Medical relevance and functional consequences of protein truncating variants

Manuel Antonio Rivas Cruz

Supervisors: Mark I. McCarthy and Peter Donnelly

Nuffield Department of Medicine and Green Templeton College

University of Oxford

A dissertation submitted in partial fulfilment of the requirements for the degree of *Doctor of Philosophy*

Hilary Term, 2015

This thesis is dedicated to my wife, Jackie, and to my family, for their unconditional love and support; and to the memory of my father Manuel A. Rivas (1960-1994), my childhood memories of your smile, sense of humor, and perseverance provide me with endless inspiration.

Acknowledgements

Thanks first to my DPhil advisors, Mark McCarthy and Peter Donnelly. Four years ago when Peter and Mark kindly suggested that I should join them at Oxford for a DPhil I was excited about the option yet nervous about going to an unfamiliar place. Over the past four years I have had the pleasure and great opportunity to learn from such amazing mentors. Mark, whose desire for openness, kindness, and respectful collaborative work has set the template for how to conduct myself when working with others. Peter, whose desire for statistical rigor and clear exposition of the intuition behind the development of statistical methods has set the bar for how I should approach scientific inquiry.

Second, I would like to thank all the past and present members of the McCarthy and Donnelly labs. Thank you for creating such a fun environment, for being great colleagues, and for offering me your friendship. Loukas Moutsianas, thank you for all the comments you provided for all the manuscripts I shared, and for your positive outlook on all aspects of life. Kyle Gaulton, thank you for being a great friend, teaching me about enhancers, open chromatin, and Chip-seq. Juan Fernandez, thank you for being a great friend from the day you started the lab, for bringing "jamoncito" from Spain, and for opening the doors to your home and welcoming me as another member of your family. Matti Pirinen, I couldn't have asked for a better young scientist to work with and learn from. Thank you for taking the time to teach me concepts in Bayesian statistics that I had not previously encountered, and for finding the problems that I wanted to tackle interesting enough to guide me and work with me on developing new computational and statistical approaches.

I would like to thank my past and present collaborators and mentors. Tuuli Lappalainen, thank you for teaching me all there is to know about RNA-seq. Daniel MacArthur, thank you for the feedback and co-leadership on the work on protein truncating variants and their functional consequences. Manolis Dermitzakis, thank you for all the support and for giving me the freedom to explore the boundaries of science. The GoT2D, T2D-GENES, and GTEx consortia members, thank you for being great collaborators. Mark Daly, thank you for your unconditional mentorship and tutelage.

I would like to acknowledge the funding from the Clarendon Scholarship, Green Templeton College, and the Nuffield Department of Medicine without which none of this work would be possible.

I would like to thank the new friends I made at Oxford. In particular, Cecilia Lindgren and Julian Maller who were there to provide support, advice, and guidance throughout my four years at Oxford. I would also like to thank my friends and family in Nicaragua, Miami, and Boston. My mother, brothers, and sisters, thank you for your sacrifice. There is not one day that passes by where I do not think about you. Don Julio and Mita, thank you for raising me, for encouraging me to embrace the sciences, and for instilling a competitive spirit.

Finally, Jackie, thank you for showing me love, sharing words of encouragement, and managing to be patient while I burned the midnight oil every day in our "Home Sweet Tiny Apartment" on Old Road, Headington. I am excited about our future together and the family we will create.

Medical relevance and functional consequences of protein truncating variants

Manuel Antonio Rivas Cruz

Nuffield Department of Medicine and Green Templeton College
University of Oxford

A dissertation submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Hilary Term, 2015

Genome-wide association studies have greatly improved our understanding of the contribution of common variants to the genetic architecture of complex traits. However, two major limitations have been highlighted. First, common variant associations typically do not identify the causal variant and/or the gene that it is exerting its effect on to influence a trait. Second, common variant associations usually consist of variants with small effects. As a consequence, it is more challenging to harness their translational impact. Association studies of rare variants and complex traits may be able to help address these limitations. Empirical population genetic data shows that deleterious variants are rare. More specifically, there is a very strong depletion of common protein truncating variants (PTVs, commonly referred to as loss-of-function variants) in the genome, a group of variants that have been shown to have large effect on gene function, are enriched for severe disease-causing mutations, but in other instances may actually be protective against disease.

This thesis is divided into three parts dedicated to the study of protein truncating variants, their medical relevance, and their functional consequences. First, I present statistical, bioinformatic, and computational methods developed for the study of protein truncating variants and their association to complex traits, and their functional consequences. Second, I present application of the methods to a number of case-control and quantitative trait studies discovering new variants and genes associated to breast and ovarian cancer, type 1 diabetes, lipids, and metabolic traits measured with NMR spectroscopy. Third, I present work on improving annotation of protein truncating variants by studying their functional consequences. Taken together, these results highlight the utility of interrogating protein truncating variants in medical and functional genomic studies.

Contents

1 Introduction **1**

2 Protein truncating variants and rare variant association studies: an Overview **5**

2.1 Common variant association studies (CVAS) 5
2.2 Advances in DNA sequencing technologies 6
2.3 Surveys of rare genetic variation and early rare variant findings 8
2.4 From common variant association studies to rare variant association studies 10
2.5 Aggregating signal across multiple variants 13
2.5.1 Biological units 13
2.5.2 *In silico* prediction of the deleterious effect of a genetic variant 13
2.5.2.1 Protein truncating variants 15
2.5.3 Models for rare variant association 18
2.6 Aims 20

3 Development of statistical, bioinformatic, and computational methods for the analysis of protein truncating variants: Association analysis **22**

3.1 Method for the analysis of protein truncating variants with dichotomous traits 23
3.1.1 Background and rationale 23
3.1.2 Intuition 24
3.1.3 Implementation 25
3.1.4 Simulations 27
3.1.5 Application of method in this thesis 29
3.1.6 Limitations and next steps 29
3.1.7 Attributions 29

3.2 Method for the analysis of protein truncating variants with quantitative traits . . . 29
3.2.1 Background and rationale . . . 29
3.2.2 Intuition . . . 30
3.2.3 Implementation . . . 31
3.2.4 Simulations . . . 35
3.2.5 Application of method in this thesis . . . 37
3.2.6 Limitations and next steps . . . 37
3.2.7 Attributions . . . 38
3.3 Methods for the analysis of protein truncating variants and multiple phenotypes . . . 38
3.3.1 Background and rationale . . . 38
3.3.2 Intuition . . . 39
3.3.3 Implementation . . . 43
3.3.4 Simulations . . . 51
3.3.5 Application of method in this thesis . . . 54
3.3.6 Limitations and next steps . . . 57
3.3.7 Attributions . . . 57
3.4 Discussion . . . 58

4 Development of statistical, bioinformatic, and computational methods for the analysis of protein truncating variants: transcriptional consequences 60
4.1 Introduction . . . 60
4.2 *In silico* annotation of protein truncating variants with RNA-seq transcript quantification data . . . 61
4.2.1 Background and rationale . . . 61
4.2.2 Intuition . . . 62
4.2.3 Implementation . . . 63
4.2.4 Results . . . 64
4.2.5 Application of the approach in this thesis, limitations, and next steps . . . 65
4.2.6 Attributions . . . 66
4.3 Assessing allele-specific expression across multiple tissues from RNA-seq read data . . . 69
4.3.1 Background and rationale . . . 69

4.3.2 Intuition . . . 71
4.3.3 Implementation . . . 72
4.3.4 Simulations . . . 75
4.3.5 Limitations and next steps . . . 76
4.3.6 Application of the method in this thesis . . . 82
4.3.7 Attributions . . . 82
4.4 Assessing impact of rare variants proximal to splice junctions: the Splice Disruption Model (SDM) . . . 82
4.4.1 Background and rationale . . . 82
4.4.2 Intuition . . . 83
4.4.3 Implementation . . . 84
4.4.4 Simulations . . . 89
4.4.5 Application of the method in this thesis . . . 89
4.4.6 Limitations and next steps . . . 90
4.4.7 Attributions . . . 90
4.5 Discussion . . . 90

5 Application of methods to the study of protein truncating variants and their relevance to medical traits: disease studies 95
5.1 Mosaic protein truncating variants in *PPM1D* contribute to breast and ovarian cancer predisposition . . . 96
5.1.1 Overview of the study . . . 96
5.1.2 Overview of my role . . . 98
5.1.3 Results . . . 99
5.1.4 Conclusion . . . 106
5.2 Cross-disorder analysis of six autoimmune diseases: Protein truncating variants in *TNFAIP3* show suggestive evidence of association to type 1 diabetes . . . 111
5.2.1 Background . . . 111
5.2.2 Results . . . 112
5.2.3 Conclusion . . . 116
5.3 Multi-ethnic exome-wide association study of protein truncating variants to type 2 diabetes . . . 122
5.3.1 Overview of the study . . . 122
5.3.2 Overview of my role . . . 123
5.3.3 Results . . . 124

5.3.4 Conclusion . . . 129
5.4 Discussion . . . 129

6 Application of methods to the study of protein truncating variants and their relevance to medical traits: quantitative traits 132
6.1 Exome sequencing analysis of protein truncating variants and the lipid profile in a multi-ethnic type 2 diabetes case-control study . . . 133
6.1.1 Overview of the study . . . 133
6.1.2 Overview of my role . . . 135
6.1.3 Results . . . 136
6.1.4 Conclusion . . . 147
6.2 Exome-wide association study of protein truncating variants and blood metabolite measurements in the Oxford Biobank Study . . . 148
6.2.1 Overview of the study . . . 148
6.2.2 Overview of my role . . . 149
6.2.3 Results . . . 153
6.2.4 Conclusion . . . 154
6.3 Discussion . . . 159

7 Improving annotation of protein truncating variants by studying their functional consequences 160
7.1 Background . . . 160
7.2 Overview of the study . . . 161
7.3 Overview of my role . . . 168
7.4 Results . . . 168
7.4.1 Transcriptional properties of PTV-containing transcripts . . . 168
7.4.2 Insights into nonsense-mediated decay . . . 171
7.4.3 Dosage compensation for heterozygous PTVs . . . 186
7.4.4 Transcriptional impact of variants proximal to splice junctions 194
7.5 Conclusions . . . 199

8 Summary and Future Work 200
8.1 Summary . . . 200
8.2 Future Work . . . 205

Appendix List of Abbreviations 210

Appendix Supplementary Tables 213

Bibliography **219**

List of Tables

3.1 Power (given in percentage) at $\alpha = 0.001$ to detect association for the two scenarios specified in the text. 1,000,000 replicates were generated to obtain a *P*-value. 37

5.1 DNA repair genes with ≥ 4 PTVs in the breast cancer pooled sequencing experiment 105
5.2 Comparison of composite PTV allelic counts in the breast cancer pooled sequencing data with the ExAC reference data set (genes with Bayes Factor > 10 and multiple (> 1) PTVs shown). 106
5.3 Results from cross-disorder analysis of PTVs and six autoimmune diseases. 116
5.4 Top 10 PTV gene association results from the type 2 diabetes multi-ethnic exome sequencing study 127

6.1 PTV gene based association results for univariate Bayesian models applied to LDL-C, TG, HDL-C, and TC levels in the multi-ethnic exome sequencing study 141
6.2 PTV gene based association results for C-alpha MRP test applied to LDL-C, TG, HDL-C, and TC levels in the multi-ethnic exome sequencing study 142
6.3 Table of monogenic dyslipidemia genes with the name of condition in OMIM 142
6.4 Grouping of metabolites into 12 metabolic clusters 150

7.1 PTV annotation flags used in the annotation pipeline. 164
7.2 Summary of PTVs in the GTEx and Geuvadis DNA sequencing data sets 165
7.3 List of 38 predictors used for modeling NMD. 179

1 Type 2 diabetes biologically motivated “premium” gene sets 215

2 Description of metabolites used in the analysis 218

List of Figures

2.1 DNA sequencing costs: 2001-2010 (2010 is the year I started my thesis work). 7
2.2 Figure adaptation of SFS presented in Coventry et al. (2010) 9
2.3 Allele frequency spectrum and its relationship to disease predisposition 11
2.4 Sample sizes required to achieve 80% power to detect association at varying allele frequencies and effect sizes 12
2.5 Diagram of variant annotation in rare variant studies. 16

3.1 Prior densities for the the PTV case control method. 26
3.2 Comparison of ranking performance between Bayesian SEMCC and Fisher's exact test 28
3.3 SEM and GEM alternative models. 33
3.4 Schematic diagram of the statistical framework proposed for rare variant association studies 40
3.5 Diagram of the expected correlation of genetic effects matrices 42
3.6 Comparison of the power to detect association in multivariate analysis with multiple quantitative trait measurements 53
3.7 Comparison of the power to detect association in cross-disorder analysis. 55
3.8 Application of MCMC algorithm to simulated data. 56

4.1 Expression levels of *SGCB* transcript isoforms across all tissues in GTEx 67
4.2 Annotation of two putative novel alleles in *SGCB* 68
4.3 Integration of personal transcriptome data for improved annotation of predicted protein truncating genetic variants. 69
4.4 Diagram of allele specific expression analysis using short read sequencing data 70
4.5 Prior density of θ for the null and alternative model 73
4.6 ASE simulations scenario I: strong ASE effects across all tissues . . . 77
4.7 ASE simulations scenario II: no ASE effects across all tissues 78

4.8 ASE simulations scenario III: heterogeneous ASE effects 79
4.9 ASE simulations scenario III: tissue-specific ASE effects 80
4.10 Diagram of splice junction quantification using FLUX Capacitor . . 83
4.11 Diagram of signal sought using the SDM algorithm 92
4.12 SDM simulations scenario I: $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(0,1)$ 93
4.13 SDM simulations scenario II: $0.5\mathcal{N}(-2,1)+0.5\mathcal{N}(0,1)$ 94

5.2 *PPM1D* PTVs in the ExAC data set 108
5.1 Somatic (mosaic) frequency estimated from MiSEQ sequencing data of *PPM1D* PTVs identified through full gene Sanger sequencing 109
5.3 Identification of mosaic PPM1D PTVs in population unselected controls. 110
5.4 *TNFAIP3* PTV association to type 1 diabetes 117
5.5 *TNFAIP3* transcript isoform expression across the GTEx tissues . . . 118
5.6 IGV snapshots I. 118
5.7 IGV snapshots II. 119
5.8 IGV snapshots III. 119
5.9 IGV snapshot of c.IVS7 + 1G>C in *TNFAIP3*. 120
5.10 Extension of C-alpha MRP test to the analysis of gene-sets and application to autoimmune targeted sequencing data set 121
5.11 Accumulation of ultra-rare protein truncating alleles amongst genes contributing to pre-specified "biologically-driven" gene-sets 128
5.12 Power to detect single variant association 130
5.13 Power to detect gene-based association 131

6.1 C-alpha MRP test applied to the dyslipidemia gene sets 143
6.2 MCMC clustering algorithm described in Chapter 3 applied to PTVs in the monogenic dyslipidemia gene set and the standard lipid profile 144
6.3 Percent triglyceride level reduction compared to disease or control population 145
6.4 Comparison of effect sizes reported in publications for variants associated to triglyceride levels to effect sizes estimated from the multi-ethnic exome sequencing study in T2D cases and controls 146
6.5 Diagram of PTV analysis in the Oxford Biobank study and NMR metabolite measurements 151
6.6 Grouping metabolomic traits for multivariate association analyses. . . 152
6.7 Manhattan plot of protein truncating variant association to blood metabolite measurements using the Bayesian SEMGEM models. . . . 155

6.8 Manhattan plot of protein truncating variant association to blood metabolite measurements using C-alpha MRP. 156
6.9 MCMC estimation algorithm applied to PTV data in exome array for *CT62* and metabolic cluster 5. 157
6.10 PTV association results for biological gene sets selected to be relevant to NMR metabolite levels. 158

7.1 Schematic diagram of study design. 162
7.2 Tissue-wide expression profile for PTV containing genes 169
7.3 Transcriptional properties of PTV containing transcripts 170
7.4 Insights into nonsense-mediated decay: performance of ASE algorithm across all iterations for five MCMC chains 174
7.5 Insights into nonsense-mediated decay: proportion of nonsense variants with allele-specific expression effects in the GTEx data set in different frequency classes 175
7.6 Characterizing allele-specific expression patterns of PTVs across multiple tissues 176
7.7 Insights into nonsense-mediated decay: modeling NMD with ASE outcome 180
7.8 Insights into nonsense-mediated decay: feature (variable) importance plots 181
7.9 ASE classification examples: no ASE and moderate ASE across all tissues 183
7.10 ASE classification examples: strong ASE across all tissues and mixture of moderate and strong ASE 184
7.11 ASE classification examples: mixture of no ASE and ASE effect and tissue-specific ASE 185
7.12 ASE data for p.S474X (rs328) in the gene *LPL* (lipoprotein lipase) . . 186
7.13 Insights into dosage compensation: examining gene expression ratios for large CNV deletion carriers (Geuvadis) 189
7.14 Insights into dosage compensation: examining gene expression ratios for large CNV deletion carriers (GTEx) 190
7.15 Insights into dosage compensation: impact of somatic mutations . . . 191
7.16 Insights into dosage compensation: examining gene expression ratios for nonsense SNV carriers with strong ASE (GEUVADIS) 192

7.17 Insights into dosage compensation: examining gene expression ratios for nonsense SNV carriers with strong ASE (GTEx) 193
7.18 Transcriptional impact of variants proximal to splice junctions 196
7.19 Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, c.IVS8+1G>C (rs35337543), in the gene *IFIH1* . . 197
7.20 Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, rs116928232, in the gene *LIPA* 198

8.1 Genome-editing for experimental quantification of the functional consequence of DNA sequence variants. 207

Chapter 1

Introduction

Genetic and environmental factors are key forces in shaping an individual's predisposition to disease. Ultimately the hope is that we will be able to translate insights from the genetic makeup of diseases to therapies and cures. In the past two decades significant strides have been made in understanding the genetic makeup of complex traits.

