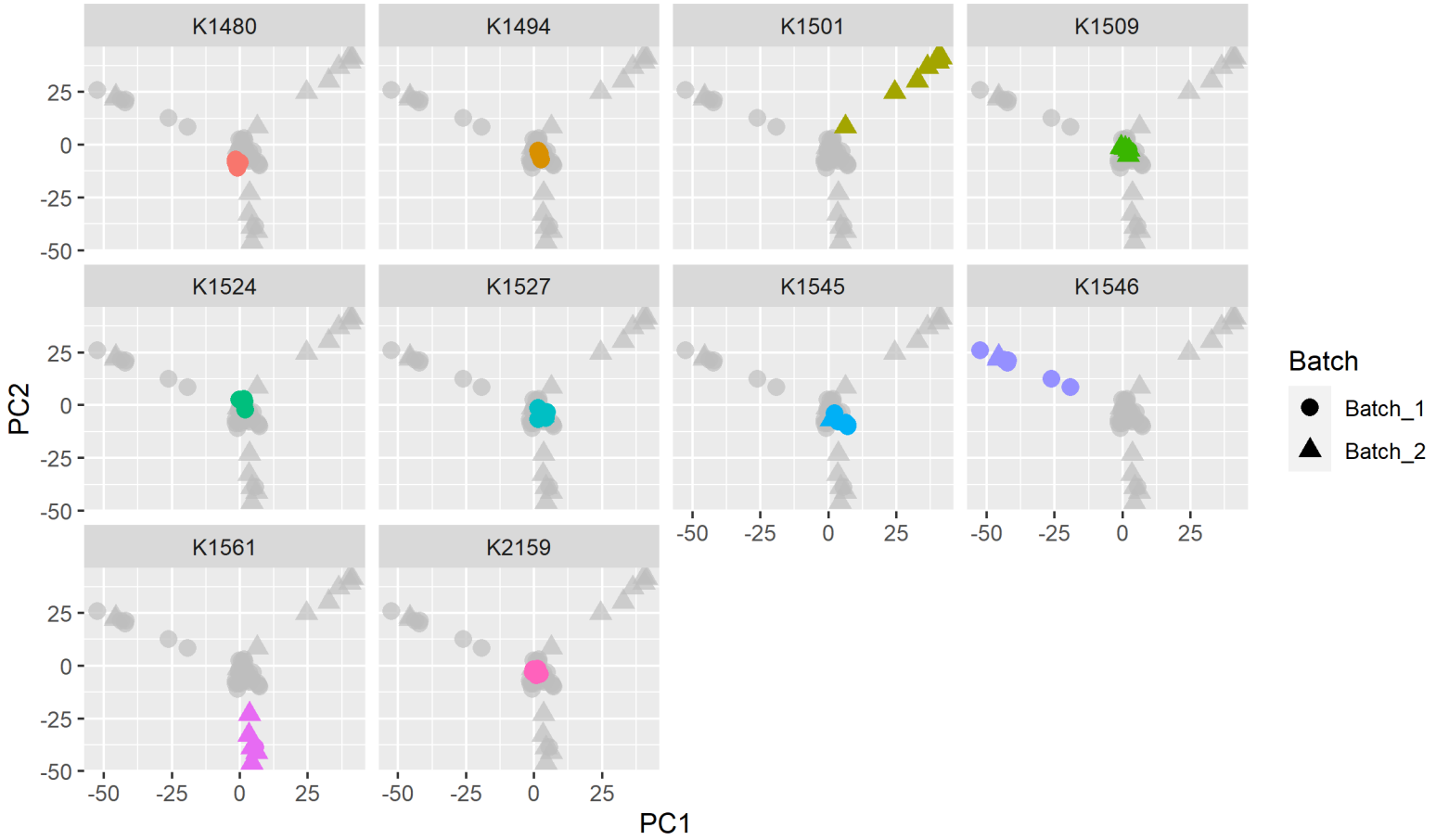
Given that we had two main sequencing batches it is possible that platform-specific biases may be present in the data, despite the variant quality control measures implemented. To identify any such biases, we applied the software `XPAT` to all samples that had not been excluded from the analysis. `XPAT` performs two rounds of PCA in an attempt to cluster samples based on sequencing platform (Yu et al., 2018). The first stage ("external PCA") is to identify genomic ancestry as standard, and the second stage ("internal PCA") aims to identify technological stratification. `XPAT` has minimum threshold requirements for the different batches, so we included samples from all 10 pedigrees to achieve this minimum. We examined the first two principal components of the `XPAT` internal PCA and did not observe any complete separation of samples by batch, although we did observe clustering by pedigree (see Figure 6.3). We note that the samples from pedigree K1501 were all sequenced in the same batch and form the right arm of the PCA plot in Figure 6.3, meaning that we cannot entirely rule out batch effects in this pedigree. Additionally, we looked at pairs of the first 10 principal components (e.g. PC1 vs PC2, PC3 vs PC4, etc.) but again only observed clustering by pedigree (see Figure 6.4). No pedigrees or samples were excluded at this stage of the analysis, leaving 41 individuals sequenced across seven pedigrees (see Table 6.2).

Pedigree	In-Family		Marry In	Total
	SCZ	UN		
K1480	4	-	1	5
K1494	4	-	1	5
K1501	4	2	1	7
K1524	4	-	-	4
K1527	5	-	1	6
K1545	5	-	1	6
K1546	5	2	1	8
Total	31	4	6	41

Table 6.2: Counts of the number of samples sequenced from each of the ten pedigrees, broken down by whether they were within the family or married in, and by their phenotype. AFF: affected (any diagnosis); BD: bipolar disorder; MDD: major depressive disorder; OTH: other related phenotype; SCZ: schizophrenia; UN: unaffected.

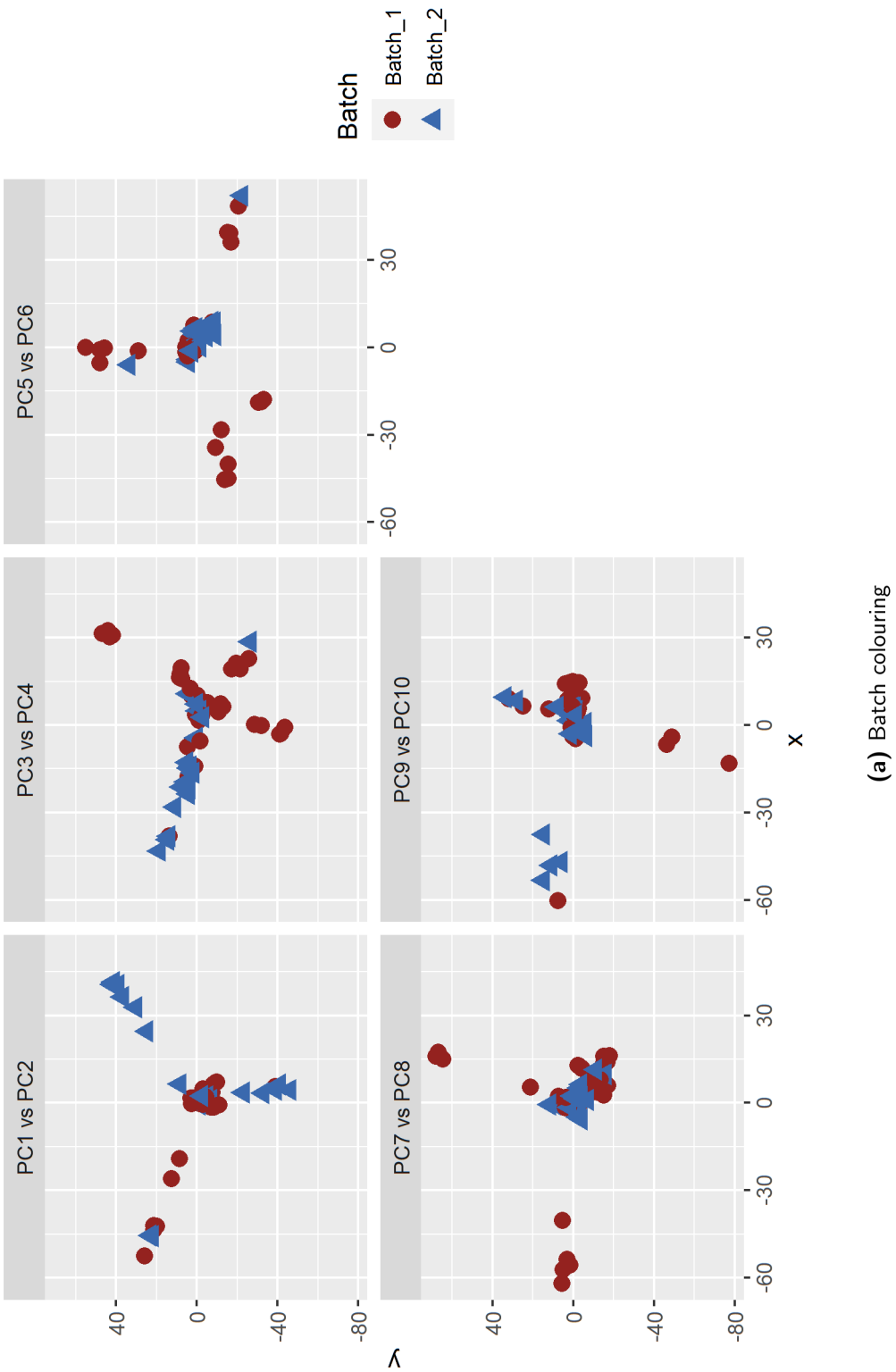


(a) Batch colouring

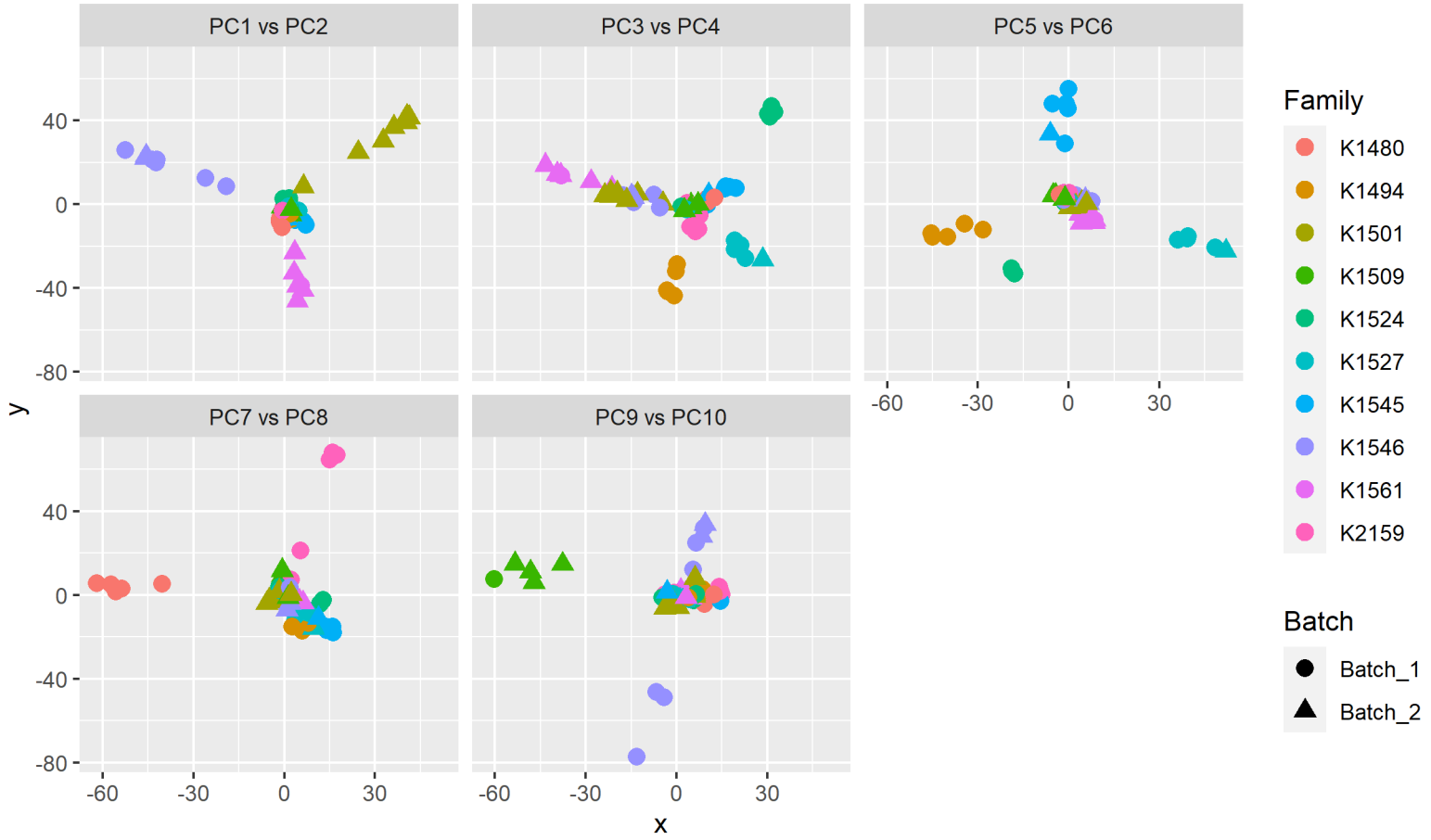


(b) Pedigree Colouring

Figure 6.3: Plot of the first two principal components from the internal PCA step of `XPAT`, aiming to identify technological stratification, with points highlighted for: **(a)** the batches; and **(b)** the pedigrees.



(a) Batch colouring



(b) Pedigree Colouring

Figure 6.4: Plot of pairs of the first 10 principal components from the internal PCA step of `XPAT`, aiming to identify technological stratification, with points coloured by: **(a)** batch; and **(b)** pedigree.

6.4 Protein-Coding Variants

Because the variants under investigation are ultra-rare, we assumed that it is highly unlikely that such variants would be present across multiple, unrelated pedigrees by chance. Therefore, we first subset to family-private variants only. All ten families were included in the family-private annotation, although only seven were carried forward for analysis as previously indicated. Additionally, any variant with missingness greater than 20% was removed. This percentage ensures that at least one individual outside every pedigree has a non-missing genotype, since the largest number of samples per pedigree is in pedigree K1546, which represents $8/41 = 19.5\%$ of the analysis cohort. This step makes it less likely that family-private variants are due to sequencing artefacts since they are confidently absent from other individuals in the cohort.

We considered all individuals with a diagnosis of schizophrenia to be cases, and all remaining individuals to be controls. Next, variants were retained if there were no Mendelian violations and they followed either a full co-segregation pattern, (carried by all in-family cases, absent from all in-family controls and absent from all marry-in samples) or a reduced co-segregation pattern (same as full co-segregation but allowing one in-family case not to carry the variant). Custom JavaScript code was added to the `FilterVcf` module from `picard` to identify the case/control status and in-family/marry-in status of samples for the co-segregation filter (see Section A.4 for pseudocode). We removed variants not present in the coding sequence of a protein-coding gene, as defined by the RefSeq `ncbiRefSeqCurated` table (O'Leary et al., 2016), downloaded from the UCSC Table Browser (Haeussler et al., 2019).

We used `vep` to annotate each variant, taking the canonical transcript of that gene. Where variants overlapped multiple genes, we examined the canonical transcript of each gene separately. As part of the annotation, we included the gnomAD v2.1.1 exome allele frequencies and dbNSFP v4.1 from which several deleteriousness metrics were extracted, namely: MPC, SIFT v2.2, PolyPhen2 v2.2.2 and CADD v1.6. Gene-based pLI scores, missense Z-scores, and loeuf scores calculated from gnomAD allele frequencies were also annotated from dbNSFP (see Subsection 2.5.3). Transcript-level information is available from dbNSFP, so where multiple scores were given for a variant, we identified the Ensembl canonical transcript ID from `vep` and extracted the scores from dbNSFP that corresponded to the appropriate transcript. To prioritize variants likely to be implicated in schizophrenia based on the SCHEMA work, we retained those that satisfied the following:

- ultra-rare in gnomAD, with minor allele count $\leqslant 5$ across all 125,748 samples;

- either PTV or predicted-deleterious missense variants (MPC > 2); and
- present in a highly LoF-intolerant gene (pLI > 0.9).

In total 15,428,001 variants passed all quality control measures following the joint genotyping of all samples, of which 2,371,087 were private to one of the seven families (see Table 6.3). After applying the main prioritization filters, no ultra-rare, functionally relevant variants were identified that had full co-segregation with cases. However, three deleterious missense URVs (Class II variants from SCHEMA) that followed a reduced co-segregation pattern were identified in three pedigrees (see Table 6.4 and Figure 6.5). We note that none of these variants were present in the K1501 pedigree, so these variants are unlikely to stem from batch effects (see Subsection 6.3.4 above). None of the three genes survived false discovery rate correction in the reported SCHEMA analysis, but there was a suggestive excess of the same class of missense variants at *ATP2B2* in the schizophrenia cases compared to controls in the SCHEMA dataset (see Table 6.5). *ATP2B2* has the highest missense Z-score of the three genes, indicating that it is the most intolerant to missense variants.

Description	**Variants**	
Quality control filters	15,428,001	
Family-private variants	2,371,087	
Co-segregation pattern	*Full*	*Reduced*
	41,644	129,721
In coding sequence	310	1,182
Ultra-rare in gnomAD	64	242
Functional relevance	0	10
LoF intolerant gene	0	3

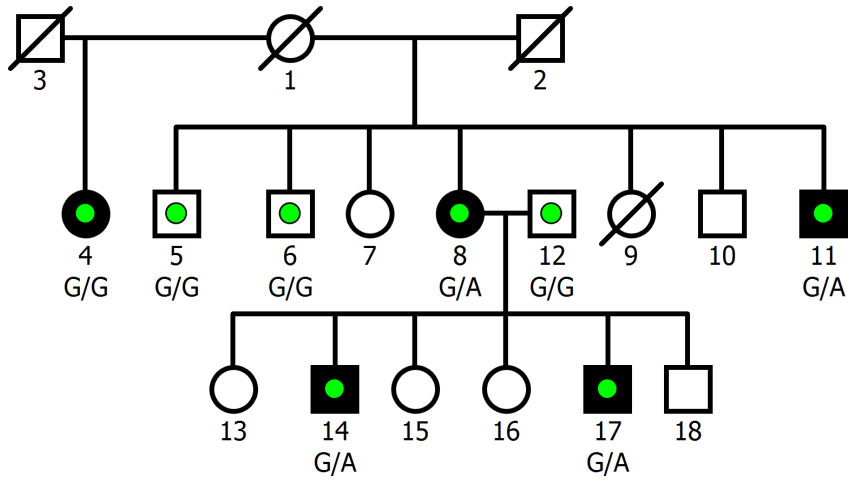
Table 6.3: The number of variants remaining after each stage of the prioritisation process across the seven pedigrees. Counts on the left are for full co-segregation and on the right are for reduced co-segregation. LoF: loss-of-function.

Pedigree	Chr	Position	Variant	Gene	Exon	HGVSp	MAC	MPC	CADD	SIFT	PolyPhen2
K1546	3	10360021	G>A	*ATP2B2*	13/23	R588C	1	2.23	31.0	D	D
K1524	10	99610923	T>C	*SLC25A28*	4/4	I341V	0	2.11	25.6	D	D
K1494	19	42232651	A>G	*GSK3A*	9/11	I377T	0	2.39	26.9	D	D

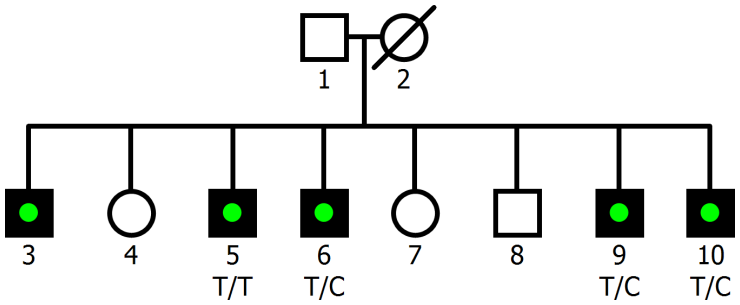
Table 6.4: Details of the three prioritized variants with reduced co-segregation. Positions are given on GRCh38. Included are the protein sequence ID for the variant (HGVSp), the minor allele count (MAC) from gnomAD exome data, and several deleteriousness prediction metrics. For SIFT and PolyPhen2, D represents "damaging" and "deleterious" respectively.

Gene	Constraint			SCHEMA (Class II)	
	pLI	mis_Z	LOEUF	OR	p-value
ATP2B2	1.00	4.55	0.15	1.920	0.000719
SLC25A28	0.93	2.92	0.37	0.617	0.744000
GSK3A	1.00	3.22	0.13	0.830	0.835000

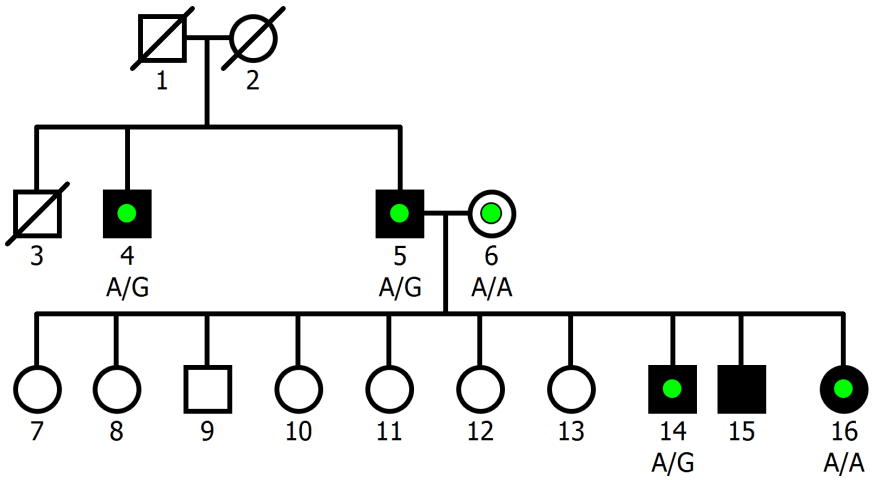
Table 6.5: Gene-level constraint information from gnomAD, including the pLI score, the missense Z-score, the LOEUF metric, and results (odds ratio (OR), p-value) for Class II variants from the SCHEMA analysis.



(a)



(b)



(c)

Figure 6.5: Pedigree images of **(a)** K1546; **(b)** K1524; and **(c)** K1494 which harbour an ultra-rare SNV with reduced co-segregation. Fully shaded boxes denote individuals with a diagnosis of schizophrenia, and sequenced individuals are marked with a coloured dot. The genotype of the identified SNV is shown beneath all sequenced individuals that were carried forward for analysis.

ATP2B2 ("ATPase plasma membrane Ca^{2+} transporting 2") is a member of the plasma membrane Ca^{2+} ATPase (PMCA) protein family which is involved in intracellular calcium homeostasis (O'Leary et al., 2016). It is found to be expressed in multiple brain tissue types in the GTEx project (Keen & Moore, 2015). In a genome-wide meta-analysis of ASD and schizophrenia, an intronic variant in this gene (rs9879311) was found to be genome-wide significant (Anney et al., 2017). Additionally, damaging *de novo* variants in *ATP2B2* have been shown to be significantly enriched in ASD cases compared to unaffected siblings in a Japanese cohort (Takata et al., 2018). A protein-protein interaction analysis of genes implicated in schizophrenia from both rare variants (CNVs and *de novo* SNVs) and common SNPs pointed to N-methyl-D-aspartate receptor (NMDAR) genes as having significant combined effects between rare and common variants (Chang et al., 2018). *ATP2B2* was found to be connected to the core members of this NMDAR interactome. A paralog of this gene is *ATP2A2*, which is a member of the sarco/endoplasmic reticulum Ca^{2+} ATPase (SERCA) protein family. Variants in this gene cause Darier's disease, which is known to increase risk for schizophrenia and bipolar disorder (Cederlöf et al., 2015). Fine-mapping of the significant loci from the PGC schizophrenia phase 3 GWAS identified an intronic variant of *ATP2A2* as highly probable of being causal (Ripke et al., 2020).

SLC25A28 ("Solute carrier Family 25 Member 28") is part of the mitochondrial carrier sub-group of the SLC gene family. It is a mitochondrial iron transporter that mediates iron uptake, and is expressed in most tissue types, including several brain tissues (Keen & Moore, 2015). There is no previous evidence of association between *SLC25A28* and schizophrenia or related disorders. *GSK3A* ("glycogen synthase kinase-3α") is one of the two isoforms of the GSK-3 protein kinase and is expressed in multiple brain tissues (Keen & Moore, 2015). Lithium, used to treat bipolar disorder, inhibits the activity of the paralog of this gene, *GSK3B* (Young, 2009). The variant in this gene was also present in the SCHEMA analysis, where one allele was observed in an individual with schizophrenia.

6.5 Copy Number Variants

CNVs were called according to Chapter 4. Within each pedigree, we removed CNV calls that were identified by only one calling method and were only found in one individual. The resulting calls have the support of either at least two calling methods or multiple individuals in the same pedigree. As in Subsection 5.5.1, for variants identified by one tool and present in multiple related samples, if the proportion of reads supporting an

event at the breakpoints was low across samples, such calls were removed. We screened CNV calls in all samples for any variants with a statistically significant association with schizophrenia (see Table 1.1). Although several such variants were initially identified in the cohort, most were excluded when examining the level of support at the breakpoints (see Table 6.6). Only one rare variant was retained, a duplication on chromosome 16p11.2 in sample K1524_5. This CNV was called by both read-depth based callers and was not observed in any other samples in this pedigree. Interestingly, this individual was the only sample in the pedigree not to carry the deleterious missense URV in *SLC25A28*.

Pedigree	Sample	Chr	Start	End	Type	Locus	Caller				LUMPY Support		
							C	E	L	M	SU	DP_S (%)	DP_E (%)
K1501	K1501_12	chr15	22775323	28847756	DUP	15q11			×		4	44 (9%)	56 (7%)
	K1501_15	chr15	22657718	28730831	DUP	15q11			×		4	40 (10%)	31 (13%)
K1524	K1524_5	chr15	30631954	32621480	DEL	15q13.3			×		4	28 (14%)	56 (7%)
	K1524_9	chr15	30631898	32621467	DEL	15q13.3			×		4	22 (18%)	23 (17%)
	K1524_5	chr16	29774001	30223000	DUP	16p11.2	×	×			-	-	-
K1546	K1546_4	chr3	195945211	197641359	DEL	3q29			×		4	68 (6%)	106 (4%)
	K1546_17	chr3	195945145	197641370	DEL	3q29			×		4	134 (3%)	107 (4%)
	K1546_8	chr3	195945163	197641349	DEL	3q29			×		4	104 (4%)	66 (6%)
	K1546_11	chr3	195945211	197641365	DEL	3q29			×		7	91 (8%)	112 (6%)

Table 6.6: Schizophrenia risk CNVs putatively identified in the cohort with a breakdown of which of the four callers identified the CNV (C - `CNVnator`; E - `ERDS`; L - `LUMPY`; M - `Manta`), and for the CNVs identified by `LUMPY` only, the number of reads that support the event (SU), the read depth at the start of the CNV (DP_S) and the read depth at the end of the CNV (DP_E). Also shown beside the read depth is the proportion of supporting reads at the start or end of the CNV. The start and end points are given for the GRCh38 reference genome.

6.6 Conclusions

We examined WGS data from 41 individuals across seven pedigrees recruited from Utah that were multiply affected by schizophrenia and performed an IBS filtering analysis to identify variants which are likely to increase disease burden. Following recent work from the SCHEMA consortium, we investigated the presence of ultra-rare, deleterious variants in LoF-intolerant genes. While no fully co-segregating pathogenic URVs were found, we did observe three missense variants with a reduced co-segregation pattern in three families. All three variants were predicted to be deleterious by additional pathogenicity metrics. In particular, *ATP2B2* has previously been implicated in schizophrenia, and the burden of URVs in this gene was nominally associated with schizophrenia in the SCHEMA dataset. Only one individual across the six families was found to carry a rare, schizophrenia risk CNV: a duplication on chromosome 16p11.2.

Chapter 7

Evaluation of Two Software Tools for Disease-Gene Prioritisation

Analysing pedigree based NGS data can be challenging, especially where sample sizes limit the power to derive significant results from linkage analysis. Two tools that aim to prioritise candidate disease-causing variants in a statistical framework are pVAAST and PERCH, which integrate novel measures of co-segregation with deleteriousness metrics. To better understand the strengths and weaknesses of these tools, we applied both to WGS data using a synthetic Mendelian phenotype in a three-generational pedigree. pVAAST performed well at identifying the pre-selected pseudo-causal variants, although PERCH did not, resulting in the removal of PERCH from subsequent analyses. We then applied pVAAST to the three pedigrees harbouring ultra-rare missense variants from Chapter 6. However, pVAAST did not score the missense variants (or their genes) favourably, likely due to the variants' low allele frequencies. Based on this, we also decided to remove pVAAST from subsequent analyses.

7.1 Introduction

In Chapter 6, we noted that genetic linkage analysis is the *de facto* methodology to identify regions of the genome that are inherited by affected individuals in family-based studies (Ott et al., 2015). A simpler, but non-statistical, IBS approach can also be applied, examining variants present in affected but absent in unaffected individuals within the family. While this strategy has been successfully employed for psychiatric disorders (Homann et al., 2016; Okayama et al., 2018; Steinberg et al., 2017), it has some limitations. Firstly, because there is no measure of co-segregation, there is no way to quantitatively rank or even combine results from different family structures. For example, we cannot know whether there is more evidence for causality from a large sibship or from a smaller but multi-generational family. Secondly, the requirement that all affected individuals carry a risk variant may be overly simplistic for complex disorders where there may be many risk variants at play, with reduced penetrance and even the presence of phenocopies. There is no obvious approach to relaxing this assumption consistently

across different family structures. In Chapter 6, we allowed one affected individual not to carry the variant of interest, but this rule does not account for the pedigree structure. Finally, the population-based filtering methods used to prioritise variants (e.g. deleteriousness metrics), even if guided by empirical work, are arbitrary and may vary from one research group to another. This hard filtering approach may remove reasonable candidate causal variants because one of their metrics is slightly less than acceptable.

Two tools which aim to address these issues are `pVAAST` and `PERCH` (described below). Both have their attractions in providing a framework for family-based next-generation sequencing analysis. However, both implement novel methodologies, and it is not immediately obvious how users could compare the output of such tools to more traditional forms of co-segregation analysis, or indeed to one another. In this Chapter we evaluate both tools, initially comparing their ability to identify variants causal for a synthetic Mendelian phenotype, representing a clear signal in a less complex genetic architecture. Since the gene-level scores are not directly comparable, we can estimate the null distribution of the scores of genes harbouring a causal variant. Next, we apply `pVAAST` and `PERCH` to the pedigrees in Chapter 6 to see how the tools behave on a complex phenotype, and to determine if they identify the same genes from the SCHEMA-based IBS filtering approach previously applied.

7.2 Software Tools

7.2.1 Description

The Pedigree Variant Annotation, Analysis and Search Tool (`pVAAST`) prioritises genes according to their evidence for association with a trait from both family and population data (Hu et al., 2014). This is an extension of the `VAAST` toolkit which compares unrelated cases and controls to identify candidate disease-causing variants (Hu et al., 2013). `pVAAST` implements a novel LOD score calculation for each variant based on classical parametric linkage analysis but designed for NGS data. This calculation incorporates the variant allele frequency and penetrance, and a grid search is used to find the values of the penetrance that maximises the underlying likelihood function. For each variant under consideration, `pVAAST` calculates a composite likelihood ratio test (CLRT) score which is the combination of the LOD score for the variant with its `VAAST` CLRT score. For each gene, a candidate variant is selected based on user-defined criteria, and the `pVAAST` CLRT score is used as the gene-based score.

pVAAST requires a set of target genomes (the pedigree data), a set of background genomes (unrelated affected or unaffected individuals), a phenotype, and a set of genomic features to be scored. Ideally, the target and background genomes should be generated under identical sequencing, alignment, and variant calling pipelines to remove technical stratification. Additionally, both cohorts should have similar genomic ancestry, so the allele frequency estimation from the background genomes is applicable to the target genomes. In the absence of pipeline- and population-matched controls, for this study a collection of 1,057 EUR-clustering exomes from a variety of sources and sequencing platforms aligned to GRCh37 was used (see "Web resources", Subsection A.2.6).

Polymorphism Evaluation, Ranking, and Classification for Heritable traits (PERCH) consists of a suite of independent Bayesian-based modules designed to evaluate a gene's relevance to a phenotype based on co-segregation, rare-variant association, and gene-gene interactions (Feng, 2017). It also includes BayesDel, a novel variant deleteriousness meta-predictor, trained using various conservation scores, other deleteriousness measures, and allele frequency. The co-segregation module (BayesSeg) is adapted from Thompson et al which calculates a log-Bayes Factor for the hypothesis that a variant is disease-causing versus neutral (Thompson et al., 2003). BayesSeg calculates a gene-based score by taking a weighted sum of the log-Bayes Factors for all variants within a gene passing quality control measures. The weight for a given variant is the BayesDel score plus the VQSLOD score derived from the VQSR process described in Subsection 2.3.1 above. Within a gene, the weights for all candidate variants are normalised which, the author states, ensures that the summation is robust to linkage disequilibrium.