Initially, genetic mapping was done using a "linkage mapping" approach. The idea was that because of meiotic recombination any marker showing correlated segregation in a family with the trait of interest must be located nearby in the genome. This approach was invented by Alfred Sturtevant in 1913 for the study of traits in fruit flies (Sturtevant, 1913). Linkage analysis successfully identified hundreds of genes associated to Mendelian diseases by the late 1980s (Altshuler et al., 2008). Nonetheless, when it was applied to common diseases it achieved very little success.

More recently, "genome-wide association mapping" in populations has been the preferred approach. This approach consists of obtaining genotype data from over half a million markers across the genome and contrasting the frequencies observed in cases to the frequencies observed in controls from the same population. These genome-wide association studies (GWAS) of common diseases have yielded over 4,000 bona-fide associations in the genome. In spite of that, the therapeutic promise has not yet been met. Possible contributions for this include: i) common variant associations only explain a small proportion of the disease heritability, ii) common variant associations do not yield the causal variant much less the gene by which it exerts its effect on to lead to disease predisposition, iii) associated common variants only have a small effect on disease predisposition thereby limiting the translational inferences that can be made, and iv) the main reason that GWAS is only seven years old.

In the past few years the field has gravitated towards association studies of rare variants which yields many opportunities in genetic mapping of human diseases, but

produces additional challenges that were not present in earlier association mapping approaches. Most notably is the limitation to perform an association of a single marker to a trait due to the lack of information available from a single rare event. Thus, it is usually required to aggregate information across multiple markers (hereinafter referred to as variants) and in order to do so some prior understanding of the functional consequences of DNA sequence variants is necessary. Focusing on protein truncating variants (PTVs) is an attractive strategy for association mapping of disease and learning about the consequences of large effect perturbation on phenotype (as these variants are assumed to have strong functional consequences).

PTVs are commonly referred to as "loss of function" (LoF) variants because it is commonly assumed that transcripts containing premature stop codons are subject to nonsense-mediated decay (NMD). NMD is a cellular mechanism that detects aberrant transcripts and prevents the expression of truncated proteins resulting in overall decreased gene expression. This biological phenomenon exists in all eukaryotes and includes at least three main proteins (called up-frameshift (UPF) proteins) that are conserved between species: UPF1, UPF2 and UPF3 (Hentze and Kulozik, 1999). The current working model of NMD in humans is that mRNAs containing premature stop codons are flagged for degradation during the first round of translation where the ribosome removes exon-exon junction complexes (EJC) - these protein complexes are bound to the mRNA after splicing occurs. If any of the EJC proteins remain bound to the mRNA, once the first round of translation is completed, NMD is subsequently activated (Maquat et al., 2010). This event occurs as a result of UPF1 (a protein involved in a complex responsible for releasing the mRNA from translation) coming into contact with UPF2 and UPF3 (proteins that bind to the exon junction complex before and after nuclear transport of the mRNA) when an exon junction complex is left on the mRNA (Popp and Maquat, 2013). To date, it is unclear how efficient NMD is at degrading mRNAs containing premature stop codons and whether or not degradation may vary from tissue to tissue. Furthermore, the efficiency of NMD may impact health and severity of disease in individuals (Holbrook et al., 2004). Thus, understanding when and how any particular predicted protein truncating variant impacts mRNA and protein levels or protein localization may become central to interpreting rare variation discovered in genome sequencing studies.

Many challenges exist: from the analysis of the raw sequencing data, to the *in silico* prediction of the functional consequences of identified DNA sequence variants, to the integration of all this information into an association testing framework, and to the downstream interpretation of association findings.

Big challenges, which are the focus of this thesis, are to develop flexible and computationally tractable methods to obtain power to detect causal protein truncating variants in rare variant association studies, to understand their contribution to disease and related traits, and to develop a comprehensive understanding of their functional consequences. The structure of this thesis is as follows.

I will begin in Chapter 2 with a brief literature review of rare variant association studies and protein truncating variants, which will include: i) a historical review of common variant association studies, ii) an examination of advances in sequencing technologies, iii) description of the results from early surveys of rare genetic variation, iv) a summary of early rare variant findings, v) a brief introduction to the transition from common variant association studies to rare variant association studies, vi) a recap of tools for *in silico* prediction of the deleterious effect of a genetic variant, vii) a sketch of main findings from the first systematic survey of protein truncating (commonly referred to as loss-of-function) variants in the human genome, and viii) a description of tests for rare variant association testing available at the time I started my thesis. I will provide some references for readers who want more details. I will conclude this chapter by presenting my approach and principal aims in addressing what I believed to be major challenges and opportunities at the start of my thesis for studying the medical relevance and functional consequences of protein truncating variants.

In Chapter 3 I will introduce statistical, bioinformatic, and computational methods that were developed for the analysis of protein truncating variants, which are implemented in the software package `MAMBA` (Modeling And Mining Big DatA). The focus of this chapter will be on methods for association analysis.

In Chapter 4 I will continue with statistical, bioinformatic, and computational methods that were developed for the analysis of PTVs. The focus of this chapter will be on methods for the interpretation of the functional consequences of PTVs.

The more significant contributions to the field of human genetics will be presented in Chapters 5-7 where I study PTVs with empirical data. In Chapter 5 I study the contribution of PTVs to disease predisposition by applying the developed methods to sequencing data in case-control study designs. First, I present a pooled sequencing study of breast and ovarian cancer patients where I contributed to the identification of signals of PTVs that were not previously detected. The PTVs were identified to result in a "gain of function" effect as opposed to the commonly assumed "loss of function" effect. This motivated interest in characterizing the functional consequences of PTVs discussed in Chapter 7. Second, I present application of the multivariate approaches

to a published targeted sequencing data set of 25 genes in over 40,000 individuals comprised of six autoimmune disease groups and one control group. I identified new signals not previously characterized by the authors of the initial publication. Third, I present application of the univariate approaches to data generated from a multi-ethnic exome sequencing study of over 12,000 individuals comprised of type 2 diabetes (T2D) patients and controls. Results from the analysis of genes with prior knowledge of association are characterized, including targeted gene-sets, and an overview of the top results is presented.

In Chapter 6, I present application of the methods to the analysis of quantitative traits. I focus on the association analysis of PTVs and lipid levels in the T2D multi-ethnic exome sequencing study where we identify signals of PTVs with large effects and characterize their impact in T2D patients. Then, I present results from the application of the methods to metabolomic measurements using the NMR platform where we identify new signals of PTVs and rare coding variants not previously identified, which warrant additional follow-up and replication.

Lastly, in Chapter 7, I present application of the methods to interpret the functional consequences of PTVs to a population-based and multi-tissue RNA sequencing (RNA-seq) data set from the Geuvadis and the Genotype Tissue Expression (GTEx) project. For Chapters 3 and 4 I provide attributions when appropriate. Furthermore, in Chapters 5, 6, and 7 where work may have been conducted as part of a consortium or a collaborative effort I present an overview of the study and an overview of my role. I will end my thesis with a summary of my work and a discussion on how it can be extended in the future.

Chapter 2

Protein truncating variants and rare variant association studies: an Overview

2.1 Common variant association studies (CVAS)

Large-scale genome-wide association studies (GWAS) have successfully identified over 4,000 loci associated to a complex trait (NHGRI catalog) (Welter et al., 2014). These studies were largely made feasible by the progression of improvements in technology and our understanding of genetic variation and the human genome circa 2001-2005.

First, in 2001, an extensive collection of genetic variation in the human genome was curated by the National Center for Biotechnology Information (NCBI) and made publicly available in the form of a database referred to as the dbSNP database (Sherry et al., 2001). This database along with the completion of the Human Genome Project (Collins et al., 1998; Lander et al., 2001; Venter et al., 2001; Consortium et al., 2004) was instrumental in guiding the study and characterization of genetic variation in four population samples in the International Haplotype Map (HapMap) project (Consortium et al., 2005).

The HapMap study laid the foundation for association studies of common variants to common diseases. First, the HapMap consortium made publicly available a resource that documents the structure of genome-wide variation and linkage disequilibrium (LD) in the human genome. Then, a group of investigators from the HapMap consortium showed that the patterns of common genetic variation could be exploited to develop genotyping technologies that could maximize the effiency and power of association studies by only using 1% of the 11 million variants in the catalogue (de Bakker et al., 2005; Pe'er et al., 2006), which later influenced the cre-

ation and design of genotyping array technologies by companies like Illumina and Affymetrix that were used for genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs) in hundreds to thousands of individuals.

The study of genetic variants in GWAS was largely limited to common variation (minor allele frequency $> 5\%$) because the tagging efficiency of low-frequency ($1\% - 5\%$) variation by the early genotyping arrays was weak ($< 30\%$ of SNPs across all the populations were captured with $r^2 \geq 0.8$, whereas for common variants the percentage was greater than 60%) (Pe'er et al., 2006). Pe'er et al. (2006) suggested that comprehensive scans for rare causal alleles would require other sets of markers, and possibly complete resequencing. However, the cost range to sequence a human genome from 2001-2007 was prohibitively expensive ($10 - 100$ million U.S. dollars, Figure 2.1).

2.2 Advances in DNA sequencing technologies

In 2004 it was clear that the state of the art methods (including the Sanger method that was mainly used) for sequencing a human genome in 2001-2005 would not be sufficient for sequencing large numbers of human genomes. The main limitation of 'first-generation' sequencing technologies was that it would require large resources to complete a human genome, e.g. time, larger workforce, and money (Shendure et al., 2004; Metzker, 2009). Thus, the National Human Genome Research Institute (NHGRI) of the United States initiated a coordinated effort to support the development of technologies to dramatically reduce the cost of DNA sequencing[1]. Interestingly, many of the projects awarded grants emphasized DNA sequencing using nanopores (5/20) or DNA sequencing by synthesis (4/20)[1].

The NHGRI effort managed to catalyze the development of 'next-generation' sequencing technologies (Voelkerding et al., 2009). Among the competitors in the commercial arena was a small company, Solexa, founded in Cambridge, UK, that focused on a patented sequencing method based on reversible dye-terminators technology, and engineered polymerases (Bentley et al., 2008). Solexa would later be acquired by Illumina, one of the main commercial companies focused on developing, manufacturing, and marketing genotyping array technologies (van Dijk et al., 2014).

[1] `http://www.genome.gov/12513162`

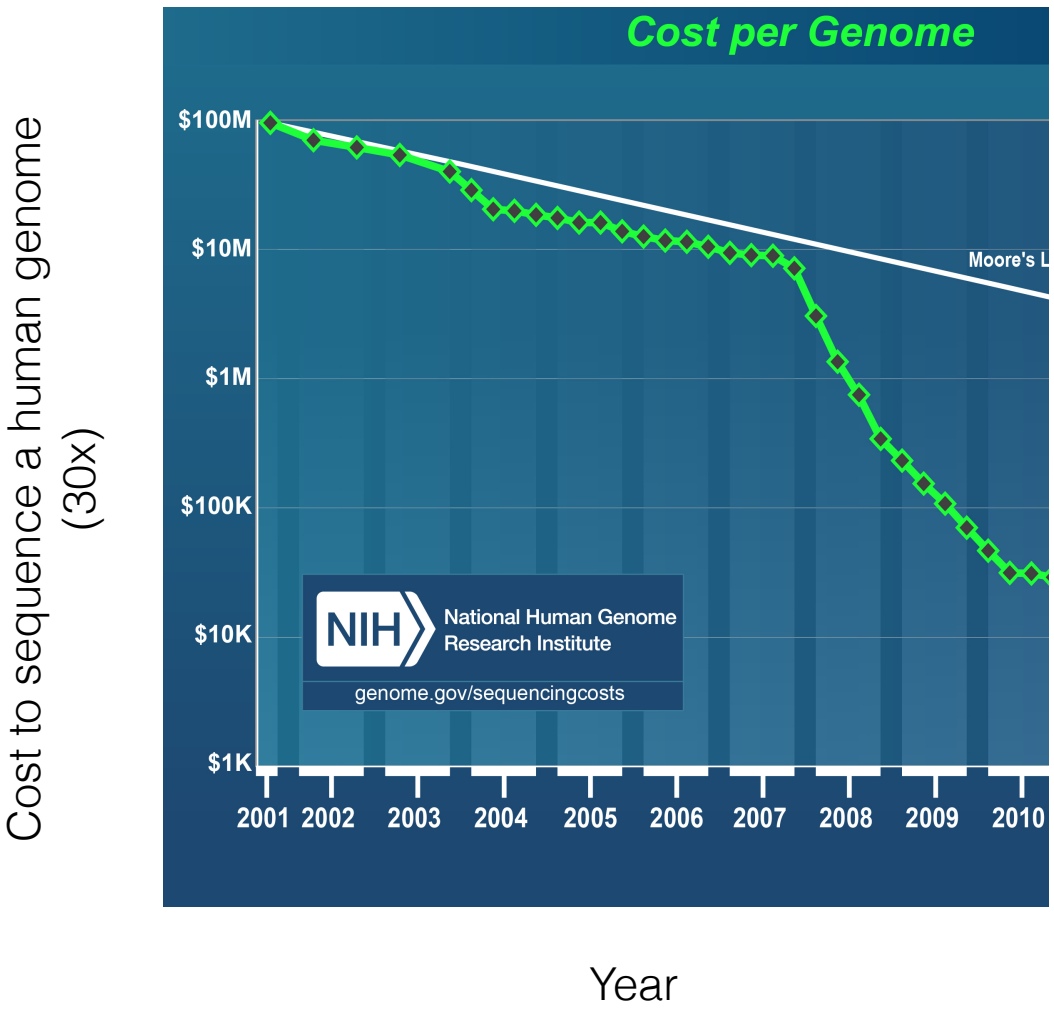


Figure 2.1: DNA sequencing costs: 2001-2010 (2010 is the year I started my thesis work). Improvements in sequencing technologies in the past decade resulted in a dramatic drop in sequencing costs. In 2007 to sequence a human genome at 30X coverage would cost over 10 million U.S. dollars. In 2008 the cost was around 500,000 U.S. dollars, and in 2010, at the start of my DPhil, the cost to sequence a human genome was about 30-40,000 U.S. dollars (over a hundred fold reduction in cost during the span of three years). Plot obtained from `http://www.genome.gov/sequencingcosts/`, and shown only to 2010 to reflect the cost of sequencing at the time of my thesis work.

In 2007, the 1000 Genomes project was proposed by Richard Durbin of the Sanger Institute with the principal aims of developing reference sequencing data sets, developing standards for the analysis and interpretation of sequencing data, and to study the nature of rare variants (which is the reason why they needed to sequence about 1000 genomes (Jocelyn Kaiser, 2007)).

Because the 'next-generation' DNA sequencing technologies were new, pilot projects

were needed to understand the raw data to generate useful information for future surveys of genetic variation, and potentially for future genetic studies of common and rare diseases. In 2008, the pilot phase of the 1000 Genomes project started (meanwhile, I had graduated from the Massachusetts Institute of Technology (MIT), and I decided that I would spend some time working at the Broad Institute. I focused on developing error models and implementing algorithms for rare variant detection from raw read sequencing data (DePristo et al., 2011)). At this time the cost of sequencing plummeted to about 500,000 U.S. dollars for a human genome (Figure 2.1). This was largely a consequence of two things: i) increased competition (454 Life Sciences, ABI Solid, and Illumina Solexa were the main competitors), and ii) improvements in sequencing methods.

Advances in DNA sequencing technologies and methods for processing raw data: for example, algorithms and software to align read data to the reference human genome (Li et al., 2008), algorithms and software to detect genetic variation and estimate genotypes (DePristo et al., 2011), and standard file formats to make data distributable (which is now referred to as the Variant Call Format [VCF], (Li et al., 2009)) stimulated the field of human genetics into adopting next-generation sequencing technologies for characterizing the nature of genetic variation.

2.3 Surveys of rare genetic variation and early rare variant findings

Between 2010 and 2012 the first surveys of rare genetic variation in tens of thousands of individuals were published (Coventry et al., 2010; Nelson et al., 2012; Keinan and Clark, 2012). In Coventry et al. (2010) Sanger sequencing of genomic PCR amplicons was used to resequence the diabetes-associated genes *KCNJ11* and *HHEX* in $13,715$ indiivduals. Coventry et al. (2010) showed that their genetic data for both *HHEX* and *KCNJ11* supported a site frequency spectra (SFS) that is consistent with recent acceleration in population growth and estimated that the acceleration occured over the past $2,000$ years. They suggested that one consequence of this observation is that future sequencing efforts will reveal an SFS with a larger proportion of rare and missense variants with possible consequences for human health. In Nelson et al. (2012), and Keinan and Clark (2012) they replicated the results of Coventry et al. (2010) in large samples and discovered a large excess of rare variants with predicted deleterious effects on gene function, further motivating the need to study the relationship between rare variants and complex traits in addition to common variants. For disease

association studies, the main challenge was (and still remains) how to interpret the frequency distribution of disease causing alleles. A counterintuitive observation is that while the SFS suggests that the majority of rare variants in a population are rare, at the level of a single individual most variant alleles observed are actually from common variants. This observation has been used to argue in favor for the common variant common disease (CVCD) hypothesis that predicts that common disease-causing alleles will be found in all human populations which manifest a given disease (Lander, 1996; Reich and Lander, 2001). It is likely that the frequency distribution of disease causing alleles will depend on selection. However, the excess of rare genetic variants in a population and an increased mutational load in an individual presents an exciting conundrum (Simons et al., 2014).