7.2.2 Implementation

For pVAAST, the background genomes include some individuals from the CEPH 1463 pedigree (which we ultimately wish to use as our testing data), so these members were removed from the dataset using the "cdr_manipulator.pl" script provided. The target genomes VCF files were converted to the "condenser" (CDR) format required by pVAAST using the "vcf2cdr.pl" script provided. For the genomic features, GFF3 files for RefSeq gene and exonic features are provided by the authors online (see "Web Resources", Subsection A.2.6). We used the "common_complex_many_fam.ctl" parameter control file supplied with pVAAST. Since we are looking for highly penetrant variants, we set the "penetrance_lower_bound" parameter to be 0.5. Additionally, we set the "max_prevalence_filter" parameter to be 0.05. These values help simplify the parameter search space for pVAAST when calculating the co-segregation score. For each pedigree, the default variant selected to generate the gene-based score is the one with

the highest VAAST CLRT score. Since we are more interested in the co-segregation aspect of pVAAST, we changed the "informative_site_selection" parameter to 2 which will select the variant with the highest LOD score as a gene representative.

During a preliminary evaluation of PERCH, several issues with the original source code were identified:

1. The KING algorithm (Manichaikul et al., 2010) is used to estimate relatedness between pairs of individuals. However, using the default scripts accompanying the tool, KING fails in its execution. Prior to running KING, we added a plink command to regenerate the input binary files with no modification. With this change, the KING command executes successfully.

2. During PERCH's quality control step, a PCA is performed using plink. However, if there is a small number of individuals sequenced, the selected founders may not carry all variants in the call set. In this instance, the PCA will fail since at least one variant will have an estimated minor allele frequency of 0. To overcome this, we added the "--nonfounders" command to the plink PCA, which treats all samples as founders. This will create bias in the allele frequency estimation due to relatedness of the samples, but since the results of the PCA are not incorporated into the BayesSeg module, this was not expected to affect downstream results.

3. During the PCA clustering using the k-means algorithm, the number of clusters is set to 15. However, if there are fewer than 15 individuals, this will fail, as the maximum number of clusters possible will be the number of individuals. We modified the number of clusters to be the minimum of 15 or one less than the number of individuals in the analysis. As above, this was not expected to affect downstream results.

4. The vGrp module is used to generate gene-level scores from variant-level information. During the execution of PERCH, this tool encountered a segmentation fault with no additional details from the execution. When re-generating the C++ source code, we included a "-g" flag to the compiler in the Makefile to allow for code profiling. When the newly compiled version of the executable was run on identical input as before, no segmentation fault was observed for vGrp.

The above points were discussed with the author of the tool (Dr Bing-Jian Feng), who accepted the proposed changes as appropriate solutions to the issues (B.J. Feng, personal communication, 30/01/2020). Additionally, as discussed with the author, we set the "--penetrance" parameter to "0.01,0.5,0.5" which represents a dominant

effect, with a prevalence in the general population of 1%. PERCH includes additional variant level filtering as part of its quality control, examining metrics such as depth of coverage, genotype quality, etc. Since these have already been incorporated into the VQSR process (described in Subsection 2.3.1 above), these filters were disabled.

7.3 Mendelian Phenotype: CEPH 1463 Pedigree

7.3.1 Data

The ideal data to compare both software tools would be publicly available NGS data from a large number of members of a pedigree with a known genetic disorder, or a pedigree harbouring a variant highly validated as pathogenic for a particular phenotype. However, this scenario is uncommon as typically such variants may be identified from targeted sequencing or genotype array data. Here we obtained data from the CEPH 1463 pedigree from the database of genotypes and phenotypes (dbGaP, accession number phs001224.v1.p1), and generated a synthetic phenotype within this family. Seventeen samples from this pedigree underwent WGS to an average of 50x coverage on a HiSeq 2000 as part of the Illumina Platinum Genomes Project (Eberle et al., 2017).

Paired-end FASTQ files for each sample were obtained using the SRA toolkit (see "Web Resources", Subsection A.2.6). However, the tool encountered several "timeout exhausted" errors, resulting in small amounts of data loss for each file. This led to some lines in the FASTQ files being truncated, resulting in malformed reads and also the read order between paired FASTQ files being out of sync. This network issue was queried with the systems administrator of our server, but it was determined that this was not resolvable locally, given that the timeout could have occurred at any stage of the network. This was echoed by the authors of the SRA toolkit on their GitHub profile (see "Web Resources", Subsection A.2.6). These sync errors were resolved by removing singleton reads using the "`repair.sh`" command from the BBMap toolkit (see "Web Resources", Subsection A.2.6) with default parameters. An average of 0.1% of reads (range 0.07 - 0.2%) were removed from the FASTQ files. Given that this proportion is small, we deemed that the downstream data should not be severely affected as long as the average depth of coverage remained sufficiently high.

pVAAST requires that any marry-in individuals in a pedigree cannot have parental genotype data, so the route of transmission of the causal variant is pre-defined. Therefore, we excluded the two paternal grandparents and decided that the synthetic phenotype

would originate from the maternal grandfather. Additionally, `pVAAST` cannot be run using the GRCh38 reference genome, only on GRCh37 or GRCh36. Given that GRCh38 is our primary reference build, we aligned the WGS data to both GRCh38 and GRCh37 for later comparison. Read alignment, data post-processing, variant calling, and VQSR for the 15 samples in the pedigree were performed as described in Section 2.2 above. No sex or relatedness issues were detected using `peddy`, and samples NA12877 and NA12878 form part of the 1000 Genomes Project EUR individuals, so this pedigree must have European genomic ancestry. Variant-level quality control measures as described in Subsection 2.3.2 above were also applied. The average depth of coverage for the 15 samples was 47× (range 43-52×) when aligned to either genome build.

7.3.2 Pseudo-Causal Variant Selection

Given that there is no known genetic disorder in the CEPH 1463 pedigree, we created a synthetic phenotype by specifying a collection of inheritance patterns we might expect from a Mendelian disorder (see Figure 7.1 below). Creating such a phenotype gives us more control over the results, allowing us to be definitive about which variants are "disease-causing". We required that the maternal grandfather (NA12891) and the mother (NA12878) were affected, and that the maternal grandmother (NA12892) and the father (NA12877) were unaffected. To include as large a number of potential causal genes, we did not specify a phenotype status in the third generation. Alternatively, we can think of this as considering all $2^{11} = 2,048$ phenotype permutations possible in the third generation and selecting those for which a variant is present in the WGS data that has a dominant inheritance pattern with this synthetic phenotype.

We retained SNVs and indels that satisfied the following properties:

- **functional impact**: alters the amino acid chain in the canonical transcript of a protein-coding exon, determined by RefSeq (O'Leary et al., 2016);
- **inheritance**: present in all affected individuals, absent from all unaffected individuals, and non-missing genotypes in all samples in the third generation; and
- **deleteriousness**: CADD v1.6 Phred-scaled score of greater than 30.0, indicating that they are ranked in the top 0.1% of all DNA variants.

Variants satisfying the above are referred to as pseudo-causal variants (PCVs). We selected these filters as they mirror how causal variants for a Mendelian phenotype may present. As both tools use measures of deleteriousness in their score calculations, a

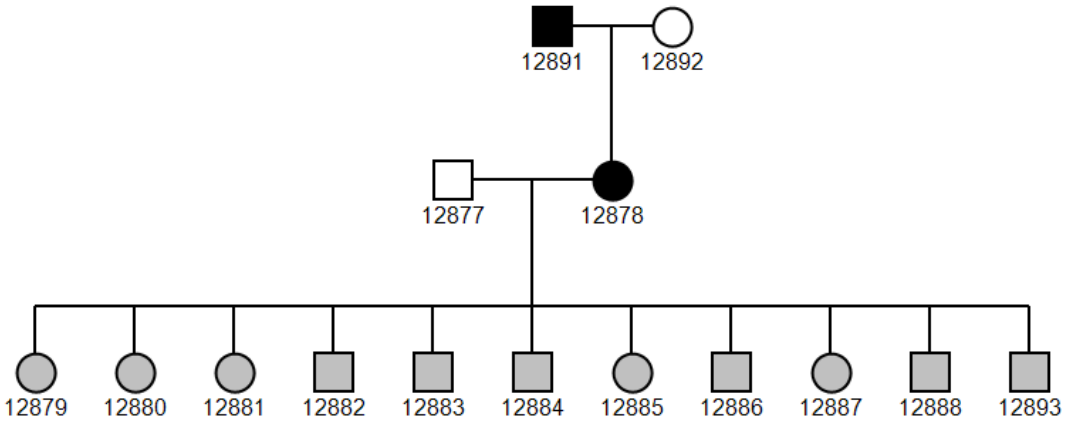


Figure 7.1: The CEPH 1463 pedigree with the paternal grandparents removed. Black boxes are affected, white boxes are unaffected and grey boxes may be either.

high CADD score should make the PCV evident when investigated. Initially, 26 such variants were identified in the pedigree. Since the data are aligned to GRCh38, we converted all variants in our original call set to GRCh37 (removing SNVs at unstable positions, as per Chapter 3) and applied the above filters to identify the PCVs. We also applied the same filters to data aligned to GRCh37 to compare the two lists of PCVs (see Table 7.1). We note that there are two PCVs which were present in the converted data list that were not present in the aligned data list. The *GPANK1* variant was not found in the GRCh37 aligned data, whereas the *MAGEB16* variant had a missing genotype for one sample in both the aligned and lifted data, and so was removed from the list.

During testing of both tools, there were some issues with the RefSeq annotation information for two variants which passed filtering. `pVAAST` takes gene feature information such as exon boundaries from a GFF3 file downloaded from the tool website (see "Web Resources", Subsection A.2.6). However, while the gene boundaries for *OR2J1* are present in this file, there is no exon boundary information, so no variant in the coding sequence can be evaluated. Similarly, `PERCH` takes gene feature information from a RefSeq-derived file provided with the tool. However, the gene boundaries for *KHDC1* in this file indicate that the prioritised PCV for this gene is actually intergenic. This is likely due to an older version of the transcript framework being used than that which identified the variant via filtering. For this reason, both *OR2J1* and *KHDC1* were removed from the list of PCVs. Thus, 24 PCVs were identified from the data converted to GRCh37, and all but two were also identified from the data aligned to GRCh37.

Chr	Position	Ref/Alt Alleles	Gene	Consequence	A/A	CADD	Align	Lift
chr1	53370357	G/T	*ECHDC2*	stop gained	Y/*	36	×	×
chr1	78392446	G/A	*NEXN*	missense	G/R	31	×	×
chr1	156314497	C/T	*TSACC*	missense	S/L	32	×	×
chr1	159410340	T/A	*OR10J1*	stop gained	C/*	33	×	×
chr2	32983480	G/A	*TTC27*	missense	R/H	31	×	×
chr2	88409984	G/A	*SMYD1*	missense	E/K	31	×	×
chr2	175292580	TTCAAATTTATCAG/T	*SCRN3*	frameshift	IQIYQ/X	33	×	×
chr3	98217178	T/A	*OR5K2*	stop gained	Y/*	33	×	×
chr3	113955187	A/C	*ZNF80*	stop gained	Y/*	31	×	×
chr4	38775922	G/T	*TLR10*	stop gained	Y/*	36	×	×
chr5	1240757	C/G	*SLC6A18*	stop gained	Y/*	34	×	×
chr6	31632134	C/A	*GPANK1*	missense	R/L	31		×
chr9	125239501	G/T	*OR1J1*	stop gained	C/*	36	×	×
chr9	131475583	G/C	*PKN3*	missense	D/H	31	×	×
chr11	55322818	G/T	*OR4C15*	stop gained	E/*	46	×	×
chr11	55339652	C/T	*OR4C16*	stop gained	Q/*	36	×	×
chr11	63057925	G/A	*SLC22A10*	stop gained	W/*	35	×	×

chr12	2943924	G/A	*NRIP2*	stop gained	R/*	35	×	×
chr12	52827608	C/T	*KRT75*	missense	A/T	31	×	×
chr13	78178550	G/A	*SCEL*	missense	R/K	35	×	×
chr16	89261482	C/A	*CDH15*	stop gained	Y/*	35	×	×
chr17	4803711	G/A	*C17orf107*	stop gained	W/*	39	×	×
chr17	74077797	C/T	*ZACN*	stop gained	Q/*	39	×	×
chrX	35821127	C/T	*MAGEB16*	stop gained	R/*	33		×

Table 7.1: The pseudo-causal variants identified by the filtering process, including: chromosome (Chr) and position (on GRCh37), variant (reference/alternate allele), gene symbol, predicted functional consequence, amino acid substitution (A/A), CADD v1.6 Phred-scaled score, and whether the variant was identified in the data aligned to (Align) or lifted to (Lift) GRCh37.

7.3.3 Results

For each of the PCVs, we ran pVAAST and PERCH on the same CEPH 1463 input data to evaluate how well both tools could identify such variants and their genes. For a given PCV, we selected the phenotypes in the third generation so that the variant would have a dominant inheritance pattern. For each tool, we determined the rank of the gene harbouring the PCV according to the respective score. Additionally, we could determine whether the PCV contributed to the gene-level score.

For pVAAST, NA12878 was designated the pedigree representative since they must be a variant carrier for all the PCVs by design. For all PCVs, each gene was ranked in the top 10 genes according to the overall pVAAST score (see Figure 7.2). The LOD score for each gene was the maximum LOD score achieved by all genes for that phenotype combination, so we can consider the pseudo-causal gene to be ranked joint first in all analyses when ranking by the LOD score alone. Additionally, for each pseudo-causal gene, the PCV was listed as one of the candidate causal variants, having achieved the maximum LOD score in that gene. From this, we see that pVAAST performs well at identifying variants following dominant inheritance patterns with high CADD scores.

For PERCH, the initial pre-processing and annotation stages were run using the default script provided with the tool. Fourteen of the genes harbouring PCVs received a positive BayesSeg score and were ranked either first or second in all genes scored (see Figure 7.3). However, nine of the genes received a BayesSeg score of 0, indicating either that the gene could not be scored, or possibly that there was equal evidence for the pathogenic model as for the neutral model. One gene (*SLC22A25*) got a negative score of -1.57, indicating that there is greater evidence that the gene does not harbour variants co-segregating with the synthetic phenotype. PERCH identified three variants from *SLC22A25* which may contribute to the final BayesSeg score: the PCV, a variant with an identical genotype pattern to the PCV and a third variant present only in the maternal grandfather. However, BayesSeg gave all three variants a co-segregation score of -1.57, which is not expected given the obvious differences in co-segregation between these three variants. We queried this unusual behaviour with the author of PERCH but did not receive a reply.

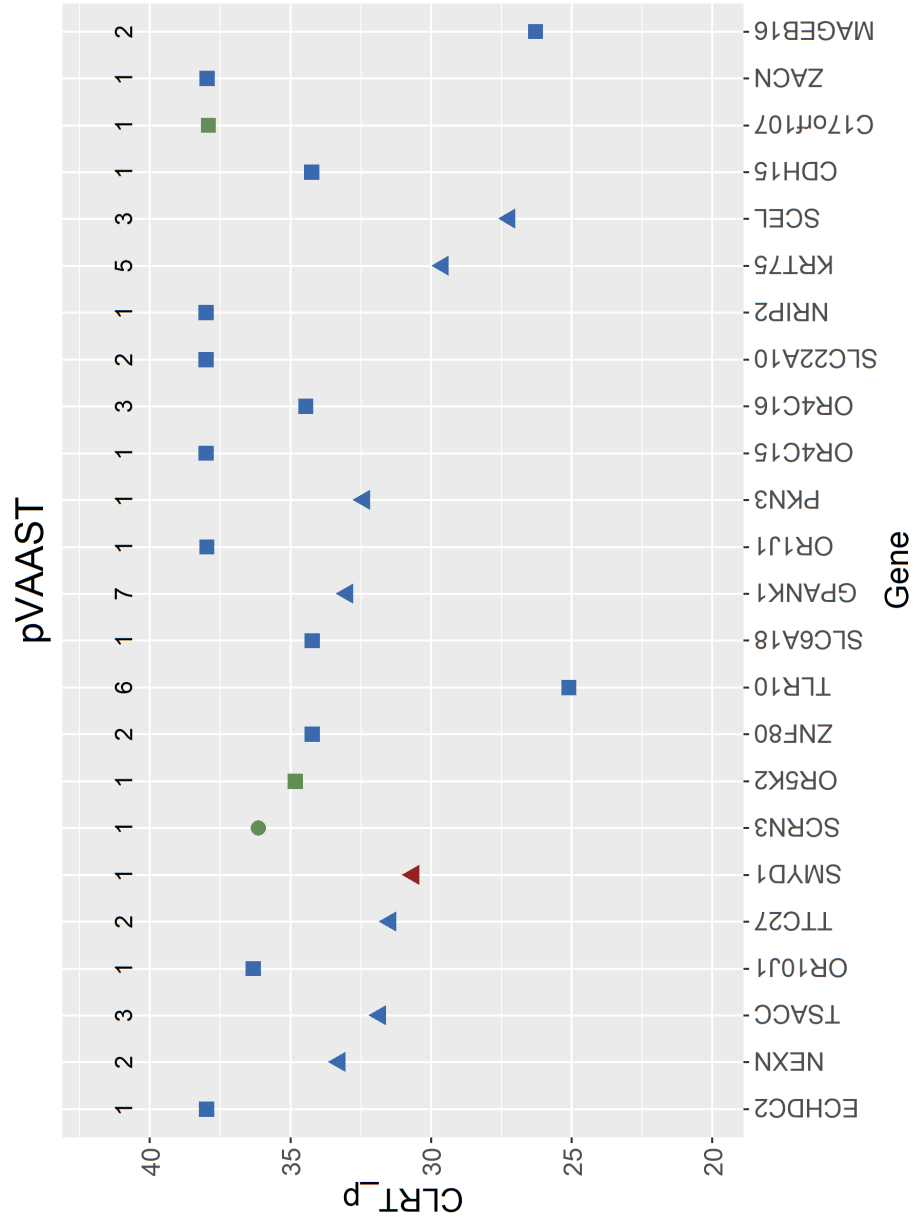


Figure 7.2: The pVAAST CLRT scores for all genes harbouring a PCV. The corresponding LOD scores are indicated by the colour, and the functional consequence of the PCV are indicated by the shape. The rank of the gene out of 19,492 genes scored is displayed at the top of the graph. Genes are ordered according to genomic position.

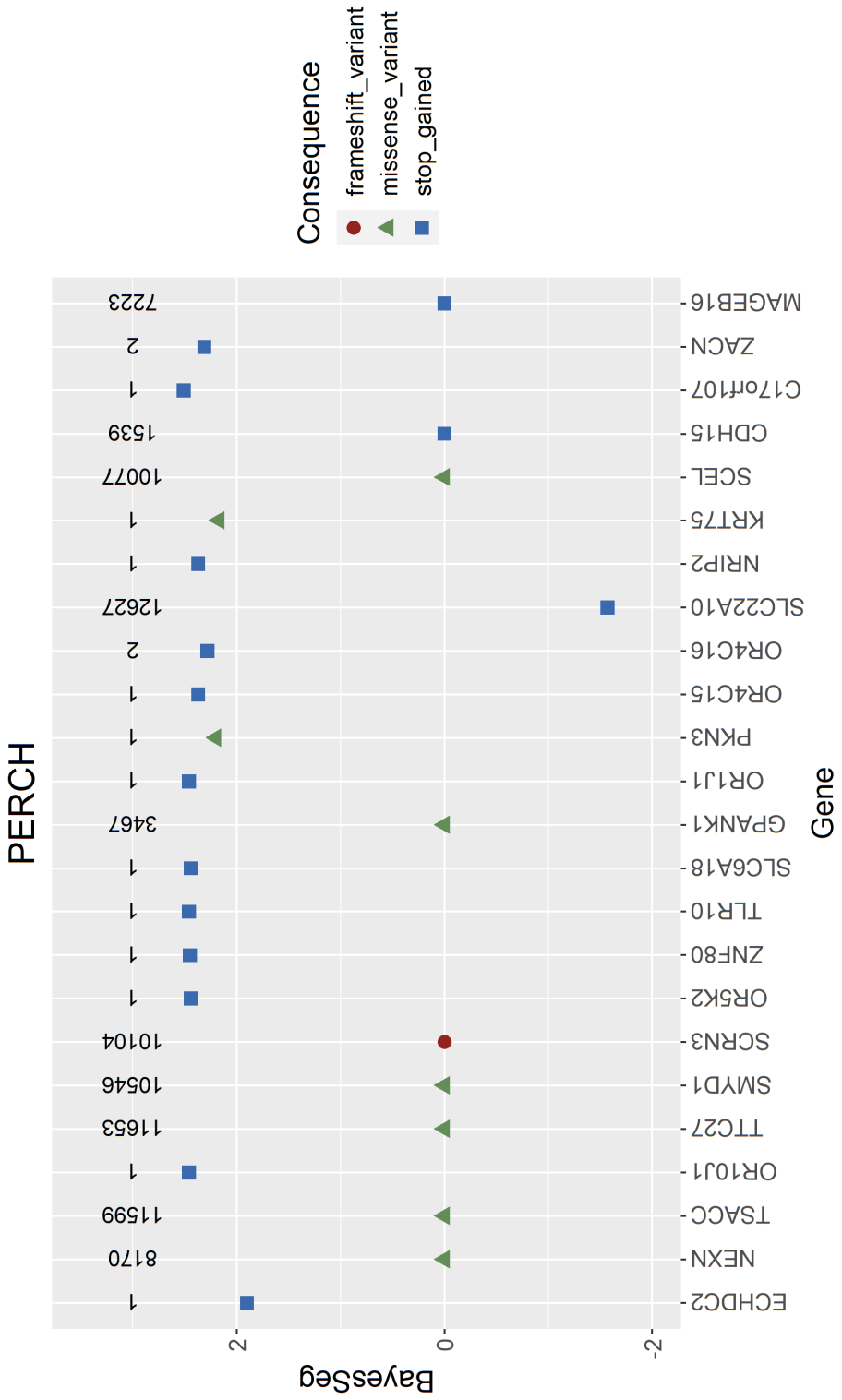


Figure 7.3: The BayesSeg scores from PERCH for all genes harbouring a PCV. Positive scores indicate evidence for causality, negative scores indicate evidence against causality and scores of zero indicate that the test was inconclusive. The functional consequence of the PCV is indicated by the shape. The rank of the gene out of 13,124 genes scored is displayed at the top of the graph. Genes are ordered according to genomic position.

7.3.4 Empirical Significance Calculation

Both the CLRT score from pVAAST and the BayesSeg score from PERCH integrate a novel measure of co-segregation variant deleteriousness to give rise to a gene-based score. To allow for comparison, we estimated the null distribution of these two scores by permuting the phenotype status of the individuals in the pedigree and running both tools on the same genetic data as before. Since there are 24 genes harbouring PCVs, we require at least $24/0.05{=}480$ phenotype permutations to accurately determine significance after Bonferroni correction for multiple testing. We chose 1,000 random permutations of the phenotypes for our estimation. The pVAAST CLRT score was found to be significant for all genes except *TLR10*, where the empirical p-value falls short of the corrected cut-off threshold, discussed below. The PERCH BayesSeg score was significant for all genes which received a positive score and was not significant for all other genes (see Figure 7.4).

For *TLR10*, three permutations generated a pVAAST CLRT score greater than the score for the true phenotype configuration. In each of these instances, the PCV was identified as one of the candidate causal variants for that gene, but the variant received a modest LOD score (LOD=0.97 for all three permutations) compared to the score for the true phenotype configuration (LOD=3.61). The pVAAST CLRT is given by the formula $\text{CLRT}_p = 2\log(10)\times\text{LOD}+\text{CLRT}_v$ where CLRT_v is the VAAST score. This VAAST CLRT score is not fixed for each variant like other deleteriousness metrics but depends on the allele frequency in the affected and unaffected cohorts. These values are estimated from the target and background data, so a change in the phenotype configuration will have an impact on the VAAST CLRT score, since the allele frequencies in the affected and unaffected cohorts will have changed. The VAAST CLRT score for the PCV in *TLR10* is 8.48, which is the lowest VAAST score across all genes harbouring a PCV. Comparatively, under the three other phenotype configurations, the VAAST CLRT score for the PCV was 20.67. This is the likely explanation for the other variants with modest LOD scores, but comparatively high VAAST scores, ranking higher than the true pVAAST score for *TLR10*.

Based on these results, we decided to continue to examine the performance of pVAAST in non-synthetic pedigree data, with a more complex genetic architecture. While the behaviour of pVAAST for *TLR10* is not precisely consistent with our expectations, pVAAST does perform well overall. However, at this point we decided to exclude PERCH from further analyses, given that we were unable to find an explanation for the unexpected behaviour of the algorithm with the genes harbouring PCVs.

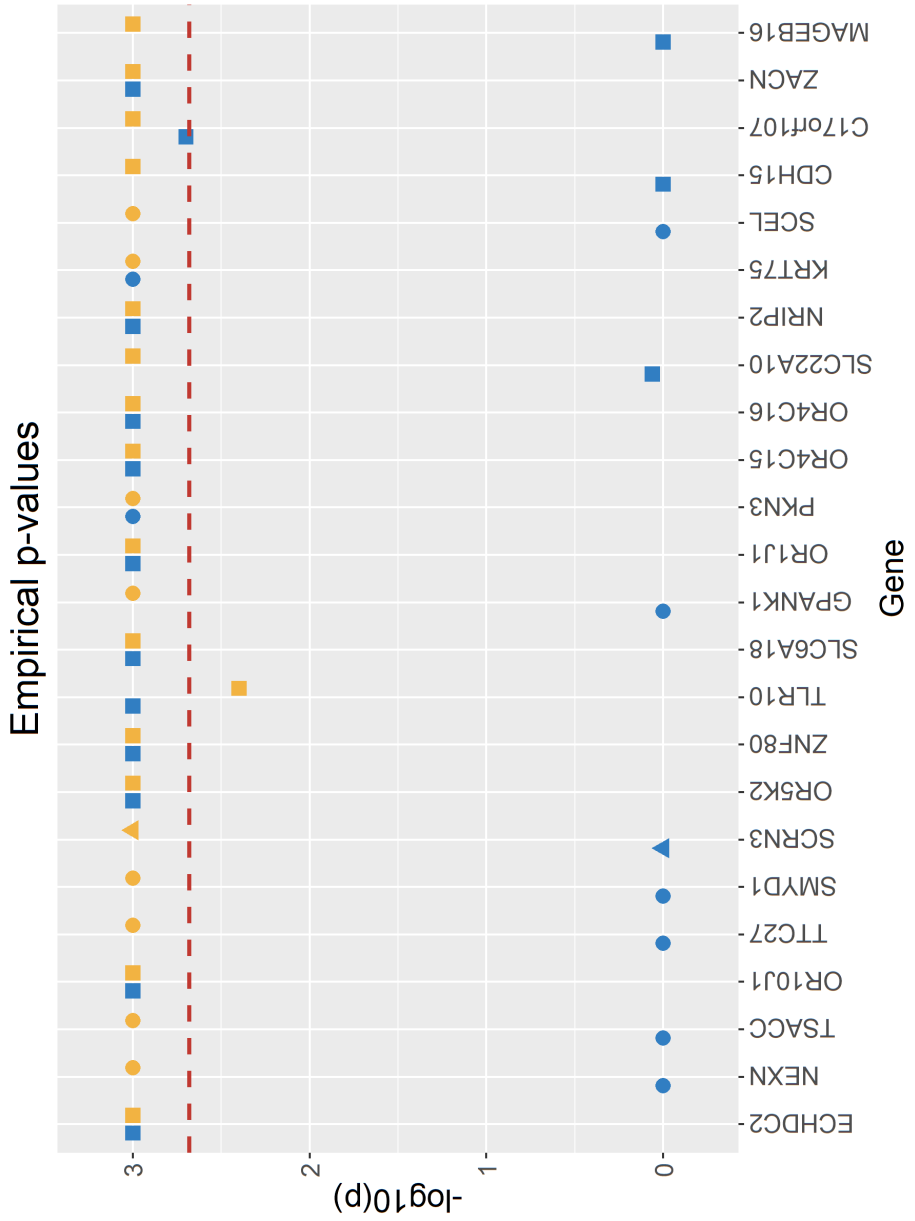


Figure 7.4: The empirical p-values of the co-segregation scores from `pVAAST` and `PERCH` plotted on the -log10 scale for the genes harbouring the PCVs. The red dashed line indicates the Bonferroni-corrected cut-off threshold. Genes are ordered according to genomic position.

7.4 Complex Phenotype: Utah Pedigrees

We examined how pVAAST would perform in the analysis of the three Utah pedigrees from Chapter 6 to see if it could identify the genes containing the three SNVs prioritised using our IBS filtering approach. The co-segregation of these variants does not follow a dominant inheritance pattern (although all sequenced carriers are affected) and the CADD v1.6 (GRCh38) scores for these variants are 25.6, 26.9, and 31.0, so this should add some complexity to the disease-gene prioritisation. However, it is expected that the three genes should score well overall.

Since the data for the Utah pedigrees were generated on GRCh38, we removed variants at unstable positions according to Chapter 3 and converted the quality-controlled data to GRCh37. We ran pVAAST on each of the three pedigrees separately to determine whether the three prioritised genes would be identified. However, both *SLC25A28* and *GSK3A* received CLRT and LOD scores of zero (see Table 7.2). We note that the variants in these two genes were not found in the background genomes file, likely because they are ultra-rare across multiple genomic ancestry groups. pVAAST includes the option to pass external allele frequencies for the background genomes, and so we extracted allele frequencies for all variants in the three pedigrees from gnomAD v2 exome data for the non-Finish European (NFE) group. However, when these external allele frequencies were provided, pVAAST stalled in its execution every time it was run and never completed its scoring.

ATP2B2 received a non-zero CLRT score (likely since the prioritised variant was found in one sample in the background genomes), but the LOD score was 0.14. While fewer samples were sequenced in K1546 than in the CEPH 1463 pedigree, the largest LOD score returned by pVAAST for K1546 was 2.16, so a near-perfectly co-segregating variant should not have so low a score. While examining the other gene and variant scores for this pedigree, we observed a variant in *INPP5K* which received a LOD score of 0.61 and ranked 7th overall. However, this variant was present in one sample in the final generation of the pedigree, and either absent or missing from all remaining samples. These LOD scores are not consistent with what we might have expected based on the observed co-segregation pattern within the pedigree. We queried this behaviour with the authors of pVAAST but did not receive a reply.

Pedigree	Gene	CADD v1.6		pVAAST		
		GRCh37	GRCh38	Rank	CLRT	LOD
K1546	*ATP2B2*	28.3	31.0	212	16.35	0.14
	*INPP5K**	35.0	36.0	7	24.15	0.61
K1524	*SLC25A28*	25.0	25.6	6,173	0.00	0.00
K1494	*GSK3A*	27.7	26.9	11,749	0.00	0.00

Table 7.2: The pVAAST scores for the three genes prioritised from Chapter 6, along with an additional gene from pedigree K1546, marked with an asterisk (*).

7.5 Conclusions

We have shown that pVAAST performs well at identifying deleterious variants with a Mendelian inheritance pattern in the CEPH 1463 pedigree but has a limited ability to detect URVs with a complex inheritance pattern in the three Utah pedigrees. We also showed that PERCH was unable to detect many of the deleterious, Mendelian variants. Both tools appear to have much to offer for disease gene prioritisation, combining multiple orthogonal data sources in a unified framework. However, the limitations described above make them unusable as comparison tools for our data, so we discounted them from further analyses.

Chapter 8

A Bayesian Framework for Pedigree-Based Causality

Identity-by-state filtering is commonly used for pedigree-based genetic analyses but suffers from limitations due to its non-statistical nature. For example, it is not clear how to quantitatively rank or combine information from multiple pedigrees or how family structure may be accounted for consistently. Here, we present a Bayesian framework to model causality of genomic variants in pedigrees. For each input variant, a Bayes factor for causality is calculated under the assumptions that a causal variant will be rare and have a dominant effect. A prior probability of causality can be calculated from population-derived metrics such as deleteriousness, allele frequency, and functional consequence. We applied this method to the CEPH 1463 pedigree and to the three pedigrees in Chapter 6 in which variants had been previously prioritised by identity-by-state filtering. The resulting metrics for the prioritised variants ranked favourably among all other variants, but our Bayesian approach now provides the advantage of a quantitative measure for prioritisation. Additionally, other genes of interest in the Utah pedigrees of Chapter 6 are identified by our model that were removed by the hard filtering strategy.

8.1 Introduction

One of the issues highlighted in the previous chapter is that the novel metrics used by `pVAAST` and `PERCH` to prioritise variants and genes may be difficult to interpret. In Bayesian modelling, a Bayes factor is used to compare the evidence for two competing models based on a specific set of data, e.g. that a variant is disease-causing verses that it is not. This is comparable to frequentist statistics where a null and alternate hypothesis are formulated, often evaluated using a p-value. However, such tests using frequentist statistics make no statement about how likely either of the hypotheses are *a priori*. Bayesian statistics can resolve this by allowing a prior distribution of the models to be incorporated.

Bayesian inference models have been used to evaluate variants based on their co-

segregation within a pedigree. To classify variants of unknown clinical significance, Petersen et al. derived an approach which calculates a Bayes factor for a variant to be disease-causing (Petersen et al., 1998). This model was originally designed for missense variants with a predicted dominant effect in *BRCA1* and *BRCA2*, which are associated with breast and ovarian cancer. Parameters underlying the model are the family-specific penetrance of the variant (the probability of the phenotype for variant carriers), the phenocopy rate (the probability of the phenotype for variant non-carriers) and the allele frequency of the variant. A proband (an affected variant carrier) is selected, and the Bayes factor is calculated using the genotypes of the affected variant carriers in the pedigree. The backbone of this term is the probability of the observed genotypes conditional on the genotype of the proband.

This model was extended by Thompson et al., who included the genotypes of unaffected variant carriers (Thompson et al., 2003). This down-weighted variants with high phenocopy rates, even if they had a high penetrance. Additionally, some individuals without genetic data may have their genotypes inferred based on the inheritance pattern, e.g. in-family parents of known carriers. These extensions allow all available genetic data to be incorporated into the co-segregation model. This is the model used to calculate a LOD score by `BayesSeg`, the co-segregation module of `PERCH` (Feng, 2017), but the gene-based LOD score is a weighted sum over multiple variants rather than a straightforward Bayes factor.

Mohammadi et al. considered a similar setup but calculated a likelihood ratio for causality also conditional on the phenotypes of the pedigree (Mohammadi et al., 2009). In addition, they assumed that the causal variant would be rare in the general population, which reduces the search space of all variants. Part of the likelihood ratio calculation involves iterating over all possibilities for the unobserved genotypes in the pedigree, and the authors describe strategies for simplifying this process with the rare-variant assumption. Additionally, Mohammadi et al. allow for age-specific penetrance and phenocopy rates, which are assumed to follow a normal distribution. This is more realistic for cancer phenotypes and adult psychiatric disorders, since an individual may carry a causal variant but may not present with the phenotype until later in life.

Despite their differences, both Thompson et al. and Mohammadi et al. methods have been shown to perform similarly at classifying simulated pathogenic and benign variants co-segregating in known cancer genes such as *BRCA1* and *MLH1* (Rañola et al., 2018). However, the likelihood ratios contain information about co-segregation only by design,

and no measure is included about the pathogenicity of the variant in question. Our aim in this chapter is to calculate a Bayes factor based on the method described in Mohammadi et al. Then, we can use population-derived metrics such as allele frequency and deleteriousness to construct a prior probability for causality, instead of considering a flat/uninformative prior as in the previous models (Petersen et al., 1998; Rañola et al., 2018). Finally, this new prior and the Bayes factor may be combined to generate a posterior probability of causality. This posterior will be a well-defined measure to quantitatively rank variants based on their contribution to the phenotype of interest.

8.2 Overview

8.2.1 Summary of Equations

There are three main components to a Bayesian inference model: the prior, the likelihood, and the posterior (Gelman et al., 1995). For a given variant, our causal model (M_1) states that the variant is a primary contributor to the phenotype. For observed data (D), the posterior is the probability of observing this model given the data, which may be calculated using Bayes Theorem:

$$\mathbb{P}(M_1 \mid D) = \frac{\mathbb{P}(D \mid M_1)\,\mathbb{P}(M_1)}{\mathbb{P}(D)}$$

The prior $\mathbb{P}(M_1)$ represents the probability that a variant is causal before we observe the pedigree data, and the likelihood $\mathbb{P}(D \mid M_1)$ is the probability of observing the data, assuming the causal model is true. The $\mathbb{P}(D)$ term is known as the normalising constant and does not depend on the model, and thus the above equation can be written as

$$\mathbb{P}(M_1 \mid D) \propto \mathbb{P}(D \mid M_1)\,\mathbb{P}(M_1)$$

or equivalently:

$$\text{posterior} \;\propto\; \text{likelihood} \;\times\; \text{prior}$$

As an alternative to the above, we can consider the neutral model (M_2) which states that the variant does not contribute to the phenotype. The corresponding posterior for this model is given by:

$$\mathbb{P}(M_2 \mid D) = \frac{\mathbb{P}(D \mid M_2)\,\mathbb{P}(M_2)}{\mathbb{P}(D)}$$

Dividing the posteriors for each model, we see that:

$$\frac{\mathbb{P}(M_1 \mid D)}{\mathbb{P}(M_2 \mid D)} = \frac{\mathbb{P}(D \mid M_1)\,\mathbb{P}(M_1)}{\mathbb{P}(D)} \frac{\mathbb{P}(D)}{\mathbb{P}(D \mid M_2)\,\mathbb{P}(M_2)} = \frac{\mathbb{P}(D \mid M_1)}{\mathbb{P}(D \mid M_2)} \times \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} \tag{8.1}$$

This is often written more succinctly as:

$$\text{Odds}_{\text{posterior}} = \text{Bayes Factor} \times \text{Odds}_{\text{prior}}$$

We can use the Bayes factor to compare the evidence for the causal model $\mathbb{P}(D \mid M_1)$ and the neutral model $\mathbb{P}(D \mid M_2)$ solely based on the co-segregation with the phenotype. These likelihoods will themselves be functions of the parameters of the model (described in Subsection 8.2.2 below). Since M_1 and M_2 are complementary to one another, we can let the prior probability of M_2 be given by $\mathbb{P}(M_2) = 1 - \mathbb{P}(M_1)$. Note that we can convert a probability p to its odds form o (and vice versa) as:

$$o = \frac{p}{1-p} \qquad \Leftrightarrow \qquad p = \frac{o}{o+1}$$

The posterior odds converted to a probability (which we refer to as the posterior probability for causality) will be our well-defined metric to prioritise variants in this Bayesian inference model.

8.2.2 Model Parameters and Assumptions

Following Mohammadi et al., we assume that the variant under consideration is rare in the general population (Mohammadi et al., 2009). This allows us to conclude that the variant originates from one founder within the pedigree, and so cannot be carried by marry-in individuals. In determining the phenotypes conditional on the genotypes, we will have different assumptions based on which model we are evaluating.

- M_1 (**Causal model**) - the variant has a dominant effect on the phenotype: this means that an individual's phenotype is determined solely by their genotype. This also reduces the search space when iterating over unobserved genotypes compared to an additive model, for example.
- M_2 (**Neutral model**) - the phenotypes are entirely independent of all genotypes and are determined by the population incidence rate.

In Bayesian inference models, we need to specify which random variables are parameters and which are data. The observed data represent fixed quantities that are specific to each pedigree. The data in our model comprises of the following:

- the observed genotypes (G_O); and
- the known phenotypes of the pedigree members (P_F). We assume that every sample has a phenotype specified.

The parameters, on the other hand, are unknown quantities in our model, and for each pedigree we can consider prior distributions for the parameters which can be determined from data-independent information about the phenotype. From this description, our parameters are:

- the unobserved genotypes (G_U), which follow a distribution determined by Mendelian segregation;
- the in-family penetrance (β), i.e. the probability of the phenotype for variant carriers, whose distribution is discussed below;
- the in-family phenocopy rate (φ), i.e. the probability of the phenotype for variant non-carriers, whose distribution is discussed below;
- the population incidence rate (α) whose distribution is discussed below; and
- the proband (p) who is selected at random from the affected carriers.

For convenience, we will let $\theta = (\beta, \varphi)$, since those two terms will often be grouped together, and both β and φ can take values between 0 and 1. We can construct a relationship between β, φ and α. If p is the allele frequency of a causal variant and $q = 1 - p$, then we have:

$$\begin{aligned} \alpha &= \mathbb{P}(P) \\ &= \mathbb{P}(P \mid \text{Geno})\,\mathbb{P}(\text{Geno}) + \mathbb{P}(P \mid \text{not Geno})\,\mathbb{P}(\text{not Geno}) \\ &= \beta(1-q^2) + \varphi(q^2) \end{aligned} \tag{8.2}$$

Here, "Geno" refers to having a genotype that confers risk for the phenotype P in a dominant fashion, so heterozygous or homozygous variant genotypes. We can see that α is a weighted average of β and φ, so we have three scenarios:

- **Scenario 1**: $\varphi < \alpha < \beta$ - recommended by Petersen et al. under the causal model (Petersen et al., 1998);
- **Scenario 2**: $\varphi = \alpha = \beta$ - the phenotype is independent of the genotype, which corresponds to the neutral model; and
- **Scenario 3**: $\beta < \alpha < \varphi$ - having the variant reduces the probability of the phenotype compared to the incidence rate and compared to not having the variant. This corresponds to a variant with a protective effect, which the causal model does not account for.

We assume Scenario 1 for the causal model and Scenario 2 for the neutral model.

8.3 Bayes Factor

The full derivation of the Bayes factor calculation is given in Appendix B, which we will summarize here.

8.3.1 Causal Model: Likelihood

To evaluate our Bayes factor, we first need to calculate the probability of the data under the causal model, i.e. $P(D|M_1)$. We can re-arrange the equations, so we are left with three terms of interest: the probability of the phenotypes, the probability of the genotypes and the prior probability of the parameters, as shown in Equation 8.3 below.

$$\mathbb{P}\left(D \mid M_1\right) = \sum_{p=1}^{k_1} \sum_{G_U} \iint_{\Omega} \underbrace{\mathbb{P}\left(P_F \mid G_F, p, \theta, M_1\right)}_{\text{phenotypes}} \underbrace{\mathbb{P}\left(G_F \mid p\right)}_{\text{inheritance}} \underbrace{\mathbb{P}\left(p, \theta \mid M_1\right)}_{\text{parameters}} d\theta \tag{8.3}$$

As described in Mohammadi et al., the assumption of a dominant effect means that the phenotypes are determined by the genotype alone, and so the "phenotypes" term in Equation 8.3 will just be a product of the penetrance values and phenocopy rates. Under the causal model, if k_1, k_2 are the number of affected/unaffected variant carriers and l_1, l_2 are the number of affected/unaffected variant non-carriers, then:

$$\mathbb{P}\left(P_F \mid G_F, p, \theta, M_1\right) = \prod_{i=1}^{n} \mathbb{P}\left(P_i \mid G_i, \theta, M_1\right) = \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2} \tag{8.4}$$

When we iterate over all unobserved genotypes, there are some configurations that we may ignore due to the rare variant assumption. To illustrate this, consider the simulated pedigree in Figure 8.1 taken from Mohammadi et al. Here, individuals 1 and 3 are known to have breast cancer and they also carry the variant of interest, and all other individuals are unaffected, with their genotypes unknown. If individual 3 is the current proband, the potential founders are individuals 6 and 7. If individual 1 is the current proband, the potential founders are individuals 2, 6 and 7. However, since individual 3 is not descended from individual 2, individual 2 cannot be a founder for the entire pedigree. So, the only pedigree founders we need consider are individuals 6 and 7, which is the same regardless of the selected proband. Since our inheritance term $\mathbb{P}\left(G_F \mid p\right)$ is now independent of the choice of proband, we will simply write it as $\mathbb{P}\left(G_F\right)$. As described in Mohammadi et al., we can calculate this term as 0.5 to the power of the number of times the variant is or isn't transmitted to a child from a parent who is a carrier. Note that this transmission probability only holds for autosomal variants. For example,

a variant on the X chromosome will have different transmission probabilities depending on the sex of the parent and child. The Bayesian model we are proposing here only considers autosomal variants.

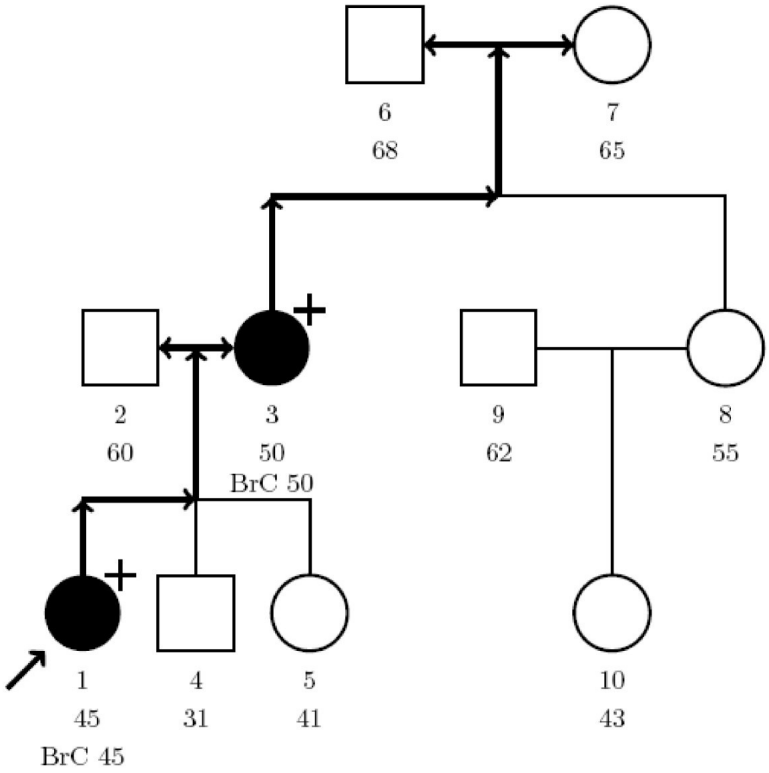


Figure 8.1: Hypothetical breast cancer pedigree taken from Mohammadi et al (Mohammadi et al., 2009). Variant carriers are denoted with a + sign, and affected individuals are coloured black.

Finally, we need to consider the prior distribution for our parameters β, φ under the causal model. While Mohammadi et al. allowed for age-specific values for these parameters, for simplicity we will assume that they are not age-dependent. One simple choice in the absence of any additional information is to let the distributions of both parameters take all values over $[0, 1]$ with equal probability, i.e. a Uniform distribution, as described in Figure 8.2a below. However, this would allow for variants with high penetrance values and high phenocopy rates, which we would like to avoid. Instead, a plausible scenario for our causal variants is that they are more likely to have high penetrance values and low phenocopy rates. Therefore, we will consider a simple prior described in Figure 8.2b and Figure 8.2c that reflects this. We can see that higher values of the penetrance are more likely than lower values, and that the reverse is true for the phenocopy rate. Since

we have the assumption that $\varphi < \beta$ for the causal model, values of the phenocopy rate that are higher than the penetrance are given a probability of zero. We refer to this prior distribution of the parameters β and φ as the Linear prior.

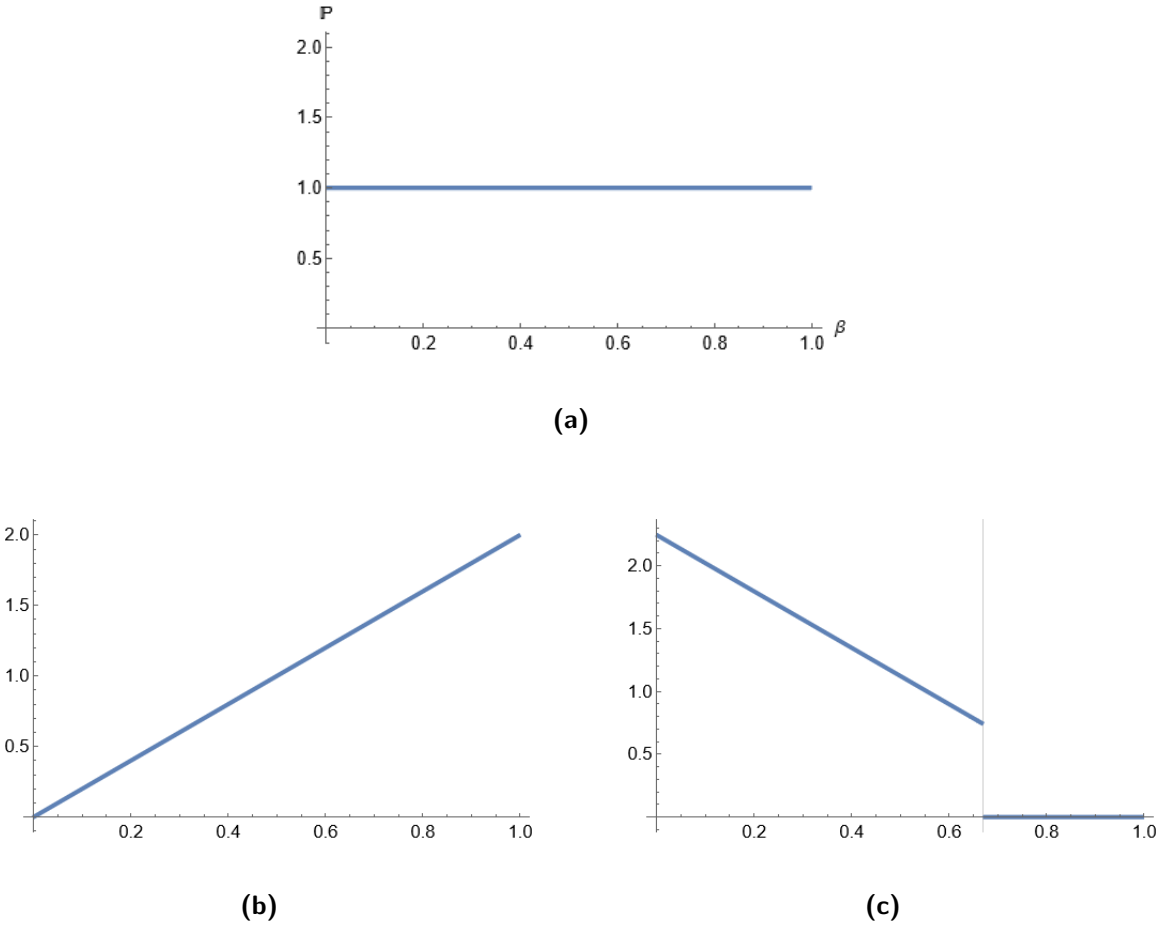


Figure 8.2: The density plots of (**a**) the Uniform prior distribution of for β; (**b**) the Linear prior distribution of for β; and (**c**) the Linear prior distribution of for φ, conditional on a fixed value of $\beta = 0.667$. Note the assumption that $\varphi < \beta$.

8.3.2 Neutral Model: Likelihood

Next, we will calculate the likelihood for the neutral model, given by:

$$\mathbb{P}(D \mid M_2) = \sum_{p=1}^{k_1+l_1} \sum_{G_U} \int_0^1 \underbrace{\mathbb{P}(P_F \mid p, \alpha, M_2)}_{\text{phenotypes}} \underbrace{\mathbb{P}(G_F \mid p)}_{\text{inheritance}} \underbrace{\mathbb{P}(p, \alpha \mid M_2)}_{\text{parameters}} \, d\alpha \tag{8.5}$$

The the main difference here from the causal model is that instead of the penetrance and phenocopy terms, this model depends on the population incidence α. Also, the

phenotypes term is simplified since the genotypes are independent of the phenotypes. We can calculate this term as:

$$\mathbb{P}\left(P_F \,|\, p, \alpha, M_2\right) = \prod_{i=1}^{n} \mathbb{P}\left(P_i \,|\, \alpha, M_2\right) = \alpha^{k_1+l_1}(1-\alpha)^{k_2+l_2} \tag{8.6}$$

The inheritance term is calculated exactly as described in the previous section. We will assume that the population incidence rate α has a Uniform prior distribution on $[0, 1]$ for simplicity.

8.3.3 Prior Sensitivity

Now that we have derived an expression for the Bayes factor from Equation 8.1 above, we will examine how it will be affected by the two different prior distributions (Uniform and Linear) for the parameter terms β and φ described above. To do this, we considered a specific phenotype combination for the CEPH 1463 pedigree (see Figure 8.3). As described in Subsection 7.3.2 above, the inheritance pattern in the first two generations is set to identify the true route of transmission. For simplicity, we set the first five children as affected, and the remaining six children as unaffected.

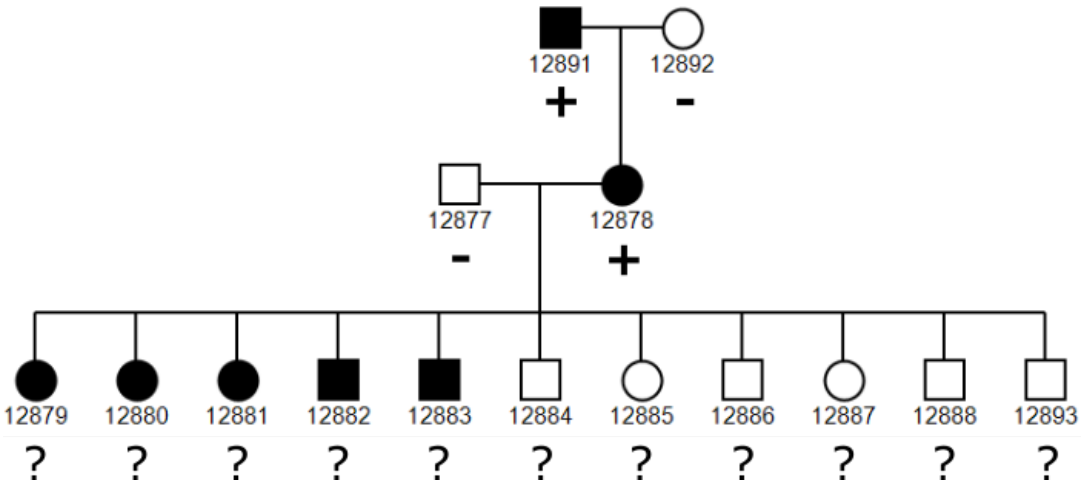


Figure 8.3: The CEPH 1463 pedigree with a specific phenotype pattern selected. Black boxes are affected, and white boxes are unaffected. Variant carriers are marked with a "+", variant non-carriers are marked with a "-", and unknown genotypes are marked with a "?".

We let the genotypes of the first two generations be pre-defined: carriers are affected and non-carriers are unaffected. For the third generation, we examined all $2^{11} = 2,048$ theoretical genotype combinations, where each child could be a carrier or non-carrier. This allows us to examine how the Bayes factors change for all potential co-segregation

configurations. For each variant, we can compute the in-family penetrance and phenocopy rates. Note that given the fixed genotypes in the first two generations, we will not be able to find a variant with completely "opposite co-segregation" (carried by all unaffected samples, absent from all affected samples). One variant had perfect co-segregation in the entire pedigree, and one other variant had opposite co-segregation among the third generation. All other variants had a mix of co-segregation types with varying penetrance and phenocopy rates.

We examined the change in Bayes factor when we swap from the Uniform prior to the Linear prior in the causal model (M_1). Note that if two different variants had the same penetrance and phenocopy rates, then they received the same Bayes factor, so none of the probability terms within the Bayes factor calculation change. This is because only the genotypes are being changed within a sibship, so the numbers of affected and unaffected carriers do not change, and the inheritance probability term in the Bayes factor term also does not change. For each pair of penetrance versus phenocopy rates, we calculated the ratio of the Bayes factors under both the prior distributions. We can view the results in Figure 8.4 below.

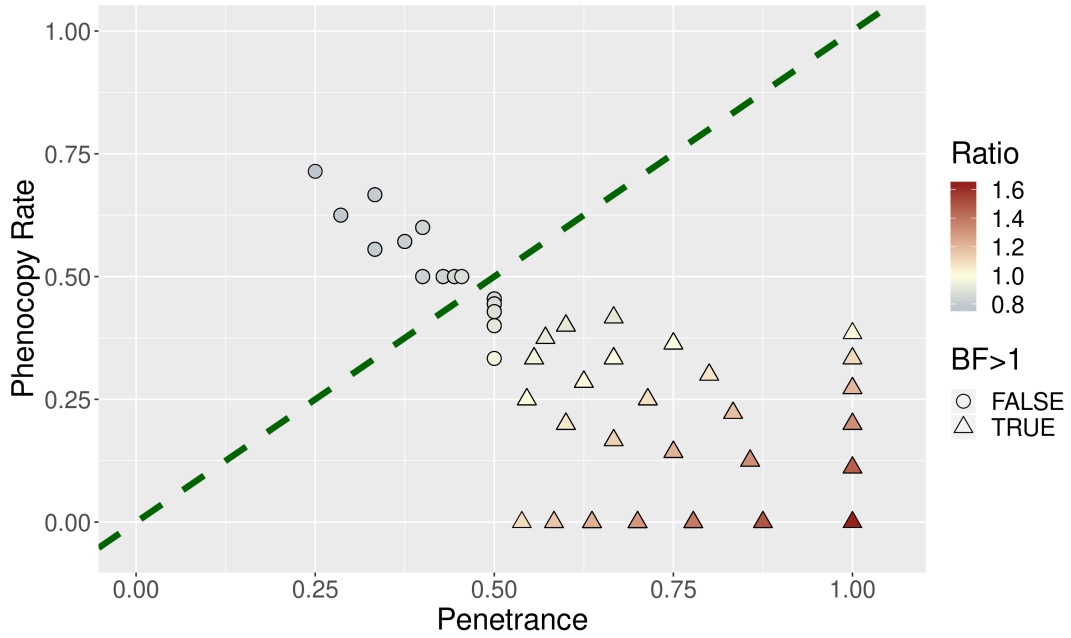


Figure 8.4: For each input variant, we plotted the in-family penetrance (β) and phenocopy rate (φ). The green dashed line represents where the penetrance equals the phenocopy rate. The shape represents whether the Bayes factor was greater than 1 or not. The colour represents the ratio of the Bayes factors under the Linear prior and the Uniform prior. Note that two variants which had the same penetrance values and the same phenocopy rates resulted in the same Bayes factors, so are represented by one point on the graph.

While the change in prior had an effect on the Bayes factors, if a variant had BF > 1 under the Uniform prior, the same was true for the Linear prior. Variants with the same penetrance values and phenocopy rates received the same Bayes factor, using either prior distribution. Figure 8.4 shows the broad inverse linear relationship between the penetrance and phenocopy rate. We see that for all variants with $\beta < \varphi$ (above the green line), the Bayes factor was decreased when we changed to the Linear prior. Only variants satisfying $\varphi < \beta$ (below the green line) had an increase in Bayes factors under the Linear prior. Some variants with $\varphi < \beta$ had their Bayes factor decreased but many of these variants represent mixed co-segregation information, with modest penetrance values. From this, we can see that the Linear prior appropriately down-weights variants with poor co-segregation and up-weights variants with good co-segregation. Based on the above, the Linear prior for β and φ was chosen for the Bayes factor calculations.

8.4 General Prior Probability for Causality

The final component in our Bayesian inference model is the prior odds that a variant is causal. To construct this prior, we will use the ClinVar database (Landrum et al., 2018) to generate a list of known pathogenic and benign variants and develop a prediction model for causality trained on these variants. Since these variants are predominantly protein-coding, we will restrict the following prior probability calculation to variants that affect the transcript of a protein-coding gene only. Non-protein-coding variants will be given a flat prior probability of causality of 0.5. We downloaded the VCF version of ClinVar (date downloaded: 29/01/22; see "Web Resources", Subsection A.2.7) containing 1,117,603 variants, and retained those that satisfied all of the following:

- an SNV or an indel;
- the clinical significance was Benign and/or Likely Benign (referred to simply as Benign), or the clinical significance was Pathogenic and/or Likely Pathogenic (referred to as Pathogenic);
- submitted by multiple submitters with no conflicts in significance;
- overlapped one gene with one functional consequence; and
- present on an autosomal chromosome.

Pathogenic variants are unlikely to be common in the general population compared to benign variants which could be rare or common, so we included allele frequency as a predictor of causality. Allele frequencies were taken from gnomAD v2.1.1 exome data,

since this has the highest number of samples covering protein-coding regions. We also included the following deleteriousness metrics as predictors: CADD, MPC, SIFT, PolyPhen2, REVEL, and FATHMM. REVEL (Ioannidis et al., 2016) and FATHMM (Shihab et al., 2013) have been shown to perform well at classifying known pathogenic and benign variants respectively (Cubuk et al., 2021; Gunning et al., 2021; Niroula & Vihinen, 2019; Tian et al., 2019). Additionally, meta-predictors have shown increased performance to classify variants over stand-alone tools (Gunning et al., 2021). Note that CADD (which is a meta-predictor) is defined for all SNVs in the genome (as well as many indels), whereas the other five predictors are specific to missense variants only.

Another predictor we considered is GERP++ conservation scores (Davydov et al., 2010), since conservation can indicate pathogenicity (Richards et al., 2015) and the scores are defined for most positions in the genome. However, GERP++ scores are included as training predictors to CADD, so including them here could introduce additional multicollinearity into the models. We evaluated this by calculating the variance inflation factors (VIF) for all selected predictors. It was found during initial testing that GERP++ had a VIF greater than 10 in almost all instances (typically indicating multicollinearity (Kutner, 2005)) and had the highest VIF of all predictors. When GERP++ was removed, the VIF for all remaining predictors fell to less than 10 (with one exception, described below). Therefore, we decided not to include GERP++ as a predictor.

ClinVar variants were used to train the MPC prediction score (Samocha et al., 2017). We identified these variants from the training set and converted them to GRCh38 with `liftOver`, removing those at unstable positions as per Chapter 3. We then removed the 344 variants in our ClinVar dataset that overlapped with the MPC training set to avoid circularity bias (Grimm et al., 2015). Each of the remaining ClinVar variants was annotated with the metrics described above, taking the missense-specific scores from dbNSFP as described in Section 2.6. Where multiple transcripts were present in dbNSFP, we took the most severe score across all transcripts as the representative score. Finally, we removed variants if the ClinVar gene or functional consequence did not match with that of `vep`. This resulted in a final list of 76,613 annotated variants (16,894 Pathogenic and 59,719 Benign), which represents 6.9% of the total variants from the original ClinVar VCF file. A breakdown of the clinical significances and functional consequences is given in Table 8.1 below. We can see that some functional consequence classes are almost exclusively Benign or Pathogenic (e.g. synonymous, frameshift, stop gained, etc.) whereas for others the split is less one-sided (missense variants, inframe deletions, etc.).

Consequence	SNV				Indel			
	BEN	PATH	Total	% PATH	BEN	PATH	Total	% PATH
3' UTR variant	1,496	2	1,498	0.1%	153	0	153	0.0%
5' UTR variant	667	10	677	1.5%	40	0	40	0.0%
Frameshift variant	-	-	-	-	37	4,156	4,193	99.1%
Inframe deletion	-	-	-	-	210	166	376	44.1%
Inframe insertion	-	-	-	-	138	21	159	13.2%
Intron variant	13,444	315	13,759	2.3%	1,744	16	1,760	0.9%
Missense variant	11,167	5,314	16,481	32.2%	-	-	-	-
Splice acceptor variant	4	1,018	1,022	99.6%	7	60	67	89.6%
Splice donor variant	7	1,470	1,477	99.5%	10	112	122	91.8%
Stop gained	19	4,112	4,131	99.5%	0	59	59	100.0%
Stop lost	1	2	3	66.7%	-	-	-	-
Synonymous variant	30,575	61	30,636	0.2%	-	-	-	-
Total	57,380	12,304	69,684	17.7%	2,339	4,590	6,929	66.2%

Table 8.1: The selected ClinVar variants broken down by variant type and functional consequence, showing counts for Benign (BEN) and Pathogenic (PATH) variants and the proportion of pathogenic (% PATH) for that functional consequence.

We used logistic regression as the basis for classifying pathogenicity using the above described predictors for each input variant. The coefficients from the regression model can be used to calculate the probability of pathogenicity, which we will use as the prior probability for causality in our Bayesian inference model. We examined five separate regression models, as detailed in Table 8.2. The simplest model is to train on all variants together using CADD scores, allele frequency, functional consequence, and variant type (Model 1). However, since the predictors may have different distributions depending on the variant type, we examined regression models for SNVs and indels separately (Model 2). Similarly, we examined regression models for missense and non-missense variants separately since we have additional missense-specific predictors (Model 3). When we added the five missense-specific predictors, the VIF of CADD was 14.96, so we decided to remove CADD as a predictor from the missense variant regerssion in Model 3. To evaluate how well the missense-specific variants performed, we examined an alternate version where CADD was the sole deleteriousness predictor for missense variants (Model 4). Finally, we also split non-missense variants by type (Model 5).

Model	**Variant**	**Predictors**
1	All variants	CADD + AF + CSQ + TYPE
2	SNV	CADD + AF + CSQ
	Indel	CADD + AF + CSQ
3	Missense	MissPred + AF
	Non-missense	CADD + AF + CSQ + TYPE
4	Missense	CADD + AF
	Non-missense	CADD + AF + CSQ + TYPE
5	Indel	CADD + AF + CSQ
	Missense	MissPred + AF
	Non-missense SNVs	CADD + AF + CSQ

Table 8.2: Description of the logistic regression models and the predictors for each. AF: gnomAD allele frequency; CSQ: functional consequence; MissPred: the five missense-specific predictors (SIFT, PolyPhen2, MPC, REVEL, FATHMM). TYPE: whether the variant is an SNV or indel.

To evaluate the performance of each regression model, we calculated the sensitivity, specificity, and Matthews Correlation Coefficient (MCC), defined in Equation 8.7 below. The MCC is a robust performance metric for unbalanced data (Boughorbel et al., 2017),

as is the case with our ClinVar variants since only 22.1% of the variants are pathogenic. Additionally, the MCC is a useful stand-alone metric since a model will only have a high MCC score if it also has both a high sensitivity and specificity.

$$\text{Sensitivity} = \frac{TP}{TP + TN} = \frac{TP}{P}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{TN}{N} \tag{8.7}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

For each regression model, we split the appropriate variant groups into training (80%) and testing (20%) datasets. The variant type and functional consequence are both categorical inputs to the models, so we converted these to a collection of dummy variables representing each category which take on either 0 or 1. To avoid perfect multicollinearity due to the dummy variable trap, the first value alphabetically in each category was dropped from the list. For the functional consequence this was 3' UTR variants, and for variant type this was SNVs. There were varying levels of missingness within the predictors, so we imputed missing predictors by taking the median value of each predictor across the training data. This method is blunt and more sophisticated imputation methods exist, but it has the advantage of simplicity, and it is easy to apply to other datasets. The exception to this was that a missing minor allele frequency is set to zero, since the median would be a poor estimate of the true frequency, which is less than the minimum frequency observed in the entire dataset. Finally, we scaled the imputed predictors to $[0, 1]$ so all of the regression coefficients could be compared.

We fit the logistic regression models to the scaled and imputed training data taking the ClinVar clinical significances as our output. We applied the same training imputation and scaling to the testing data. During scaling, values in the testing data that fell outside $[0, 1]$ were clipped to ensure the testing would be comparable to the training data. We then applied the regression model to the scaled, imputed testing data to classify each variant. Finally, where variants were split by type within a model, we combined the individual predictions and calculated the performance metrics of the model on the training and testing dataset. To try to minimise overfitting to the training data, we used the metrics on the testing data only to compare the models. We used 1,000 bootstrap resamples on the training and testing data to estimate a central 95% confidence interval for each of the performance metrics. The results are shown in Figure 8.5 below.