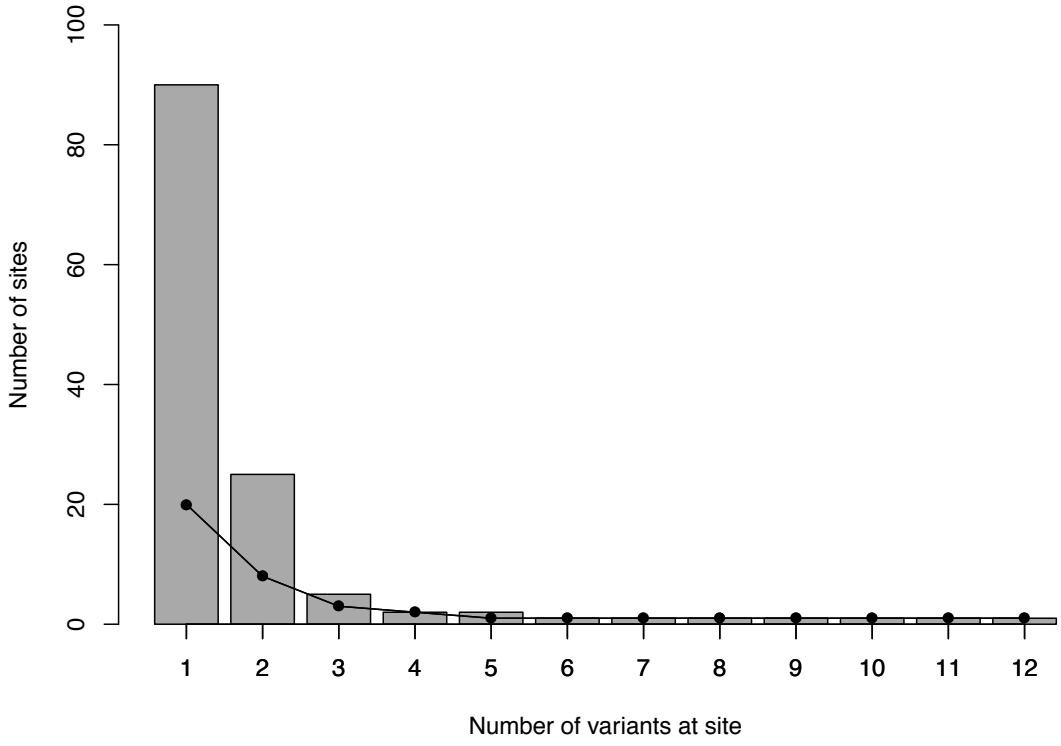


Figure 2.2: Figure adaptation of SFS presented in Coventry et al. (2010). In Coventry et al. (2010); Nelson et al. (2012); Keinan and Clark (2012) they found that the number of rare variant sites detected in experimental data was about 4x larger than expected by using the Wright-Fisher constant population size model and mutation rate Θ estimated by Watterson's method (black line, (Hartl et al., 1997)).

With the dramatic drop in sequencing costs it was increasingly possible to assess

the contribution of rare genetic variants to disease predisposition. At the time I started my thesis work there were only a limited number of rare variants identified to be associated to a complex trait and most of them restricted to genes/regions that were previously implicated either through genome-wide association study (Nejentsev et al., 2009; Johansen et al., 2010; Momozawa et al., 2011; Rivas et al., 2011) or genes identified through familial studies of related monogenic disease (Cohen et al., 2006; Lusis and Pajukanta, 2008) (Figure 2.3).

2.4 From common variant association studies to rare variant association studies

In 2010, at the start of my thesis work, the cost of DNA sequencing had plummeted to the point where targeted, exome, and/or genome sequencing in tens of thousands of individuals was feasible. A major interest in the field was to embark on genome-wide association studies of rare variants (Gibson, 2012).

Rare variant association studies (RVAS) rely on the analysis of genetic variants identified in population-scale DNA sequencing studies (Zuk et al., 2014). In 2010 it was clear that rare variants were going to be extremely abundant (Coventry et al., 2010). Furthermore, it was later shown that rare variants discovered in sequencing studies were going to be mainly private and population-specific (Fu et al., 2013).

In order to successfully identify rare variant associations it was suggested that effects would need to be stronger than that observed with common variants (McCarthy et al., 2008; Manolio et al., 2009). In Figure 2.4 I highlight the sample sizes required to achieve 80% power to detect association at the significance threshold of $P < 5 \times 10^{-8}$ that is commonly used for common variant association studies across varying allele frequencies and effect sizes. For example, $10,378$ samples (with equal case-control ratio and an assumed disease prevalence of 5%) would be required to achieve 80% power to detect association at genome-wide significance for a variant with 10% minor allele frequency (MAF) and an odds ratio of 1.3, whereas over a million samples would be required to achieve 80% power to detect association signal at genome-wide significance for a variant of the same effect size with 0.1% MAF (Purcell et al., 2003). If sample sizes were fixed to $10,378$ then it would require an odds ratio over 6 to achieve 80% power to detect association for a 0.1% variant. Both requirements to reach equivalence appear to be unrealistic for binary traits from an epidemiological point of view. From an epidemiological point of view, sample collection of over half a million case samples for any single disease would require investment in additional

healthcare infrastructure across the world. Of course, for quantitative traits such numbers are not unrealistic (Ollier et al., 2005; Chen et al., 2011).

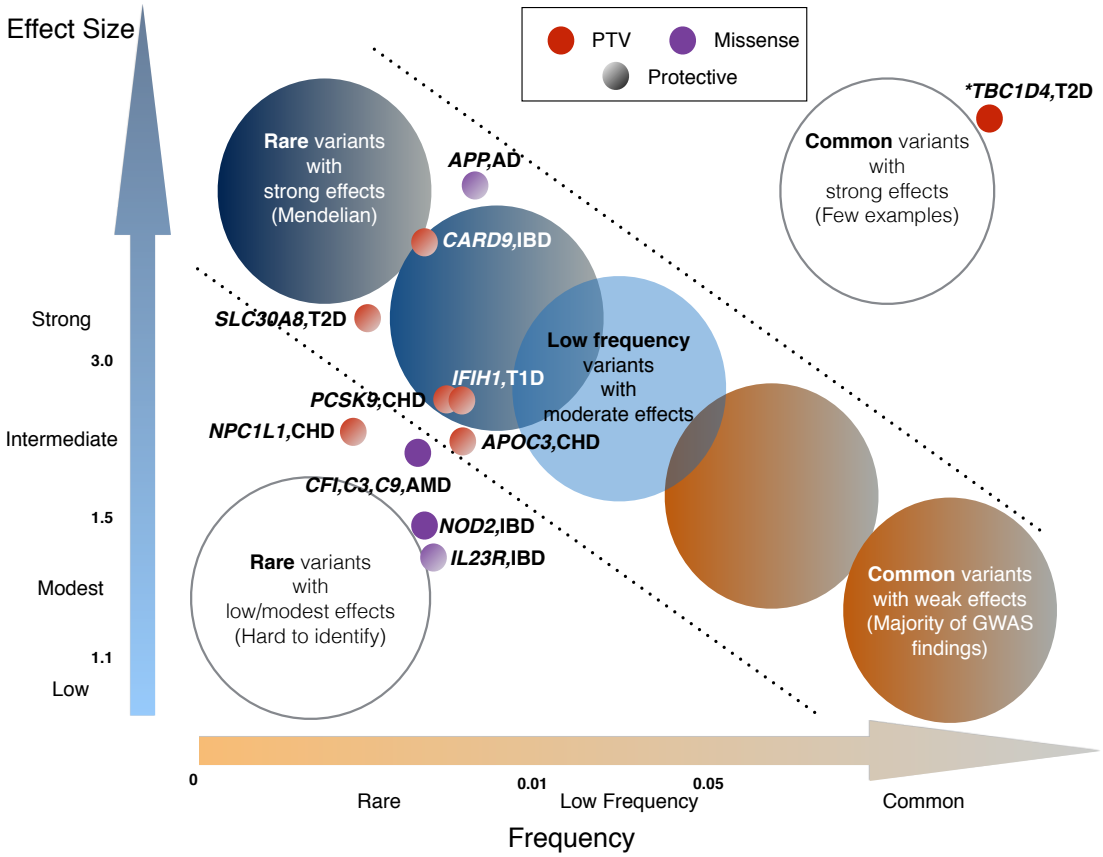


Figure 2.3: Allele frequency spectrum and its relationship to disease predisposition. The expected relationship in the field of human genetics between allele frequency and effect size for feasible identification of variants impacting disease susceptibility as we moved from common variant association studies to rare variant association studies (McCarthy et al., 2008). Some examples of recently identified rare variants associated to disease are overlayed including some conferring protection to disease (gradient fill). $*$ The *TB1CD4* nonsense variant associated to type 2 diabetes is common in the population of Greenland, very rare in other parts of the world. Figure adapted from Manolio et al. (2009), which was adapted from McCarthy et al. (2008).

Because rare (MAF $\leq 1\%$) variants are extremely abundant and the majority of rare variants are extremely rare (MAF $\leq 0.01\%$) rare variant association studies are not amenable to standard genome-wide association approaches applied to common genetic variants (Zuk et al., 2014).

A proposed approach has been to aggregate evidence of association across multiple variants in a gene (Li and Leal, 2008). The idea is that allelic heterogeneity, the presence of multiple distinct disease alleles in close proximity (Pritchard and Cox,

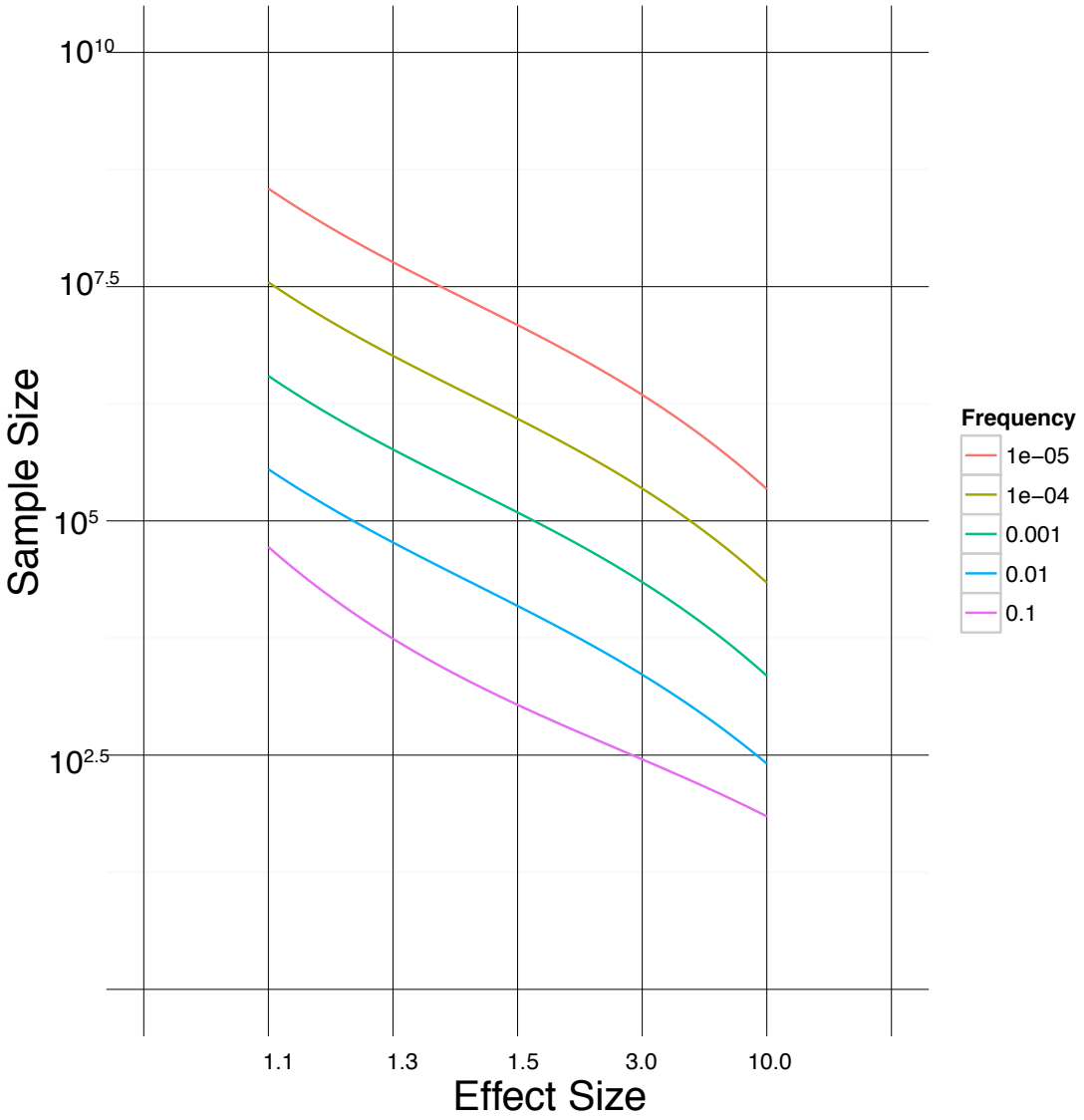


Figure 2.4: Sample sizes required to achieve 80% power to detect association at varying allele frequencies and effect sizes. The number of samples required to achieve 80% power to detect association at an α level equal to 5×10^{-8} for rare variants compared to common variants follows a logarithmic relationship. For example, at $\alpha = 5 \times 10^{-8}$, for a variant with minor allele frequency of 0.01% in the population with an effect size of 3 would require $348,082$ to achieve 80% power to detect association, whereas for a common 10% variant it would only require 450 samples. The curve showing the sample sizes required to achieve 80% power to detect association for a common 10% variant is not parallel to the others (purple). This is due to an error in the Genetic Power Calculator when performing penetrance calculations, i.e. the genotype relative risk for homozygous genotype for a common 10% variant with an odds ratio of 10 exceeds the limit allowed. Hence, I obtained an estimate with a lower genotype relative risk (20) for a homozygous genotype.

2002), can be present in a gene. Early empirical results suggested that this may be a common observation. For example, in Rivas et al. (2011) 6 independent coding variants in *NOD2* were identified to confer risk to Crohn's disease. These results, of allelic heterogeneity in a gene, were also corroborated by additional independent coding variant associations to inflammatory bowel disease in *IL23R* and *CARD9*, further suggesting that allelic heterogeneity may be a common observation in the allelic architecture of human diseases and that aggregating signal across multiple variants in a locus may be an attractive strategy for association testing.

2.5 Aggregating signal across multiple variants

Approaches that aggregate signal across multiple variants are commonly referred to as 'aggregation tests' (Asimit and Zeggini, 2010). When designing an 'aggregation test', there are three main questions to consider:

- Across which biological units should variants be combined (subsection 2.5.1)?
- Which variants mapping within those units should be included (subsection 2.5.2)?
- Which statistical models should be used (subsection 2.5.3)?

2.5.1 Biological units

As for the first question, an obvious choice of biological unit across which to combine variants is the gene, considering all the coding sequence of the gene as the relevant unit.

An alternative would be to combine variants across all genes within a pathway, but this adds several complications, not least because of the rather imprecise knowledge of biological pathways.

Another strategy would be to combine variants across genes in sets of genes with prior established relationship to a disease. For example, genes causing Mendelian forms of the trait undergoing analysis (Kathiresan et al., 2008).

2.5.2 *In silico* prediction of the deleterious effect of a genetic variant

As for the second question, i.e. "which variants mapping within those units should be included?", within a gene one could include all observed variants, but this would be likely to include many neutral variants and/or variants with opposite direction

of effects, both of which can lead to loss of power under many statistical methods. The ideal approach would be to combine only those variants which affect the trait of interest. This is, however, difficult in practice because these will not be known in advance. One approximation would be to include only the variants with functional effects. Even this still remains challenging because of the limited knowledge of the function of coding variants in the human genome: commonly used predictors of the function of missense variants (Kumar et al., 2009; Adzhubei et al., 2010) can often be unreliable (Flanagan et al., 2010).

The solution of predicting the functional consequences of rare variants is referred to as "variant annotation". There are many ways that predictions about the functional consequences of DNA sequence variants can be made including incorporating measures of constraint (Samocha et al., 2014) to predicting their impact on protein structure (Kumar et al., 2009; Adzhubei et al., 2010). However, a basic property of variant annotation is assigning functional consequences to a variant based on its relationship to coding sequence in the genome and how it may change the coding sequence and affect the gene product (McCarthy et al., 2014).

At the time I started my thesis most of the software programs that currently exist for variant annotation were under development. Some of the widely used programs today include VEP (Variant Effect Predictor), PLINK/SEQ, ANNOVAR, and SNPEFF. The main difference between variant annotation programs is which variant annotation they choose to report and what labels of variant annotation they choose to support. For example, PLINK/SEQ supports annotation of non-canonical sites predicted to disrupt splicing, and others, like VEP, currently support regulatory feature annotation. Nonetheless, these programs largely address the problem of variant annotation in a similar fashion: i) a choice of a reference transcript set (examples include GENCODE (Harrow et al., 2012), REFSEQ, CCDS, and ENSEMBL) is made, ii) a set of DNA sequence variants is uploaded, and iii) the software program performs *in silico* translation for polyadenylated mRNAs that encode a protein and overlap the position of the DNA sequence variants. For any given gene multiple transcripts (commonly referred to as "isoforms") may exist. As a consequence a software program usually reports multiple annotations for any DNA sequence variant. It is common practice for the software program to pick the "worst" annotation (ranking is based on some common criteria for "deleteriousness" of a coding variant, i.e. protein truncating variants > missense > synonymous) and reports that choice as the annotation of the DNA sequence variant. The availability of multiple annotations for any given DNA sequence variant makes it complicated to choose the one that is relevant and

the choice of the "worst" annotation may be inappopriate as it is entirely possible that the transcript supporting that annotation is hardly ever expressed. Thus, understanding the functional consequences of genetic variants is critical for improving variant annotation.