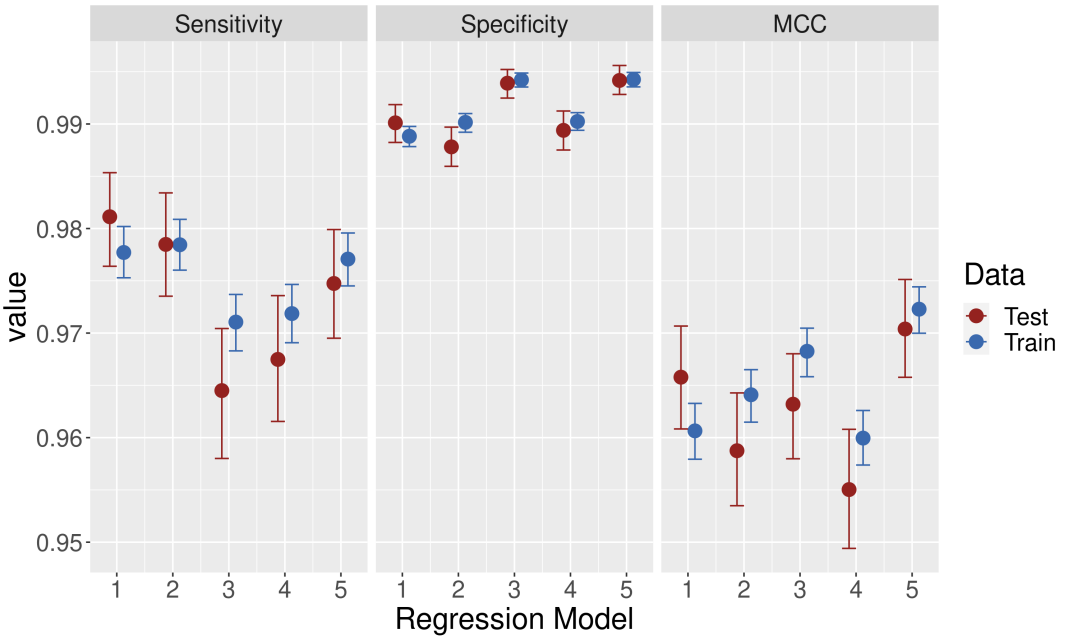


Figure 8.5: The performance metrics (sensitivity, specificity and Matthews correlation coefficient (MCC)) of the five regression models on the ClinVar training and testing datasets.

We can see that Models 1, 2, and 5 perform comparably in terms of sensitivity, and Models 3 and 5 perform comparably in terms of specificity. Based on the overall performance scores, we decided to use Model 5 as our main regression model to calculate the prior probability of causality. However, it is worth noting that Model 1 performs well overall and is a reasonable alternative if a simpler model is desired. The relative difference in performance between Models 3 and 4 indicates that the selected missense-specific predictors add value compared to using CADD alone. Relative to Model 1, splitting by variant type (Model 2) does not appear to improve performance as much as splitting into missense verses non-missense (Model 3). Finally, we note that given that the metrics for the training and testing are broadly the same for all metrics, there does not appear to be evidence of overfitting of each of the models to the training data.

The regression coefficients from Model 5 are shown in Table 8.3 below. Note that these coefficients are the natural logarithm of their corresponding effect sizes, so the larger a coefficient is in absolute value, the more important to the prediction model it is. We can see that the predictor with the strongest effect on the output is the allele frequency. Variants with high allele frequencies will have lower prior probabilities of pathogenicity due to the fact that the vast majority of common variants in ClinVar are Benign. For indels, the variant functional consequence is more important than CADD scores. For missense variants, REVEL and FATHMM scores do well at classifying

Pathogenic variants, compared to the well-known SIFT and PolyPhen scores. For non-missense SNVs, CADD scores are important, but so too are some of the functional consequences such as being a frameshift, splice-site, or stop-gained variant.

Predictor	Indel	Missense	Non-Missense SNV
Intercept	-407.268	-5.374	-34.553
Allele frequency	-1404.457	-1970.702	-644.398
CADD	2.509	-	21.938
MPC	-	1.549	-
PolyPhen2	-	-0.024	-
REVEL	-	9.412	-
SIFT	-	-1.115	-
FATHMM	-	3.296	-
5' UTR variant	-460.211	-	25.719
Frameshift variant	412.750	-	-
Inframe deletion	407.479	-	-
Inframe insertion	405.489	-	-
Intron variant	403.397	-	27.438
Splice acceptor variant	434.589	-	28.690
Splice donor variant	409.648	-	33.457
Stop gained	854.191	-	29.206
Stop lost	-	-	29.105
Synonymous variant	-	-	25.150

Table 8.3: List of the coefficients of regression Model 5, split into indels, missense variants and non-missense SNVs.

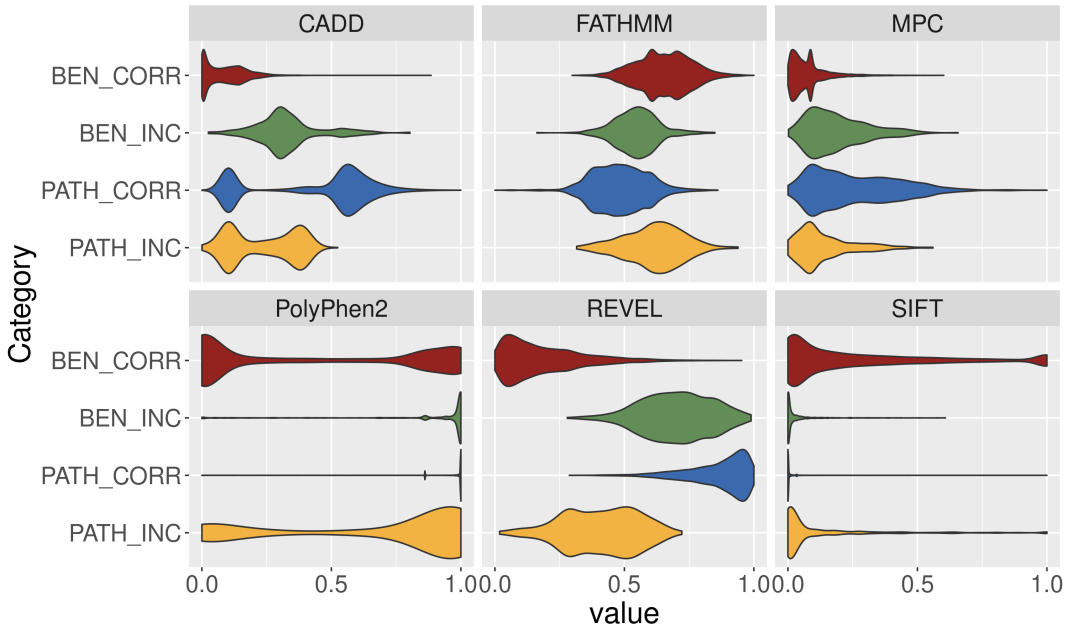
While the selected regression model performs well, we wished to examine why variants in the Benign or Pathogenic sets were mis-classified. Benign variants with a prior of 0.5 or greater and Pathogenic variants with a prior less than 0.5 were labelled as "incorrect". All other variants were labelled as "correct". Firstly, we calculated the proportion of correctly and incorrectly classified variants that fell into each functional consequence category, shown in Table 8.4 below. We can see that the majority (79.94%) of incorrectly classified Benign variants are missense variants. However, a further 8.15% of variants are in-frame deletions, compared to 0.27% of correctly classified Benign variants. Similarly,

in the incorrectly classified Pathogenic variants, there are virtually no frameshift, splice-site and stop gained variants (which are strongly present in the correctly classified group), and the majority are missense (66.32%) or intronic (19.91%).

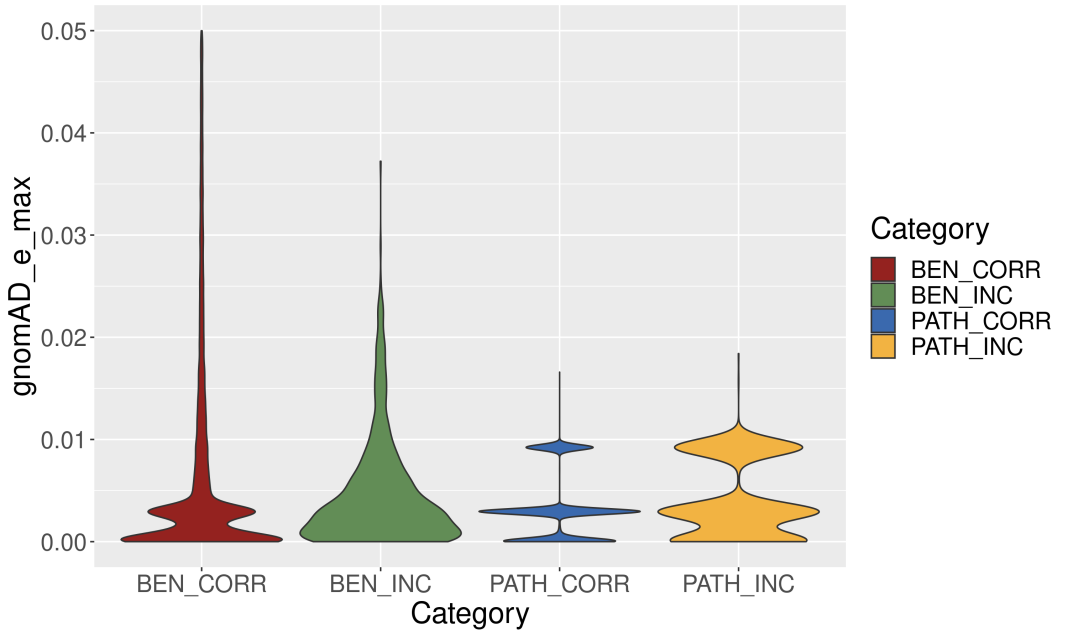
Consequence	Benign		Pathogenic		% INC
	CORR	INC	CORR	INC	
3' UTR	2.79%	0.00%	0.00%	0.30%	0.12%
5' UTR	1.20%	0.00%	0.00%	1.50%	1.39%
Frameshift	0.05%	1.57%	25.60%	0.30%	0.29%
Inframe deletion	0.27%	8.15%	1.02%	0.00%	13.83%
Inframe insertion	0.23%	0.00%	0.00%	3.14%	13.21%
Intron	25.63%	6.90%	1.22%	19.91%	1.14%
Missense	18.04%	79.94%	30.02%	66.32%	5.78%
Splice acceptor	0.01%	0.47%	6.64%	0.15%	0.37%
Splice donor	0.02%	1.25%	9.74%	0.15%	0.56%
Stop gained	0.02%	1.41%	25.69%	0.30%	0.26%
Stop lost	0.00%	0.16%	0.01%	0.15%	66.67%
Synonymous	51.75%	0.16%	0.06%	7.78%	0.17%

Table 8.4: The proportion of each functional consequence in the correctly and incorrectly classified Benign and Pathogenic variants, as well as the total proportion of incorrectly classified variants per consequence. CORR: correctly classified; INC incorrectly classified.

Based on the above, we examined the distribution of the scaled, imputed deleteriousness predictors, as shown in Figure 8.6a below. We know from the regression coefficients that REVEL and FATHMM scores are important for the missense model and that CADD scores are important for the indel and non-missense models. We can see that for all three scores, the distribution of incorrectly classified Pathogenic variants is shifted down compared to correctly classified Pathogenic variants. Similarly, the distribution of incorrectly classified Benign variants is shifted up compared to correctly classified Benign variants. Violin plots for the allele frequencies less than 5% are shown in Figure Figure 8.6b. The densities for the incorrectly classified categories largely overlap the densities for their corresponding correctly classified categories, so allele frequency is not likely to play as large a role in the misclassification as the other predictors. Based on the above, the deleteriousness metrics are likely what is resulting in the misclassification of missense variants.



(a)



(b)

Figure 8.6: Violin plots for Benign and Pathogenic variants split by whether they were misclassified or not. (**a**) The deleteriousness metrics; (**b**) The allele frequency less than 5%. BEN: Benign; PATH: Pathogenic; CORR: correctly classified; INC incorrectly classified.

As a positive control, we also examined all variants of unknown significance (VUS) from ClinVar. For these variants, there is no clear accumulation of evidence for or against pathogenicity, so we expect a more mixed distribution of the prior probabilities for causality for VUS than those of Benign or Pathogenic variants. As before, we selected variants on the autosomal chromosomes, but relaxed the requirement for multiple submitters to have supported the significance. This resulted in 405,808 VUS in total. Since we are no longer evaluating the performance of the regression model, we do not need the training and testing subsets of the Pathogenic/Benign variants, so we re-trained Model 5 with the entire Pathogenic/Benign data. Using the resulting regression coefficients, we calculated prior probabilities of causality for all Benign variants, Pathogenic variants, and variants of unknown significance.

Boxplots of these prior probabilities are shown in Figure 8.7. We can see that the vast majority of Benign variants receive a low prior probability of causality, and the majority of Pathogenic variants receive a high prior probability of causality, as expected. The distribution of prior probability for VUS lies between those of Benign and Pathogenic variants. Additionally, the majority of VUS have a prior probability less than 0.5, indicating that they are less likely to be pathogenic. This is consistent with what we might expect, given that variants with no clear evidence for pathogenicity are in general more numerous than known pathogenic variants.

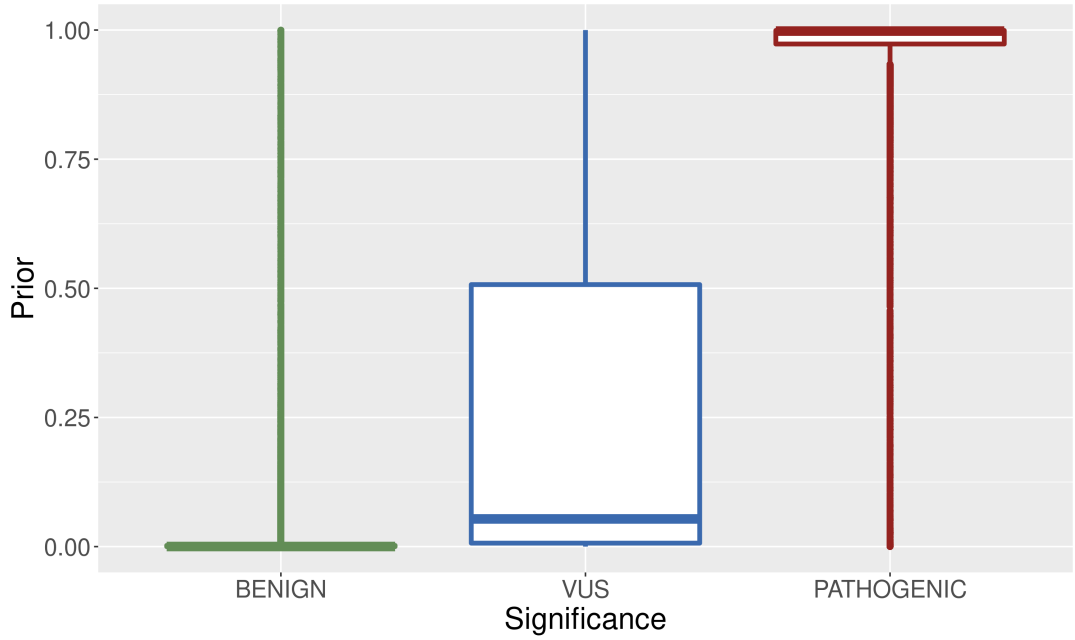


Figure 8.7: Boxplots of the prior probabilities of causality for Benign variants (green), variants of unknown significance (blue), and Pathogenic variants (red) from ClinVar.

8.5 Mendelian Phenotype: CEPH 1463 Pedigree

To test the full Bayesian inference model, we examined the CEPH 1463 pedigree, and the pseudo-causal variants as defined in Subsection 7.3.2 above. Using the same input data as for the evaluation of pVAAST and PERCH above, we calculated Bayes factors for each variant using the 20 unique phenotype configurations matching the 23 autosomal pseudo-causal variants. Given that these data had been converted to GRCh37, we re-generated the ClinVar dataset that had been curated on this reference genome (date downloaded: 29/01/22, see "Web Resources", Subsection A.2.7) to construct the prior probability of causality. Using the same selection criteria as before, we retained 75,458 variants (16,651 Pathogenic and 58,807 Benign). Variants also present in the CEPH 1463 pedigree were removed prior to training. The ClinVar variants were annotated using the GRCh37 versions of CADD, gnomAD and dbNSFP. The regression model was applied exactly as described above (using the entire dataset to train the regression model), and prior probabilities for causality were calculated for every variant examined in the pedigree using the coefficients of Model 5.

For each phenotype configuration we calculated the posterior probability of causality based on Equation 8.1 above and used this to rank all variants in the CEPH 1463 pedigree. The results of this are shown in Table 8.5 below. We can see that the Bayes factors for each of the pseudo-causal variants ranked joint first amongst all variants evaluated. This is consistent with the fact that the pseudo-causal variants were selected to have perfect co-segregation, so there should be no variants with a better Bayes factor. The prior probabilities for causality, on the other hand, are mixed with both high and low values. This is expected since the pseudo-causal variants were not selected based on an allele frequency filter, so we will be able to examine how the model behaves on for rare and common variants.

Variants with a prior of at least 0.5 have an allele frequency of 0.2% or lower. Conversely, variants with a prior less than 0.5 had an allele frequency of at least 1.6% (with the exception of the variant in *SMYD1*, discussed below). This behaviour is consistent with what we expect from the regression model, given that Pathogenic variants in ClinVar are rare in the general population. The missense variant in *SMYD1* has an allele frequency of 0.22%. In the missense regression model, REVEL and FATHMM scores are the next best predictors after allele frequency. However, the variant's REVEL score was low at 0.18, and the FATHMM score was 1.8, indicating that the variant is not deleterious. So, although the variant is rare in the general population and ranked highly by CADD, there

is disagreement with the other deleteriousness predictors, resulting in this variant having a low prior.

The variability in the priors is reflected in the posterior probability for causality. The rare pseudo-causal variants (with the exception of the *SMYD1* variant) ranked in the top 10 variants, with a posterior probability of causality higher than 99.999%. This shows that our model performs well at identifying rare, deleterious variants that follow a dominant Mendelian inheritance pattern. The prior (and posterior) probabilities for common pseudo-causal variants are close to zero, indicating that even variants with high co-segregation scores will receive low posterior probabilities if their priors indicate that they are not relevant to pathogenicity. This is important, as we expect to observe low-impact common variants that co-segregate with a phenotype by chance, given the density of variants generated by NGS data.

Gene	Consequence	CADD	AF	BF		Prior	Posterior	
				Value	Rank		Value	Rank
OR4C15	stop gained	31	0.204%	4,650.537	1*	99.995%	100.000%	4
ECHDC2	stop gained	31	0.002%	2,314.488	1*	99.970%	100.000%	7
TLR10	stop gained	33	0.002%	4,666.715	1*	99.970%	100.000%	6
NRIP2	stop gained	31	0.084%	4,650.537	1*	99.871%	100.000%	6
PKN3	missense	31	0.001%	3,689.634	1*	97.466%	99.999%	9
SMYD1	missense	31	0.223%	1,167.091	1*	0.263%	75.451%	370,159
OR1J1	stop gained	31	1.608%	4,666.715	1*	0.001%	2.380%	1,637,140
TSACC	missense	36	2.258%	4,666.715	1*	0.000%	0.000%	2,480,335
OR10J1	stop gained	35	15.426%	4,666.715	1*	0.000%	0.000%	2,486,995
OR5K2	stop gained	32	18.811%	3,729.780	1*	0.000%	0.000%	2,470,523
CDH15	stop gained	36	19.100%	2,314.488	1*	0.000%	0.000%	2,585,340
ZNF80	stop gained	33	29.877%	4,666.715	1*	0.000%	0.000%	2,516,683
KRT75	missense	39	19.080%	3,689.634	1*	0.000%	0.000%	2,575,685
GPANK1	missense	35	19.898%	4,650.537	1*	0.000%	0.000%	2,546,616
SCRN3	frameshift	36	35.002%	3,729.780	1*	0.000%	0.000%	2,475,800
TTC27	missense	46	21.296%	4,650.537	1*	0.000%	0.000%	2,551,188

C17orf107	stop gained	35	37.736%	3,729.780	1*	0.000%	0.000%	2,443,860
ZACN	stop gained	35	39.629%	4,650.537	1*	0.000%	0.000%	2,593,918
SLC6A18	stop gained	31	39.958%	4,666.715	1*	0.000%	0.000%	2,553,387
OR4C16	stop gained	33	40.878%	4,650.537	1*	0.000%	0.000%	2,609,627
NEXN	missense	36	39.523%	4,666.715	1*	0.000%	0.000%	2,570,431
SLC22A10	stop gained	39	57.542%	4,650.537	1*	0.000%	0.000%	2,614,966
SCEL	missense	34	57.313%	3,689.634	1*	0.000%	0.000%	2,632,456

Table 8.5: The results for the pseudo-causal variants on applying the fully Bayesian inference model to the CEPH 1463 pedigree. Variants are ordered by their posterior probability for causality. Shown for each variant is: the gene, the functional consequence, the CADD v1.6 score, the gnomAD v2.1.1 population maximum allele frequency (AF), the Bayes factor (BF), the prior and posterior probabilities of causality. An asterisk (*) beside a rank indicates that the rank is joint with other variants in that specific phenotype configuration. Note that the probabilities are rounded to three decimal places.

8.6 Complex Phenotype: Utah Pedigrees

To examine the model on a complex phenotype, we applied the full Bayesian inference model to the three Utah pedigrees harbouring a variant prioritised from the SCHEMA-based IBS filtering from Chapter 6. Bayes factors were calculated using the Linear prior for the penetrance and phenocopy rates, as described above. While the General prior probability of causality described in the previous section classifies ClinVar variants well, there is no guarantee that it will perform as well at identifying variants that are implicated in schizophrenia. To mitigate this, we generated a new prior probability of causality based on the SCHEMA filtering analysis. We took Model 5 in the previous section as the basis for our regression model but removed any predictor that was not present in SCHEMA. Note that in the SCHEMA filtering analysis, the minor allele count (MAC) was used to prioritise ultra-rare variants instead of the minor allele frequency (MAF). However, for common and low frequency variants, the MAC may be less representative of the true population MAF, especially as the total observed alleles (the denominator of the MAF calculation) is not specified. For example, a variant may have a high MAC, but still be rare across all genomic ancestry groups. For this reason, we opted to use the maximum MAF across genomic ancestry groups as before.

For the "indel" and "non-missense SNV" regression models we used allele frequency and functional consequence as predictors, and for the "missense" model we used allele frequency and the MPC score. The only SCHEMA prioritisation step that is not accounted for is the pLI scores. However, since pLI is a gene-based score, a benign and pathogenic variant in the same gene would get the same score, so it is not appropriate to include pLI scores as part of this regression model. All other aspects of the model (training vs testing, imputation, scaling, etc.) were implemented as described for the General prior probability of causality above.

The performance metrics for the regression models underlying the General prior and the SCHEMA prior are shown in Figure 8.8 below. The SCHEMA prior regression does not perform as well as the General prior regression, which is to be expected as we have removed some predictors which are known to be useful to the model e.g. CADD and REVEL. That said, the regression model for the SCHEMA prior performs reasonably well overall, and the coefficients should suffice for generating a prior probability of causality that reflects the prioritisation paradigm in the SCHEMA filtering analysis. The regression coefficients using the entire ClinVar dataset (i.e. not split into training/testing, with the overlapping variants from the pedigrees removed) are shown in Table 8.6.

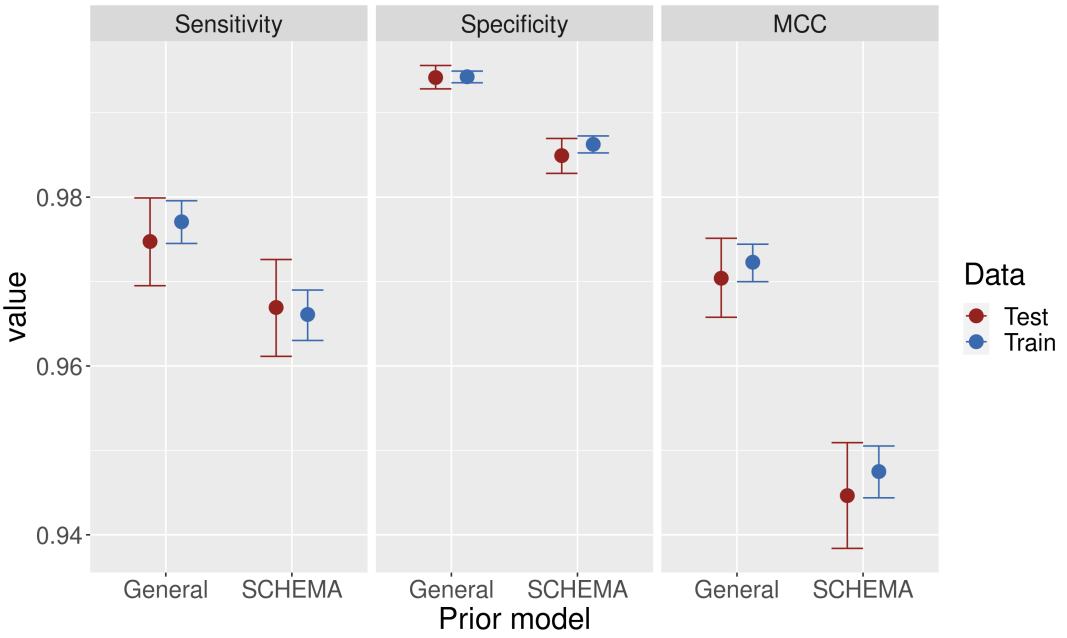


Figure 8.8: The performance metrics (sensitivity, specificity and Matthews correlation coefficient (MCC)) of the regression model underlying the SCHEMA prior compared to the regression model underlying the General prior at classifying ClinVar variants.

Predictor	Indel	Missense	Non-Missense SNV
Intercept	-428.689	1.083	-22.407
Allele frequency	-1232.208	-2728.476	-947.059
MPC	-	5.669	-
5' UTR variant	-376.139	-	19.298
Frameshift variant	435.214	-	-
Inframe deletion	429.839	-	-
Inframe insertion	427.746	-	-
Intron variant	425.312	-	19.906
Splice acceptor variant	452.399	-	28.573
Splice donor variant	432.194	-	29.158
Stop gained	819.443	-	30.580
Stop lost	-	-	24.971
Synonymous variant	-	-	16.892
Synonymous variant	-	-	25.150

Table 8.6: List of the coefficients of the regression models underlying the SCHEMA prior, split into indels, missense variants, and non-missense SNVs.

As with the General prior, we examined the variants that had been misclassified by the SCHEMA prior. The proportion of correctly and incorrectly classified variants that fell into each functional consequence cattegory is shown in Table 8.7. These proportions are broadly the same as those calculated under the General prior Table 8.4.

Consequence	Benign		Pathogenic		% INC
	CORR	INC	CORR	INC	
3' UTR	2.73%	0.00%	0.00%	0.35%	0.13%
5' UTR	1.17%	0.00%	0.00%	1.75%	1.5%
Frameshift	0.05%	0.99%	25.47%	0.35%	0.24%
Inframe deletion	0.27%	6.16%	1.02%	0.00%	13.59%
Inframe insertion	0.23%	0.00%	0.00%	3.66%	13.91%
Intron	25.66%	0.00%	0.00%	57.42%	2.23%
Missense	17.19%	91.26%	31.67%	24.96%	5.63%
Splice acceptor	0.01%	0.37%	6.60%	0.17%	0.37%
Splice donor	0.02%	0.49%	9.69%	0.00%	0.25%
Stop gained	0.02%	0.62%	25.53%	0.70%	0.22%
Stop lost	0.00%	0.12%	0.01%	0.00%	33.33%
Synonymous	52.65%	0.00%	0.00%	10.65%	0.21%

Table 8.7: The proportion of each functional consequence in the correctly and incorrectly classified Benign and Pathogenic variants, as well as the total proportion of incorrectly classified variants per consequence under the SCHEMA prior. CORR: correctly classified; INC incorrectly classified.

We generated posterior probabilities for causality for the family private variants in pedigrees K1546, K1494, and K1524. To ensure our results were consistent with the SCHEMA filtering analysis, we considered variants in genes with pLI > 0.9. For ease of presentation, we have shown the top 10 variants ranked by their posterior probability of causality in Table 8.8. The variants that had been prioritised from Chapter 6 are shown in green. The IBS-prioritised variants were ranked at least 4^{th} in each pedigree, and the posterior probability for causality was over 98% for all three variants. The ranking of the Bayes factors across all variants scored is moderate, which is consistent with the fact that the variants have a reduced co-segregation pattern. However, for each of the three

IBS-prioritised variants, the prior probability of causality is high (greater than 97%), which results in the large posterior probabilities of causality.

For the posterior ranking, six variants ranked higher than the IBS-prioritised variants overall in their respective pedigees. Of these six, all were rare or absent from gnomAD, five were LoF variants, and the remaining variant was missense. Ultra-rare LoF variants tend to receive a high prior due to the fact that ClinVar LoF variants are almost exclusively pathogenic. Missense variants require additional evidence from the MPC score to be causal, and this is likely why the five LoF variants are ranking higher according to the posterior ranking than the three IBS-prioritised missense variants. Based on the posterior probabilities of the three prioritised variants in their respective pedigrees, we can see that the missense variant in *ATP2B2* has the most evidence for causality. Additionally, the Bayes factor for the *ATP2B2* variant is also the largest out of the three, showing that it has the most evidence for causality based solely on co-segregation within the pedigrees.

While it is reassuring that the model gave high posterior probabilities to the IBS-prioritised variants, we were also interested in the other variants prioritised by the Bayesian model as shown in Table 8.8. One finding of particular interest is a frameshift variant in *KDM2B* that perfectly co-segregates with schizophrenia in pedigree K1546. This variant has a Bayes factor of 46.981 (the highest achieved in the pedigree), a prior probability of causality of 78.259% and a posterior probability of causality of 99.412%. This frameshift variant is rare in the general population (maximum allele frequency of 0.4%) but was not sufficiently rare to have been included in the SCHEMA-based analysis in Chapter 6 (minor allele count of 709).

KDM2B (lysine demethylase 2B) is involved in histone demethylation and has been linked to defects in embryonic development (Boulard et al., 2015). A recent study identified an ultra-rare missense variant in this gene perfectly co-segregating with schizophrenia in a Japanese pedigree, using linkage analysis and targeted exome sequencing (Yokotsuka-Ishida et al., 2021). The expression of *KDM2B* was found to be halved in the proband of this pedigree compared to five unrelated controls, and the psychiatric symptoms were more severe in the variant carriers compared to their non-carrier relatives. The authors suggest that haploinsufficiency of this gene is likely contributing to schizophrenia in this pedigree. The frameshift variant we observed in pedigree K1546 is present in the first exon of *KDM2B*, so if haploinsufficiency of this gene is implicated in schizophrenia, it is possible that this loss-of-function variant at the start of the amino acid chain may be having a similar effect in pedigree K1546.

An ultra-rare missense variant in *PBRM1* was observed in pedigree K1494 that had a reduced co-segregation pattern with schizophrenia. This variant has a Bayes factor of 5.102 (the second highest achieved in the pedigree), a prior probability of causality of 79.955% and a posterior probability of causality of 95.317%. The variant wasn't given an MPC score, so was removed from the previous analysis in Chapter 6. *PBRM1* (polybromo 1) is known to play a role in transcriptional activation (O'Leary et al., 2016), and is found to be expressed in many tissue types, including the brain (Lonsdale et al., 2013). GWAS for schizophrenia (Ikeda et al., 2019) and bipolar disorder (Mullins et al., 2021) have shown associations with *PBRM1*, and recently a SNP on chr3p21.1 (which contains *PBRM1*) was found to be associated with schizophrenia from a trio-based WES study (M. Li et al., 2020).

Finally, an ultra-rare missense variant in *TTBK1* in K1524 co-segregated perfectly with schizophrenia. This variant has a Bayes factor of 6.039 (the highest achieved in the pedigree), a prior probability of causality of 87.199% and a posterior probability of causality of 97.627%. This variant had a low MPC score of 1.04 and so was not included in the analysis in Chapter 6. *TTBK1* (tau tubulin kinase 1) is involved in the regulation of the phosphorylation of the tau protein (O'Leary et al., 2016) and is found to be expressed predominantly in brain tissue types (Lonsdale et al., 2013). A *de novo* missense variant in this gene has been implicated in childhood-onset schizophrenia (Ambalavanan et al., 2016).