In Figure 2.5 a diagram of commonly used variant annotations in rare variant studies is shown. Protein truncating variants, the focus of my thesis, is a label given to a group of variants that are predicted to have severe disruptive consequences on gene function. The members of the group include:

1. Single nucleotide variants (SNVs) that introduce a premature stop called nonsense variants,

2. SNVs that are predicted to disrupt splicing referred to as splice donor and splice acceptor (depending on the location of the disruption),

3. SNVs that are predicted to remove a stop codon referred to as "stop lost" (although these normally enlarge, rather than truncate the protein, these variants are usually grouped in PTVs because transcripts containing stop lost alleles may under degradation by the NMD machinery resulting in complete loss of function (Hentze and Kulozik, 1999)),

4. SNVs that are predicted to remove a start codon referred to as "start lost", and

5. indels that are predicted to shift the reading frame of a gene commonly referred to as "frameshift indels".

Protein altering variants is common nomenclature for variants that result in any alteration in the protein coding sequence of a gene. This is a more inclusive category that consists of: PTVs, SNVs that introduce a missense substition, and indels that remove or introduce an inframe codon, see Figure 2.5.

2.5.2.1 Protein truncating variants

One class of variants where functional prediction is assumed to be more straightforward are those which are predicted to truncate the resulting protein product. The common assumption is that PTVs are typically subject to nonsense-mediated decay (NMD), a cellular mechanism that detects nonsense mutations and prevents the expression of truncated proteins, resulting in loss of function of that copy of the protein (MacArthur et al., 2012). Hence, PTVs are commonly referred to as loss of

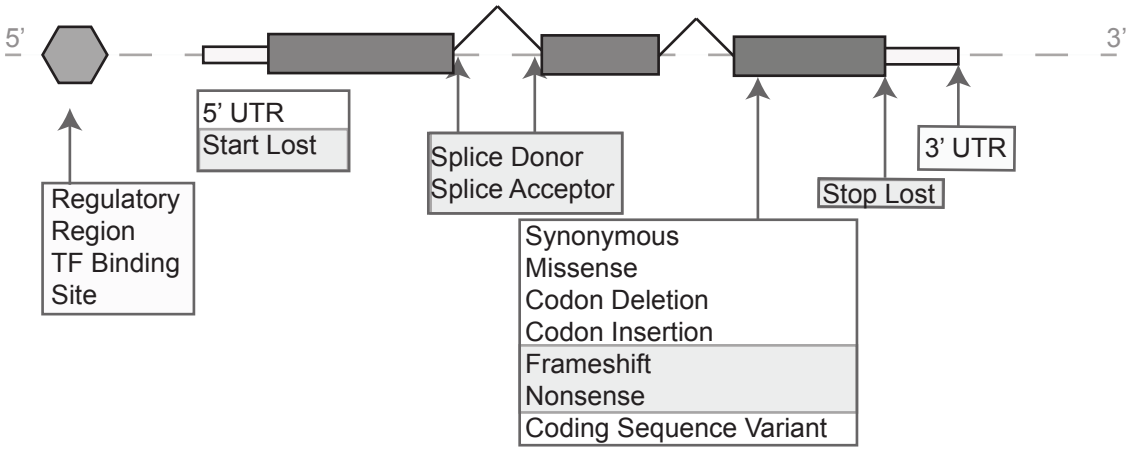


Figure 2.5: Diagram of variant annotation in rare variant studies. Protein truncating variants (light gray annotation box): Nonsense, Splice Donor, Splice Acceptor, Stop Lost, Frameshift, Start Lost. Protein altering variants: PTVs + Missense, Codon Deletion, Codon Insertion, Stop Lost, Start Lost. Dark boxes in figure are protein coding exons of a gene, white boxes of figure are the untranslated exons of a gene and light gray hexagon represents the regulatory region of a gene. Adapted from ENSEMBL.

function variants in the literature (MacArthur et al., 2012; Flannick et al., 2014). To a first approximation PTVs are likely to have the same functional consequence, so that variants in this class within a gene may naturally be combined in assessing their effect on a trait or phenotype of interest. However, a major problem persists in that it is not known in advance which variant will result in a truncating event and/or if the variant will result in loss of function of that copy of the protein. Furthermore, the functional consequences of predicted PTVs has not been systematically characterized nor quantified. As a result, the performance of the commonly used approaches to predict PTVs is not known.

A systematic survey of predicted proten truncating (loss of function) variants in the human genome. In 2012, in the middle of my thesis work, MacArthur et al. (2012) reported results from the first systematic survey of predicted protein truncating (loss of function) variants in the human genome. They found that a "typical" individual human genome is expected to carry about 100 protein truncating variants:

1. 24 large deletions[2],

[2]The number for large deletions and small indels are likely to be underestimated due to the contemporary state of indel and large deletion calling.

2. 38 small indels[2],

3. 12 splice site single nucleotide variants, and

4. 23 nonsense SNVs.

Moreover, MacArthur et al. (2012) estimated that 20 of the 100 PTVs exist in a homozygous state suggesting that 20 genes in the genome are completely dysfunctional in any given individual. At first, these numbers appear to indicate that a shockingly number of protein-coding genes are tolerant to truncation. However, careful examination shows that protein-coding genes containing homozygous PTV carriers are largely redundant, enriched for olfactory receptors, and appear to be protein-coding genes with low protein-protein interaction connectivity (MacArthur et al., 2012). In fact, from an updated analysis of unpublished data the average number of homozygous PTVs (that exist at a frequency $< 1\%$ in the population) in an individual appear to be less than 1 and not 20, which can be attributed to common PTVs in protein-coding genes with low evolutionary constraint (personal communication with Daniel MacArthur).

PTVs and their contribution to disease predisposition in 2010. In 2010, only a handful of PTVs had been associated to a common disease (excluding cancer).

In 2001, prior to the application of next generation sequencing technologies to association studies of common diseases, a frameshift indel in *NOD2* (common in the Ashkenazi Jewish population) had been found to be associated to Crohn's disease (Ogura et al., 2001).

Then, in 2006, Cohen et al. (2006) reported the discovery of loss of function variants in *PCSK9* associated with reduced levels (compared to the generation population) of low-density lipoprotein cholesterol (LDL-C) and protection against coronary heart disease. This result has piqued the interest of the field to study the association of PTVs to complex traits and has led to the development of new therapeutics (Stein et al., 2012; Crunkhorn, 2012). Today, over three drugs are in clinical trials including monoclonal antibody drugs: evolocumab (Amgen), 1D05-IgG2 (Merck & Co.), and alirocumab (Aventis/Regeneron). Preliminary results from clinical trials suggest that alirocumab cut roughly in half the number of heart attacks and strokes and could reach the market by 2015 (Reuters, 2014).

In 2009, the first genetic discovery of PTV association to a common disease using NGS was reported (Nejentsev et al., 2009). In Nejentsev et al. (2009) they found three PTVs in *IFIH1* that conferred protection to type 1 diabetes.

The second genetic discovery of PTV association to a common disease using NGS was reported in 2011. In Rivas et al. (2011), I and other colleagues, reported the association of a predicted splice disrupting variant in *CARD9* that conferred strong protection (OR = 0.29) to inflammatory bowel disease risk.

Germline PTVs and their contribution to cancer predisposition. A substantial (> 5) number of PTVs have been identified to be associated to cancer predisposition. These findings largely emerged prior to the development of next generation sequencing technologies (Rahman, 2014) where efforts were carried out to classify cancer-predisposing genes as tumor suppressors (Hanahan and Weinberg, 2000) or oncogenes (Stehelin et al., 1976). For example, in *BRCA1* and *BRCA2* our current understanding is that protein truncating variants largely predispose to breast and ovarian cancer disease risk (Hopper et al., 1999). However, not all PTVs confer the same risk. PTVs in the central part of *BRCA2* confer significantly higher relative risks of ovarian cancer compared with breast cancer than PTVs at either end of the gene (Rahman, 2014; Thompson and Easton, 2001). The exact cause of this observation is not known. Today, international efforts like The Cancer Genome Atlas and the International Cancer Genome Consortium have set out to obtain a comprehensive catalogue of variants responsible for cancer using NGS (Hudson et al., 2010; Network et al., 2011). However, these large projects have mainly focused on somatic variants. Our catalogue of germline PTVs contributing to cancer predisposition remains incomplete and despite the identification of some cancer predisposing genes (such as *TP53*) our knowledge of the associated cancer risks for germline PTV carriers of a substantial number of the identified genes is largely unknown (Rahman, 2014).

2.5.3 Models for rare variant association

As for the third question, i.e."which statistical models should be used?", at the time I started my thesis work, the problem of assessing rare variant association was addressed in two ways. The first approach involved aggregating genotypes or trait values across individuals that carry a rare variant in a functional unit. These approaches are commonly referred to as "collapsing" or "burden" tests. In Madsen and Browning (2009) they proposed a weighted sum statistic where weights are assigned to be inversely proportional to the minor allele frequency of the variant. In Price et al. (2010) the authors proposed a test that would choose the optimal minor allele frequency threshold cutoff for including a variant in the grouping.

More formally, let N_1 be the number of cases and N_2 be the number of controls. Define p_0, the probability observing a rare variant allele in the cases, to be $\frac{N_1}{N_1+N_2}$. Let m be the number of variants in a unit (e.g. gene), let w_i be the weight assigned to variant i, let y_i be the number of rare variant allelic copies in cases for variant i, and let n_i be the total number of rare variant allelic copies for variant i, then the burden test can be obtained with the following formulation

$$T_{\text{burden}} = \left(\sum_{i=1}^{m} w_i \left[y_i - n_i p_0 \right] \right)^2, \tag{2.1}$$

and a P value is obtained by comparing the T_{burden} statistic to a chi-square distribution with 1 degree of freedom (Lee et al., 2014). The burden test is especially powerful in the scenario where all variants confer risk to disease or where they all have similar effects. The major limitation of the test is that power can quickly decrease when null variants are introduced or when a mixture of effects exist (Neale et al., 2011; Basu and Pan, 2011).

Another approach is to scan for signal of dispersion for rare variants in a testing unit. In Neale et al. (2011) we presented a rare variant testing framework (we referred to it as the C-alpha test) that would allow a mixture of effects (risk or protective) to be present. We also presented a weighted version of the statistic in the spirit of the rare variant testing framework of Madsen and Browning (2009). However, the approach was not designed for quantitative traits; it was simply designed to handle case-control data.

More formally, I use the same setup as the burden test: let N_1 be the number of cases and N_2 be the number of controls. Define p_0, the probability observing a rare variant allele in the cases, to be $\frac{N_1}{N_1+N_2}$. Let m be the number of variants in a unit (e.g. gene), let y_i be the number of rare variant allelic copies in cases for variant i, and let n_i be the total number of rare variant allelic copies for variant i, then the C-alpha test can be obtained by aggregating the signal with the following formulation

$$T_{\text{calpha}} = \sum_{i=1}^{m} w_i^2 \left[(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0) \right]^2, \tag{2.2}$$

where w_i is a weight assigned to each of the variants using the same approaches described for the burden test. The T_{calpha} statistic is then standardized using the variance of T_{calpha} and the standardized statistic is used to obtain a P value using a normal approximation. However, in practice, Neale et al. (2011) suggested to evaluate a P value using permutation-based procedures because the statistic did not follow

asymptotic behavior (a common problem with statistical tests for rare variants given the small number of observations).

Since 2010 the development of new statistical frameworks for rare variant association studies has been an active area of research (Wu et al., 2011; Lee et al., 2014). None of the approaches available at the start of my thesis focused on the problem of testing for association by concentration on a group of variants with functional effects of interest like PTVs.

2.6 Aims

Throughout my thesis work my main objective was to develop approaches for improving our understanding of the medical relevance and functional consequences of PTVs, and to identify PTVs associated to disease or health related quantitative traits.

In the field of human genetics, at the start of my thesis work, neither software tools nor statistical methods for association analysis of PTVs across multiple study designs (case-control, single quantitative trait, cross-disorder, or multiple quantitative traits) were available. In the work presented in Chapter 3 my aim was to contribute to the development of new statistical, bioinformatic, and computational methods for association analysis of protein truncating variants across a broad range of study design scenarios. I developed the software package `MAMBA`; we designed and implemented the Bayesian similar effects model for assessing association between protein truncating variants and dichotomous traits, and the Bayesian similar effects and grouped effects model for assessing association between protein truncating variants and quantitative traits; and we designed and implemented the C-alpha multiple rare variants and phenotypes (C-alpha MRP) test for testing for association between protein truncating variants across a broad range of functional units and study design scenarios.

In the work presented in Chapter 4 my aim was to develop new statistical, bioinformatic, and computational methods to interpret the functional consequences of PTVs. To achieve that aim we designed and implemented the framework for variant annotation using empirical transcript isoform quantification with RNA-seq data, we designed and implemented the independent tissue model for assessing allele-specific expression across multiple tissues, and we developed and implemented the splice disruption model for assessing the impact of variants proximal to splice junctions on splicing - all of which were made available in the software package `MAMBA`.

The aim of the work presented in Chapter 5 was to identify PTVs conferring risk or protection to disease. To achieve that aim we applied the statistical models and

the sofware presented in Chapter 3 to a pooled targeted sequencing data set from a breast cancer study, a published targeted sequencing data set from a genetic study of six autoimmune diseases, and an exome sequencing data set from a genetic study of type 2 diabetes.

The aim of the work presented in Chapter 6 was to identify PTVs contributing to quantitative trait variation. To achieve that aim I applied the statistical models and the sofware presented in Chapter 3 to genotype and lipid phenotype data from a multi-ethnic type 2 diabetes case-control exome sequencing study, and to exome array genotype data and metabolomics phenotypes from the Oxford Biobank cohort.

The aim of the work presented in Chapter 7 was to improve annotation of protein truncating variants by studying their functional consequences. To achieve that aim we focused on the transcriptional consequences and applied the statistical models and software presented in Chapter 4 to RNA sequencing data sets from the Geuvadis RNA sequencing project and the Genotype Tissue Expression project.

I hope that the work presented in this thesis will be a step towards meeting the challenge of understanding the functional consequences of PTVs, understanding their role in disease predisposition and phenotype variation, and in developing statistical methods and bioinformatic tools for testing association in rare variant association studies.

Chapter 3

Development of statistical, bioinformatic, and computational methods for the analysis of protein truncating variants: Association analysis

In Chapter 2 I briefly reviewed the state of statistical methods for rare variant association analysis when I started my thesis in 2010. In this chapter I present statistical, bioinformatic, and computational methods that we developed to enable assessment of protein truncating variant association in rare variant association study designs. First, I present statistical methods for assessing association between protein truncating variants and disease, or a single quantitative trait, in a Bayesian framework. Then, I present methodology for assessing association between PTVs and multiple phenotypes for either cross-disorder analysis of multiple disease groups or cross-phenotype analysis of multiple quantitative traits. For each method I present the background and rationale, the intuition motivating the design of the statistical approach, the implementation, a series of simulations to assess the method's performance relative to other published methods, a description of the application of the method in this thesis, a summary of the method's limitations and foreseeable next steps, and provide attributions. In addition, for each method, I attempt to address these four questions:

1. What methods were available and what did they not do?
2. What did I do to solve it?
3. How did I address the opportunity?

4. What did I find?

3.1 Method for the analysis of protein truncating variants with dichotomous traits

3.1.1 Background and rationale

PTVs are known to be important for human disease as they are enriched for severe disease-causing mutations (Holbrook et al., 2004). However, the extent to which they contribute to common disease predisposition is unknown. To assess association between PTVs and dichotomous traits three main options were available at the beginning of my thesis work: 1) to perform case-control sequencing and apply a formal test by collapsing the counts of PTV alleles in a gene and test for association by applying Fisher's exact test or a burden test that gives higher weight to allelic variants with lower population frequencies (Madsen and Browning, 2009); 2) to perform case-control sequencing and to apply a formal test by treating the PTVs as separate variants and combining signals by applying a variance-component test to the data (Neale et al., 2011); and 3) to apply an informal approach, and one that has been shown to work in cancer studies, which is to rank the initial case only sequencing data set as a screen, to screen in other datasets that are publicly available, and then to apply a formal test in additional follow-up case-control sequencing datasets. Although the informal approach seems pretty obvious and merits consideration there are very limited number of studies pursuing this approach. There are many reasons for this: 1) the approach does not guarantee that other PTV copies will be observed in independent sample set (because these variants are usually private and appear in different settings, e.g. CNVs, complex indels); and 2) few researchers have the experimental setup to first perform case only sequencing, and then to apply formal tests in follow-up case-control resequencing data. Thus, we sought to develop a statistical method that would allow me and other investigators to apply a more quantitative approach to formally describe the PTV data in a case-control sequencing experiment, or to formally rank genes from a case only sequencing experiment when reference data sets were available for comparison.

A more quantitative way of describing PTV data and ranking genes is via a Bayesian approach. In a Bayesian framework the null and alternative models are compared via the Bayes factor. Using the frequentist tests, one potential strategy to prioritize the genes for follow-up and replication is to rank by P values. In such a

ranking exercise, the use of Bayes factors has considerable advantages over the use of P values. The interpretation of a P value obtained in a particular experiment depends on both the alternative hypothesis and the power of the statistical method used. In the current context, power will change with n, the number of PTVs observed at the gene: there will be more power to reject the null hypothesis (for a specific alternative hypothesis) at genes which happen to have more PTVs in the sample. As a consequence, a low P value will be more informative with more PTVs than the same P value would be for a gene with fewer PTVs. For this reason simple ranking of genes by P value is unlikely to be optimal. In contrast, BFs quantify the evidence for the alternative hypothesis as compared to the null (P values measure tail probabilities under the null), so can be compared naturally even for genes with different numbers of PTVs.