CHR:POS:REF:ALT	Gene	CSQ	MAC	MPC	AF	BF		Prior	Posterior	
						Value	Rank		Value	Rank
chr1:35189207:G:A	*SFPQ*	stop gained	0	-	0.000%	1.550	6,349*	99.937%	99.959%	1
chr2:72135150:G:T	*CYP26B1*	stop gained	**76**	-	0.083%	4.116	1,276*	99.821%	99.956%	2
chr4:68322867:C:CG	*YTHDC1*	frameshift	0	-	0.000%	1.318	14,063*	99.883%	99.912%	3
chr3:10360021:G:A	*ATP2B2*	missense	1	2.23	0.003%	6.277	719*	97.372%	99.572%	4
chr6:130120959:G:T	*L3MBTL3*	splice donor	**322**	-	0.410%	1.295	17,727*	97.372%	99.504%	5
chr12:121580827:CTG:C	*KDM2B*	frameshift	**709**	-	0.421%	46.981	1*	78.259%	99.412%	6
chr19:41986225:T:G	*ATP1A3*	missense	0	3.10	0.000%	1.055	17,797*	99.106%	99.152%	7
chr19:41986226:A:G	*ATP1A3*	missense	0	2.91	0.000%	1.055	17,797*	98.894%	98.951%	8
chr6:75117532:C:T	*COL12A1*	missense	0	**1.23**	0.000%	6.277	719*	92.609%	98.745%	9
chr3:39410844:G:A	*RPSA*	missense	0	**N/A**	0.000%	6.277	719*	92.401%	98.707%	10

(a)

CHR:POS:REF:ALT	Gene	CSQ	MAC	MPC	AF	BF		Prior	Posterior	
						Value	Rank		Value	Rank
chr5:136172478:AT:A	*SMAD5*	frameshift	**11**	-	0.013%	0.981	11,437*	99.883%	99.881%	1
chr14:71724684:A:G	*SIPA1L1*	missense	0	**1.26**	0.000%	10.608	1*	92.902%	99.285%	2
chr19:42232651:A:G	*GSK3A*	missense	0	2.39	0.000%	2.056	3,696*	97.989%	99.012%	3
chr16:30669556:G:A	*FBRS*	missense	0	**0.34**	0.000%	10.608	1*	81.748%	97.939%	4
chr15:87929240:A:G	*NTRK3*	missense	4	**1.49**	0.003%	2.056	3,696*	94.047%	97.014%	5
chr11:8712724:G:A	*ST5*	missense	2	**1.95**	0.003%	0.981	11,437*	96.365%	96.297%	6
chr22:18083561:C:T	*PEX26*	missense	**25**	**0.68**	0.014%	5.102	2,199*	81.066%	95.623%	7
chr3:52634810:G:A	*PBRM1*	missense	**61**	**N/A**	0.051%	5.102	2,199*	79.955%	95.317%	8
chr16:53462597:A:C	*RBL2*	missense	**54**	**0.91**	0.051%	10.608	1*	65.417%	95.253%	9
chr11:67250601:G:T	*KDM2A*	missense	1	**0.92**	0.001%	2.056	3,696*	89.525%	94.616%	10

(b)

CHR:POS:REF:ALT	Gene	CSQ	MAC	MPC	AF	BF		Prior	Posterior	
						Value	Rank		Value	Rank
chr16:70557024:C:CTA	*SF3B3*	frameshift	1	-	0.001%	0.724	9,508*	99.883%	99.839%	1
chr10:99610923:T:C	*SLC25A28*	missense	0	2.12	0.000%	1.608	2,576*	97.251%	98.272%	2
chr20:51674280:T:C	*ATP9A*	missense	0	**0.93**	0.000%	6.039	1*	89.917%	98.177%	3
chr6:43257902:C:T	*TTBK1*	missense	**8**	**1.04**	0.013%	6.039	1*	87.199%	97.627%	4
chr7:5313814:A:T	*TNRC18*	missense	**7**	**0.87**	0.009%	6.039	1*	86.443%	97.469%	5
chr2:233198115:C:T	*INPP5D*	missense	**11**	**N/A**	0.010%	6.039	1*	82.282%	96.557%	6
chr10:75398805:G:A	*ZNF503*	missense	0	**1.51**	0.000%	1.608	2,576*	94.566%	96.550%	7
chr6:157200828:T:G	*ARID1B*	missense	**23**	**N/A**	0.016%	6.039	1*	77.832%	95.496%	8
chr3:170178789:T:C	*PHC3*	missense	2	2.00	0.002%	0.724	9,508*	96.678%	95.472%	9
chr2:86466529:A:G	*KDM3A*	missense	0	**0.10**	0.000%	6.039	1*	77.199%	95.337%	10

(c)

Table 8.8: The top 10 variants ranked by their posterior probability of causality for pedigrees: **(a)** K1546; **(b)** K1494; and **(c)** K1524. The variants identified from the filtering approach in Chapter 6 are shown in green. Other variants of interest are marked in yellow. Metrics displayed in red would not have passed the SCHEMA hard-filtering thresholds. Ranks marked with an asterisk (*) are joint with other variants in that pedigree. Variants with MPC marked as "N/A" had no score in dbNSFP for the vep canonical transcript. CHR: chromosome; POS: genomic position on GRCh38; REF: reference allele; ALT: alternate allele; CSQ: consequence; MAC: minor allele count in gnomAD; AF: population maximum allele frequency in gnomAD; BF: Bayes factor.

8.7 Conclusions

Overall, our novel Bayesian inference model performs well at identifying candidate disease-causing variants that have been prioritised using commonly applied filtering approaches, both for Mendelian and complex phenotypes. The model also discovers additional genes that are of interest to schizophrenia that were missed by the strict filtering approach applied in Chapter 6, highlighting the benefits of a more nuanced, integrated approach. The analysis is not limited by the frequency of the variants as we found with `pVAAST`, and the various scores are consistent with our expectations. Finally, the resulting metric used to prioritise variants is a well-defined probability, which allows for a more readily interpretable comparison of the evidence for causality of different variants across pedigrees.

Chapter 9

Conclusions and Future Work

This chapter summarises the results and conclusions of the primary analysis chapters of my thesis. I highlight some of the strengths and weaknesses of the individual analyses and suggest some future directions for this research.

9.1 Converting Single Nucleotide Variants Between Genome Builds

Chapter 3 investigated the known instabilities in converting SNVs between builds GRCh37 and GRCh38 of the human reference genome with the commonly used tools `liftOver` and `CrossMap`. I identified every position between the builds that either did not have a uniquely invertible mapping or mapped to a different chromosome, approximately 0.5% of either build. These instabilities were primarily located in either segmental duplications and/or sections of the genome that were updated between builds. I applied a similar methodology to two WGS samples and showed that the unstable positions were the same as previously characterised. Pre-filtering SNVs at these unstable positions prior to conversion resulted in variants that were entirely stable to the conversion process. This pre-filtering would have removed the missense variant we described in Section 3.1 which swapped to a different chromosome following conversion and whose CADD score changed from 20.8 to 0.009. The methodology implemented here is easily generalisable to any pairs of genome builds, requiring only a chain file. I have included the source code to run the algorithm as well as BED files describing the novel CUPs on GitHub (see "Web Resources", Subsection A.2.2).

Since two independent conversion tools generate identical CUPs, I concluded that these regions are determined by the chain files. This is important to note, as alternative chain files exist for converting between GRCh37 and GRCh38, and so the full algorithm would need to be applied for different chain files. However, the chain files used here are the only ones supplied by the authors of `liftOver` and `CrossMap`. While the full-genome data give insight into the behaviour of SNVs under build conversion, this does not account

for regions spanning multiple base pairs, as conversion tools are typically sensitive to this (Luu et al., 2020). A simple resolution to this might be to exclude regions that overlap any unstable positions, but this could potentially remove a large proportion of the input data. Recently, the Telomere-to-Telomere (T2T) consortium released a new build of the human reference genome, predominantly assembled using ultra-long read sequencing (Nurk et al., 2022). This new build (T2T-CHM13) uncovers the remaining 8% of the genome that has eluded sequencing technologies since the first release of the human genome over 20 years ago. Identifying conversion instabilities between this new build and the two examined here would be a straightforward extension of my work, and will no doubt prove crucial as T2T-CHM13 becomes more widely used.

9.2 CNV Calling Pipeline for Family-Based WGS Data

In Chapter 4, I present a novel CNV calling pipeline for family-based WGS data. This work was conceived and implemented jointly with a member of our research team (Dr Niamh Ryan). This pipeline implements a consensus of four CNV calling methods, comprising two RD-based methods, and two PR/SR-based methods. CNVs were required to have been called by at least two methods, but by considering calls from close relatives, we were able to recover CNVs that had no consensus call. Additionally, we benchmarked this pipeline and a method designed for the same data (Khan et al., 2018) against a set of curated "Gold Standard" CNV calls from a reference sample. We found that our pipeline outperformed all other comparison pipelines at calling CNVs of length at least 1kbp. Given the wide detection ability of CNV calling methods, both individually and as a pair (Kosugi et al., 2019), this benchmarking evaluation is important when selecting a consensus, and is not often performed prior to CNV analysis.

We decided to maximise the ability of our pipeline at detecting true positives, but by design this will lead to the inclusion of false positives, which our pipeline does not address. One resolution to this would be to consider CNV genotyping software to provide probabilistic estimates of the veracity of the CNV calls. Such methods often require additional tuning and may have an impact on the performance metrics but have the ability to remove putative CNV calls that are the result of sequencing artefacts. Another limitation of this work is the curated "Gold Standard" CNV calls. This collection was generated from five different CNV call sets on the same reference sample, but a low overlap was observed between these call sets. Some CNV calls were identified using long-read sequencing, which may not be detectable using short-read sequencing. Additionally, it was not possible to determine the type for each CNV (i.e. deletion or duplication), so

only CNV regions were examined. CNV callers are known to have variable performance on CNV type and length, so resolving this issue would provide greater insights into the performance of the benchmarked pipelines.

9.3 WGS Study of Discordant Identical Twins

In Chapter 5, I examined WGS data in a cohort of 17 monozygotic twins discordant for psychiatric illness. I examined various classes of common and rare discordant variants, i.e. not present in the co-twin. This analysis identified seven genes containing rare, family-private, predicted-deleterious, missense variants present in affected individuals. One such variant was found in a sample with major depressive disorder in *POLG*, which has previously been implicated in mood disorders and psychosis. Seven rare, gene-disrupting CNVs in affected individuals were also observed, one of which (a duplication at 3q29 in a sample with bipolar disorder) was predicted to be pathogenic in ClinGen. Deletions of this region are associated with schizophrenia, but duplications of this region have been observed in individuals with autism and intellectual disabilities (Rehm et al., 2015). I also examined other likely sources of schizophrenic genomic risk (regulatory variants, somatic CNVs, multi-nucleotide expansions) but observed no conclusive findings from these discordant variants. To the best of our knowledge this is the largest NGS study of discordant, psychiatric MZ twins. A previous study on eight MZ twins discordant for schizophrenia did not report any discordant *de novo* variants in protein-coding regions (Tang et al., 2017).

Somatic mosaicism has been implicated as a likely contributor to schizophrenia risk in discordant MZ twins using WES data (Nishioka et al., 2018). However, the depth of coverage for WGS data is likely not sufficient to accurately detect somatic variants, which may explain the negative findings of our somatic CNV analysis. Additionally, we have identified variants from blood-derived DNA, but there is no guarantee that these *de novo* variants will also be present in brain tissue, which has been shown to be important for schizophrenia (Fullard et al., 2019). Other plausible biological factors such as DNA methylation may be driving the discordance between the twins, which has been examined previously for schizophrenia (Castellani et al., 2015; Q. Li et al., 2021). Long-read sequencing technologies from Oxford Nanopore allow for the simultaneous examination of DNA sequences as well as methylation changes (Jain et al., 2016), so this could be an attractive option for targeted re-sequencing of known risk genes.

On a phenotype level, it is still possible that the "unaffected" twin within a pair may go on to receive a psychiatric diagnosis, although this is unlikely for the twin-pairs that were sampled after the typical age-of-onset. Given the age profile of some of the twin pairs, some of them may have children old enough to have a reliable psychiatric diagnosis. A follow up study including offspring of the twin pairs could allow for the examination of transmission and segregation of the prioritised variants in the next generation. While I have identified post-zygotic variants in blood tissue, almost 10% of *de novo* variants are estimated to occur post-fertilization and prior to progenitor germ cell specification and are likely to be present in both germ and blood cells (Sasani et al., 2019). Additionally, 2.1% of *de novo* variants are estimated to occur after the twinning event, but prior to progenitor germ cell specification (Jonsson et al., 2021). Given these relatively low transmission rates, if the variants identified here were found in affected offspring of the MZ twins, this would provide additional support for these variants as disease-causing within that pedigree.

9.4 WGS Study of Multiplex Utah Pedigrees

In Chapter 6, I examined WGS data from a cohort of 41 individuals from seven Utah pedigrees multiply affected with schizophrenia. In the absence of sufficient sample numbers to perform linkage analysis, I applied an IBS filtering approach to prioritise family-private variants based on the results of the recent SCHEMA analysis. This analysis did not identify any ultra-rare variants that perfectly co-segregated with schizophrenia in any pedigree, but I did identify three ultra-rare missense variants in three pedigrees that had a reduced co-segregation pattern. The most compelling evidence was from the gene *ATP2B2* in pedigree K1546 which has been implicated in a meta-analysis of common variants from schizophrenia and autism samples (Anney et al., 2017). In pedigree K1524, I observed one schizophrenia risk CNV (16p11.2 proximal duplication), which was present in the only affected sample not carrying a prioritised ultra-rare missense variant. The methodology implemented here is robust and follows practices established by large-scale genomics consortia.

The IBS filtering method implemented here is non-statistical, so cannot quantitatively compare the amount of evidence for causality for the three prioritised variants. Additionally, it is possible that some of the un-sequenced unaffected individuals may carry the prioritised variant, which would then exclude the prioritised variants based on the co-segregation criteria. While I did not allow for variants to be present in the unaffecteds (i.e. reduced in-family penetrance), this may be overly strict given that no Mendelian

sub-types of schizophrenia are known (Giusti-Rodríguez & Sullivan, 2013). Relaxing this restriction may reveal other variants of interest, although the interpretation may be challenging given the low sample numbers. This work is an exploratory analysis of these seven pedigrees, and while the three missense variants are interesting, there may be additional genetic factors contributing to schizophrenia risk in these pedigrees. Other classes of rare DNA variants such as CNVs or multi-nucleotide repeats may also be co-segregating with schizophrenia in these pedigrees, which warrant further investigation.

9.5 Evaluation of Two Software Tools for Disease-Gene Prioritisation

Chapter 7 evaluates the ability of `pVAAST` and `PERCH` to detect candidate causal variants based on co-segregation within a pedigree. Using a synthetic Mendelian phenotype from the three-generational CEPH 1463 pedigree, I showed that `pVAAST` performs well at identifying the pre-selected pseudo-causal variants. `PERCH`, on the other hand, correctly identified only 14 of the 24 pre-selected pseudo-causal variants, and the remaining 10 variants received either no score or a score indicating evidence against causality. I attempted to explore the components of `PERCH`'s underlying co-segregation model but could not explain this unexpected behaviour. I estimated the null distribution of the scores for each pseudo causal gene using both tools using a permutation test, and the significance values broadly reflected the above observations about the tools. Next, I applied `pVAAST` to the three pedigrees from Chapter 6, but the three variants prioritised from the IBS filtering strategy did not score well. The co-segregation scores for two of the variants were zero, while the third was not consistent with other co-segregation scores in the same pedigree on visual inspection.

One major limitation with this benchmarking is the time taken to run the individual tools. `PERCH` took approximately 1-2 minutes to complete scoring on the CEPH 1463 pedigree, compared to 30-45 minutes for `pVAAST` on the 24 pseudo-causal genes only. This means that generating a sufficient number of permutations to accurately estimate the null distribution can be time-consuming, (e.g. over one month for 1,000 permutations for `pVAAST`). It is likely that `pVAAST` performs poorly on the three Utah pedigrees because the prioritised variants are ultra-rare in gnomAD, and `pVAAST` estimates allele frequencies from the exome aggregation consortium (ExAC), a predecessor of gnomAD. I was unable to supply external allele frequencies to `pVAAST` but resolving this issue might allow `pVAAST` to score a larger number of variants in the pedigrees. The limited ability of the tools to detect the variants prioritised from my IBS filtering make them unusable as

comparison tools for my data, so I removed them from further analysis.

9.6 Bayesian Inference Model to Measure Co-Segregation in Pedigrees

In Chapter 8, I developed a Bayesian model to measure pedigree-based causality using NGS data. The Bayes factor calculation extended previous work (Mohammadi et al., 2009) and compares the likelihood that a variant is causal for a phenotype in a pedigree to the likelihood that the variant is not causal. The prior probability for causality for a variant is determined by functional consequence, allele frequency and deleterious measures. These probabilities were calculated from coefficients derived from a logistic regression model, trained using benign and pathogenic variants taken from ClinVar. The resulting posterior probability for causality was used to rank variants in a given pedigree. This novel Bayesian model has the advantage over tools like `pVAAST` and `PERCH` that the resulting metric is a well-defined probability that incorporates prior information about the models in a more coherent format. This final metric is more readily interpreted and enables a more consistent comparison across pedigrees.

I applied this model to the CEPH 1463 pedigree described in Chapter 7, to determine how well the model could identify the pseudo-causal variants from a synthetic Mendelian phenotype. The Bayes factor for each pseudo-causal variant was the joint highest in the pedigree, as expected. The prior probabilities for causality were negatively correlated with the allele frequency, which was expected given that pathogenic variants in ClinVar are rare in the general population. Next, I applied the model to the three pedigrees from Chapter 6, and the three prioritised variants were in the top four of all family private variants in constrained genes for each pedigree. Finally, I extended our search to other highly scoring variants in these three pedigrees. This uncovered a frameshift variant in *KDM2B* observed in pedigree K1546 that perfectly co-segregated with schizophrenia. This variant was rare in the general population, but not sufficiently rare to have been included in the IBS filtering in Chapter 6. A missense variant in this gene has been reported to co-segregate with schizophrenia in a Japanese pedigree (Yokotsuka-Ishida et al., 2021). This shows the benefit of a more integrated approach to measuring co-segregation compared to a blunt IBS filtering method.

Despite its advantages over other approaches studied in this thesis, the novel Bayesian model has its limitations. Since the prior probability for causality is heavily dominated by allele frequency, rare variants can expect to have high posterior probabilities unless

they have low Bayes factors. The minimum value a Bayes factor can take will be influenced by the number of unaffected individuals in the pedigree carrying the variant. Since there were low numbers of unaffected individuals, the model may not be able to correctly down-weight ultra-rare variants due to a lack of noteworthy co-segregation scores. However, information from multiple pedigrees can be combined by multiplying Bayes factors for the same variant, so expanding the search space to non-family private variants may help alleviate this issue. This will depend on polygenicity of the phenotype being examined, and how likely it is that many rare variants have an influence. The prior distributions for the parameter terms were selected for ease of integration into the model. Facilitating other distributions (e.g. the Beta distribution) could allow for greater flexibility for modelling phenotypes, although this would result in a non-closed-form calculation for the Bayes factors, which may come at the cost of the runtime of the code. An additional modification would be to allow for age-specific distributions, which would be particularly useful for schizophrenia.

9.7 Final Remarks

In this thesis, I identified and characterised rare DNA variants likely implicated in schizophrenia and related disorders using WGS data in various family structures. My work highlighted some pitfalls and inconsistencies often encountered when working with various classes of variants using NGS data and provided robust solutions to avoid them. Finally, I have examined the strengths and weaknesses of several strategies for disease-gene prioritisation in family based NGS studies. Based on this, I have created a novel Bayesian framework to resolve some of the issues observed, which should help researchers further explore the aetiology of complex genetic disorders in pedigree data.

Appendix A

Supplementary Information

A.1 WGS Details

The following is adapted from a report provided by Edinburgh Genomics on the release of data.

EGCG utilises Illumina SeqLab, which integrates Illumina TruSeq library preparation, Illumina cBot2 cluster generation, Illumina HiSeqX sequencing, Hamilton Microlab STAR integrative automation, and Genologics Clarity LIMS X Edition.

Sample QC: Genomic DNA (gDNA) samples were evaluated for quantity and quality using an AATI Fragment Analyzer and the DNF-487 Standard Sensitivity Genomic DNA Analysis Kit. The AATI ProSize 2.0 software provides a quantification value and a quality (integrity) score for each gDNA sample. gDNA samples failed sample QC if they were found to have i) a quality score < 7 and ii) little or no high molecular weight material. For such samples, replacement samples were requested. Based on the quantification results, gDNA samples were pre-normalised to fall within the acceptable range of the Illumina SeqLab TruSeq Nano library preparation method using the Hamilton MicroLab STAR.

Library Preparation: Next Generation sequencing libraries were prepared using Illumina SeqLab specific TruSeq Nano High Throughput library preparation kits in conjunction with the Hamilton MicroLab STAR and Clarity LIMS X Edition. The gDNA samples were normalised to the concentration and volume required for the Illumina TruSeq Nano library preparation kits, then sheared to a 450bp mean insert size using a Covaris LE220 focused-ultrasonicator. The inserts were ligated with blunt ended, Atailed, size selected, TruSeq adapters and enriched using 8 cycles of PCR amplification.

Library QC: The libraries were evaluated for mean peak size and quantity using the Caliper GX Touch with a HT DNA 1k/12K/HI SENS LabChip and HT DNA HI SENS Reagent Kit. The libraries were normalised to 5nM using the GX data and the actual

concentration was established using a Roche LightCycler 480 and a Kapa Illumina Library Quantification kit and Standards.

Sequencing: The libraries were normalised, denatured, and pooled in eights for clustering and sequencing using a Hamilton MicroLab STAR with Genologics Clarity LIMS X Edition. Libraries were clustered onto HiSeqX Flow cell v2.5 on cBot2s and the clustered flow cell was transferred to a HiSeqX for sequencing using a HiSeqX Ten Reagent kit v2.5.

Yield and Coverage: Yield was calculated as the number of bases provided in the FASTQ files, expressed in gigabases (Gb). Coverage was defined as the average number of bases covering each position of the reference genome. The expected yield and coverage for all samples is 120 Gb and 30x respectively. Samples that fell below both thresholds were resequenced.

Bioinformatics Demultiplexing was performed using `bcl2fastq` (2.17.1.14), allowing 1 mismatch when assigning reads to barcodes. Adapters were trimmed during the demultiplexing process. `BCBio-Nextgen` (0.9.7) was used to perform alignment, BAM file preparation and variant detection. BCBio uses `bwa mem` (0.7.13) to align the raw reads to the GRCh38 (with alt, decoy and HLA sequences) genome, then `samblaster` (0.1.22) to mark the duplicated fragments, and the `Genome Analysis ToolKit` (3.4-0-g7e26428) for the indel realignment and base recalibration. The genotype likelihoods were calculated using Genome Analysis Toolkit (3.4-0-g7e26428) `HaplotypeCaller` creating a final gVCF file.

A.2 Web Resources

The URLs supplied in this section were accessible on 30/06/22.

A.2.1 Chapter 2

- The FastQC tool:
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Read alignment and variant calling pipeline:
  https://github.com/cathaloruaidh/WGS_Alignment_Calling
- picard tools
  https://broadinstitute.github.io/picard/
- VQSR and variant filtering pipeline:
  https://github.com/cathaloruaidh/WGS_JointGeno_VQSR
- GATK advice on hard filtering:
  https://gatk.broadinstitute.org/hc/en-us/articles/360037499012-I-am-unable-to-use-VQSR-recalibration-to-filter-variants
- peddy open issue regarding half-siblings:
  https://github.com/brentp/peddy/issues/21
- peddy open issue regarding consanguinity:
  https://github.com/brentp/peddy/issues/56

A.2.2 Chapter 3

- UCSC Genome Browser user guide on build conversion:
  https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Convert
- UCSC Genome Browser support forum on liftOver errors, with variants swapping chromosomes:
  https://groups.google.com/a/soe.ucsc.edu/g/genome/c/P3M1Q5baozM/m/Slyjdco5BwAJ
- The online implementation of liftOver:
  https://genome.ucsc.edu/cgi-bin/hgLiftOver
- The executable file for the version of liftOver used here:
  http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64.v369/liftOver

- The online manual for CrossMap:
  https://crossmap.readthedocs.io/en/latest/

- Chain file for GRCh37 to GRCh38, provided by the UCSC Genomics Institute:
  http://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz

- Chain file for GRCh38 to GRCh37, provided by the UCSC Genomics Institute:
  http://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/hg38ToHg19.over.chain.gz

- The UCSC Genomics Institute recommendation on the minimum proportion of sequence identity:
  http://genomewiki.ucsc.edu/index.php/LiftOver_Howto

- The Illumina Platinum Genomes project:
  https://www.illumina.com/platinumgenomes.html

- List of URLs for the VCF files used in the WGS section ("small_variants" for NA12877 and NA12878):
  https://github.com/Illumina/PlatinumGenomes/blob/master/files/2017-1.0.files

- Source code for the algorithm to identify conversion unstable positions, as well as the BED files for all positions on GRCh37 and GRCh38:
  https://github.com/cathaloruaidh/genomeBuildConversion

A.2.3 Chapter 4

- The Database of Genomic Variants on the difficulty of determining whether a CNV is a deletion or duplication:
  http://dgv.tcag.ca/dgv/app/faq?ref=GRCh37/hg19#q2

A.2.4 Chapter 5

- RegulomeDB:
  https://regulomedb.org/regulome-search/

- MoChA (Mosaic Chromosomal Alterations):
  https://github.com/freeseek/mocha

A.2.5 Chapter 6

- BAM to FASTQ revertion:
  https://github.com/cathaloruaidh/BAMtoFASTQ

A.2.6 Chapter 7

- SRA toolkit:
  https://hpc.nih.gov/apps/sratoolkit.html
- Discussion on the "network timeout" issue from the authors of the SRA toolkit:
  https://github.com/ncbi/sra-tools/issues/139#issuecomment-405562470
- BBMap:
  https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/
- Background exome data supplied with pVAAST:
  http://www.yandell-lab.org/software/VAAST/data/VAAST2/GRCh37/Background_CDR/
- RefSeq genomic feature GFF3 file:
  http://www.yandell-lab.org/software/VAAST/data/VAAST2/GRCh37/Features/

A.2.7 Chapter 8

- The GRCh37 version of ClinVar (released 29/01/2022):
  https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/weekly/clinvar_20220129.vcf.gz
- The GRCh38 version of ClinVar (released 29/01/2022):
  https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/weekly/clinvar_20220129.vcf.gz

A.3 Software Versions

Software	Version	Software	Version
BBMap	38	MoChA	2019-10-16
bcftools	1.11	peddy	0.4.3
bedtools	2.28.0	PERCH	1.0
bgzip	1.4.1	picard	2.9.2
BWA-MEM	0.7.13	plink	1.9
cn.mops	3.8	pVAAST	2.2.0
CNVnator	0.3.3	R	2.4.1
CrossMap	0.4.2	samblaster	0.1.22
DELLY	0.8.1	samplot	1.3.0
ERDS	1.1	samtools	1.4.1
ExpansionHunter	3.0.0	shuf	8.22
FastQC	0.11.5	SnpSift	5.0
GATK	3.4/3.8	split	8.22
IGV	2.8.9	tabix	1.4.1
kinship2	1.8.5	vcftools	0.1.15
LUMPY	0.2.13	vep	97.0
Manta	1.4.0	XPAT	1.0

Table A.1: Versions of the software used in this thesis.

A.4 Novel Javascript Code

```
Function Accept(genotypes)
    fam_aff ← [...] # IDs of affected family members
    fam_unaff ← [...] # IDs of unaffected family members
    fam_req ← [...] # IDs of family members required to carry the variant
    fam_marry ← [...] # IDs of marry-in family members
    num_aff ← 0
    for i ← 0 to numSamples do
        if i in fam_aff and i is not HomRef then
            num_aff ← num_aff + 1
        end
        else if i in fam_unaff and i is not HomRef then
            return false
        if i in fam_req and i is HomRef then
            return false
        end
        else if i in fam_marry and i is not HomRef then
            return false
        else if i in fam_unaff and i is not HomRef then
            return false
        if num_aff == length(fam_aff) then
            return true
        end
    end
end
```

Algorithm 1: `Accept` compares the genotypes of a variant to the pedigree phenotypes to ensure full co-segregation. Reduced co-segregation is achieved by changing the number of affected individuals required.

Appendix B

Mathematical Details for the Bayesian Inference Model

B.1 Definitions and Theorems

Define the following variables/expressions for a given variant in a family:

p	the proband
G_i	the genotype of the variant for individual i, equal to 0 or 1
G_p	the genotype of the proband (assumed to carry the variant)
G_O	the genotypes observed in the family (sequenced individuals)
G_U	the genotypes of the individuals not observed in the family
G_F	the genotypes of all individuals in the family
P_F	the phenotypes of the family
β	the family-based penetrance, i.e. $\mathbb{P}(P_i = 1 \mid G_i = 1)$
φ	the family-based phenocopy rate, i.e. $\mathbb{P}(P_i = 1 \mid G_i = 0)$
α	the population incidence rate, i.e. $\mathbb{P}(P = 1)$
D	the data in the model
M_1	the causal model for the variant
M_2	the neutral model for the variant
k_1	number of affected carriers $(P_i = 1, G_i = 1)$
k_2	number of unaffected carriers $(P_i = 0, G_i = 1)$
k	number of carriers $(k_1 + k_2)$
l_1	number of affected non-carriers $(P_i = 1, G_i = 0)$
l_2	number of unaffected non-carriers $(P_i = 0, G_i = 0)$
l	number of non-carriers $(l_1 + l_2)$
n	number of individuals in the family $(k + l)$

Theorem B.1 (Binomial Theorem)**.** *Given any two real numbers $x, y \in \mathbb{R}$ and a natural*

number $n \in \mathbb{N}^+$, we have:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

where $\binom{n}{k} = \frac{n!}{(n-k)!k!}$

Corollary B.2. *Given any real number $a \in \mathbb{R}$ and a natural number $n \in \mathbb{N}^+$, we have:*

$$(1-a)^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k a^k$$

B.2 Overview

B.2.1 Data and Parameters

The data D in our model refer to the observed genotypes (G_O), and the known phenotypes in the family (P_F). We assume that every sample has a phenotype specified, but samples may have missing genotypes.

The unobserved genotypes (G_U) are a discrete random variable with probabilities determined by Mendelian segregation. The in-family penetrance (β) and the in-family phenocopy rate (φ) are probabilities, and so are continuous random variables. For convenience, let $\theta = (\beta, \varphi)$, and let $\Omega = [0,1] \times [0,1]$ be the domain of θ. The population prevalence (α) is also a continuous variable in $[0,1]$. p is the proband, which should be selected at random from the affected variant carriers.

B.2.2 Models

For a given variant the **causal model** (M_1) states that the variant is a primary contributor to the phenotype. As an alternate to this, the **neutral model** (M_2) states that the variant under examination does not contribute to the phenotype of interest. Ultimately we will calulcate a Bayes factor for these two models based on our data, which is the ratio of the likelihood for the causal model and the likelihood of the neutral model:

$$BF = \frac{\mathbb{P}(D \mid M_1)}{\mathbb{P}(D \mid M_2)}$$

B.2.3 Assumptions

Following Mohammadi et al., we assume that the variant is rare in the general population (Mohammadi et al., 2009). This allows us to conclude that the variant originates from one founder within the pedigree, and so cannot be carried by marry-in individuals. In determining the phenotypes conditional on the genotypes, we will have different assumptions based on which model we are evaluating:

- M_1 (**Causal Model**) - the variant has a dominant effect on the phenotype: this means that an individual's phenotype is determined solely by their genotype
- M_2 (**Neutral Model**) - the phenotypes are entirely independent of all genotypes and are determined by the population incidence rate

B.2.4 Variables

We can construct a relationship between φ and β. If α is the population incidence of a phenotype, p is the allele frequency of a causal variant and $q = 1 - p$, then we have:

$$\begin{aligned} \alpha &= \mathbb{P}(P) \\ &= \mathbb{P}(P \mid \text{Geno})\,\mathbb{P}(\text{Geno}) + \mathbb{P}(P \mid \text{not Geno})\,\mathbb{P}(\text{not Geno}) \\ &= \beta(1-q^2) + \varphi(q^2) \end{aligned} \tag{B.1}$$

Here, "Geno" refers to having a genotype that confers risk for the phenotype in a dominant fashion, so heterozygous or homozygous variant genotypes. We can see that α is a weighted average of β and φ, so we have three scenarios:

- **Scenario 1:** $\varphi < \alpha < \beta$ - recommended by Petersen et al. under the causal model (Petersen et al., 1998);
- **Scenario 2:** $\varphi = \alpha = \beta$ - here, the phenotype is independent of the genotype, which corresponds to the neutral model; and
- **Scenario 3:** $\beta < \alpha < \varphi$ - having the variant reduces the probability of the phenotype compared to the incidence rate, and compared to not having the variant. This corresponds to a variant with a protective effect, which the causal model does not account for.

We assume **Scenario 1** for the causal model and **Scenario 2** for the neutral model.