3.1.2 Intuition

We chose a Bayesian framework to assess association between protein truncating variants and dichotomous traits because it would allow us to address the three aims in the informal approach, i.e. to rank the initial data as a screen, then to incorporate publicly available datasets and reassess association, and finally to apply the models in additional datasets. Among the features we considered in the design of the method was the ability to easily incorporate summary level data such as PTV allelic counts. Furthermore, because prior density of effects can be incorporated under the alternative and the null model we wanted to incorporate a prior density that reflected our belief of effect sizes that we would prefer to detect in empirical data sets, i.e. PTVs with strong effects (odds ratios ≥ 4). This is of particular interest when attempting to identify signal of PTVs associated to rare diseases where effects may be especially strong. In addition, we wanted researchers to have the ability to detect protective PTV signals, which are of potential interest to pharmaceutical companies looking for drug targets. As a consequence, we considered two alternative prior densities in a single alternative model allowing for the detection of combined risk or combined protective PTV signals. In summary, we would assume similar effects between multiple protein truncating variants (either all protective or all risk), assume that under the alternative model effects were likely to be strong effects, and that the presence of risk and protective effects in PTVs in a single gene was very unlikely. To address the last assumption we simply assumed effects were the same between PTVs. However, in Section 3.2 I give a more nuanced treatment of PTVs.

3.1.3 Implementation

In this subsection I present the mathematical derivation of the statistical approach we developed for assessing association between protein truncating variants and dichotomous traits.

Setup. Let the gene be the biological unit, S be the number of case samples, R the number of control (or reference) samples, N the total number of samples, y the number of combined PTV allelic observations in a particular gene in cases, n the number of total PTV allelic observations in a particular gene in cases or controls.

Since $n \ll \min(S, R)$ we assumed that PTV alleles among cases and controls will be distributed according to a binomial distribution with n trials and success probability θ. Then the Bayes factor (BF) is obtained by comparing the marginal likelihoods for the two models:

$$BF = \frac{\int \binom{n}{y} \theta^y (1-\theta)^{n-y} g(\theta)\, d\theta}{\int \binom{n}{y} \theta^y (1-\theta)^{n-y} f(\theta)\, d\theta}, \tag{3.1}$$

where $f(\theta)$ is the prior density for θ under the null model and $g(\theta)$ is the prior density for θ under the alternative model.

Null model. Under the null model, the prior density for θ is a point mass at $\theta = \frac{S}{N}$ because that is the probability we expect to observe a PTV allele in cases, which is dependent on the case to control ratio in the rare variant study. If some bias was expected, maybe as a result of population stratification or familial sharing then a different prior density would be appropriate that reflects the uncertainty regarding the probability of observing a PTV allele in cases (in this implementation we have not taken any type of bias into consideration).

The marginal likelihood for y is given by

$$\int P(y|\theta) P(\theta)\, d\theta = \binom{n}{y} \left(\frac{S}{N}\right)^y \left(1 - \frac{S}{N}\right)^{n-y}. \tag{3.2}$$

Alternative model. Under the alternative model, we assumed that PTVs will all have the same function. This assumption implies that the PTVs will have the same direction of effect.

In order to allow either protective PTVs or risk PTVs in a gene we assumed that the prior on θ is a mixture of beta distributions:

$$\theta \sim \frac{1}{2}\text{Beta}\left(\alpha_1, \beta_1\right) + \frac{1}{2}\text{Beta}\left(\alpha_2, \beta_2\right). \tag{3.3}$$

Then, the marginal likelihood for y is given by

$$\begin{aligned}\int P\left(y|\theta\right) P\left(\theta\right) d\theta = \frac{1}{2}\binom{n}{y}\Bigg[&\frac{\Gamma\left(\alpha_1+\beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)}\frac{\Gamma\left(y+\alpha_1\right)\Gamma\left(n-y+\beta_1\right)}{\Gamma\left(n+\alpha_1+\beta_1\right)} \\ &+ \frac{\Gamma\left(\alpha_2+\beta_2\right)}{\Gamma\left(\alpha_2\right)\Gamma\left(\beta_2\right)}\frac{\Gamma\left(y+\alpha_2\right)\Gamma\left(n-y+\beta_2\right)}{\Gamma\left(n+\alpha_2+\beta_2\right)}\Bigg].\end{aligned} \tag{3.4}$$

As indicated in Equation 3.1, the Bayes Factor is then given by the ratio of the marginal likelihoods so that

$$BF_{\text{SEMCC}} = \frac{\frac{1}{2}\binom{n}{y}\left[\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\frac{\Gamma(y+\alpha_1)\Gamma(n-y+\beta_1)}{\Gamma(n+\alpha_1+\beta_1)} + \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)}\frac{\Gamma(y+\alpha_2)\Gamma(n-y+\beta_2)}{\Gamma(n+\alpha_2+\beta_2)}\right]}{\binom{n}{y}\left(\frac{S}{N}\right)^y\left(1-\frac{S}{N}\right)^{n-y}}. \tag{3.5}$$

We refer to this approach as the Bayesian similar effects model for case-control (SEMCC) data.

By setting the hyperparameters $\alpha_1, \alpha_2 = c\frac{\text{OR}}{\text{OR}+\frac{R}{S}}, \beta_1, \beta_2 = c\frac{\frac{R}{S}}{\text{OR}+\frac{R}{S}}$, we centered the mean of the prior on the point mass that corresponds to PTVs in aggregate having effect size OR.

We let OR= 4 (modeling risk PTVs) and OR= $\frac{1}{4}$ (modeling protective PTVs). Then, we let $c = 5$ (parameter controls the dispersion of the prior), the choice for $c = 5$ allows a symmetric prior centered on 4 or $\frac{1}{4}$ with tails having strong prior

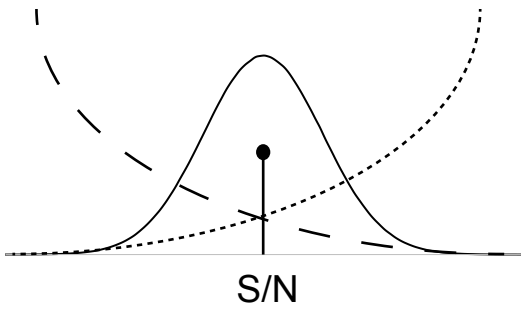


Figure 3.1: The density of the null distribution (solid curve) and a point mass reflecting the prior density of the null model with a point mass at $\theta = \frac{S}{N}$ (solid line). Mixture of alternative prior densities (dotted and dashed lines) when $R = S$: risk prior density (dotted) with hyperparameters $\alpha_1 = 4$ and $\beta_1 = 1$, and protective prior density (dashed) with hyperparameters $\alpha_2 = 1$ and $\beta_2 = 4$.

probability, which can be useful for prioritizing signals when studying rare diseases where strong effects may likely be present. The hyperparameters, $\alpha_1, \alpha_2, \beta_1$, and β_2 are calculated. The shape of the alternative prior densities is only centered at 4 and $\frac{1}{4}$, this does not mean that we believe all PTVs will have an effect size of 4. Throughout this thesis I use these hyperparameters to reflect our prior assumption that under the alternative model PTVs will have strong effects.

Software implementation. The SEMCC approach for assessing association between protein truncating variants and dichotomous traits is implemented in the software `MAMBA` using the module option `--module SEMCC`. It is implemented using the PyPLINK/SEQ python bindings and works directly from PLINK/SEQ project files (`https://atgu.mgh.harvard.edu/plinkseq/`). In the implementation variants annotated as `frameshift`, `esplice`, `start-lost`, `nonsense`, or `readthrough` are considered PTVs, and the allelic counts in a gene are counted across cases and controls. Alternatively, the method could be imported in Python using the command `from mamba.SEMCC import bfcasecon` where count data can be used.

3.1.4 Simulations

To study the behavior of ranking genes in a Bayesian model comparison framework compared to a frequentist framework I conducted a simulation study. First, I let S, the total number of case samples, equal R, the total number of control (or reference) samples, which is equal to $10,000$. Then, I simulated PTV allelic counts across a total of 20,000 genes where I introduced i) moderate effects (odds ratio (OR)=2; combined PTV allele frequency (AF) in controls = 0.03%) to 100 genes, ii) strong effects (OR=4; combined PTV AF = 0.03%) to a separate 100 genes, and iii) no effects (OR=1; combined PTV AF = 0.3%) to the remaining 19,800 genes.

For each gene I applied Fisher's exact test (Fisher, 1925) and the SEMCC model. I sought to compare the ranking of genes with introduced effects compared to the genes with no effects: for each rank, $k = 1, \ldots, 19900$, I computed a statistic, π, which is equal to the proportion of strong effect genes (or moderate effect genes) that achieve a ranking higher than rank k. I found that genes simulated under the alternative model of strong effects (or moderate effects) were more likely to be ranked higher by the Bayesian SEMCC approach than Fisher's exact test (Figure 3.2). I performed one additional simulation experiment where I increased the variance explained in liability scale by increasing the combined PTV allele frequency to 0.05%. The additional simulation experiment demonstrated similar results (Figure 3.2).

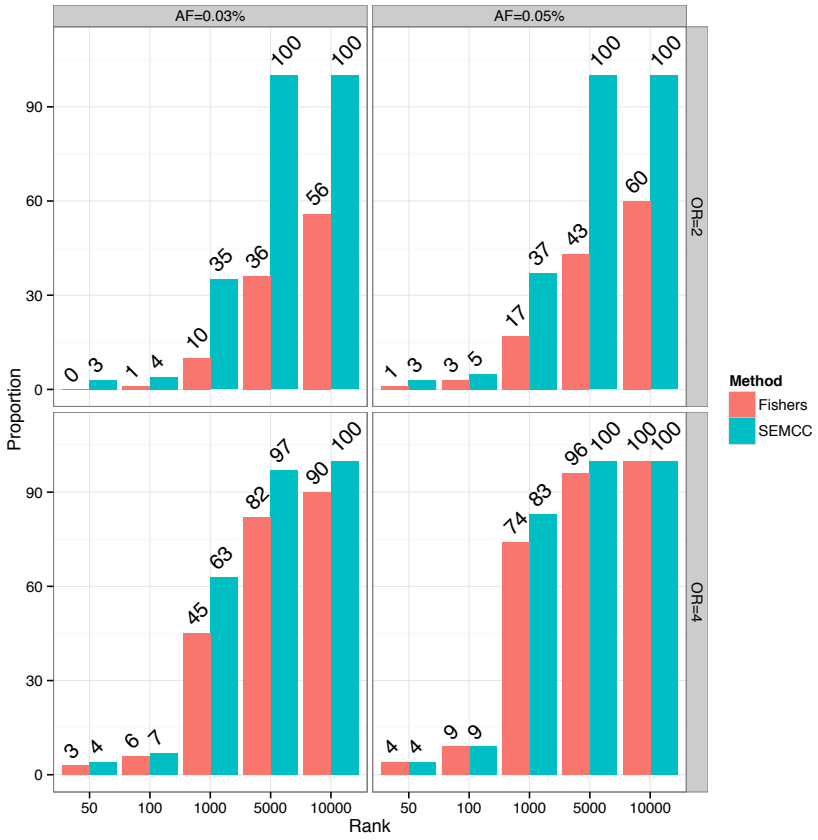


Figure 3.2: Comparison of ranking performance between Bayesian SEMCC (aqua) and Fisher's exact test (red). For each rank, $k = 1, \ldots, 19900$, I compute a statistic, π, equal to the proportion of genes simulated with strong effects that have a rank less than or equal to k. In this figure the simulation experiment is as follows. First, I let S, the total number of case samples, equal to R, the total number of control (or reference) samples, which is equal to $10,000$. I simulated PTV allelic counts across a total of 20,000 genes where I introduced i) moderate effects (odds ratio (OR)=2; combined PTV allele frequency (AF) in controls = 0.03%) to 100 genes, ii) strong effects (OR=4; combined PTV AF = 0.03%) to a separate 100 genes, and iii) no effects (OR=1; combined PTV AF = 0.3%) to the remaining 19,800 genes. One additional simulation experiment was conducted where I increased the variance explained in liability scale by increasing the combined PTV allele frequency to 0.05%. Barplots are shown of the statistic, π, for ranks $k = 50, 100, 1000, 5000$, and 1000 for the moderate effect genes (OR=2) compared to the no effect genes, and for the strong effect genes (OR=4) compared to the no effect genes.

3.1.5 Application of method in this thesis

In this thesis I applied the method, that I refer to as the Bayesian SEMCC model, to a breast cancer targeted sequencing study of PTVs in genes involved in the DNA repair pathway presented in Section 5.1. Then, I applied it to a multi-ethnic exome sequencing data set to assess association between PTVs and type 2 diabetes in Section 5.3.

3.1.6 Limitations and next steps

The Bayesian approach that we developed for assessing association between protein truncating variants and dichotomous traits has several limitations including: 1) it does not handle covariates such as principal components, age, or gender; 2) it does not handle relatedness; 3) it does not treat the recessive model explicitly; and 4) it does not allow variants to be grouped based on their predicted effects on nonsense-mediated decay, which is incorporated in the quantitative trait models presented in the next section. I envisage that these limitations will be addressed in future work.

3.1.7 Attributions

Together with my advisors and Matti Pirinen, I participated in designing the method. I developed the analytical framework, implemented it in the software `MAMBA`, and generated the simulations. Matti Pirinen provided guidance, and advised me in the development of the statistical method and derivation of the statistic.

3.2 Method for the analysis of protein truncating variants with quantitative traits

3.2.1 Background and rationale

At the beginning of my thesis work, in 2010, a small number of methods were available for the analysis of rare variants and quantitative traits (Bansal et al., 2010). The major area of development in the field dealt with the analysis of rare variants and dichotomous traits. In the setting of quantitative traits (QTs) a straightforward approach for the analysis of PTVs is to collapse the trait values of PTV carriers and compare against a standardised distribution, such as the standard normal. Another approach, is to assess association between PTVs and a single QT using a variance-component test like the Sequence Kernel Association Test (SKAT) (Wu et al., 2011).

The SKAT test was designed to be a generalization of the binomial C-alpha approach I co-developed prior to the beginning of my thesis work (Neale et al., 2011), which I later contributed to the development of the formal extension in Clarke et al. (2013), but is not introduced in this thesis. Similar to the Bayesian approach presented in Section 3.1 we sought to develop an approach that could easily be applied to obtain a quantitative way of describing the data and ranking genes in the two phases of a study, discovery and replication, and that was specifically designed for the analysis of protein truncating variants.

Some of the properties that we wanted the approach to have, that were not addressed in any of the published methods, included: i) the ability to incorporate prior belief about the effects that I expected to observe in empirical data sets, and ii) the ability to group variants based on their predicted effects on nonsense-mediated decay. This approach was published in *Bioinformatics* as an 'Original Paper' titled *Assessing association between protein truncating variants and quantative traits* (Rivas et al., 2013).

3.2.2 Intuition

In the QT setting, the Bayes factor reduces to a ratio of densities involving only trait values of the carriers of PTVs ($Y_{PTV} = \{Y_1, \ldots, Y_n\}$), because under both models the data on the non-carriers of PTVs has the same distribution. The main intuition behind the prior densities chosen for the alternative model is that the quantitative trait value of individuals carrying a PTV in a gene that contributes to phenotypic variation in the population will be towards the tails of the phenotype distribution. Furthermore, we wanted to consider the straightforward case in which all PTVs are assumed to have the same effect, which would be appropriate, for example, if all caused loss of function. However, this assumption is known not to hold (further research to improve the functional annotation of PTVs is presented in Chapter 7). As a consequence, we wanted to extend the straightforward approach to allow for the possibility that some PTVs have one effect and some have another, for example if most PTVs are subject to nonsense-mediated decay (NMD), and hence are loss of function, while some (for example in the final exon) escape NMD and can act as loss of function, gain of function (by deleting an inhibitory domain), or neutral variants. The approach could be extended to allow more than two groups of PTVs, with PTVs in each group having the same effect.

3.2.3 Implementation

In this subsection I present the mathematical derivation of the statistical approach we developed for assessing association between protein truncating variants and quantative traits in a single gene.

Setup. Assume that among N individuals studied, n individuals carried one of the k protein truncating variants (PTVs) observed in a gene considered. Typically $n \ll N$. Since PTVs are typically rare, we assumed that individuals carry at most one such variant. (Two PTVs on the same chromosome are likely to have the same effect as one PTV, and individuals with PTVs on both chromosomes could either be treated as a separate class in a recessive model (Lim et al., 2014a), or in the same class if a dominant model were thought appropriate.) The same PTV might be carried by several individuals.

We let $Y_1, \ldots, Y_N$ be the standardised quantitative trait values of the individuals and we assumed that the trait values $Y_1, \ldots, Y_n$ correspond to the carriers of PTVs and the values $Y_{n+1}, \ldots, Y_N$ correspond to the non-carriers of PTVs. We assumed that standardised trait values across the whole sample follow a standard normal distribution, which can be achieved by applying quantile normalisation. If the PTV has a big effect on the trait, individuals carrying the PTV may have extreme trait values, which would be moved closer to the other values under quantile normalisation. In this setting there is a potential loss of power through quantile normalisation. However, by doing so it allows the method to be more robust to potential outlier measurements, which may occur for many reasons beside genetics, for example measurement errors.

Null model. Under the null model, the gene does not affect the trait and the trait values of PTV carriers and the PTV non-carriers follow the standard normal distribution:

$$\text{NULL}: \quad Y_i \sim \mathcal{N}(0, 1^2), \text{ for } i = 1, 2, \ldots, n \tag{3.6}$$

$$Y_j \sim \mathcal{N}(0, 1^2), \text{ for } j = n+1, n+2, \ldots, N. \tag{3.7}$$

The statistical challenge is to look for strong evidence against the null model. If PTVs at the gene under consideration affect trait values this will cause a deviation from normality for $Y_1, \ldots, Y_n$.

The Bayesian approach requires specification of the alternative hypothesis. I present two alternate specifications of the alternative hypothesis: 1) the similar effects

model (SEM) and 2) the grouped effects model (GEM) with an extension where we use information about predictions of nonsense-mediated decay (GEM-NMD).

Similar effects model. Under the similar effects model (SEM), we assumed that the effect of a PTV is to shift the mean of the distribution of trait values, so that the trait values of the PTV carriers follow a normal distribution with mean μ and standard deviation s, whereas the trait values for the remaining individuals follow a standard normal distribution:

$$\text{SEM}: \quad Y_i \sim \mathcal{N}(\mu, s^2), \text{ for } i = 1, 2, \ldots, n \tag{3.8}$$

$$Y_j \sim \mathcal{N}(0, 1^2), \text{ for } j = n+1, n+2, \ldots, N. \tag{3.9}$$

Here, we fix the value of s, to $s = 1$, but more general approaches could also allow a change under the alternative hypothesis in the variance of trait values, or potentially in their distribution.