B.3 Causal Model

To calculate $\mathbb{P}(D \mid M_1)$, we first marginalise over the parameters:

$$\begin{aligned}\mathbb{P}(D \mid M_1) &= \int \mathbb{P}(D \mid \Theta, M_1)\,\mathbb{P}(\Theta \mid M_1)\; d\Theta \\ &= \sum_{p=1}^{k_1} \sum_{G_U} \iint_{\Omega} \mathbb{P}(G_O, P_F \mid p, G_U, \theta, M_1)\mathbb{P}(p, G_U, \theta \mid M_1)\; d\theta\end{aligned} \tag{B.2}$$

We can re-write the probability terms of the integrand in Equation B.2 as follows:

$$\begin{aligned}\mathbb{P}(G_O, P_F \mid p, G_U, \theta, M_1) &= \overbrace{\mathbb{P}(P_F \mid G_O, p, G_U, \theta, M_1)}^{\text{phenotypes}}\,\mathbb{P}(G_O \mid p, G_U, \theta, M_1) \\ \mathbb{P}(G_O \mid p, G_U, \theta, M_1) &= \frac{\overbrace{\mathbb{P}(G_O, G_U \mid p, \theta, M_1)}^{\text{inheritance}}}{\mathbb{P}(G_U \mid p, \theta, M_1)} \\ \mathbb{P}(p, G_U, \theta \mid M_1) &= \mathbb{P}(G_U \mid p, \theta, M_1)\,\overbrace{\mathbb{P}(p, \theta \mid M_1)}^{\text{parameters}}\end{aligned} \tag{B.3}$$

For convenience, we will write $G_F = G_O, G_U$. In the inheritance term, once the proband, the observed and the unobserved genotypes are specified, this probability is independent of β, φ and the model, so we can write it as $\mathbb{P}(G_F \mid p)$. Combining these into Equation B.2 we get:

$$\mathbb{P}(D \mid M_1) = \sum_{p=1}^{k_1} \sum_{G_U} \iint_{\Omega} \underbrace{\mathbb{P}(P_F \mid G_F, p, \theta, M_1)}_{\text{phenotypes}} \underbrace{\mathbb{P}(G_F \mid p)}_{\text{inheritance}} \underbrace{\mathbb{P}(p, \theta \mid M_1)}_{\text{parameters}}\; d\theta \tag{B.4}$$

When we iterate over all unobserved genotypes and calculate the inheritance probability term $\mathbb{P}(G_F \mid p)$, there are some genotype configurations that we may ignore due to the rare variant assumption. For example: given a proband, their married-in parent is a theoretical source for the variant (i.e. a founder). However, if there are other candidate probands in the pedigree that are unrelated to the marry-in individual, the marry-in individual should be ignored since they cannot carry the variant when considering the other probands. This means that the only founders we need consider are those that are ancestors of all carriers for a given variant. Such founders will give rise to an identical collection of permissible unobserved genotypes regardless of the proband selection. Pseudocode for the algorithms to enumerate the permissible genotypes is

shown in Section B.6. As an example to illustrate this, consider the pedigree in Figure B.1 taken from Mohammadi et al.:

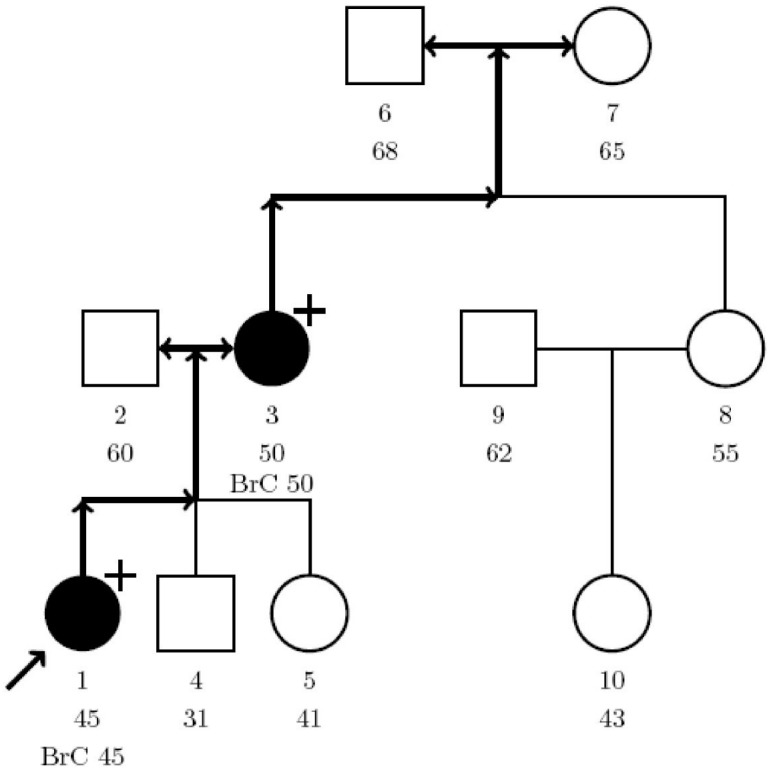


Figure B.1: A simulated breast cancer (BrC) pedigree showing affected individuals (shaded) and theoretical *BRCA1* carriers (+) (Mohammadi et al., 2009). The first level labels are the individuals' IDs, the second level are their ages at sampling, and the third level are the ages and diagnoses, if applicable.

Here, individuals 1 and 3 are known to have breast cancer and they also carry the variant of interest. All other individuals are unaffected and their genotypes are unknown. If individual 3 is the current proband, the potential founders are individuals 6 and 7. If individual 1 is the current proband, the potential founders are individuals 2, 6 and 7. However, since individual 3 is not descended from individual 2, individual 2 cannot be a founder for the entire pedigree. So, the only pedigree founders we need consider are individuals 6 and 7, which is the same regardless of the proband selection. Since $\mathbb{P}(G_F \mid p)$ is now independent of the choice of proband, we will simply write it as $\mathbb{P}(G_F)$.

Given that we have a fixed collection of founders, this will give rise to the same set of unobserved genotypes regardless of the proband selection. Hence, we can simplify

Equation B.4:

$$\begin{aligned}\mathbb{P}(D \mid M_1) &= \sum_{p=1}^{k_1} \sum_{G_U} \iint_{\Omega} \mathbb{P}(P_F \mid G_F, p, \theta, M_1)\, \mathbb{P}(G_F)\, \mathbb{P}(p, \theta \mid M_1)\ d\theta \\ &= k_1 \sum_{G_U} \iint_{\Omega} \mathbb{P}(P_F \mid G_F, p, \theta, M_1)\, \mathbb{P}(G_F)\, \mathbb{P}(p, \theta \mid M_1)\ d\theta\end{aligned} \tag{B.5}$$

Following Mohammadi et al., since the variant is assumed to have a dominant effect, the phenotypes are determined solely by an individual's genotype, and so are independent within the family. This means that the phenotype term can be expressed as a product of β and φ:

$$\mathbb{P}(P_F \mid G_F, p, \theta, M_1) = \prod_{i=1}^{n} \mathbb{P}(P_i \mid G_i, \theta, M_1) = \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2} \tag{B.6}$$

This gives:

$$\begin{aligned}\mathbb{P}(D \mid M_1) &= k_1 \sum_{G_U} \iint_{\Omega} \mathbb{P}(P_F \mid G_F, p, \theta, M_1)\, \mathbb{P}(G_F)\, \mathbb{P}(p, \theta \mid M_1)\ d\theta \\ &= k_1 \sum_{G_U} \mathbb{P}(G_F) \int_0^1 \int_0^1 \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\mathbb{P}(p, \beta, \varphi \mid M_1)\ d\beta\, d\varphi\end{aligned} \tag{B.7}$$

The choice of proband is independent of β and φ, and we will let the prior probability for this selection be uniform over the k_1 potential probands, giving $\mathbb{P}(p, \beta, \varphi \mid M_1) = \frac{1}{k_1}\mathbb{P}(\beta, \varphi \mid M_1)$.

It is worth noting that the prior probability for the choice of proband is dependent on the data, in particular the number of probands for a given variant in the pedigree. Strictly speaking, this breaks the principle in Bayesian inference modelling that the prior distributions should be determined independently from the data. However, as discussed by Andrew Gelman on his blog (Gelman, 2016), we can consider this data-dependent prior as an approximation to the true prior distribution for the parameter. If the proband selection has a prior distribution that depends on some hyperparameter N, we can approximate N by a simple point estimate of k_1.

Returning to Equation B.7, we have:

$$\begin{aligned}\mathbb{P}\left(D \mid M_1\right) &= k_1 \sum_{G_U} \mathbb{P}\left(G_F\right) \frac{1}{k_1} \int_0^1 \int_0^1 \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\mathbb{P}\left(\beta,\varphi \mid M_1\right)\, d\beta\, d\varphi \\ &= \sum_{G_U} \mathbb{P}\left(G_F\right) \underbrace{\int_0^1 \int_0^1 \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\mathbb{P}\left(\beta,\varphi \mid M_1\right)\, d\beta\, d\varphi}_{I}\end{aligned} \tag{B.8}$$

B.3.1 Uniform Prior

In the absence of additional information, we'll let the prior joint distribution for β and φ be uniform over $\beta \in [0,1], \varphi \in [0,1]$ with the assumption that $\varphi < \beta$. This means that the joint density function is given by:

$$f(\beta,\varphi) = \left\{ \begin{array}{ll} 2 & \text{if } 0 \leqslant \varphi < \beta \leqslant 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{B.9}$$

We can confirm that this is a valid density function since:

$$\int_0^1 \int_0^1 f(\beta,\varphi)\, d\varphi\, d\beta = \int_0^1 \int_0^\beta 2\, d\varphi\, d\beta = 2\int_0^1 \beta\, d\beta = 2\left[\frac{\beta^2}{2}\right]_0^1 = 2\left[\tfrac{1}{2}\right] = 1$$

Alternatively, swapping the order of integration:

$$\int_0^1 \int_0^1 f(\beta,\varphi)\, d\beta\, d\varphi = \int_0^1 \int_\varphi^1 2\, d\beta\, d\varphi = 2\int_0^1 (1-\varphi)\, d\varphi = 2\left[\varphi - \frac{\varphi^2}{2}\right]_0^1 = 2\left[1-\tfrac{1}{2}\right] = 1$$

We can simplify the integral I in Equation B.8 above as follows:

$$\begin{aligned} I &= \int_0^1 \int_0^1 \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\mathbb{P}\left(\beta,\varphi \mid M_1\right)\, d\varphi\, d\beta \\ &= \int_0^1 \int_0^\beta \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}(2)\, d\varphi\, d\beta \\ &= 2\int_0^1 \beta^{k_1}(1-\beta)^{k_2} \underbrace{\left(\int_0^\beta \varphi^{l_1}(1-\varphi)^{l_2}\, d\varphi\right)}_{I_\varphi} d\beta \end{aligned} \tag{B.10}$$

We can evaluate I_φ by using Corollary B.2 on $(1-\varphi)^{l_2}$, which is valid since $l_2 \geqslant 0$.

$$\begin{aligned} I_\varphi &= \int_0^\beta \varphi^{l_1}(1-\varphi)^{l_2}\, d\varphi \\ &= \int_0^\beta \varphi^{l_1} \left\{ \sum_{i=0}^{l_2} \binom{l_2}{i} (-1)^{l_2-i} (\varphi)^{l_2-i} \right\} d\varphi \\ &= \sum_{i=0}^{l_2} \binom{l_2}{i} (-1)^{l_2-i} \left\{ \int_0^\beta \varphi^{l_1+l_2-i}\, d\varphi \right\} \\ &= \sum_{i=0}^{l_2} \binom{l_2}{i} (-1)^{l_2-i} \left[\frac{\varphi^{l-i+1}}{l-i+1} \right]_{\varphi=0}^{\varphi=\beta} \\ &= \sum_{i=0}^{l_2} \binom{l_2}{i} (-1)^{l_2-i} \left(\frac{\beta^{l-i+1}}{l-i+1} \right) \end{aligned} \tag{B.11}$$

The exponent in the final integrand is positive since $l_2 \geqslant i$, and so $l = l_1 + l_2 \geqslant i$, giving $l - i + 1 \geqslant 1$. Now, we can substitute the above into Equation B.10:

$$\begin{aligned} I &= 2\int_0^1 \beta^{k_1}(1-\beta)^{k_2} I_\varphi\, d\beta \\ &= 2\int_0^1 \beta^{k_1}(1-\beta)^{k_2} \left[\sum_{i=0}^{l_2} \binom{l_2}{i} (-1)^{l_2-i} \left(\frac{\beta^{l-i+1}}{l-i+1} \right) \right] d\beta \\ &= 2\sum_{i=0}^{l_2} \binom{l_2}{i} \frac{(-1)^{l_2-i}}{l-i+1} \int_0^1 \beta^{k_1+l-i+1}(1-\beta)^{k_2}\, d\beta \\ &= 2\sum_{i=0}^{l_2} \binom{l_2}{i} \frac{(-1)^{l_2-i}}{l-i+1} \frac{(k_1+l-i+1)!(k_2)!}{(n-i+2)!} \\ &= 2\sum_{i=0}^{l_2} \binom{l_2}{i} \frac{(-1)^{l_2-i}}{l-i+1} \frac{1}{(n-i+2)\binom{n-i+1}{k_2}} \end{aligned} \tag{B.12}$$

where we have used the Beta and Gamma functions for positive integers x and y to simplify:

$$\begin{aligned} B(x+1, y+1) &:= \int_0^1 t^x(1-t)^y\, dt = \frac{\Gamma(x+1)\Gamma(y+1)}{\Gamma(x+y+2)} = \frac{x!y!}{(x+y+1)!} \\ \Gamma(x+1) &= x! \end{aligned}$$

Finally, substituting this into Equation B.8 we get

$$\mathbb{P}(D \mid M_1) = 2 \sum_{G_U} \mathbb{P}(G_F) \sum_{i=0}^{l_2} \binom{l_2}{i} \frac{(-1)^{l_2-i}}{l-i+1} \frac{1}{(n-i+2)\binom{n-i+1}{k_2}} \tag{B.13}$$

B.3.2 Beta Prior

Given the form Equation B.8, a sensible option for the prior distribution of β and φ is the Beta distribution. This distribution is a conjugate prior to the integrand above and based on two hyper-parameters x and y. The Beta distribution is often used to model the random behaviour of probabilities, which fits with the definition of β and φ. Additionally, when $x = y = 1$, the Beta distribution becomes the uniform distribution.

Suppose β and φ have Beta prior distributions with hyperparameters x_β, y_β and x_φ, y_φ respectively, with $x_\beta, y_\beta, x_\varphi, y_\varphi \in \mathbb{N}^+$. We require the hyperparameters to be integers to apply the Binomial Theorem later on. It is possible to have non-integer hyperparameters, but this would involve evaluating the Beta function at non-integer values, which does not have a closed form. Such integrals would likely only be solved by numerical methods, which are typically computationally expensive.

For a given $\beta \in [0, 1]$ and with the assumption that $\varphi < \beta$, we have the following probability density function for φ:

$$\mathbb{P}(\varphi \mid \beta, M_1) = \begin{cases} \dfrac{1}{f(\beta, x_\varphi, y_\varphi)} \dfrac{\varphi^{x_\varphi - 1}(1-\varphi)^{y_\varphi - 1}}{B(x_\varphi, y_\varphi)} & \text{if } 0 \leqslant \varphi < \beta \\ 0 & \text{otherwise} \end{cases}$$

where f is some normalisation function that is constant with respect to φ. This function is required since φ is only non-zero on $[0, \beta]$, so we need to scale the new density function appropriately. For this to be a valid probability density function, it must integrate to 1, so:

$$\begin{aligned} 1 &= \int_0^1 \mathbb{P}(\varphi \mid \beta, M_1) \; d\varphi \\ &= \frac{1}{f(\beta, x_\varphi, y_\varphi)} \int_0^\beta \frac{\varphi^{x_\varphi - 1}(1-\varphi)^{y_\varphi - 1}}{B(x_\varphi, y_\varphi)} \, d\varphi \\ &= \frac{I(\beta; x_\varphi, y_\varphi)}{f(\beta, x_\varphi, y_\varphi)} \end{aligned}$$

This gives us the relation $f(\beta, x_\varphi, y_\varphi) = I(\beta; x_\varphi, y_\varphi)$, where $I(x; a, b)$ is the regularized incomplete Beta function. When $x_\varphi = 1$ or $y_\varphi = 1$, this function takes specific values, so we have:

$$f(\beta, x_\varphi, y_\varphi) = \begin{cases} 1 - (1-\beta)^{y_\varphi} & \text{if } x_\varphi = 1 \\ \beta^{x_\varphi} & \text{if } y_\varphi = 1 \end{cases}$$

Other values for x_φ or y_φ are possible, but this function is defined iteratively. The prior distribution for β is also a Beta distribution, but no adjustment is needed since it is defined for $0 \leqslant \beta \leqslant 1$.

Alternatively, we can fix φ and with the assumption that $\varphi < \beta$, we get the following probability density function for β:

$$\mathbb{P}\left(\beta \,|\, \varphi, M_1\right) = \begin{cases} \dfrac{1}{g(\varphi, x_\beta, y_\beta)} \dfrac{\beta^{x_\beta - 1}(1-\beta)^{y_\beta - 1}}{B(x_\beta, y_\beta)} & \text{if } 0 \leqslant \varphi < \beta \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

where g is some normalisation function that is constant with respect to β. Again, this density function must integrate to 1, so:

$$\begin{aligned} 1 &= \int_0^1 \mathbb{P}\left(\beta \,|\, \varphi, M_1\right)\, d\beta \\ &= \frac{1}{g(\varphi, x_\beta, y_\beta)} \int_\varphi^1 \frac{\beta^{x_\beta - 1}(1-\beta)^{y_\beta - 1}}{B(x_\beta, y_\beta)}\, d\beta \\ &= \frac{1}{g(\varphi, x_\beta, y_\beta)} \left(\int_0^1 \frac{\beta^{x_\beta - 1}(1-\beta)^{y_\beta - 1}}{B(x_\beta, y_\beta)}\, d\beta - \int_0^\varphi \frac{\beta^{x_\beta - 1}(1-\beta)^{y_\beta - 1}}{B(x_\beta, y_\beta)}\, d\beta \right) \\ &= \frac{1}{g(\varphi, x_\beta, y_\beta)} \left(1 - I(\varphi; x_\beta, y_\beta)\right) \end{aligned}$$

This gives us the relation $g(\varphi, x_\beta, y_\beta) = 1 - I(\varphi; x_\beta, y_\beta)$, which reduces to:

$$g(\varphi, x_\beta, y_\beta) = \begin{cases} (1-\varphi)^{y_\beta} & \text{if } x_\beta = 1 \\ 1 - \varphi^{x_\beta} & \text{if } y_\beta = 1 \end{cases}$$

We will use the former method ($\varphi \in [0, \beta]$ and $\beta \in [0, 1]$) to evaluate part of the integral term in Equation B.8 now:

$$
\begin{aligned}
I &= \int_0^1 \int_0^1 \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\mathbb{P}\left(\beta,\varphi \mid M_1\right)\, d\beta\, d\varphi \\
&= \int_0^1 \int_0^\beta \beta^{k_1}(1-\beta)^{k_2}\varphi^{l_1}(1-\varphi)^{l_2}\frac{1}{f(\beta,x_\varphi,y_\varphi)}\frac{\beta^{x_\beta-1}(1-\beta)^{y_\beta-1}}{B(x_\beta,y_\beta)}\frac{\varphi^{x_\varphi-1}(1-\varphi)^{y_\varphi-1}}{B(x_\varphi,y_\varphi)}\, d\varphi\, d\beta \\
&= \frac{1}{B(x_\beta,y_\beta)B(x_\varphi,y_\varphi)}\int_0^1 \frac{\beta^{k_1+x_\beta-1}(1-\beta)^{k_2+y_\beta-1}}{f(\beta,x_\varphi,y_\varphi)} \underbrace{\left(\int_0^\beta \varphi^{l_1+x_\varphi-1}(1-\varphi)^{l_2+y_\varphi-1}\, d\varphi\right)}_{I_\varphi} d\beta
\end{aligned}
\tag{B.14}
$$

I_φ can be evaluated almost identically as with the uniform prior in Equation B.12, since the hyperparameters are absorbed into the other exponents in the integrand:

$$
\begin{aligned}
I_\varphi &= \int_0^\beta \varphi^{l_1+x_\varphi-1}(1-\varphi)^{l_2+y_\varphi-1}\, d\varphi \\
&= \sum_{i=0}^{l_2+y_\varphi-1} \binom{l_2+y_\varphi-1}{i}(-1)^{l_2+y_\varphi-1-i}\left(\frac{\beta^{l+x_\varphi+y_\varphi-i-1}}{l+x_\varphi+y_\varphi-i-1}\right)
\end{aligned}
\tag{B.15}
$$

When we re-arrange Equation B.14, we will have to evaluate the following integral:

$$
\int_0^1 \frac{\beta^{k_1+l+x_\beta+x_\varphi+y_\varphi-i-2}(1-\beta)^{k_2+y_\beta-1}}{f(\beta,x_\varphi,y_\varphi)}\, d\beta
$$

This integral will be solved differently depending on the form of $f(\beta,x_\varphi,y_\varphi)$. For example, if $x_\beta=2$ and $y_\beta=1$, the integral may involve a logarithm term, such as:

$$
\int_0^1 \frac{\beta^3(1-\beta)^1}{1-(1-\beta)^2}\, d\beta = \frac{17}{6} - 4\log(2)
$$

However, if $x_\beta=3$ and $y_\beta=1$, the integral may result in an arc-tangent and a logarithm term, such as:

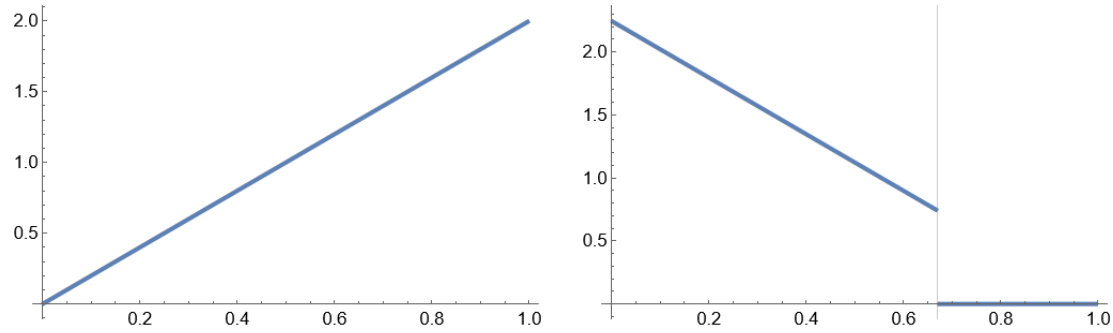
$$
\int_0^1 \frac{\beta^4(1-\beta)^1}{1-(1-\beta)^3}\, d\beta = -\frac{13}{3} + \frac{\sqrt{3}}{2}\pi + \frac{3}{2}\log(3)
$$

Since there is no general solution for this, we will consider a special case of the Beta prior which will help with the calculations.

B.3.3 Linear Prior

A plausible scenario for our causal variants is that they are more likely to have higher penetrance and lower phenocopy rates. Therefore, we will let $x_\beta=2, y_\beta=1$ and $x_\varphi=1, y_\varphi=2$, which give rise to the prior probability distributions in Figure B.2 below.

We will refer to this as the Linear Prior for β and φ, with the understanding that the condition $\varphi < \beta$ also holds.



(a) **(b)**

Figure B.2: Linear prior distribution for **(a)** β; **(b)** φ conditional on a given β.

In Equation B.15, we can substitute the known values for the hyperparameters:

$$\begin{aligned} I_\varphi &= \sum_{i=0}^{l_2+y_\varphi-1} \binom{l_2+y_\varphi-1}{i} (-1)^{l_2+y_\varphi-1-i} \left(\frac{\beta^{l+x_\varphi+y_\varphi-i-1}}{l+x_\varphi+y_\varphi-i-1} \right) \\ &= \sum_{i=0}^{l_2+1} \binom{l_2+1}{i} (-1)^{l_2+1-i} \left(\frac{\beta^{l+2-i}}{l+2-i} \right) \end{aligned} \tag{B.16}$$

Returning to the original I, we have

$$\begin{aligned} I &= \frac{1}{B(x_\beta, y_\beta)B(x_\varphi, y_\varphi)} \int_0^1 \frac{\beta^{k_1+x_\beta-1}(1-\beta)^{k_2+y_\beta-1}}{f(\beta, x_\varphi, y_\varphi)} \left[I_\varphi\right] \, d\beta \\ &= \frac{1}{B(2,1)B(1,2)} \int_0^1 \frac{\beta^{k_1+1}(1-\beta)^{k_2}}{f(\beta,2,1)} \left[\sum_{i=0}^{l_2+1} \binom{l_2+1}{i} (-1)^{l_2+1-i} \left(\frac{\beta^{l+2-i}}{l+2-i} \right) \right] \, d\beta \\ &= 4 \sum_{i=0}^{l_2+1} \binom{l_2+1}{i} \frac{(-1)^{l_2+1-i}}{l+2-i} \int_0^1 \frac{\beta^{k_1+l+3-i}(1-\beta)^{k_2}}{1-(1-\beta)^2} \, d\beta \\ &= 4 \sum_{i=0}^{l_2+1} \binom{l_2+1}{i} \frac{(-1)^{l_2+1-i}}{l+2-i} \underbrace{\int_0^1 \frac{\beta^{k_1+l+2-i}(1-\beta)^{k_2}}{2-\beta} \, d\beta}_{I_\beta} \end{aligned} \tag{B.17}$$

We will apply a change of variables of $t = 2 - \beta$ and the Binomial Theorem to evaluate I_β:

$$
\begin{aligned}
I_\beta &= \int_0^1 \frac{\beta^{k_1+l+2-i}(1-\beta)^{k_2}}{2-\beta}\, d\beta \\
&= \int_1^2 \frac{(2-t)^{k_1+l+2-i}(t-1)^{k_2}}{t}\, dt \\
&= \int_1^2 \left\{ \frac{1}{t} \left(\sum_{q=0}^{k_1+l+2-i} \binom{k_1+l+2-i}{q} 2^q (-1)^{k_1+l+2-i-q} t^{k_1+l+2-i-q} \right) \times \cdots \right. \\
&\quad \left. \cdots \times \left(\sum_{r=0}^{k_2} \binom{k_2}{r} t^r (-1)^{k_2-r} \right)\, dt \right\} \\
&= \sum_{q=0}^{k_1+l+2-i} \binom{k_1+l+2-i}{q} 2^q (-1)^{k_1+l+2-i-q} \sum_{r=0}^{k_2} \binom{k_2}{r} (-1)^{k_2-r} \int_1^2 t^{k_1+l+1-i-q+r}\, dt
\end{aligned}
\tag{B.18}
$$

If $q = k_1+l+2-i$ and $r = 0$, then the exponent in the integrand of I_β will be -1, giving rise to a logarithm. Otherwise, the exponent will be positive, and thus the integrand will be a polynomial. For brevity in the following summations, we will implicitly ignore the term that gives rise to a logarithm and add it separately. We use $\sum'$ to denote this.

$$
\begin{aligned}
I_\beta &= {\sum_{q=0}^{k_1+l+2-i}}' \binom{k_1+l+2-i}{q} 2^q (-1)^{k_1+l+2-i-q} {\sum_{r=0}^{k_2}}' \binom{k_2}{r} (-1)^{k_2-r} \frac{2^{k_1+l+2-i-q+r}-1}{k_1+l+2-i-q+r} \\
&\quad + 2^{k_1+l+2-i}(-1)^{k_2} \log(2)
\end{aligned}
\tag{B.19}
$$

B.4 Neutral Model

Here, we assume that the phenotypes and genotypes for the current variant are independent. However, the phenotypes are not independent of the choice of proband, seeing as the proband is a case. Additionally, since the variant is not causal, the general probability of having the phenotype does not depend on the genotype status, so $\beta = \varphi$, which should be equal to the population incidence rate α. We will use $\alpha \in [0, 1]$ instead of $\theta = (\beta, \varphi)$ for convenience.

We can follow the same idea to derive Equation B.4 above, giving us:

$$\begin{aligned}
\mathbb{P}(D \mid M_2) &= \int \mathbb{P}(D \mid \Theta, M_2)\, \mathbb{P}(\Theta \mid M_2)\ d\Theta \\
&= \sum_{p=1}^{k_1+l_1} \sum_{G_U} \int_0^1 \mathbb{P}(P_F \mid G_F, p, \alpha, M_2)\, \mathbb{P}(G_F \mid p)\, \mathbb{P}(p, \alpha \mid M_2)\ d\alpha \qquad \text{(B.20)} \\
\mathbb{P}(D \mid M_2) &= \sum_{p=1}^{k_1+l_1} \sum_{G_U} \int_0^1 \underbrace{\mathbb{P}(P_F \mid p, \alpha, M_2)}_{\text{phenotypes}} \underbrace{\mathbb{P}(G_F \mid p)}_{\text{inheritance}} \underbrace{\mathbb{P}(p, \alpha \mid M_2)}_{\text{parameters}}\ d\alpha
\end{aligned}$$

Note: if we started by splitting the $\mathbb{P}(G_O, P_F \mid \Theta, M_2)$ term due to independence, we would end up with the same final equation. As before, the selection of the proband is independent of the population incidence rate. Note that there are $k_1 + l_1$ affected individuals with genomic data for the variant, so this determines the prior for the proband selection, which gives us

$$\mathbb{P}(D \mid M_2) = \sum_{p=1}^{k_1+l_1} \sum_{G_U} \frac{\mathbb{P}(G_F \mid p)}{k_1 + l_1} \int_0^1 \mathbb{P}(P_F \mid p, \alpha, M_2)\, \mathbb{P}(\alpha \mid M_2)\ d\alpha \qquad \text{(B.21)}$$

As before, the unobserved genotypes will be the same regardless of the proband selection due to the rare variant assumption. Therefore, we can simplify the above:

$$\begin{aligned}
\mathbb{P}(D \mid M_2) &= \sum_{p=1}^{k_1+l_1} \sum_{G_U} \frac{\mathbb{P}(G_F \mid p)}{k_1 + l_1} \int_0^1 \mathbb{P}(P_F \mid p, \alpha, M_2)\, \mathbb{P}(\alpha \mid M_2)\ d\alpha \\
&= (k_1 + l_1) \sum_{G_U} \frac{\mathbb{P}(G_F)}{k_1 + l_1} \int_0^1 \mathbb{P}(P_F \mid p, \alpha, M_2)\, \mathbb{P}(\alpha \mid M_2)\ d\alpha \qquad \text{(B.22)} \\
&= \sum_{G_U} \mathbb{P}(G_F) \int_0^1 \mathbb{P}(P_F \mid p, \alpha, M_2)\, \mathbb{P}(\alpha \mid M_2)\ d\alpha
\end{aligned}$$

We assumed that the phenotype are determined by the population incidence rate, giving:

$$\mathbb{P}(P_F \mid p, \alpha, M_2) = \prod_{i=1}^{n} \mathbb{P}(P_i \mid \alpha, M_2) = \alpha^{k_1+l_1}(1-\alpha)^{k_2+l_2}$$

Substituting this into Equation B.22, we get:

$$\mathbb{P}(D \mid M_2) = \sum_{G_U} \mathbb{P}(G_F) \int_0^1 \alpha^{k_1+l_1}(1-\alpha)^{k_2+l_2} \mathbb{P}(\alpha \mid M_2)\ d\alpha \qquad \text{(B.23)}$$

B.4.1 Uniform Priors

We will let α have a uniform prior, giving us $\mathbb{P}(\alpha \mid M_2) = 1$ on $[0,1]$, and so:

$$\begin{aligned}
\mathbb{P}(D \mid M_2) &= \sum_{G_U} \mathbb{P}(G_F) \int_0^1 \alpha^{k_1+l_1}(1-\alpha)^{k_2+l_2}\, d\alpha \\
&= \sum_{G_U} \mathbb{P}(G_F)\, \frac{(k_1+l_1)!(k_2+l_2)!}{(n+1)!} \\
&= \sum_{G_U} \mathbb{P}(G_F)\, \frac{1}{(n+1)\binom{n}{k_1+l_1}}
\end{aligned} \tag{B.24}$$

B.5 Summary of Formulae

B.5.1 Causal Model

Uniform prior for β and φ:

$$\mathbb{P}(D \mid M_1) = 2 \sum_{G_U} \mathbb{P}(G_F) \sum_{i=0}^{l_2} \binom{l_2}{i} \frac{(-1)^{l_2-i}}{l-i+1} \frac{1}{(n-i+2)\binom{n-i+1}{k_2}} \tag{B.25}$$

Linear Prior, i.e. $\beta \sim \text{Beta}(2,1)$ and $\varphi \sim \text{Beta}(1,2)$ with $\varphi < \beta$:

$$\begin{aligned} \mathbb{P}(D \mid M_1) &= 4 \sum_{G_U} \mathbb{P}(G_F) \sum_{i=0}^{l_2+1} \binom{l_2+1}{i} \frac{(-1)^{l_2+1-i}}{l+2-i} (I) \\ I &= {\sum_{q=0}^{k_1+l+2-i}}' \binom{k_1+l+2-i}{q} 2^q (-1)^{k_1+l+2-i-q} {\sum_{r=0}^{k_2}}' \binom{k_2}{r} (-1)^{k_2-r} \frac{2^{k_1+l+2-i-q+r}-1}{k_1+l+2-i-q+r} + 2^{k_1+l+2-i} \log(2) \end{aligned} \tag{B.26}$$

where the $\sum'$ indicates that the case that $q = k_1 + l + 2 - i$ and $r = 0$ is ignored, since it is integrated separately.

B.5.2 Neutral Model

Uniform prior for α:

$$\mathbb{P}(D \mid M_2) = \sum_{G_U} \mathbb{P}(G_F) \frac{1}{(n+1)\binom{n}{k_1+l_1}} \tag{B.27}$$

B.6 Algorithms

Function SetGenerations(*vector*, *list*)
  minGen $\leftarrow \min\{$ vector[i] : vector[i] $\geqslant 1\}$
  minI $\leftarrow \min\{$ i : vector[i] $==$ minGen $\}$
  if *minGen* $== 0$ *or undefined* **then**
    return
  end
  if *minGen* $== 1$ **then**
    add vector to list
    return
  end
  tmp1 $\leftarrow$ vector
  tmp1[minI] $\leftarrow$ 0
  SetGenerations (tmp1, list)
  if *individual minI has one parent p with vector[p]* $== 1$ **then**
    tmp2 $\leftarrow$ vector
    tmp2[minI] $\leftarrow$ 1
    SetGenerations (tmp2, list)
  end
end

Algorithm 2: SetGenerations finds all permissible genotypes given a genotype vector by setting the generations of genotypes $\geqslant 2$.

```
Function FindGenerations(vector, list)
    find all probands for the variant
    if number of probands == 0 then
        return
    end
    find the founders of all carriers
    if number of founders == 0 then
        return
    end
    for founder in founders do
        mark the founder as a carrier (if their genotype is unknown)
        if another founder is a carrier then
            return
        end
        mark all other founders as non-carriers
        mark all descendants of pairs of non-carriers as non-carriers
        mark all descendants of the founder who are ancestors of a carrier as
         carriers
        if only one parent p has vector[p] ⩾ 1 then
            vector[i] ← vector[p] + 1
        end
        SetGenerations (vector, list)
    end
end
```

Algorithm 3: `FindGenerations` finds all permissible genotypes given a genotype vector by setting the generations of genotypes $\geqslant 2$.

Appendix C

Published Material

Here we show for completeness the material based on Chapter 3 published in *Briefings in Bioinformatics* (Ormond et al., 2021).

OXFORD

Briefings in Bioinformatics, 22(5), 2021, 1–7

https://doi.org/10.1093/bib/bbab069
Problem Solving Protocol

Converting single nucleotide variants between genome builds: from cautionary tale to solution

Cathal Ormond, Niamh M. Ryan, Aiden Corvin and Elizabeth A. Heron

Corresponding author. Elizabeth A. Heron, Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity Centre for Health Sciences, Trinity College Dublin, James' Street, Dublin 8, Ireland. Tel.: +353 1 896 4589; Fax: +353 1 896 3405; E-mail: eaheron@tcd.ie

Abstract

Next-generation sequencing studies are dependent on a high-quality reference genome for single nucleotide variant (SNV) calling. Although the two most recent builds of the human genome are widely used, position information is typically not directly comparable between them. Re-alignment gives the most accurate position information, but this procedure is often computationally expensive, and therefore, tools such as *liftOver* and *CrossMap* are used to convert data from one build to another. However, the positions of converted SNVs do not always match SNVs derived from aligned data, and in some instances, SNVs are known to change chromosome when converted. This is a significant problem when compiling sequencing resources or comparing results across studies. Here, we describe a novel algorithm to identify positions that are unstable when converting between human genome reference builds. These positions are detected independent of the conversion tools and are determined by the chain files, which provide a mapping of contiguous positions from one build to another. We also provide the list of unstable positions for converting between the two most commonly used builds GRCh37 and GRCh38. Pre-excluding SNVs at these positions, prior to conversion, results in SNVs that are stable to conversion. This simple procedure gives the same final list of stable SNVs as applying the algorithm and subsequently removing variants at unstable positions. This work highlights the care that must be taken when converting SNVs between genome builds and provides a simple method for ensuring higher confidence converted data. Unstable positions and algorithm code, available at https://github.com/cathaloruaidh/genomeBuildConversion

Key words: genome build conversion; liftOver; CrossMap; GRCh37; GRCh38

INTRODUCTION

The human reference genome is a fundamental and essential resource for next-generation sequencing studies, aiding in tasks such as genome assembly and variant calling. Without a reference, *de novo* assembly of each sequenced genome would need to take place, which is computationally intensive and in certain scenarios may result in a poor quality assembly [1]. The most frequently used human reference genomes are those constructed by the Genome Reference Consortium (GRC) [2], who to date have released 38 iterative reference builds. The two most recent builds of the genome are GRCh37 (released in 2009) and GRCh38 (released in 2013). The UCSC Genomics Institute have also released analogous versions of these builds, referred to as hg19 and hg38, respectively [3].

Both GRCh37 and GRCh38 were generated by sequencing DNA from a collection of human donors, predominantly using Sanger sequencing [4, 5]. DNA sequences were combined to form high-confidence contiguous segments known as contigs, which were joined to form a *de novo* assembly of the reference genome. One of the major updates in GRCh38 was the closing

Cathal Ormond is a PhD Student in the Neuropsychiatric Genetics Research Group in the Department of Psychiatry, Trinity College Dublin.
Niamh M. Ryan is a Postdoctoral Research Fellow in the Neuropsychiatric Genetics Research Group in the Department of Psychiatry, Trinity College Dublin.
Aiden Corvin is a Professor in the Neuropsychiatric Genetics Research Group and the Head of the Department of Psychiatry, Trinity College Dublin.
Elizabeth A. Heron is an Assistant Professor in Biostatistical Genomics in the Neuropsychiatric Genetics Research Group in the Department of Psychiatry, Trinity College Dublin.
Submitted: 2 November 2020; **Received (in revised form):** 27 January 2021

of numerous gaps where sequencing had previously not been possible [6]. GRCh38 also contains a much larger collection of unlocalized (known sequence and chromosome but position unknown) and unplaced (known sequence, but chromosome and position unknown) contigs, as well as alternate contigs (known alternate representations of specific regions of the genome to account for population differences) [6]. Users need to be aware that different builds result in different genome assemblies and subsequently can impact genomic analyses, including single nucleotide variant (SNV) analyses [7].

Despite the improvements that the latest build brings, updates to the base-pair coordinates typically mean that not all positions are comparable between builds. Researchers are sometimes hesitant to switch to GRCh38, as there exists a wealth of annotation information available for GRCh37 and many pipelines and tools are still based on the older, GRCh37, version [7]. A similar problem arises when comparing new sequencing data to data aligned to an older build, as both data sets must be aligned to the same build to be comparable. Although re-alignment of the original sequence data to the new build typically provides the most accurate base-pair position information, this can be quite computationally expensive [7]. Also, the raw sequence data required for alignment, if available, can be large, so long-term storage may not be feasible. An alternative approach to re-alignment is to convert between genome builds using tools such as *liftOver* (provided as part of the Genome Browser tool [3] hosted by the UCSC Genomics Institute), *CrossMap* [8] or *Remap* (hosted by the National Centre for Biotechnology Information [9]). This process is aided by a chain file, which provides a mapping of contiguous positions from one build to another. The ability to convert between builds using these tools has proved vital, allowing the integration of a wide range of SNV annotation databases and sequence data, regardless of how they were originally aligned, for example gnomAD [10], CADD [11] and dbNSFP [12].

For those who do choose to convert between GRCh37 and GRCh38, there are known problems with this conversion process, particularly for SNVs. In the online user guide for the UCSC Genome Browser, the authors note that 'occasionally, a chunk of sequence may be moved to an entirely different chromosome' (see Web resources in Methods section). This is echoed in Liu *et al.* [12], where the authors note that after converting the dbNSFP database to other builds using *liftOver*, 'there are a few SNVs whose coordinates in hg38 and hg19 ... have inconsistent chromosome numbers'. This phenomenon can prove problematic for downstream analyses if, for example, annotation information from converted data is not consistent with annotation information from re-aligned data. For example, suppose we wish to examine variants in protein coding regions of the genome, prioritized using the CADD deleteriousness score [11]. Consider the T > A substitution at position 15690247 on chromosome 22 of GRCh38 (chr22:c.15690247 T > A), contained in the first exon of the *POTEH* gene. CADD v1.6 gives the variant a C-score of 20.8, indicating that it is in the top percentile of all ranked deleterious variants. If we convert the position to GRCh37 (using either *liftOver* or *CrossMap*), this variant maps to position 19553586 on chromosome 14, where the reference allele is still T (chr14:c.19553586 T > A) but the variant is now in the first exon of *POTEG*. CADD v1.6 for GRCh37 gives this variant a C-score of 0.009, indicating that it is now in the bottom percentile of all ranked deleterious variants in the genome. This inconsistency shows how downstream results can be negatively impacted by instabilities in the conversion process.

Pan *et al.* (2019) [13] examined SNVs from data aligned under a range of bioinformatics pipelines to data converted between GRCh37 and GRCh38 using both *liftOver* and *CrossMap*. The authors noted that on average, 1% of SNVs did not convert from GRCh37 to GRCh38, and an average of 5% of SNVs did not convert from GRCh38 to GRCh37. Furthermore, on average, 1.5% of SNVs which were successfully converted were not found in the corresponding aligned data, a trend that was more pronounced when converting from GRCh38 to GRCh37. Such discordant sites were noted to be low-confidence calls, have lower average read depth and have a higher than average GC content. The authors urged caution when converting SNVs between builds.

Recently, Luu *et al.* (2020) [14] benchmarked six tools (including *liftOver*, *CrossMap* and *Remap*) for converting multi-base-pair regions derived from epigenetic data from GRCh37 to GRCh38. The authors found a high degree of correlation between the six tools but noted that gapped regions in both chain files can result in conversion failure, or even regions mapping to incorrect locations. A guideline to improve conversion is offered, which involves removing input data which overlap with the gapped regions, as well as removing input data which map to multiple regions or alternate contigs. However, if this strategy were applied to SNV data, some variants (such as those in un-gapped regions which also change chromosome under conversion) may not necessarily be removed.

Here, we present a novel algorithm to identify base-pair positions in the human genome which exhibit unstable behaviour when converting between genome reference builds. In addition, we are providing the list of these unstable positions for the two most recent builds (GRCh37 and GRCh38). This list can be used to pre-exclude SNVs prior to conversion to remove potentially problematic variants, resulting in stable SNVs and improving the quality of sequencing data post-conversion.

METHODS

Full-genome data

As genome build conversion tools are primarily based on base-pair position information only, it is possible to examine all base-pair positions in the genome. This allows the behaviour of all potential SNVs to be examined when converting between builds, rather than just a subset that might be found on an individual sample's genome. To this effect, a BED entry was created for each base-pair position in both the GRCh37 (GCA_000001405.1) and GRCh38 (GCA_000001405.15) reference genomes, which we refer to as the full-genome data. This includes positions that are not typically amenable to short-read whole genome sequencing (WGS), such as known gaps in the genome assembly. Positions on the unplaced, unlocalized and alternate contigs were not included in the input data, and so only the standard 23 pairs of chromosomes were considered. Each entry was given a label containing the original chromosome and start position for unique identification, and the input BED file was split by chromosome for parallelization [15]. This generated 3 095 677 412 positions for GRCh37 and 3 088 269 832 positions for GRCh38.

Algorithm to identify novel conversion-unstable positions

To identify base-pair positions that are unstable in the conversion process (defined below), each input file was converted from the source build to the target build and then back to the source

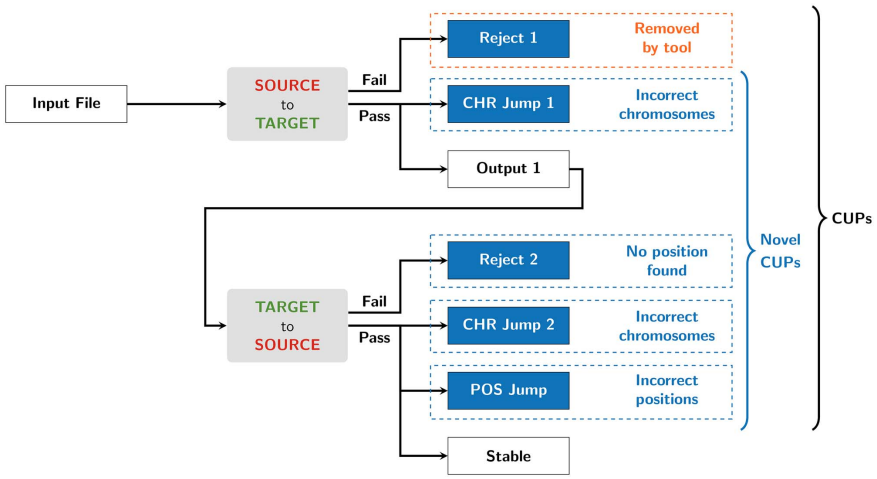


Figure 1. Flow chart of the algorithm to identify novel CUPs.

build again (Figure 1). Entries in the output files were extracted if they satisfied one of the following conditions:

- positions which failed on the first conversion (Reject_1),
- positions which mapped to a different chromosome on the first conversion (CHR_Jump_1),
- positions which failed on the second conversion (Reject_2),
- positions which did not map back to the original chromosome on the second conversion (CHR_Jump_2) and
- positions which did not map back to the original position on the second conversion (POS_Jump).

We refer to these collectively as conversion-unstable positions (CUPs), and all other positions are referred to as stable. Note that entries in the Reject_1 category are typically identified by the conversion tool, so the latter four entries are what we refer to from here on collectively as novel CUPs. Reject_1 and CHR_Jump_1 positions were removed prior to the second conversion (from the target build back to the source build). Despite not being included in the input data, entries that mapped to the unplaced, unlocalized, and alternate contigs were retained in the CHR_Jump_1 and CHR_Jump_2 categories to ensure that each base-pair position had an accurate category designation.

Both *liftOver* and *CrossMap* (v0.4.2) were used for the conversion. *Remap* was not considered as its input file is limited to 250 000 entries, which is much smaller than the lengths of the input chromosomes. Integral to this conversion process is a build-specific chain file, allowing for small-scale differences, e.g. discrepancies arising from fixing base-pair position errors between builds. Chain files mapping between GRCh37 and GRCh38 (one for each direction) were obtained from the *liftOver* website hosted by the UCSC Genomics Institute (see Web resources in Methods section) and the same chain files are used by both *liftOver* and *CrossMap*, allowing us to also check the robustness of CUP identification, as a consensus between tools would give higher confidence in the output. This algorithm was run twice, once for the GRCh37 build as the source and once for the GRCh38 build as the source.

Comparison with assembly annotation sets

To better understand the possible reasons for CUPs occurring, we also identified where these positions originated. Given the reconstruction of some contigs in the development of GRCh38 [6], one explanation for base-pair positions being rejected during a conversion is that the position is not in the target build. In the online support forum for the UCSC Genome Browser, it is noted that variants may change chromosomes between builds because they lie in repetitive regions or segmental duplications (see Web resources in Methods section). In an attempt to isolate the source of each CUP, the following assembly annotation sets were obtained from the UCSC Table browser [3] for both genome builds:

- Gaps in the build (gap): regions that are not present in the build, including telomeres, the short arms of specific chromosomes and gaps between known contigs. The centromeres are present in the GRCh37 gap set (as they did not form part of the assembly) but are not in the GRCh38 gap set. In the interests of a fairer comparison, the centromeres were removed from the GRCh37 gap set prior to comparison.
- Differences between contigs (hg38ContigDiff): regions that are different in the GRCh38 and GRCh37 builds due to updates in individual contigs.
- Segmental duplications (genomicSuperDups): regions longer than 1 kb that have a high degree of similarity with other regions.

Given the overlap between these sets, positions unique to each of the three sets, as well as positions which were present in more than one set (multiple) or no set (other), were considered (Supplementary Figure S1, Supplementary Data available online

at http://bib.oxfordjournals.org/). For the CUPs identified above, contiguous entries were collapsed into multi-base-pair regions using *bedtools* [16], to allow for quicker comparison with the assembly annotation sets. The proportion of overlap in CUP category A of assembly annotation set B is defined as $|A \cap B|/|A|$ and was computed using *bedtools*.

Whole genome sequence data

The well-characterized NA12877 and NA12878 samples for the Coriell Institute CEPH 1463 family were used to examine the behaviour of SNVs from WGS data when converting between builds. High-confidence, pedigree-validated variant calls for both samples were obtained from the Illumina Platinum Genomics project in VCF format on both GRCh37 and GRCh38 [17]. As we are only considering the behaviour of SNVs and aim to compare the WGS data with the full-genome data, only biallelic SNVs were extracted for both samples. A slightly modified version of the above algorithm was implemented using the *LiftoverVcf* module from *picard* rather than *liftOver*, as *liftOver* does not handle VCF file format. *CrossMap* can accommodate VCF file format. The *LiftoverVcf* module is based on *liftOver* but additionally checks the reference allele of each variant with the target reference genome, removing any sites where there is a mismatch. For VCF files, *CrossMap* updates the reference allele to that of the target build where there is a discrepancy and returns a failure if the alternate allele on the source build is the same as the updated reference allele on the target build. If a reference allele was updated to an ambiguous base (denoted by IUPAC codes), these were removed and considered a mismatch. For the WGS data, two additional output categories were included for variants which failed due to reference-allele mismatches on the first conversion (Mismatch_1) or on the second conversion (Mismatch_2). Position and genotype discordance rates between the converted and the aligned data were computed using *bedtools* and *GenotypeConcordance* (from *picard*), respectively. These were calculated as the proportion of variants in the converted data where the position/genotype did not match that of a variant in the aligned data. Genotype discordance rates are calculated as a proportion of variants whose position matched a variant in the aligned data.

Since individual base-pair positions are converted independently of one another, variants, which are present in any of the novel CUPs can also be excluded prior to conversion to ensure all variants, are stable and data are of high quality. These filtered data were compared with the output from the algorithm on the original data to confirm that both methods are equivalent. In addition to the VCF data files, BED files were generated using position information extracted from the VCF data. This allowed us to apply our original position-based algorithm (that used the *liftOver* and *CrossMap* tools) for comparison.

Web resources

- UCSC Genome Browser user guide on build conversion: https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Convert
- UCSC Genome Browser support forum on *liftOver* errors, with variants swapping chromosomes: https://groups.google.com/a/soe.ucsc.edu/g/genome/c/P3M1Q5baozM/m/SlyjdcoSBwAJ
- The online implementation of *liftOver*: https://genome.ucsc.edu/cgi-bin/hgLiftOver
- The online manual for *CrossMap*: https://crossmap.readthedocs.io/en/latest/
- Chain files for GRCh37 to GRCh38, provided by the UCSC Genomics Institute: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz
- Chain files for GRCh38 to GRCh37, provided by the UCSC Genomics Institute: http://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/hg38ToHg19.over.chain.gz
- Illumina Platinum Genomes project: https://www.illumina.com/platinumgenomes.html

RESULTS

Full-genome data

We examined every base-pair position in both builds of the human reference genome to identify positions that are unstable to conversion. Both *liftOver* and *CrossMap* gave identical output for the same input data (Table 1; Supplementary Table S1, Supplementary Data available online at http://bib.oxfordjournals.org/). On GRCh37, ~11.3 Mb of novel CUPs were identified (representing 0.37% of the build), and on GRCh38, ~20 Mb of novel CUPs were identified (0.65% of the build). For both builds, a successive application of the algorithm on the stable positions using either tool did not identify any additional base-pair positions for any of the CUP categories.

We compared each of the CUPs with three assembly annotation sets (gaps in the assembly, contig differences and segmental duplications). For both builds, the proportion of overlap for each CUP category across all the assembly annotation sets was at least 97.5% for all except the Reject_1 category on GRCh37, where the proportion was 69.2% (Figure 2). However, the centromeres that were removed from the gap set (which do not overlap with the other assembly annotation sets) account for an additional 29.4% of the Reject_1 category, giving a total overlap proportion explained of 98.6% (Supplementary Table S2, Supplementary Data available online at http://bib.oxfordjournals.org/). A large proportion (~70%) of Reject_1 categories on both builds are composed of the gap set, whereas the novel CUPs are heavily dominated by the contig differences and segmental duplications.

WGS data

As a proof of principle, we also examined the presence of CUPs in WGS data for two individuals from the CEPH 1463 family. Sample NA12877 had 3 518 008 SNVs on GRCh37 and 3 576 396 SNVs on GRCh38. Sample NA12878 had 3 523 638 SNVs on GRCh37 and 3 594 064 SNVs on GRCh38. Each of these represents ~0.1% of the full genome data for their respective build. For both samples, the CUPs identified from the VCF data were contained within the CUPs identified from the corresponding BED data, as expected. The only positions from the VCF data that were not contained in the BED data were the mismatch categories (due to reference allele mismatches). Furthermore, the CUPs identified from the BED positions from the WGS data were contained within the respective full-genome CUPs. *liftOver* and *CrossMap* broadly agreed on the CUPs derived from the VCF data, with differences arising purely due to how each tool treats the reference allele in the target build, including ambiguous bases (Mismatch_1, Mismatch_2).

The number of stable SNVs was the same for the filtered data (novel CUPs excluded) as for the unfiltered original WGS data when the algorithm was applied (Table 2; Supplementary

Table 1. Details of the stable positions and CUPs for the full-genome data for GRCh37 and GRCh38, including the number of base-pairs (bps) for each category and the proportion of the genome build covered (%). Novel CUPs are highlighted in grey

Category	GRCh37				GRCh38			
	GRCh37 to GRCh38 (bps)	% of Source	GRCh38 to GRCh37 (bps)	% of Source	GRCh38 to GRCh37 (bps)	% of Source	GRCh37 to GRCh38 (bps)	% of Source
All	3 095 677 412	100.000	2 859 470 792	92.370	3 088 269 832	100.000	2 862 067 878	92.675
Reject_1	234 712 067	7.582	–	–	218 510 733	7.076	–	–
CHR_Jump_1	1 494 553	0.048	–	–	7 691 221	0.249	–	–
Reject_2	–	–	100 180	0.003	–	–	73 770	0.002
CHR_Jump_2	–	–	799 922	0.026	–	–	292 083	0.009
POS_Jump	–	–	8,907,439	0.288	–	–	12 038 774	0.390
Stable	2 859 470 792	92.370	2 849 663 251	92.053	2 862 067 878	92.675	2 849 663 251	92.274
Novel CUPs	–	–	11 302 094	0.365	–	–	20 095 848	0.651

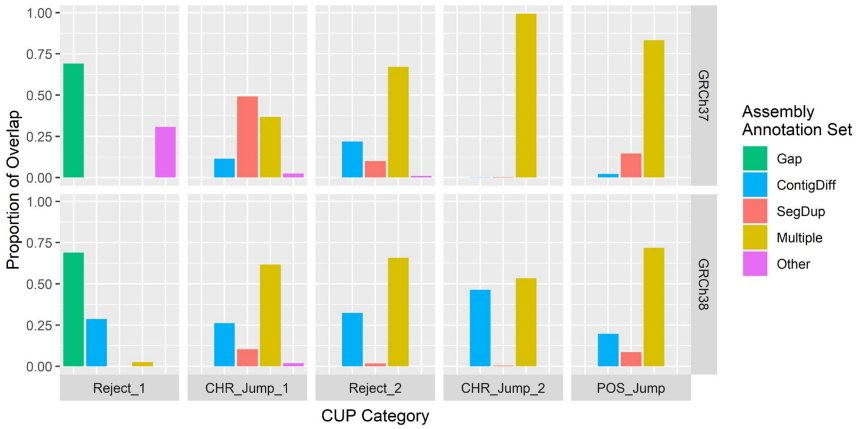


Figure 2. The proportion of CUPs that overlaps with the assembly annotation sets, for the GRCh37 (top) and GRCh38 (bottom) builds. Here, 'Multiple' represents positions present in one or more of the assembly annotation sets and 'Other' represents positions present in none of the assembly annotation sets (this includes the centromeres for GRCh37). Gap: gaps in the assembly; ContigDiff: differences in contigs between builds; SegDup: segmental duplications.

Tables S3 and S4, Supplementary Data available online at http://bib.oxfordjournals.org/). As expected, no additional variants in the CUP categories were identified on a successive application of the algorithm to either the original data or to the filtered data. The SNVs at novel CUPs represented ~0.13% of SNVs on either build. The position and genotype discordance metrics between the converted and aligned data are given in Supplementary Table S5, Supplementary Data available online at http://bib.oxfordjournals.org/.

DISCUSSION

Here, we have replicated the previously observed phenomenon whereby a small proportion of SNVs change chromosome when they are converted to another genome build [12]. Additionally, we have identified novel sites where base-pair position information does not behave as expected or where a one-to-one mapping between positions on both builds is not present. The novel CUPs represent 0.37% of the GRCh37 build and 0.65% of the GRCh38 build. This is important, as annotation data rely heavily on position information and downstream analysis can be negatively impacted by inaccuracies during the conversion process, as evidenced by our motivating example above.

The CUPs show a high degree of overlap with the three assembly annotation sets. For both builds, the Reject_1 positions (failure of the first conversion between builds) are dominated by the gap and contig differences sets. This is a highly plausible explanation for these base-pair positions as the conversion tools will fail when regions of the genome are not present, or have been updated, in the target build. For example, on GRCh37, the centromeres make up ~30% of the Reject_1 category (appearing in the 'Other' set in Figure 2), which is to be expected as the centromeres were broadly reconstructed during the assembly of GRCh38. The intersection between the contig differences and segmental duplications accounts for less than 6% of all the assembly annotation sets (Supplementary Figure S1, Supplementary Data available online at http://bib.oxfordjournals.org/);

Table 2. Counts of base-pair positions (bps) and proportions (%) of all SNVs present in WGS data for sample NA12878 broken down by genome build (GRCh37, GRCh38), conversion tool (*liftOver* or *CrossMap*) and whether the original or filtered data were considered

Source	Category	liftOver				CrossMap			
		Original		Filtered		Original		Filtered	
		Count (bps)	%	Count (bps)	%	Count (bps)	%	Count (bps)	%
GRCh37	All	3 523 638	100.000	3 518 229	100.000	3 523 638	100.000	3 518 229	100.000
	Reject_1	4947	0.140	4947	0.141	4947	0.140	4947	0.141
	Mismatch_1	20 533	0.583	19 976	0.568	20 510	0.582	19 959	0.567
	Mismatch_2	128	0.004	0	0.000	123	0.003	0	0.000
	Novel CUPs	4724	0.134	0	0.000	4735	0.134	0	0.000
	Stable	3 493 306	99.139	3 493 306	99.292	3 493 323	99.140	3 493 323	99.292
GRCh38	All	3 594 064	100.000	3 588 396	100.000	3 594 064	100.000	3 588 396	100.000
	Reject_1	25 852	0.719	25 852	0.720	25 852	0.719	25 852	0.720
	Mismatch_1	16 772	0.467	15 741	0.439	16 740	0.466	15 726	0.438
	Mismatch_2	85	0.002	0	0.000	81	0.002	0	0.000
	Novel CUPs	4552	0.127	0	0.000	4573	0.127	0	0.000
	Stable	3 546 803	98.685	3 546 803	98.841	3 546 818	98.685	3 546 818	98.841

All novel CUPs have been combined into one entry in the table (novel CUPs, highlighted in grey), see Supplementary Tables S3 and S4, Supplementary Data available online at http://bib.oxfordjournals.org/, for a full breakdown of the individual novel CUP categories.

however, the novel CUPs are largely composed of this intersection. If a region is contained in both a segmental duplication and a contig difference, this may indicate that the region is better placed in another part of the genome, which would explain the conversion instability. There remains a small proportion of each of the unstable regions that is not covered by at least one of the three assembly annotation sets (Figure 2, Supplementary Table S2, Supplementary Data available online at http://bib.oxfordjournals.org/).

In our study, both conversion tools identified the same unstable regions, which accords with the findings of Luu *et al.* 2020 [14], in their study of six conversion tools (including *liftOver* and *CrossMap*). Additionally, once the novel CUPs are removed from the full-genome data, a successive application of the algorithm on the stable positions does not identify any further novel CUPs, meaning that there is a one-to-one mapping for all stable base-pair positions between builds. Finally, the WGS data fully agree with the theoretical full-genome data. The comparison between filtered and original WGS data shows that pre-excluding variants at novel CUPs results in the same list of stable variants as applying the full-genome algorithm to the original WGS data. We provide a list of regions to exclude so that the user may remove any variants in novel CUPs prior to conversion.

Pan *et al.* 2019 [13] reported conversion failure rates for WGS data of on average 1% from GRCh37 to GRCh38 and 5% from GRCh38 to GRCh37, noting that the SNVs that failed tended to have much lower depth of coverage and may represent false-positive variant calls. Here, we observe much lower tool conversion failure rates of 0.14% from GRCh37 to GRCh38 and 0.72% from GRCh38 to GRCh37 for the WGS data. We note that the SNVs used in the analysis here were detected by multiple calling algorithms and have been pedigree-validated by confirming a Mendelian inheritance pattern in the samples' children, suggesting that this dataset is a particularly clean and accurate set of SNVs [17]. This may account for the decrease in conversion failure rates compared with the previous study. However, we note that the trend in performance is in the same direction and that converting from GRCh37 to GRCh38 is more accurate than GRCh38 to GRCh37. While Pan *et al.* (2019) show that read depth and variant quality may have an impact on discordance rates, the variants examined here did not have this information available, and thus, we were unfortunately not able to assess these aspects of the novel CUPs.

The combined position and genotype discordance rates were on average 3.07% when converting from GRCh38 to GRCh37 and 1.68% when converting from GRCh37 to GRCh38 (Supplementary Table S5, Supplementary Data available online at http://bib.oxfordjournals.org/). When variants in the novel CUPs were pre-excluded, these rates reduced to 2.97 and 1.61%, respectively. This is higher than the average discordance rate observed by Pan *et al.* (2019) of 1.5%; however, these rates are not directly comparable. Pan *et al.*'s average discordance rate is taken across all bioinformatics pipelines, across both builds and across both tools. Although Pan *et al.* (2019) do not provide the exact rates to compare, our discordance rates are broadly in line with those observed in their Figure 6A [13]. As with the conversion failure rates, both this study and Pan *et al.* (2019) found that converting from GRCh38 to GRCh37 yields higher discordance rates. We note that the genotype discordance rates are quite low at an average of 0.0011% for both builds (Supplementary Table S5, Supplementary Data available online at http://bib.oxfordjournals.org/). This indicates that when the position of a variant has been correctly converted, the genotype is also highly likely to be correct.

This study has some limitations. Firstly, since two independent conversion tools generate identical CUPs, we conclude that these regions are determined by the chain files as both tools utilize the same chain files. This is important to note, as alternative chain files exist for converting between GRCh37 and GRCh38, and therefore, the full algorithm will need to be applied if different chain files are used. A link to the source code used to generate the CUPs is provided for this purpose. However, it is worth noting that the chain files used here are the only ones supplied by the authors of *liftOver* and *CrossMap*. Secondly, while the full-genome data give insight into the behaviour of SNVs under build conversion, this does not account for regions spanning multiple base-pairs, as conversion tools are typically sensitive to this [14]. Finally, we have used aligned WGS data as a gold standard for evaluating the accuracy of converted

data, but it is important to note that although the SNVs are pedigree-validated, they may still contain false positive variant calls.

Here, we have clearly highlighted the care that must be taken when converting between genome builds to ensure high-quality data. Although we have shown results for the two most recent builds of the human genome, the same argument can be applied when converting between any other build pair, or indeed for non-human genomes. Unless the user is familiar with the instabilities we have described, we recommend following the simple strategy devised here of removing variants at novel CUPs to ensure high-confidence data when converting SNVs between the two most recent builds of the human genome.

Key Points

- When using tools such as *liftOver* and *CrossMap* to convert SNVs between the two most recent builds of the human reference genome (GRCh37 and GRCh38), some base-pair positions map to different chromosomes.
- Additionally, when converting from target build back to the source build, there are base-pair positions which do not map back to the same original position. This means that for these base-pair positions, a one-to-one correspondence between builds does not exist.
- These CUPs are predominantly comprised of regions with known annotation: gaps in the assembly, contig differences between builds and segmental duplications.
- The CUPs identified for the full-genome data were the same regardless of the conversion tool used, indicating that they are determined by the chain files.
- Pre-excluding SNVs at these CUPs prior to conversion results in SNVs that are stable to conversion.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

The authors acknowledge the support of the Trinity Centre for High Performance Computing (ResearchIT). The authors also express their gratitude to the reviewers for the helpful comments and suggestions provided.

Funding

National Institute of Health (5U01MH 109499-04); Science Foundation Ireland (16/SPP/3324).

Data availability

The data underlying this article (including BED files for the CUPs as well as the source code for the algorithm to generate them) are available at https://github.com/cathaloruaidh/genomeBuildConversion/.

Conflict of interest

The authors have no conflict of interests to declare.

References

1. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;**13**(1):36–46.
2. Church DM, Schneider VA, Graves T, *et al.* Modernizing reference genome assemblies. *PLoS Biol* 2011;**9**(7):e1001091.
3. Haeussler M, Zweig AS, Tyner C, *et al.* The UCSC genome browser database: 2019 update. *Nucleic Acids Res* 2019;**47**(D1):D853–d858.
4. Genome Reference Consortium. Announcing GRCh38. 2013. http://genomeref.blogspot.com/2013/12/announcing-grch38.html (08 October 2020, date last accessed).
5. E pluribus unum. *Nature Methods* 2010;**7**(5):331.
6. Schneider VA, Graves-Lindsay T, Howe K, *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;**27**(5):849–64.
7. Guo Y, Dai Y, Yu H, *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 2017;**109**(2):83–90.
8. Zhao H, Sun Z, Wang J, *et al.* Cross Map: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2013;**30**(7):1006–7.
9. NCBI. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2018;**46**(D1): D8–D13.
10. Karczewski KJ, Francioli LC, Tiao G, *et al.* Variation across 141, 456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes, bioRxiv 2019; 531210.
11. Rentzsch P, Witten D, Cooper GM, *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**(D1):D886–94.
12. Liu X, Wu C, Li C, *et al.* Db NSFP v3.0: a one-stop database of functional predictions and annotations for human non-synonymous and splice-site SNVs. *Hum Mutat* 2016;**37**(3): 235–41.
13. Pan B, Kusko R, Xiao W, *et al.* Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* 2019;**20**(Suppl 2):101.
14. Luu P-L, Ong P-T, Dinh T-P, *et al.* Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genomics and Bioinformatics* 2020; **2**(3).
15. Tange O. GNU parallel-the command-line power tool. ;Login: The USENIX Magazine. 2011;**36**:42–7.
16. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**(6): 841–2.
17. Eberle MA, Fritzilas E, Krusche P, *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* 2017;**27**(1):157–64.

Bibliography

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011, Jun). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, *21*(6), 974–984.
50

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013, Jan). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*, *Chapter 7*, Unit7.20.
27

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., ... Zbicz, K. (2018, 01). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, *46*(D1), D8-D13.
30

Alkan, C., Coe, B. P., & Eichler, E. E. (2011, May). Genome structural variation discovery and genotyping. *Nat Rev Genet*, *12*(5), 363–376.
47

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990, Oct). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403–410.
32

Ambalavanan, A., Girard, S. L., Ahn, K., Zhou, S., Dionne-Laporte, A., Spiegelman, D., ... Rouleau, G. A. (2016, 06). De novo variants in sporadic cases of childhood onset schizophrenia. *Eur J Hum Genet*, *24*(6), 944–948.
155

Anney, R. J. L., Ripke, S., Anttila, V., Grove, J., Holmans, P., Huang, H., ... Daly, M. J. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism*, *8*, 21.
106, 164

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., ... Murray, R. (2018, 06). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395).