The distribution of the trait mean μ is specified under the alternative hypothesis. Since it will not be known in advance whether PTVs will increase or decrease trait values, as is the case for case-control analysis where it is not known if PTVs will confer risk or protection in advance, we used a 50:50 mixture of two normal distributions as a prior for μ:

$$\mu \sim \frac{1}{2}\mathcal{N}(-a, t^2) + \frac{1}{2}\mathcal{N}(a, t^2),$$

We let the hyperparameters be $a = 1.5$ and $t^2 = 0.5$. With these values 95% of the prior mass for μ lies in the set $(-2.89, -0.12) \cup (0.12, 2.89)$ following the signal the method is tailored for, i.e. trait values of individuals carrying PTVs in genes contributing to trait variation stronly deviate from normality.

In a Bayesian framework the null and alternative models are compared via the Bayes factor. The data on the non-carriers of PTVs has the same distribution under both models. As a result, a comparison of the densities involving the trait values of the PTV carriers is sufficient.

$$BF_{\text{SEM}} = \frac{Pr(Y|\text{SEM})}{Pr(Y|\text{NULL})} = \frac{Pr(Y_{PTV}|\text{SEM})}{Pr(Y_{PTV}|\text{NULL})} \tag{3.10}$$

$$= \frac{\int \prod_i f(Y_i|\mu, s^2)(0.5 f(\mu|-a, t^2) + 0.5 f(\mu|a, t^2))d\mu}{\prod_i f(Y_i|0, 1^2)} \tag{3.11}$$

$$= 0.5\, BF(a, t, s, \{1, \ldots, n\}) + 0.5\, BF(-a, t, s, \{1, \ldots, n\}), \tag{3.12}$$

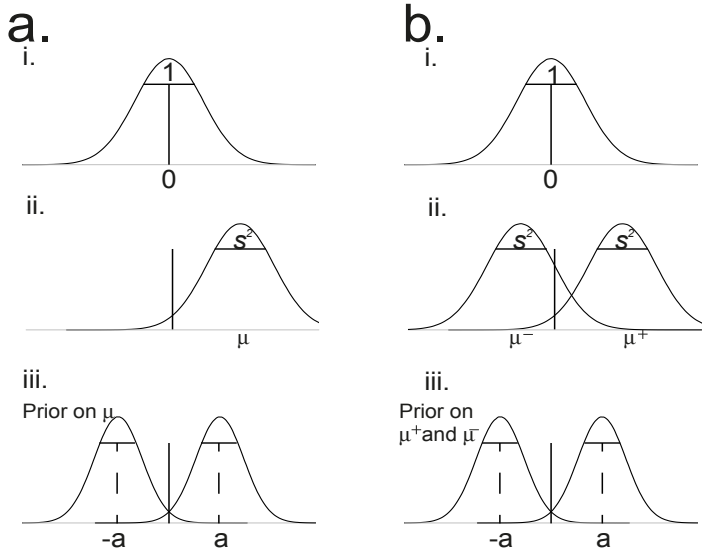


Figure 3.3: a) Prior and sampling distribution for the similar effects model (SEM): i) distribution of the trait values under the null model, $Y_i \sim \mathcal{N}(0,1)$; ii) distribution of the trait values under the alternative model, $Y_i \sim \mathcal{N}(\mu, s^2)$; iii) 50:50 mixture of two normal distributions as prior for $\mu \sim \frac{1}{2}\mathcal{N}(-a, t^2) + \frac{1}{2}\mathcal{N}(a, t^2)$. (b) Prior and sampling distribution for the grouped effects model (GEM): i) distribution of the trait values under the null model; ii) under the alternative model, trait values are grouped around μ^+ and μ^-; iii) priors for $\mu^+ \sim \mathcal{N}(a, t^2)$ and $\mu^- \sim \mathcal{N}(-a, t^2)$.

where $f(\cdot|d, v)$ is the density function of the normal distribution with mean d and variance v and

$$BF(a,t,s,I) = s^{-|I|}t^{-1}\left(\frac{|I|}{s^2}+\frac{1}{t^2}\right)^{-0.5} \tag{3.13}$$

$$\times \exp\left(-0.5a^2t^{-2} - 0.5\left(s^{-2}-1\right)\sum_i Y_i^2\right) \tag{3.14}$$

$$\times \exp\left(\frac{\left(\frac{\sum_i Y_i}{s^2}+\frac{a}{t^2}\right)^2}{2\left(\frac{|I|}{s^2}+\frac{1}{t^2}\right)}\right), \tag{3.15}$$

where the index i runs through the index set I whose size is $|I|$, i.e. the number of PTV carriers.

Grouped effects model. In some situations different PTVs can have different effects on the trait (Holbrook et al., 2004). For example, different PTVs can effect different transcript isoforms, either trigger or escape nonsense-mediated decay, or act as dominant gain of function variants, all of which have been observed in genes

predisposing to cancer (Futreal et al., 2004; Isidor et al., 2011; Ruark et al., 2012).

One way to approach the problem is to have an alternative model where PTVs that are predicted to escape NMD are placed in one of the groups with all other PTVs in the other group. A limitation of this approach is that variants that escape NMD may have a similar effect on trait values as variants that trigger NMD or that the predictions are incorrect (which will be discussed in Chapter 7) thereby resulting in loss of power to detect association.

The grouped effects model is an extension of the similar effects model that uses biologically informed weights to reflect the added information given by NMD predictions:

Let G be the set of all 2^k possible assignments of the k PTVs into two groups labelled $+$ and $-$. For any grouping $g \in G$, then let $I_+(g)$ and $I_-(g)$ be the sets of indexes of the individuals who carry PTVs that g assigns to groups $+$ and $-$, respectively. The model

$$\begin{aligned} \text{GEM}(g): \quad & Y_i \sim \mathcal{N}(\mu_+, s^2), \text{ for } i \in I_+(g) && (3.16)\\ & Y_i \sim \mathcal{N}(\mu_-, s^2), \text{ for } i \in I_-(g) && (3.17)\\ & Y_j \sim \mathcal{N}(0, 1^2), \text{ for } j > n. && (3.18) \end{aligned}$$

The priors are $\mu_+ \sim \mathcal{N}(a, t^2)$ and $\mu_- \sim \mathcal{N}(-a, t^2)$,with $a = 1.5$ and $t^2 = 0.5$, and again we keep $s = 1$ fixed. In other words, we assumed that the phenotype of an individual with a protein truncating variant in a relevant gene will be towards the tails of the distribution.

The Bayes factor between GEM(g) and the null model is

$$\begin{aligned} BF_{\text{GEM}}(g) &= \frac{Pr(Y|\text{GEM}(g))}{Pr(Y|\text{NULL})} = \frac{Pr(Y_{PTV}|\text{GEM}(g))}{Pr(Y_{PTV}|\text{NULL})} && (3.19)\\ &= \frac{\int \prod_{i\in I_+(g)} f(Y_i|\mu_+, s^2) f(\mu_+|a, t^2) d\mu_+}{\prod_{i\in I_+(g)} f(Y_i|0, 1^2)} && (3.20)\\ &\quad \times \frac{\int \prod_{i\in I_-(g)} f(Y_i|\mu_-, s^2) f(\mu_-|-a, t^2) d\mu_-}{\prod_{i\in I_-(g)} f(Y_i|0, 1^2)} && (3.21)\\ &= BF(a, t, s, I_+(g)) \times BF(-a, t, s, I_-(g)) && (3.22) \end{aligned}$$

The final GEM is a linear combination of submodels GEM(g) and we chose a weight $w_g \geq 0$ for each grouping $g \in G$ with the constraint $\sum_g w_g = 1$.

The corresponding Bayes factor is

$$BF_{\text{GEM}} = \sum_{g \in G} w_g BF_{\text{GEM}}(g). \tag{3.23}$$

We let GEM-NMD be the approach that uses biologically informed weights based on NMD predictions. We let $I_1(I_0)$ be the indexes of individuals carrying variants that are predicted to trigger (escape) NMD. We defined GEM-NMD by using $w_g \propto 1$ for the groupings $g \in G$ in which individuals in I_1 and individuals in I_0 belong to different groups and by using $w_g \propto 2^{-\min\{|I_+(g)|,|I_-(g)\}}$ for other groupings. Thus, for similar effects $w_g \propto 1$ as $\min\{|I_+(g)|, |I_-(g)\} = 0$ and the farther away the groupings are removed from the predictions the smaller the value of w_g.

Software implementation. To analyze exome and/or genome sequencing data sets with the Bayesian models I made a friendly implementation in the software `MAMBA` where hyperparameter priors `--a [1] --t [0.71]` can be specified by the user. For example, if the expected genetic effect is quite uncertain an analyst may want to increase the prior variance introduced or if the analyst expects small effects then he/she could decrease the prior shift in mean that is expected. The implementation allows generation of gene-based manhattan plots, output with annotation of genetic variants that were used, and variants may be annotated with any reference transcript set as long as it is properly loaded to the PLINK/SEQ project[1]; examples of reference transcript set include RefSEQ, Gencode, and ENSEMBL. The implementation also gives the option to the user to annotate their variants based on transcript isoforms that are expressed in their tissue of interest (and exclude those that are not) by using the option `--tissue`. Tissues included for filtering are updated with the latest release of the GTEx project[2]. In Section 4.2 I present results as to how this information may be incorporated for variant annotation. The module for the model presented in 3.2 is `--module SEMGEM`.

3.2.4 Simulations

I performed power comparisons among the commonly used frequentist approaches (SKAT (Wu et al., 2011); SKAT-O (Lee et al., 2012); a weighted version of SKAT where for variants predicted to trigger NMD I assigned a weight of $\sqrt{2}$, and for the variants predicted to escape NMD I assigned a weight of 1 [SKATw]; multiple linear regression as implemented in R; and collapse) and the Bayesian models described in

[1] `https://atgu.mgh.harvard.edu/plinkseq/locdb.shtml`
[2] `http://gtexportal.org`

this section (SEM, GEM, GEM-NMD). The collapse approach is simply to use the statistic

$$T = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2, \tag{3.24}$$

which follows a χ^2_1 distribution under the null. For SKAT I used the default weight parameters of Beta(1,25) for the linear weighted kernel and Davies' method to compute the P value (Davies, 1980), SKAT-O (Lee et al., 2012) using method="optimal". To be able to compare the methods I included the use of Bayes factors as frequentist test statistics.

I generated simulations for two scenarios:

1. I consider the scenario where all the PTVs in a gene have the same effect on the trait,

2. I consider the scenario where we assumed that PTVs in a gene contribute to trait variance and can have an effect in opposite directions with the effect specified by impact of NMD (here we assume that a variant predicted to trigger NMD decreases trait value, whereas a variant predicted to escape NMD increases trait value).

I simulated $3-5$ PTVs each seen in a single individual. Trait values for PTV carriers were drawn from an $\mathcal{N}(2,1)$ distribution, and the remaining trait values for a total of $2,000$ trait observations were drawn from an $\mathcal{N}(0,1)$. In the second scenario, trait values were drawn from an $\mathcal{N}(-2,1)$ distribution for individuals with a copy of a variant predicted to trigger NMD, and $\mathcal{N}(2,1)$ for individuals with a copy of a variant predicted to escape NMD, and $\mathcal{N}(0,1)$ otherwise. I selected only one variant to escape NMD.

To evaluate power I repeated the procedure 100 times, and for the Bayesian model I generated $1,000,000$ replicates under the null to evaluate a P-value. I evaluated power (given in percentage) at an α level equal to 0.001.

The simulation model was chosen to assess the performance of the developed method compared to known frequentist approaches in scenarios we thought it would be most suitable to employ, i.e. rare PTVs with strong effects and the possible scenario where PTVs in a gene may have opposite effects as a result of differences in their functional consequences. In practice it is likely that not all PTVs within a gene will

$\alpha = 0.001$	i) similar			ii) grouped		
NPTV	3	4	5	3	4	5
SKAT	42	49	69	44	55	66
SKAT-O	53	65	86	37	45	60
Collapse	55	81	86	2	0	1
SEM	55	81	86	2	0	1
GEM-NMD	-	-	-	51	65	79
SKATw	-	-	-	28	46	56
Multiple linear regression	42	49	69	44	55	66

Table 3.1: Power (given in percentage) at $\alpha = 0.001$ to detect association for the two scenarios specified in the text. 1,000,000 replicates were generated to obtain a P-value.

have an impact on trait variation. In this scenario both the frequentist and Bayesian model comparison approach will lose power to detect association.

For the first scenario, where all PTVs in a gene have the same effect on the trait, SEM and the collapse test were the most powerful: $55\%, 81\%$, and 86% power for 3, 4, and 5 PTVs, respectively (Table 3.1). GEM-NMD, which uses predictions of NMD was the most powerful model for the grouped simulated scenario (Table 3.1).

3.2.5 Application of method in this thesis

In this thesis I applied the Bayesian models for assessing association between protein truncating variants and quantitative traits in Chapter 6. In Section 6.1 I applied the models to a multi-ethnic exome sequencing data set and the standard lipid profile. Then, in Section 6.2, I applied the models to an exome array genotyping data set and 123 metabolite measurements.

3.2.6 Limitations and next steps

The approach that we developed for assessing association between protein truncating variants and quantiatative traits has several limitations including: 1) it does not handle covariates directly, although I envisage that when applied the trait measurements used could reflect adjustment for covariates such as age, gender, and principal components, and 2) it does not handle relatedness. I envisage that these limitations may be addressed in future work.

3.2.7 Attributions

Together with my advisors and Matti Pirinen, I participated in designing the method. I participated in developing the analytical framework, implemented it in the software `MAMBA`, and generated the simulations. Matti Pirinen provided the analytical derivation for the similar effects model, advised me in the additional extensions of the method including the GEM and GEM-NMD models, and evaluated how this Bayesian approach compares to frequentist approaches in ranking of genes (presented in Rivas et al. (2013), but not discussed in this section). Kyle Gaulton and Loukas Moutsianas provided comments and feedback.

3.3 Methods for the analysis of protein truncating variants and multiple phenotypes

3.3.1 Background and rationale

In Sections 3.1 and 3.2 I considered how to assess association between protein truncating variants and a single phenotype. As far as I was aware, at the time I started my thesis work, there were no approaches that addressed the problem of assessing association between rare variants and multiple phenotypes. In this section I consider the problem of assessing association between multiple rare variants and their effects across multiple, possibly correlated phenotypes. In Sections 3.1 and 3.2 I presented Bayesian approaches for assessing association between protein truncating variants and a phenotype. In this section I present a frequentist approach - the main reason being that at the time I decided to work on this problem I had just completed collaborating in developing the Generalized C-alpha test presented in Clarke et al. (2013). Hence, it was a matter of convenience to show proof of principle in the approach. However, as shown in the latter section and in Future directions, extending it in a Bayesian framework is a worthwhile area of future research.

This opportunity, i.e. assessing association between multiple rare variants and multiple phenotypes, arises in sequencing association study designs of related disorders or population biobanks where a large number of health related measurements are made. Analyzing multiple phenotypes simultaneously may be useful for improving power to detect genetic associations (Stephens, 2013). However, a major challenge that remains, even after detecting association, is how to interpret the associations and estimate the effects of multiple variants and their impact across multiple phenotypes. The estimation exercise can highlight, for example, if a protein truncating

variant that confers protection to one disease may confer risk to another, which may be interesting to those in understanding potential side effects of inhibiting a putative drug target. Thus, I set out to address both opportunities for the analysis of rare variants and multiple phenotypes.

In this section I present an alternative to current widely-used statistical tests, a test, that I referred to as the C-alpha test for multiple rare variants and phenotypes (MRP), that exploits correlation of genetic effects for association in case-control, cross-disorder, single-trait, and multi-trait rare variant association study designs, Figure 3.4. We formulated the approach as an instance of homogeneity C-alpha tests of Neyman and Scott (1966) and Zelterman and Chen (1988). Derivation of the test combines the information over multiple variants residing in a biologically defined genomic unit as well as over dichotomous and quantitative phenotypes providing a general framework for association testing in rare variant association studies. This test exploits knowledge about the correlation of genetic effects across 1) a group of genetic variants and 2) a group of traits. Properties of the approach, as implemented, include that it is computationally efficient, making the analysis of large study designs practical; it is flexible and extensible making the analysis of gene-sets, pathways, and networks feasible; and it includes standard univariate and multivariate gene-based tests as a special case. As a consequence, this approach can be employed even where estimates of the expected correlation of genetic effects are not available. In addition, we present a clustering algorithm to estimate how variants are impacting the multiple traits profile.

We studied the behavior of the test and algorithm through simulations, and in Chapters 5 and 6 I will illustrate its utility through applications to a targeted sequencing data set of autoimmune cases (Section 5.2), exome sequencing study of lipid traits (Section 6.1), and an exome array study of metabolic traits measured using NMR spectroscopy (Section 6.2). The hope is that the methods presented in this section will be valuable tools for studies of complex diseases, particularly where high-dimensional genetic and phenotype data is available and that it will initiate further research into extensions of this test.

3.3.2 Intuition

I describe the intuition behind the test and present additional details including derivation of the test in the next subsection.

Let's assume we are studying k diseases or quantitative traits and m variants. Following Zelterman and Chen (1988), a score test statistic is derived, T, to test

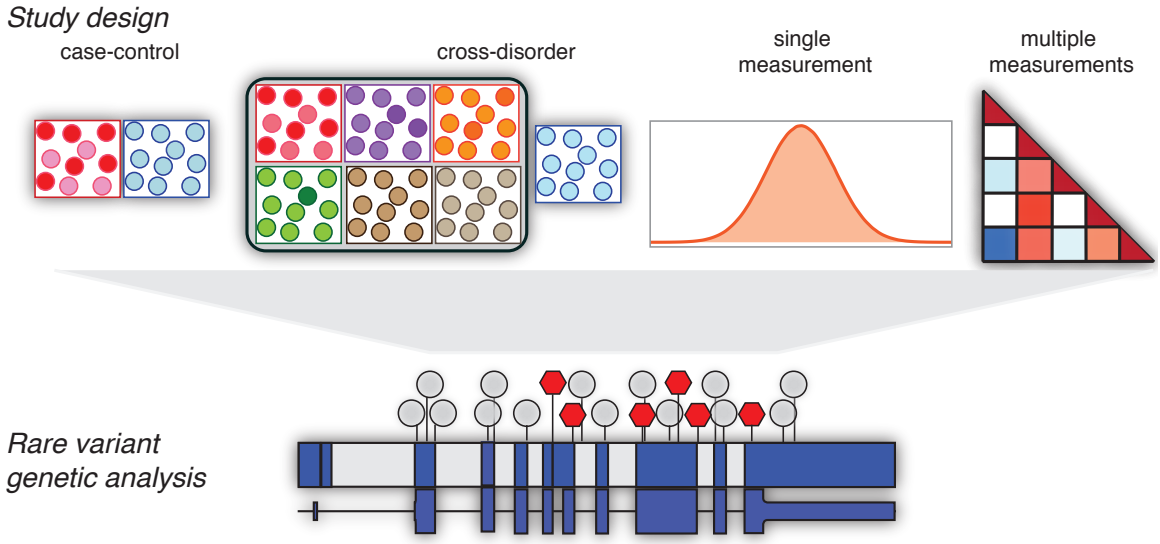


Figure 3.4: Schematic diagram of the statistical framework proposed for rare variant association studies. Family of C-alpha tests presented in this thesis can be applied to a broad range of rare variant association study designs including: case-control (top left); cross-disorder (e.g. autoimmune diseases, top second left); single continuous trait (e.g. BMI, top third from left); and cross-phenotype (e.g. metabolomics, top far right) studies. The statistical tests can incorporate information on expected correlation of effects across phenotypes and/or across rare variants, in particular, rare protein truncating where variants may be assumed to have correlated effects across a phenotype (red hexagons, bottom).

whether the distribution of effect sizes has non-zero variance. Previously, with other colleagues, we used this framework to derive a statistic for rare variant analysis of case-control data, referred to as Binomial C-alpha, or rare variant analysis of a single quantitative trait, referred to as Generalized C-alpha, and here we extend the test to high-dimensional settings $k > 1$ (Neale et al., 2011; Clarke et al., 2013). The multivariate scenario allows the investigators to choose the correlation structure of the effect sizes against which the test is most powerful, denoted by a km-by-km matrix $\mathbf{U}$. In practice, this matrix $\mathbf{U}$ can be defined as a Kronecker product of two sub-matrices[3]: an m-by-m matrix $\mathbf{R}_{\text{var}}$ defining the expected correlations of genetic effects across the variants on any single trait, and a k-by-k matrix $\mathbf{R}_{\text{phen}}$ defining the expected correlations of genetic effects of a single variant across the traits (Figure 3.5).

The intuition is that if these prescribed correlations are approximately valid in practice, it can increase the power of the test by appropriately combining information across variants and/or across traits. In $\mathbf{R}_{\text{var}}$ the expected correlations between the effects of different variants on any one trait is assigned. For example, for multiple protein truncating variants in the same gene we may set the pairwise correlations of genetic effects to 1 to reflect the assumption that all the PTVs in a gene have the same biological consequence. In some other setting, we may indicate no prior knowledge about the pairwise correlation of genetic effects by setting the expected correlation of genetic effects to 0 and leaving $\mathbf{R}_{\text{var}}$ as a diagonal matrix. It is easy to see that to mimic the grouped effects model (presented in Section 3.2) we can set the expected correlation of genetic effects between a PTV predicted to trigger NMD and a PTV predicted to escape NMD to 0. An interesting alternative scenario is to assume negative genetic correlations, which assume that a pair of variants has opposite effect on gene activity. Such information can be obtained through integration of additional data sources, such as functional assay data (Findlay et al., 2014; Majithia et al., 2014).

In $\mathbf{R}_{\text{phen}}$ the expected correlations between the effects of a single variant on multiple traits is assigned. For instance, in the analysis of two disease groups, we may expect that a variant will likely have a similar effect on both diseases. The extreme case is to set the correlation to 1, which is equivalent to treating the two diseases as a single disease (not presented in this thesis). Alternatively, pairwise correlation of genetic effects can be estimated from common variant data and the intuition is that

[3] `http://en.wikipedia.org/wiki/Kronecker_product`

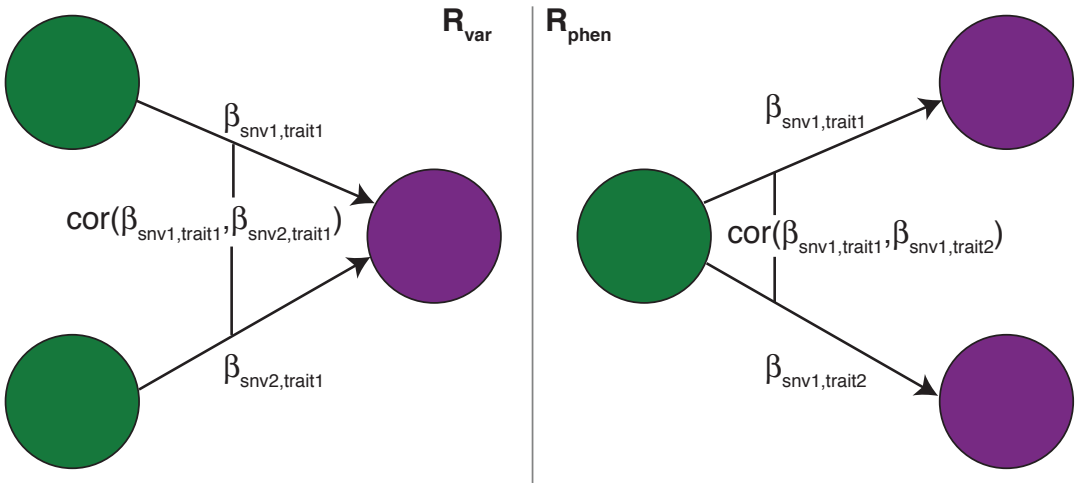


Figure 3.5: Diagram of the expected correlation of genetic effects matrices: $\mathbf{R}_{\text{var}}$ expected correlation of genetic effects between a pair of variants on a trait and $\mathbf{R}_{\text{phen}}$ expected correlation of genetic effects from a single variant on a pair of traits.

we can improve power to detect association incorporating this information (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). If no additional information is available, like the estimates of the genetic correlation of effects, we propose setting $\mathbf{R}_{\text{phen}}$ to the diagonal matrix in the cross-disorder analysis where we may actually gain power to detect association over the single case group versus control analysis, or to the total correlation estimated from the data in continuous trait analysis (as these usually correlate with the pairwise estimates of correlation of genetic effects (Pirinen, 2015)).

After detecting signal of association there are two key questions we are interested in addressing:

1. What variants are driving the association signal?

2. Which phenotypes are associated to the variant?

In Neale et al. (2011) an EM algorithm was introduced to address these questions with case-control data. The EM algorithm results can be used to identify which variants are most likely detrimental or protective in the case-control setting. In the multiple phenotypes setting we developed a Markov Chain Monte Carlo algorithm to identify which variants are driving the signal and the relationship between the variants and the multiple phenotypes collected. This algorithm is a clustering algorithm where the aim was to cluster variants based on the phenotypic profile of individual carriers of the variants.

3.3.3 Implementation

In this subsection I present the mathematical derivation of the C-alpha MRP test for cross-disorder and cross-phenotype analysis of quantitative traits. In addition, I present the MCMC algorithm for clustering variants based on the phenotype profile of individual carriers. Finally, I give a description of the bioinformatic implementation.

C-alpha MRP test for cross-phenotype association analysis. Let n be the number of individuals and k be the number of quantitative trait measurements on each individual, for example, it is common to measure HDL-C, LDL-C, and triglycerides as part of a lipid profile in which case $k = 3$. Let m be the number of variants in a testing unit $\boldsymbol{G}$, where $\boldsymbol{G}$ can be, for example, a gene, pathway, or a network.

A multivariate linear regression model in matrix form is

$$\underset{(n\times k)}{\mathrm{Y}} = \underset{(n\times m)}{\mathrm{X}}\,\underset{(m\times k)}{\mathrm{B}} + \underset{(n\times k)}{\mathrm{E}}, \tag{3.25}$$

where the matrices $\mathbf{Y} = [y_{is}]$, $\mathbf{X} = [x_{i\ell}]$, $\mathbf{B} = [\beta_{\ell s}]$ and $\mathbf{E} = [e_{is}]$ describe the trait values (y_{is}), copies of minor allele ($x_{i\ell}$), variant-trait effects ($\beta_{\ell s}$), and residual errors (e_{is}), for individual i, trait s and variant ℓ. Assuming that each trait has been transformed to a standard normal distribution and that the mean trait value among the individuals who carry the major allele (the allele that has the greatest frequency among all the alleles in a given population) at every locus studied is zero we drop the intercept terms from the model.

Let

$$\mathbf{V_X} = (\mathbf{X}^\intercal\mathbf{X})^{-1} \tag{3.26}$$

in which case the least-squares estimate of β coefficients was given by

$$\hat{\mathbf{B}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\,\mathbf{X}^\intercal\mathbf{Y} = \mathbf{V_X}\mathbf{X}^\intercal\mathbf{Y}, \tag{3.27}$$

which is the ordinary least-squares estimate of β coefficients[4] (more complex models can be derived with generalized least squares or instrumental variables). An unbiased estimator of the residual variance-covariance matrix (Rice, 2006) is

$$\mathbf{V_Y} = \frac{1}{n-m}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)^\intercal\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right). \tag{3.28}$$

Asymptotically the variance-covariance matrix of the $m \times k$ elements of $\hat{\mathbf{B}}$ is the Kronecker product between $\mathbf{V_X}$ and $\mathbf{V_Y}$ (Rice, 2006; Heij et al., 2004):

$$\hat{\mathbf{\Sigma}} = \mathrm{Var}\left(\hat{\mathbf{B}}\right) = \mathbf{V_X} \otimes \mathbf{V_Y}, \tag{3.29}$$

[4] http://en.wikipedia.org/wiki/Linear_regression

which can be represented as an $m \times m$ block matrix where each block is a scalar multiple of the $k \times k$ matrix $\mathbf{V_Y}$:

$$\hat{\mathbf{\Sigma}} = \mathbf{V_X} \otimes \mathbf{V_Y} = \left[\begin{array}{c|c|c|c} v_{11}^X \mathbf{V_Y} & v_{12}^X \mathbf{V_Y} & \cdots & v_{1m}^X \mathbf{V_Y} \\ \hline v_{21}^X \mathbf{V_Y} & v_{22}^X \mathbf{V_Y} & \cdots & v_{2m}^X \mathbf{V_Y} \\ \hline \vdots & & \ddots & \vdots \\ \hline v_{m1}^X \mathbf{V_Y} & v_{m2}^X \mathbf{V_Y} & \cdots & v_{mm}^X \mathbf{V_Y} \end{array}\right]. \tag{3.30}$$

Each $k \times k$ diagonal block is a covariance matrix of effects of one variant on k traits, whereas the off-diagonal blocks contain the covariances between the estimated regression coefficients for different variants. We used the asymptotic covariance matrix in deriving a test statistic but since we considered rare variants we did not rely on large sample results in statistical inference; instead we used permutation based methods throughout.

Let $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^\intercal, \ldots, \hat{\boldsymbol{\beta}}_m^\intercal\right)^\intercal$ be a vector of point estimates of all mk effects, ordered in m blocks of variant-specific k-dimensional effect estimates. Asymptotically, under the null model, $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(0, \hat{\mathbf{\Sigma}}\right)$ and under the alternative model, $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \hat{\mathbf{\Sigma}}\right)$ for some $\boldsymbol{\beta}$ for which $\beta_{\ell s} \neq 0$ for some ℓ and s. We considered a multivariate normal density because it would allow us to describe the set of (possibly) correlated random variables β (effect sizes), and plenty of literature exists on estimating parameters of a multivariate distribution. The multivariate normal probability density function[5] is given by

$$f\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\right) \propto \hat{\mathbf{\Sigma}}^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\intercal \hat{\mathbf{\Sigma}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]. \tag{3.31}$$

Under the alternative model the gradient (a vector operator denoted by ∇[6]), which gives the first derivative of the density, is

$$\nabla f\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\right) = f\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\right) \hat{\mathbf{\Sigma}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right), \tag{3.32}$$

and the matrix of the second derivatives (Hessian) is

$$H\left(f\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\right)\right) = f\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}\right) \mathbf{A}, \tag{3.33}$$

where

$$\mathbf{A} = \left(\left(\hat{\mathbf{\Sigma}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)\left(\hat{\mathbf{\Sigma}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^\intercal - \hat{\mathbf{\Sigma}}^{-1}\right). \tag{3.34}$$

The first and second derivatives for the probability density function of a multivariate normal is provided in "The matrix cookbook" (Petersen et al., 2008).

[5] http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Estimation_of_parameters

[6] http://en.wikipedia.org/wiki/Gradient

In Zelterman and Chen (1988), the tests of central mixtures over several parameters is presented in Section 4. In that paper they denote $Y_1, \ldots, Y_n$ as the independent random variables with respective probability density functions $f_i\left(y_i|\boldsymbol{\lambda}_i\right)$, conditional on a k dimensional parameter λ_i. Under the null hypothesis H_0, all $\boldsymbol{\lambda}_i$ are equal to a fixed vector $\boldsymbol{\lambda}_0$. Given their notation, the score-test statistic derived is of the form

$$\begin{aligned} T_{\text{Zelterman and Chen (1988)}} = & \frac{1}{2}\sum_{i=1}^{n}\sum_{s=1}^{k}\sigma_{ss}\frac{\partial^2 f_i\left(y_i|\boldsymbol{\lambda}_0\right)}{\partial\lambda_s^2}\Big/ f_i\left(y_i|\boldsymbol{\lambda}_0\right) \\ & + \sum_{i=1}^{n}\sum_{s<t}\sigma_{st}\frac{\partial^2 f_i\left(y_i|\boldsymbol{\lambda}_0\right)}{\partial\lambda_s\lambda_t}\Big/ f_i\left(y_i|\boldsymbol{\lambda}_0\right), \end{aligned} \tag{3.35}$$

where σ_{ss} and σ_{st} denote expected correlation of genetic effects for all k dimensions within each independent random variable. In our formulation $n = 1$ since we treat all mk effects (across variants and traits) in a single density, i.e. we only have one independent random variable: an mk vector $\hat{\boldsymbol{\beta}}$. Testing against the null hypothesis, i.e. $\boldsymbol{\beta} = \mathbf{0}$, using the score-test derivation in Zelterman and Chen (1988) is then

$$T_{\text{Zelterman and Chen (1988)}} = \frac{1}{2}\sum_{i=1}^{1}\sum_{s=1}^{mk}\sigma_{ss}\frac{f\left(\hat{\beta}|\beta\right)\mathbf{A}_{(s,s)}}{f\left(\hat{\beta}|\beta\right)} + \sum_{i=1}^{1}\sum_{s<t}\sigma_{st}\frac{f\left(\hat{\beta}|\beta\right)\mathbf{A}_{(s,t)}}{f\left(\hat{\beta}|\beta\right)}. \tag{3.36}$$

This test statistic can be represented as a sum over all elements of the matrix $\mathbf{U}\odot\mathbf{A}$, where $\odot$ denotes the Hadamard product of two matrices (element wise multiplication) and $(mk)\times(mk)$ symmetric matrix $\mathbf{U}$ collects the correlations between the random effects $\boldsymbol{\beta}$ under the alternative model against which the test is most powerful. This gives the derivation of the test statistic that I use throughout the thesis in the applications:

$$T = \text{sum}\left(\mathbf{U}\odot\mathbf{A}\right) = 2T_{\text{Zelterman and Chen (1988)}}. \tag{3.37}$$

The intuition behind the choice of $\mathbf{U}$ was presented in Section 3.3.2. To remind the readers I repeat how the matrix $\mathbf{U}$ may be defined: one way to define the correlation matrix $\mathbf{U}$ is through a $m\times m$ correlation matrix $\mathbf{R}_{\text{var}}$ that describes the expected correlations between the effects of different variants on any one trait and an $k\times k$ correlation matrix $\mathbf{R}_{\text{phen}}$ that describes the correlations of the effects of a single genetic variant on the traits : $\mathbf{U} = \mathbf{R}_{\text{var}}\otimes\mathbf{R}_{\text{phen}}$. In this case the test statistic is

$$T = \text{sum}\left(\left(\mathbf{R}_{\text{var}}\otimes\mathbf{R}_{\text{phen}}\right)\odot\mathbf{A}\right). \tag{3.38}$$

In the implementation P values are assessed by permuting the trait vectors among the individuals.

C-alpha MRP tests for cross-disorder association analysis. Suppose that we have case collections from k diseases (groups labeled $1, \ldots, k$, with $n_1, \ldots, n_k$ individuals, respectively) and one group of controls (group 0 with n_0 individuals). As with quantitative traits, we are interested in studying m rare variants within a testing unit $\boldsymbol{G}$. This approach generalizes the C-alpha test of Neale et al. (2011) to multiple case collections and to correlated effect sizes between different variants.