14

Association, A. (2013). *Diagnostic and statistical manual of mental disorders (dsm-5®)*. American Psychiatric Publishing.
14

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., ... Abecasis, G. R. (2015, Oct). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.
26

Bandelt, H. J., Kloss-Brandstätter, A., Richards, M. B., Yao, Y. G., & Logan, I. (2014, Feb). The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *J Hum Genet*, *59*(2), 66–77.
33

Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., & Layer, R. M. (2021, 05). Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol*, *22*(1), 161.
49

Bloom, R. J., Kähler, A. K., Collins, A. L., Chen, G., Cannon, T. D., Hultman, C., & Sullivan, P. F. (2013, May). Comprehensive analysis of copy number variation in monozygotic twins discordant for bipolar disorder or schizophrenia. *Schizophr Res*, *146*(1-3), 289–290.
63

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*, *12*(6), e0177678.
140

Boulard, M., Edwards, J. R., & Bestor, T. H. (2015, May). FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat Genet*, *47*(5), 479–485.
154

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., ... Snyder, M. (2012, Sep). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, *22*(9), 1790–1797.
74

Bryois, J., Garrett, M. E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G. D., ... Crawford, G. E. (2018, 08). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat Commun*, *9*(1), 3121.
75

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, *8*(12), e1002822.
10

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet*, *11*, 424.
10

Castellani, C. A., Awamleh, Z., Melka, M. G., O'Reilly, R. L., & Singh, S. M. (2014, Apr). Copy number variation distribution in six monozygotic twin pairs discordant for schizophrenia. *Twin Res Hum Genet*, *17*(2), 108–120.
63

Castellani, C. A., Melka, M. G., Gui, J. L., Gallo, A. J., O'Reilly, R. L., & Singh, S. M. (2017, Nov). Post-zygotic genomic changes in glutamate and dopamine pathway genes may explain discordance of monozygotic twins for schizophrenia. *Clin Transl Med*, *6*(1), 43.
63, 70

Castellani, C. A., Melka, M. G., Gui, J. L., O'Reilly, R. L., & Singh, S. M. (2015, Dec). Integration of DNA sequence and DNA methylation changes in monozygotic twin pairs discordant for schizophrenia. *Schizophr Res*, *169*(1-3), 433–440.
163

Cederlöf, M., Bergen, S. E., Långström, N., Larsson, H., Boman, M., Craddock, N., ... Lichtenstein, P. (2015, May). The association between Darier disease, bipolar disorder, and schizophrenia revisited: a population-based family study. *Bipolar Disord*, *17*(3), 340–344.
106

Chang, X., Lima, L. A., Liu, Y., Li, J., Li, Q., Sleiman, P. M. A., & Hakonarson, H. (2018). Common and Rare Genetic Risk Factors Converge in Protein Interaction Networks Underlying Schizophrenia. *Front Genet*, *9*, 434.
106

Chapman, B., Kirchner, R., Pantano, L., Naumenko, S., Smet, M. D., Beltrame, L., ... Turner, S. (2021, January). *bcbio/bcbio-nextgen: v1.2.5.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.4429770` doi: 10.5281/zenodo.4429770
50

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., ... Saunders, C. T. (2016, 04). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222.
50, 57

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., ... Hubbard, T. (2011, Jul). Modernizing reference genome assemblies. *PLoS Biol*, *9*(7), e1001091.
12, 29

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet*, *3*, 35.
28

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeno-Tárraga, A., Cleland, I., ... Zalunin, V. (2013, Jan). Facing growth in the European Nucleotide Archive. *Nucleic Acids Res*, *41*(Database issue), D30–35.
20

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010, Apr). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, *38*(6), 1767–1771.
19

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W., Vulto-van Silfhout, A. T., Bosco, P., ... Eichler, E. E. (2014, 10). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet*, *46*(10), 1063–1071.
14

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., ... Xavier, R. J. (2020, 05). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451.
26

Consortium, G. R. (2013). *Announcing GRCh38.* `http://genomeref.blogspot.com/2013/12/announcing-grch38.html`. (Accessed: 2022-02-07)
12

Coughlin, C. R., Scharer, G. H., & Shaikh, T. H. (2012). Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Med*, *4*(10), 80.
13

Crossley, B. M., Bai, J., Glaser, A., Maes, R., Porter, E., Killian, M. L., ... Toohey-Kurth, K. (2020, Nov). Guidelines for Sanger sequencing and molecular assay monitoring. *J Vet Diagn Invest*, *32*(6), 767–775.
11

Cubuk, C., Garrett, A., Choi, S., King, L., Loveday, C., Torr, B., ... CanVIG-Uk (2021, Nov). Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genet Med*, *23*(11), 2096–2104.
138

Dahoun, S., Gagos, S., Gagnebin, M., Gehrig, C., Burgi, C., Simon, F., ... Blouin, J. L. (2008, Aug). Monozygotic twins discordant for trisomy 21 and maternal 21q inheritance: a complex series of events. *Am. J. Med. Genet. A*, *146A*(16), 2086–2093.
62

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Wang, J. (2011, Aug). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.
20, 25

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... Li, H. (2021, 02). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2).
24

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010, Dec). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, *6*(12), e1001025.
138

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., ... Rademakers, R. (2011, Oct). Expanded GGGGCC hexanucleotide

repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, *72*(2), 245–256.
85

de la Torre-Ubieta, L., Stein, J. L., Won, H., Opland, C. K., Liang, D., Lu, D., & Geschwind, D. H. (2018, 01). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell*, *172*(1-2), 289–304.
75

DeLisi, L. E. (2016, May). A Case for Returning to Multiplex Families for Further Understanding the Heritability of Schizophrenia: A Psychiatrist's Perspective. *Mol Neuropsychiatry*, *2*(1), 15–19.
10

Devenney, E. M., Landin-Romero, R., Irish, M., Hornberger, M., Mioshi, E., Halliday, G. M., ... Hodges, J. R. (2017). expansion. *Neuroimage Clin*, *13*, 439–445.
85

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., ... Eberle, M. A. (2017, 11). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, *27*(11), 1895–1903.
86

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Birney, E. (2012, Sep). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.
75

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., ... Bentley, D. R. (2017, 01). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*, *27*(1), 157–164.
41, 42, 50, 115

Endicott, J., & Spitzer, R. L. (1978, Jul). A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch Gen Psychiatry*, *35*(7), 837–844.
90

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, *29*(1), 51-76.
9

Faust, G. G., & Hall, I. M. (2014, Sep). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, *30*(17), 2503–2505.
22

Feng, B. J. (2017, 03). PERCH: A Unified Framework for Disease Gene Prioritization. *Hum. Mutat.*, *38*(3), 243–251.
14, 17, 113, 128

Feuk, L., Carson, A. R., & Scherer, S. W. (2006, Feb). Structural variation in the human genome. *Nat. Rev. Genet.*, *7*(2), 85–97.
12

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., ... Carter, N. P. (2009, Apr). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.*, *84*(4), 524–533.
26, 80

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl*, *1*.
70

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., ... Flicek, P. (2019, 01). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*, *47*(D1), D766-D773.
75

Freed, D., Aldana, R., Weber, J. A., & Edwards, J. S. (2017). The sentieon genomics tools - a fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2017/05/12/115717` doi: 10.1101/115717
94

Friedrich, S., Barbulescu, R., Helleday, T., & Sonnhammer, E. L. L. (2020, 06). MetaCNV - a consensus approach to infer accurate copy numbers from low coverage data. *BMC Med Genomics*, *13*(1), 76.
48

Fullard, J. F., Charney, A. W., Voloudakis, G., Uzilov, A. V., Haroutunian, V., & Roussos, P. (2019, 01). Assessment of somatic single-nucleotide variation in brain tissue of cases with schizophrenia. *Transl Psychiatry*, *9*(1), 21.
75, 163

Fullard, J. F., Giambartolomei, C., Hauberg, M. E., Xu, K., Voloudakis, G., Shao, Z., ... Roussos, P. (2017, 05). Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. *Hum Mol Genet*, *26*(10), 1942–1951.
75

Gelman, A. (2016). *Data-dependent prior as an approximation to hierarchical model.* `https://statmodeling.stat.columbia.edu/2016/03/25/28321/`. (Accessed: 2021-11-20)
182

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Ratan, Florida: Chapman and Hall/CRC.
129

Genome Reference Consortium. (2010). E pluribus unum. *Nat Methods*, *7*(5), 331.
12

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., ... McCarroll, S. A. (2016, 11). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci*, *19*(11), 1433–1441.
71

Gibson, G. (2012, Jan). Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, *13*(2), 135–145.
9, 89

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annu Rev Genet*, *45*, 203–226.
47

Giusti-Rodríguez, P., & Sullivan, P. F. (2013, Nov). The genomics of schizophrenia: update and implications. *J Clin Invest*, *123*(11), 4557–4563.
165

Glahn, D. C., Nimgaonkar, V. L., Raventos, H., Contreras, J., McIntosh, A. M., Thomson, P. A., ... Blangero, J. (2019, 04). Rediscovering the value of families for psychiatric genetics research. *Mol. Psychiatry*, *24*(4), 523–535.
10, 90

Gong, T., Hayes, V. M., & Chan, E. K. F. (2020). Shiny-SoSV: A web-based performance calculator for somatic structural variant detection. *PLoS One*, *15*(8), e0238108.

50

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016, 05). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333–351.
11

Green, E. K., Rees, E., Walters, J. T., Smith, K. G., Forty, L., Grozeva, D., ... Kirov, G. (2016, Jan). Copy number variation in bipolar disorder. *Mol Psychiatry*, *21*(1), 89–93.
14

Griffiths, A., Wessler, S., Lewontin, R., & Carroll, S. (2008). *Introduction to genetic analysis* (No. v. 10). W. H. Freeman.
62

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., ... Borgwardt, K. M. (2015, May). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat*, *36*(5), 513–523.
138

Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2021, 08). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet*, *58*(8), 547–555.
138

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017, 03). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90.
12, 29, 30

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., ... Kent, W. J. (2019, 01). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*, *47*(D1), D853-D858.
12, 17, 30, 37, 39, 60, 102

Hakonen, A. H., Heiskanen, S., Juvonen, V., Lappalainen, I., Luoma, P. T., Rantamaki, M., ... Suomalainen, A. (2005, Sep). Mitochondrial DNA polymerase W748S mutation: a common cause of autosomal recessive ataxia with ancient European origin. *Am J Hum Genet*, *77*(3), 430–441.
71

Hanafusa, K., & Hayashi, N. (2019, 08). The Flot2 component of the lipid raft changes localization during neural differentiation of P19C6 cells. *BMC Mol Cell Biol*, *20*(1), 38.
78

Haraksingh, R. R., Abyzov, A., & Urban, A. E. (2017, 04). Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*, *18*(1), 321.
13, 50

Heather, J. M., & Chain, B. (2016, Jan). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.
11

Helderman-van den Enden, A. T., Maaswinkel-Mooij, P. D., Hoogendoorn, E., Willemsen, R., Maat-Kievit, J. A., Losekoot, M., & Oostra, B. A. (1999, Mar). Monozygotic twin brothers with the fragile X syndrome: different CGG repeats and different mental capacities. *J. Med. Genet.*, *36*(3), 253–257.
62

Homann, O. R., Misura, K., Lamas, E., Sandrock, R. W., Nelson, P., McDonough, S. I., & DeLisi, L. E. (2016, 12). Whole-genome sequencing in multiplex families with psychoses reveals mutations in the SHANK2 and SMARCA1 genes segregating with illness. *Mol. Psychiatry*, *21*(12), 1690–1695.
111

Howrigan, D. P., Rose, S. A., Samocha, K. E., Fromer, M., Cerrato, F., Chen, W. J., ... Neale, B. M. (2020, 02). Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat Neurosci*, *23*(2), 185–193.
71

Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., & Yandell, M. (2013, Sep). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.*, *37*(6), 622–634.
112

Hu, H., Roach, J., Coon, H., & et al. (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature Biotechnology*, *32*(7), 663-9. doi: 10.1038/nbt.2895

14, 17, 112

Huang, Y., Zhao, Y., Ren, Y., Yi, Y., Li, X., Gao, Z., ... Wu, L. (2019, Mar). Identifying Genomic Variations in Monozygotic Twins Discordant for Autism Spectrum Disorder Using Whole-Genome Sequencing. *Mol Ther Nucleic Acids*, *14*, 204–211.
63

Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., ... Iwata, N. (2019, 06). Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. *Schizophr Bull*, *45*(4), 824–834.
155

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... Sieh, W. (2016, Oct). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*, *99*(4), 877–885.
138

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016, 11). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*, *17*(1), 239.
163

Johansson, V., Hultman, C. M., Kizling, I., Martinsson, L., Borg, J., Hedman, A., & Cannon, T. D. (2019, 02). The schizophrenia and bipolar twin study in Sweden (STAR). *Schizophr Res*, *204*, 183–192.
63

Jonsson, H., Magnusdottir, E., Eggertsson, H. P., Stefansson, O. A., Arnadottir, G. A., Eiriksson, O., ... Stefansson, K. (2021, 01). Differences between germline genomes of monozygotic twins. *Nat Genet*, *53*(1), 27–34.
62, 63, 164

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2019/01/30/531210` doi: 10.1101/531210
30, 80

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... Xavier, R. J. (2020, 05). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443.
26, 28

Kasahara, T., Ishiwata, M., Kakiuchi, C., Fuke, S., Iwata, N., Ozaki, N., ... Kato, T. (2017, Aug). Enrichment of deleterious variants of mitochondrial DNA polymerase gene (POLG1) in bipolar disorder. *Psychiatry Clin Neurosci*, *71*(8), 518–529.
71

Kasahara, T., Takata, A., Kato, T. M., Kubota-Sakashita, M., Sawada, T., Kakita, A., ... Kato, T. (2016, Jan). Depression-like episodes in mice harboring mtDNA deletions in paraventricular thalamus. *Mol Psychiatry*, *21*(1), 39–48.
71

Kato, T., Iwamoto, K., Kakiuchi, C., Kuratomi, G., & Okazaki, Y. (2005, Jul). Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders. *Mol. Psychiatry*, *10*(7), 622–630.
62

Keen, J. C., & Moore, H. M. (2015, Feb). The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *J Pers Med*, *5*(1), 22–29.
71, 106

Kendall, K. M., Rees, E., Bracher-Smith, M., Legge, S., Riglin, L., Zammit, S., ... Walters, J. T. R. (2019, 08). Association of Rare Copy Number Variants With Risk of Depression. *JAMA Psychiatry*, *76*(8), 818–825.
14

Kent, W. J. (2002, Apr). BLAT–the BLAST-like alignment tool. *Genome Res*, *12*(4), 656–664.
32

Khan, F. F., Melton, P. E., McCarthy, N. S., Morar, B., Blangero, J., Moses, E. K., & Jablensky, A. (2018, 07). Whole genome sequencing of 91 multiplex schizophrenia families reveals increased burden of rare, exonic copy number variation in schizophrenia probands and genetic heterogeneity. *Schizophr Res*, *197*, 337–345.
48, 55, 56, 162

Kim, S. Y., Kim, J. H., & Chung, Y. J. (2012, Sep). Effect of Combining Multiple CNV Defining Algorithms on the Reliability of CNV Calls from SNP Genotyping Data. *Genomics Inform*, *10*(3), 194–199.
13

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., & Hochreiter, S. (2012, May). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res*, *40*(9), e69.
56

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., ... Murray, J. C. (2002, Oct). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat. Genet.*, *32*(2), 285–289.
62

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019, Jun). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, *20*(1), 117.
13, 48, 50, 162

Kutner, M. (2005). *Applied linear statistical models*. McGraw-Hill Irwin.
138

Kuzniar, A., Maassen, J., Verhoeven, S., Santuari, L., Shneider, C., Kloosterman, W. P., & de Ridder, J. (2020). sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data. *PeerJ*, *8*, e8214.
50

LaFramboise, T. (2009, Jul). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res*, *37*(13), 4181–4193.
10

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... Szustakowki, J. (2001, Feb). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.
20

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2018, 01). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, *46*(D1), D1062-D1067.

137

Laplana, M., Royo, J. L., Aluja, A., López, R., Heine-Sunyer, D., & Fibla, J. (2014). Absence of substantial copy number differences in a pair of monozygotic twins discordant for features of autism spectrum disorder. *Case Rep Genet*, *2014*, 516529.
63

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019, 03). Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, *177*(1), 70–84.
70

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014, Jun). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, *15*(6), R84.
50, 57

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Williams, A. L. (2016, 08). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.
28

Li, H. (2011, Mar). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, *27*(5), 718–719.
19

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. 22

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009, Aug). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.
20, 50, 95

Li, M., Shen, L., Chen, L., Huai, C., Huang, H., Wu, X., ... Qin, S. (2020, 01). Novel genetic susceptibility loci identified by family based whole exome sequencing in Han Chinese schizophrenia patients. *Transl Psychiatry*, *10*(1), 5.
155

Li, Q., Wang, Z., Zong, L., Ye, L., Ye, J., Ou, H., ... Zhao, C. (2021, 11). Allele-specific DNA methylation maps in monozygotic twins discordant for psychiatric disorders reveal that disease-associated switching at the EIPR1 regulatory loci modulates neural function. *Mol Psychiatry*, *26*(11), 6630–6642.

163

Lichtenstein, P., Björk, C., Hultman, C. M., Scolnick, E., Sklar, P., & Sullivan, P. F. (2006, Oct). Recurrence risks for schizophrenia in a Swedish national cohort. *Psychol Med*, *36*(10), 1417–1425.
63

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009, Jan). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, *373*(9659), 234–239.
14

Liu, X., Li, C., Mou, C., Dong, Y., & Tu, Y. (2020, Dec). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*, *12*(1), 103.
27

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016, Mar). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat*, *37*(3), 235–241.
12, 30, 46

Loh, P. R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., ... Price, A. L. (2018, 07). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, *559*(7714), 350–355.
83

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... Moore, H. F. (2013, Jun). The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, *45*(6), 580–585.
155

Luu, P. L., Ong, P. T., Dinh, T. P., & Clark, S. J. (2020, Sep). Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genom Bioinform*, *2*(3), lqaa054.
31, 162

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., & Scherer, S. W. (2014, Jan). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, *42*(Database issue), D986–992.
26, 57, 80

Malhotra, D., & Sebat, J. (2012, Mar). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, *148*(6), 1223–1241.
14

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010, Nov). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873.
25, 66, 114

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... Visscher, P. M. (2009, Oct). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.
10

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., ... Sebat, J. (2017, 01). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.*, *49*(1), 27–35.
14, 15, 78

McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019, 11). Next-Generation Sequencing Technologies. *Cold Spring Harb Perspect Med*, *9*(11).
11

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... Cunningham, F. (2016, 06). The Ensembl Variant Effect Predictor. *Genome Biol*, *17*(1), 122.
28

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., ... Schafer, A. J. (2011, Feb). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65.
58

Mitchell, K. J. (2012, Jan). What is complex about complex disorders? *Genome Biol*, *13*(1), 237.
9

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., ... Gymrek, M. (2021, 01). Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, *589*(7841), 246–250.
85

Mohammadi, L., Vreeswijk, M. P., Oldenburg, R., van den Ouweland, A., Oosterwijk, J. C., van der Hout, A. H., ... van Houwelingen, H. C. (2009, Jun). A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. *BMC Cancer*, *9*, 211.
128, 130, 133, 166, 179, 181

Mojarad, B. A., Engchuan, W., Trost, B., Backstrom, I., Yin, Y., Thiruvahindrapuram, B., ... Yuen, R. K. C. (2022, May). Genome-wide tandem repeat expansions contribute to schizophrenia risk. *Mol Psychiatry*.
85

Morris, N., Elston, R. C., Barnholtz-Sloan, J. S., & Sun, X. (2015). Novel approaches to the analysis of family data in genetic epidemiology. *Front Genet*, *6*, 27.
11

Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R. I., Qiao, Z., ... Andreassen, O. A. (2021, 06). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet*, *53*(6), 817–829.
155

Ng, P. C., & Henikoff, S. (2003, 07). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812-3814.
26

Niroula, A., & Vihinen, M. (2019, 02). How good are pathogenicity predictors in detecting benign variants? *PLoS Comput Biol*, *15*(2), e1006481.
138

Nishioka, M., Bundo, M., Ueda, J., Yoshikawa, A., Nishimura, F., Sasaki, T., ... Iwamoto, K. (2018, Apr). Identification of somatic mutations in monozygotic twins discordant for psychiatric disorders. *NPJ Schizophr*, *4*(1), 7.
163

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., ... Phillippy, A. M. (2022, 04). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.
162

Okayama, T., Hashiguchi, Y., Kikuyama, H., Yoneda, H., & Kanazawa, T. (2018, 10). Next-generation sequencing analysis of multiplex families with atypical psychosis. *Transl Psychiatry*, *8*(1), 221.

111

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... Pruitt, K. D. (2016, Jan). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, *44*(D1), D733–745.
70, 102, 106, 116, 155

Ormond, C., Ryan, N. M., Corvin, A., & Heron, E. A. (2021, 09). Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief Bioinform*, *22*(5).
29, 195

Orr, H. T., & Zoghbi, H. Y. (2007). Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, *30*, 575–621.
85

Ott, J., Wang, J., & Leal, S. M. (2015, May). Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet*, *16*(5), 275–284.
13, 111

Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., ... Hong, H. (2019, Mar). Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*, *20*(Suppl 2), 101.
31, 42, 44

Pearson, K. (1901). III. Mathematical contributions to the theory of evolution.;VIII. On the inheritance of characters not capable of exact quantitative measurement.;Part I. Introductory. Part II. On the inheritance of coat-colour in horses. Part III. On the inheritance of eye-colour in man. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*(262-273), 79-150. doi: 10.1098/rsta.1900.0024
9

Pedersen, B. S., & Quinlan, A. R. (2017, Mar). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am. J. Hum. Genet.*, *100*(3), 406–413.
25, 64, 95

Petersen, G. M., Parmigiani, G., & Thomas, D. (1998, Jun). Missense mutations in disease genes: a Bayesian approach to evaluate causality. *Am J Hum Genet*, *62*(6), 1516–1524.

128, 129, 131, 179

Pirooznia, M., Goes, F. S., & Zandi, P. P. (2015). Whole-genome CNV analysis: advances in computational approaches. *Front Genet*, *6*, 138.
48

Polymeropoulos, M. H., Xiao, H., Torrey, E. F., DeLisi, L. E., Crow, T., & Merril, C. R. (1993, Jul). Search for a genetic event in monozygotic twins discordant for schizophrenia. *Psychiatry Res*, *48*(1), 27–36.
62

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., ... Sklar, P. (2014, Feb). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, *506*(7487), 185–190.
89

Purcell, S. M., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007, Sep). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559–575.
65

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., ... Sklar, P. (2009, Aug). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752.
14

Quinlan, A. R., & Hall, I. M. (2010, Mar). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.
37, 51

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`
68, 96

Rañola, J. M. O., Liu, Q., Rosenthal, E. A., & Shirts, B. H. (2018, 04). A comparison of cosegregation analysis methods for the clinical setting. *Fam Cancer*, *17*(2), 295–302.
128, 129

Rao, J., Peng, L., Liang, X., Jiang, H., Geng, C., Zhao, X., ... Mu, F. (2020, Nov). Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms. *BMC Bioinformatics*, *21*(1), 518.

50

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012, Sep). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333-i339.
57

Rees, E., & Kirov, G. (2021, 06). Copy number variation and neuropsychiatric illness. *Curr Opin Genet Dev*, *68*, 57–63.
89

Rees, E., Walters, J. T., Chambert, K. D., O'Dushlaine, C., Szatkiewicz, J., Richards, A. L., ... Spencer, C. C. (2014, Mar). CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum. Mol. Genet.*, *23*(6), 1669–1676.
14, 15

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., ... Zellner, S. (2015, 06). ClinGen–the Clinical Genome Resource. *N Engl J Med*, *372*(23), 2235–2242.
80, 163

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet*, *16*, 133–151.
11

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019, 01). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, *47*(D1), D886-D894.
27, 30

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... Rehm, H. L. (2015, May). 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.*, *17*(5), 405–424.
138

Ripke, S., Walters, J. T., & O'Donovan, M. C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*. Retrieved from `https://www.medrxiv.org/content/early/2020/09/13/2020.09.12.20192922` doi: 10.1101/2020.09.12.20192922
106

Risch, N., & Merikangas, K. (1996, Sep). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516–1517.
10

Robertson, S. P., Thompson, S., Morgan, T., Holder-Espinasse, M., Martinot-Duquenoy, V., Wilkie, A. O., & Manouvrier-Hanu, S. (2006, May). Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. *Eur. J. Hum. Genet.*, *14*(5), 549–554.
62

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011, Jan). Integrative genomics viewer. *Nat Biotechnol*, *29*(1), 24–26.
71

Sakuntabhai, A., Ruiz-Perez, V., Carter, S., Jacobsen, N., Burge, S., Monk, S., ... Hovnanian, A. (1999, Mar). Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nat. Genet.*, *21*(3), 271–277.
62

Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., ... Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2017/06/12/148353` doi: 10.1101/148353
27, 138

Sanders, S. J. (2015, Aug). First glimpses of the neurobiology of autism spectrum disorder. *Curr. Opin. Genet. Dev.*, *33*, 80–92.
14

Sanders, S. J., Neale, B. M., Huang, H., Werling, D. M., An, J. Y., Dong, S., ... Freimer, N. B. (2017, 12). Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat Neurosci*, *20*(12), 1661–1668. ([PubMed Central:PMC7785336] [DOI:10.1038/s41593-017-0017-9] [PubMed:28712454])
10

Sasani, T. A., Pedersen, B. S., Gao, Z., Baird, L., Przeworski, M., Jorde, L. B., & Quinlan, A. R. (2019, 09). Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*, *8*.
164

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., ... Church, D. M. (2017, 05). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*, *27*(5), 849–864.
12, 29, 37

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009, Jun). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, *19*(3), 212–219.
9

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., ... Gaunt, T. R. (2013, Jan). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*, *34*(1), 57–65.
138

Singh, S. M., Castellani, C. A., & Hill, K. A. (2020). Postzygotic Somatic Mutations in the Human Brain Expand the Threshold-Liability Model of Schizophrenia. *Front Psychiatry*, *11*, 587162.
62, 83

Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., ... Barrett, J. C. (2016, Apr). Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat Neurosci*, *19*(4), 571–577.
89

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., ... Daly, M. J. (2022, 04). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, *604*(7906), 509–516.
16, 71, 89

Sinnwell, J. P., Therneau, T. M., & Schaid, D. J. (2014). The kinship2 R package for pedigree data. *Hum. Hered.*, *78*(2), 91–93.
25

Smit, A., Hubley, R., & Green, P. (2015). *Repeatmasker.* Retrieved from `https://www.repeatmasker.org/`
51

Spitzer, R. L., Endicott, J., & Robins, E. (1978, Jun). Research diagnostic criteria: rationale and reliability. *Arch Gen Psychiatry*, *35*(6), 773–782.
91

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med*, *61*, 437–455.
47

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., ... Myin-Germeys, I. (2009, Aug). Common variants conferring risk of schizophrenia. *Nature*, *460*(7256), 744–747.
14

Steinberg, S., Gudmundsdottir, S., Sveinbjornsson, G., Suvisaari, J., Paunio, T., Torniainen-Holm, M., ... Stefansson, K. (2017, Aug). Truncating mutations in RBM12 are associated with psychosis. *Nat. Genet.*, *49*(8), 1251–1254.
111

Stilo, S. A., & Murray, R. M. (2019, 09). Non-Genetic Factors in Schizophrenia. *Curr Psychiatry Rep*, *21*(10), 100.
14

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012, Jul). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.*, *13*(8), 537–551.
11

Sullivan, P. F., & Geschwind, D. H. (2019, 03). Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell*, *177*(1), 162–183.
14

Swanwick, C. C., Shapiro, M. E., Yi, Z., Chang, K., & Wenthold, R. J. (2009, Apr). NMDA receptors interact with flotillin-1 and -2, lipid raft-associated proteins. *FEBS Lett*, *583*(8), 1226–1230.
78

Takata, A., Miyake, N., Tsurusaki, Y., Fukai, R., Miyatake, S., Koshimizu, E., ... Matsumoto, N. (2018, 01). Integrative Analyses of De Novo Mutations Provide Deeper Biological Insights into Autism Spectrum Disorder. *Cell Rep*, *22*(3), 734–747.
106

Tang, J., Fan, Y., Li, H., Xiang, Q., Zhang, D. F., Li, Z., ... Chen, X. (2017, Jun). Whole-genome sequencing of monozygotic twins discordant for schizophrenia indicates multiple genetic risk factors for schizophrenia. *J Genet Genomics*, *44*(6), 295–306.
63, 74

Tange, O. (2011, Feb). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, *36*(1), 42-47. Retrieved from `http://www.gnu.org/s/parallel`
33

Taylor, D. M., Thum, M. Y., & Abdalla, H. (2008, Nov). Dichorionic triamniotic triplet pregnancy with monozygotic twins discordant for trisomy 13 after preimplantation genetic screening: case report. *Fertil. Steril.*, *90*(5), 5–9.
62

Thompson, D., Easton, D. F., & Goldgar, D. E. (2003, Sep). A full-likelihood method for the evaluation of causality of sequence variants from family data. *Am J Hum Genet*, *73*(3), 652–655.
113, 128

Tian, Y., Pesaran, T., Chamberlin, A., Fenwick, R. B., Li, S., Gau, C. L., ... Qian, D. (2019, 09). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Sci Rep*, *9*(1), 12752.
138

Tiihonen, J., Lönnqvist, J., Wahlbeck, K., Klaukka, T., Niskanen, L., Tanskanen, A., & Haukka, J. (2009, Aug). 11-year follow-up of mortality in patients with schizophrenia: a population-based cohort study (FIN11 study). *Lancet*, *374*(9690), 620–627.
14

Treangen, T. J., & Salzberg, S. L. (2011, Nov). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, *13*(1), 36–46.
29

Trost, B., Engchuan, W., Nguyen, C. M., Thiruvahindrapuram, B., Dolzhenko, E., Backstrom, I., ... Yuen, R. K. C. (2020, 10). Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*, *586*(7827), 80–86.
85

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J. R., Sung, W. W. L., ... Scherer, S. W. (2018, 01). A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am. J. Hum. Genet.*, *102*(1), 142–155.
50, 51, 58

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., ... van Os, J. (2022, 04). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508.

14

Vadgama, N., Pittman, A., Simpson, M., Nirmalananthan, N., Murray, R., Yoshikawa, T., ... Nasir, J. (2019, Jul). De novo single-nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes. *Eur. J. Hum. Genet.*, *27*(7), 1121–1133.
63

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, *43*, 1–33.
11, 21, 24

Verhoeven, W. M., Egger, J. I., Kremer, B. P., de Pont, B. J., & Marcelis, C. L. (2011). Recurrent major depression, ataxia, and cardiomyopathy: association with a novel POLG mutation? *Neuropsychiatr Dis Treat*, *7*, 293–296.
71

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., ... Abyzov, A. (2018, 12). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, *362*(6420).
75

Wang, K., & Bucan, M. (2008, Jun). Copy Number Variation Detection via High-Density SNP Genotyping. *CSH Protoc*, *2008*, pdb.top46.
12

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013, Feb). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, *14*(2), 125–138.
47

Weiss, M. M., Hermsen, M. A., Meijer, G. A., van Grieken, N. C., Baak, J. P., Kuipers, E. J., & van Diest, P. J. (1999, Oct). Comparative genomic hybridisation. *MP, Mol. Pathol.*, *52*(5), 243–251.
13

Wilfert, A. B., Turner, T. N., Murali, S. C., Hsieh, P., Sulovari, A., Wang, T., ... Chung, W. K. (2021, 08). Recent ultra-rare inherited variants implicate new autism candidate risk genes. *Nat Genet*, *53*(8), 1125–1134.
90

Xiao, X., Zhang, C. Y., Zhang, Z., Hu, Z., Li, M., & Li, T. (2021, Oct). Revisiting tandem repeats in psychiatric disorders from perspectives of genetics, physiology, and brain evolution. *Mol Psychiatry*.
85

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009, Nov). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865–2871.
57

Yokotsuka-Ishida, S., Nakamura, M., Tomiyasu, Y., Nagai, M., Kato, Y., Tomiyasu, A., ... Sano, A. (2021, Jun). Positional cloning and comprehensive mutation analysis identified a novel KDM2B mutation in a Japanese family with minor malformations, intellectual disability, and schizophrenia. *J Hum Genet*, *66*(6), 597–606.
154, 166

Young, W. (2009). Review of lithium effects on brain and blood. *Cell Transplant*, *18*(9), 951–975.
106

Yu, Y., Hu, H., Bohlender, R. J., Hu, F., Chen, J. S., Holt, C., ... Huff, C. D. (2018, Apr). XPAT: a toolkit to conduct cross-platform association studies with heterogeneous sequencing datasets. *Nucleic Acids Res.*, *46*(6), e32.
97

Zarate, S., Carroll, A., Mahmoud, M., Krasheninina, O., Jun, G., Salerno, W. J., ... Sedlazeck, F. J. (2020, 12). Parliament2: Accurate structural variant calling at scale. *Gigascience*, *9*(12).
48, 50

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015, 03). A copy number variation map of the human genome. *Nat. Rev. Genet.*, *16*(3), 172–183.
12, 47

Zech, N. H., Wisser, J., Natalucci, G., Riegel, M., Baumer, A., & Schinzel, A. (2008, Aug). Monochorionic-diamniotic twins discordant in gender from a naturally conceived pregnancy through postzygotic sex chromosome loss in a 47,XXY zygote. *Prenat. Diagn.*, *28*(8), 759–763.
62

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., & Wang, L. (2014, Apr). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007.
30

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, *14 Suppl 11*, S1.
12, 13, 17, 47

Zhou, B., Ho, S. S., Zhang, X., Pattni, R., Haraksingh, R. R., & Urban, A. E. (2018, 11). Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet*, *55*(11), 735–743.
47

Zhu, M., Need, A. C., Han, Y., Ge, D., Maia, J. M., Zhu, Q., ... Goldstein, D. B. (2012, Sep). Using ERDS to infer copy-number variants in high-coverage genomes. *Am. J. Hum. Genet.*, *91*(3), 408–421.
50

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., ... Salit, M. (2016, Jun). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*, *3*, 160025.
50

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., ... Lander, E. S. (2014, Jan). Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U.S.A.*, *111*(4), E455–464.
10