Let $p_s = n_s/n$ be the proportion of samples in group s, where $n = n_0 + \ldots + n_k$. Denote by $\beta_{\ell s}$ the log of the ratio of frequencies of the rare allele at locus ℓ between the case group s and the controls. (For rare variants this is approximately the commonly-used log-odds ratio parameter.) Suppose that at locus ℓ, $y_{\ell s}$ rare variants are observed in group s with $y_\ell = y_{\ell 0} + \ldots, y_{\ell k}$. Then the density function for the observations is

$$f_\ell(\boldsymbol{y}_\ell|\boldsymbol{\beta}_\ell) = \frac{y_\ell!}{y_{\ell 0}! \cdots y_{\ell k}!} \pi_{\ell 0}^{y_{\ell 0}} \cdots \pi_{\ell k}^{y_{\ell k}}, \tag{3.39}$$

where $\pi_{\ell 0} = c_\ell p_0$ and $\pi_{\ell s} = c_\ell p_s \exp(\beta_{\ell s})$ for $s = 1, \ldots, k$ with a normalizing constant

$$c_\ell = \left(p_0 + p_1 \exp(\beta_{\ell 1}) + \ldots + p_k \exp(\beta_{\ell k})\right)^{-1}.$$

If we assume that the rare variants are in linkage equilibrium then the density over all variants is $f(\boldsymbol{Y}|\boldsymbol{\beta}) = \prod_\ell f_\ell(\boldsymbol{y}_\ell|\boldsymbol{\beta}_\ell)$.

To derive the test statistic of Zelterman and Chen (1988) for the multivariate case, we let

$$\mathbf{A} = \frac{H_0(f(\boldsymbol{Y}|\boldsymbol{\beta}))}{f(\boldsymbol{Y}|\boldsymbol{\beta} = 0)} = \left[\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1m} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2m} \\ \hline \vdots & & \ddots & \vdots \\ \hline \mathbf{A}_{m1} & \mathbf{A}_{m2} & \cdots & \mathbf{A}_{mm} \end{array}\right]. \tag{3.40}$$

where H_0 denotes the Hessian, i.e., the matrix of the second derivatives of f with respect to mk parameters $\beta_{\ell s}$, evaluated at $\boldsymbol{\beta} = 0$, and divided $\mathbf{A}$ into $m \times m$ blocks of $k \times k$ matrices.

Then,

$$[\mathbf{A}_{\ell\ell}](s,s) = (y_{\ell s} - y_\ell p_s)^2 - y_\ell p_s(1 - p_s), \tag{3.41}$$

$$[\mathbf{A}_{\ell\ell}](s,t) = (y_{\ell s} - y_\ell p_s)(y_{\ell t} - y_\ell p_t) + y_\ell p_s p_t, \text{ for } s \neq t \tag{3.42}$$

$$[\boldsymbol{A}_{\ell\ell'}](s,t) = (y_{\ell s} - y_\ell p_s)(y_{\ell' t} - y_{\ell'} p_t), \text{ for } \ell \neq \ell', \tag{3.43}$$

where $\ell, \ell' = 1, \ldots, m$ and $s, t = 1, \ldots, k$.

As with quantitative traits, a score test statistic T for testing whether $\boldsymbol{\beta} = \mathbf{0}$ is a sum over all elements of the matrix $\mathbf{U} \odot \mathbf{A}$, where $\odot$ denotes the Hadamard product

of two matrices (element wise multiplication) and $(mk) \times (mk)$ symmetric matrix $\mathbf{U}$ collects the correlations between the effect sizes $\boldsymbol{\beta}$ under the alternative model against which the test is most powerful:

$$T = \text{sum}\left(\mathbf{U} \odot \mathbf{A}\right). \tag{3.44}$$

Again P values are assessed by permuting the trait vectors among the individuals.

Extension to the analysis of gene-sets, pathways, networks. Suppose that we are interested in the analysis of a targeted gene set, for examples genes in GWAS regions and/or genes involved in monogenic disorders related to the phenotypes studied and only consider putative protein truncating variants. If we assume that each PTV at one gene has the same effect on the phenotype(s), we may combine all PTVs in the gene and consider a version of the C-alpha MRP test where each gene has taken the role of a single variant. However, genes in a set may have a mixture of positive and negative effects (risk conferring or protective), so we would like to be able to use dispersion across the genes rather than a burden test approach. Such a gene set model is a special case of the full C-alpha MRP model where the correlations between all variants within the same gene have been set to 1 in the $\mathbf{R}_{\text{phen}}$ matrix and otherwise set to 0. This approach may increase power for testing whether any of the genes have non-zero effects on the phenotypes studied over the approach where each gene is tested separately or when all PTVs are collapsed in a gene-set.

Estimation algorithm. There are many ways that we could characterize which variants are driving the association signal. As stated in the previous subsection in Neale et al. (2011) an EM algorithm was introduced to identify which variants were driving the association signal. That algorithm can be used to identify, after declaring association, which variants are most likely detrimental or protective and to estimate the proportion of variants contributing to the signal. In the multiple phenotypes setting I was interested in characterizing the rare variants driving the association across multiple quantitative traits by addressing the following questions:

1. What variants are driving the association signal?
2. How are the phenotypes affected?

The algorithm we conceived of was to cluster the variants based on the phenotype profile of the PTV carriers. This led to the development of the MCMC algorithm that follows.

Let k be the number of clusters of variants, where cluster 1 is the null model cluster; l be the number of quantitative traits; S be the empirical covariance matrix of the quantitative traits (= correlation for standardised traits); $\mathbf{R}_{\text{phen}}$ be the expected correlation of effects of the quantitative traits; n_c be the number of variants in cluster c, for $c = 1, \ldots, k$; N_c be the number of individuals in cluster c, for $c = 1, \ldots, k$; $\boldsymbol{y}$, the collection of quantitative trait vectors across all m variants; $\boldsymbol{y}_m$, the collection of quantitative trait vectors for individuals with variant m; and $\boldsymbol{y}_i$ the quantitative trait vector for individual i.

We wanted to estimate, for each cluster c, for $c = 1, \ldots, k$, the parameters: b_c, which denotes the mean of cluster c, where $b_1 = 0$; and p_c, the probability of cluster c, with $p_1 + \cdots + p_k = 1$. For variant m we have the parameter $t_m \in \{1, \ldots, k\}$ describing its cluster membership. Additional parameters include: Σ_c, the variance-covariance matrix of individuals in cluster c, where $\Sigma_1 = S$; Θ_c, the variance-covariance matrix of the effects in cluster c.

Model. The joint probability of all variables is given by

$$P\left(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{p}, \boldsymbol{t}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}\right) = P\left(\boldsymbol{y}|t, \boldsymbol{b}, \boldsymbol{\Sigma}\right) P\left(t|\boldsymbol{p}\right) P\left(\boldsymbol{p}\right) P\left(\boldsymbol{b}|\boldsymbol{\Theta}\right) P\left(\boldsymbol{\Theta}\right) P\left(\boldsymbol{\Sigma}\right),$$

where the individual factors are:

$$P\left(\boldsymbol{y}|t, \boldsymbol{b}, \boldsymbol{\Sigma}\right) = \prod_k \prod_{m \in c} \prod_{i \in m} f\left(\boldsymbol{y}_i | \text{mean} = b_c, \text{var} = \Sigma_c\right);$$

$$P\left(\boldsymbol{b}|\boldsymbol{\Theta}\right) = \prod_{c=2}^{k} P\left(b_c|\Theta_c\right), P\left(b_c|\Theta_c\right) = f(b_c|\text{mean} = b_0, \text{var} = \Theta_c),$$

where f is the multivariate normal density; and

$$P\left(t_m|\boldsymbol{p}\right) \text{is sampled from } \{1, \ldots, k\} \text{ with probability } \boldsymbol{p} = (p_1, \ldots, p_k).$$

The priors are:

$$P\left(\boldsymbol{\Sigma}\right) = \prod_{c=2}^{k} \mathcal{W}^{-1}\left(\Sigma_c|\Omega_c, \eta_0\right),$$

where Ω_0 can be the empirical covariance matrix of the phenotypes;

$$P\left(\boldsymbol{\Theta}\right) = \prod_{c=2}^{k} \mathcal{W}^{-1}\left(\Theta_c|\Psi_c, \nu_0\right),$$

where ν_0 can be set to be large, i.e. much greater than l if we are confident that the true covariance matrix is near some covariance matrix Θ_0, e.g. the expected

correlation of genetic effects, the empirical covariance matrix, or the identity matrix, Ψ_0 can be set to be $(\nu_0 - l - 1)\Theta_0$, making the distribution of Θ concentrated around Θ_0, or choosing $\nu_0 = l + 2$ and $\Psi_0 = \Theta_0$, making Θ only loosely centred around Θ_0 (the choice of the prior is discussed in Chapter 7 of Hoff (2009)); and

$$P(\boldsymbol{p}) = \text{Dirichlet}(1, \ldots, 1). \tag{3.45}$$

Posterior approximation with the Gibbs sampler. For multiparameter models the joint posterior distribution is difficult to sample from directly. However, it can be easier to sample from the full conditional distribution for each parameter (Hoff, 2009). The Gibbs sampler is a Markov Chain Monte Carlo (MCMC) algorithm that constructs a dependent sequence of parameter values whose distribution approximates (when it converges) the target joint posterior distribution (Geman and Geman, 1984; Casella and George, 1992; Hoff, 2009). In this clustering algorithm we are interested in estimating five parameters $\boldsymbol{b}, \boldsymbol{p}, \boldsymbol{t}, \boldsymbol{\Sigma}$, and $\boldsymbol{\Theta}$. Gibbs updates are from the conditionals and because of conjugacy one only needs the hyperparameters to define them. The Gibbs update can then be obtained by doing a lookup of the posterior hyperparameters using the conjugate prior table in the Wikipedia page `http://en.wikipedia.org/wiki/Conjugate_prior`:

1. Initialize $\boldsymbol{t}^{(0)}$ for all m and $\boldsymbol{p}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$ for all clusters c.

2. Repeat for $\ell = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$

   (a) Update $\boldsymbol{b}$. For each cluster $c = 2, \ldots, k$, update

   $$b_c^{(\ell)} \sim \mathcal{N}\left(\left(\Theta_0^{-1} + N_c\left(\boldsymbol{\Theta}_c^{-1}\right)^{(\ell-1)}\right)^{-1}\left(N_c\left(\boldsymbol{\Theta}_c^{-1}\right)^{(\ell-1)}\bar{\boldsymbol{y}}_c\right), \left(\Theta_0^{-1} + N_c\left(\boldsymbol{\Theta}_c^{-1}\right)^{(\ell-1)}\right)^{-1}\right);$$

   (b) Update $\boldsymbol{\Theta}$. For each cluster $c = 2, \ldots, k$, update

   $$\boldsymbol{\Theta}_c^{(\ell)} \sim \mathcal{W}^{-1}\left(\left(\Psi_0 + \left(\boldsymbol{b}_c^{(\ell)}\right)\left(\boldsymbol{b}_c{}^{(\ell)}\right)^{\intercal}\right)^{-1}, 1 + \nu_0\right);$$

   (c) Update $\boldsymbol{\Sigma}$. For each cluster $c = 2, \ldots, k$, update

   $$\boldsymbol{\Sigma}_c^{(\ell)} \sim \mathcal{W}^{-1}\left(\left(\Omega_0 + \sum_{m \in c}\sum_{i \in m}\left(\boldsymbol{y}_i - \boldsymbol{b}_c^{(\ell)}\right)\left(\boldsymbol{y}_i - \boldsymbol{b}_c{}^{(\ell)}\right)^{\intercal}\right)^{-1}, N_c + \eta_0\right);$$

(d) For $\boldsymbol{p}$

$$\boldsymbol{p}^{(\ell)} \sim \text{Dirichlet}\left(1 + n_1, \ldots, 1 + n_k\right).$$

(e) For $\boldsymbol{t}$

For each variant m,

$$u_c^{(\ell)} = p_c \prod_{i \in m} f\left(\boldsymbol{y}_i | \text{mean} = \boldsymbol{b}_c^{(\ell)},\ \text{var} = \boldsymbol{\Sigma}_c^{(\ell)}\right) \text{ and } w_c^{(\ell)} = \frac{u_c}{u_1 + \cdots + u_k}$$

where

$$t_m \sim \text{Discrete}\left(\{1, \ldots, k\},\ \text{with probabilities } (w_1, \ldots, w_k)\right).$$

Software implementation and additional comments. To analyze exome and/or genome sequencing data sets with the C-alpha MRP test and/or the MCMC algorithm (after detecting association) I implemented a couple of modules in the software `MAMBA`. The modules for the C-alpha MRP test include `--module crossdisorder` for cross-disorder analysis and `--module MRPQT` for cross-phenotype analysis of quantitative traits. An adaptive permutation scheme is applied. Thus, for scenarios where I cannot reject the null at a small α I am able to quickly process the data and improve the computational performance of the approach. In order to run the MCMC algorithm I added a module option `--montecarlo` to the `MRPQT` module, and module option `--pvalthr` for the P value threshold to use to run the MCMC algorithm.

A concern that may arise when applying the C-alpha MRP test is the wide variety of different parameters that one can set in choosing the matrix $\mathbf{U}$. I proposed a correlation structure of the effect sizes against which the test is most powerful which can be obtained in a principled manner. The choice of correlation matrix $\mathbf{R}_{\text{var}}$ may be based on annotation of variants. For example, it seems reasonable to assume that the effects of all protein truncating variants of a particular gene are highly similar (Rivas et al., 2013), or the expected correlation of genetic effect can be set based on the predictions of NMD. For the correlation matrix $\mathbf{R}_{\text{phen}}$ we proposed to use information about the genetic correlations estimated from twin or population studies (Pirinen, 2015) or about shared genetic effects estimated from genome-wide association studies (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). When that information is not available, as may often be the case, then for the cross-disorder analysis we proposed to use the identity matrix (assuming no correlation of genetic effects since that information is not available) or for cross-phenotype analysis of quantitative

traits we propose to use the variance-covariance matrix of the phenotypes, as this is usually consistent with the expected correlation of genetic effects (Pirinen, 2015).

Here, I have presented a gene-based test of association for rare variants and multiple phenotypes that has as special case the univariate single marker tests, the multivariate single marker tests, and the aggregate single trait tests. In practice, I would recommend that these complementary approaches be applied as they may improve power to detect association in different alternative scenarios.

3.3.4 Simulations

To evaluate the performance of the C-alpha MRP test I conducted a series of simulation studies. I compared the C-alpha MRP test to univariate approaches, and in the quantitative trait setting I also compared it to the standard multivariate Wald test. I begin with quantitative traits and move on to dichotomous traits. Finally, I performed some simulations to show that the presented MCMC algorithm can recapitulate effects introduced in the simulations.

Cross-phenotype alternative simulations. I undertook a series of simulation studies to demonstrate the performance of the statistical test presented in this section compared to the commonly used univariate approach SKAT and the standard multivariate Wald test.

Simulated genotypes were generated for 40 variants drawn from the allele frequency distribution estimated from the Exome Chip Design file available from the Exome Array design website for 2,500; 5,000; 10,000; and 20,000 individuals[7].

Phenotypic effects were introduced in the following manner. Let $V_{p\ell}$ denote the phenotypic variance explained by marker ℓ, n the number of samples being analyzed, m the number of markers simulated in a gene $\boldsymbol{G}$, δ the proportion of markers contributing to phenotypic variance and f_ℓ denote the frequency of marker ℓ. For a single marker $V_{p\ell} = 2\beta_\ell^2 f_\ell (1 - f_\ell)$.

First, I simulated a single phenotype such that $V_p = \sum_{\ell=1}^{m} V_{p\ell} = .0025$. I fixed $m = 40$ and $\delta = 0.5$ and 1, f_ℓ is sampled from the allele frequency derived from the exome array, and $\beta_\ell = \sqrt{\frac{V_p}{2m\delta f_\ell(1-f_\ell)}}$. For each individual i, $Y_i = \sum_{\ell=1}^{m} \beta_\ell x_{i\ell} + \varepsilon_i$ where $\varepsilon_i \sim N(0, 1 - V_p)$.

[7] `http://genome.sph.umich.edu/wiki/Exome_Chip_Design`

Next, I let k be the number of phenotypes collected for an individual. I simulated multiple phenotypes,

$$\mathbf{Y_i} = \sum_{\ell=1}^{m} X_{il}\boldsymbol{\beta}_l + \mathbf{e}_i,$$

such that $V_p = .0025$ for any phenotype that is associated with the variants and the errors $\mathbf{e}_i \sim N(0, \Sigma)$ are distributed according to the total correlation estimated from a population based data set for triglyceride, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol. I introduced effects such that the expected correlation of effects between a variant on a pair of traits was directionally consistent with the expected correlation of genetic effects, i.e. the variance-covariance matrix of the effect sizes is given by $\begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$.

In the simulated scenarios the C-alpha MRP test performed better than the univariate approaches and the standard multivariate Wald test. For example power to detect association at $\alpha = 0.001$ at sample size of 2,500 with 100% of the variants contributing to phenotypic variance is 47.5%, 0%, and 5.5% for C-alpha MRP, SKAT, and Wald test respectively (Figure 3.6).

Cross-disorder alternative simulations. To compare power to detect association for the C-alpha MRP test and the univariate SKAT test I performed a series of simulation studies for two scenarios: 1) variant effects are disease-specific, and 2) variant effects are shared. I used simulated datasets of 6,000 individuals (1,500 individuals for each disease group [three total disease groups] and a collection of 1,500 controls) generated by Loukas Moutsianas.

The simulated genotypes were generated using HAPGEN2 (Su et al., 2011). HAPGEN2 generates case-control data using a haplotype reshuffling approach based on the Li & Stephens model (Li and Stephens, 2003). Under this model, simulated (unobserved) haplotypes are assumed to be an imperfect mosaic of actual (observed) haplotypes, and are simulated using a Hidden Markov Model with recombination and mutation rates as parameters. Case-control samples are generated by over-sampling haplotype segments which contain alleles at which phenotypic effects are introduced (based on the relative risks assigned to them). In the haplotypes available from the original, 379 sample from the 1000 Genomes Project of European ancestry as a reference panel, observed variation is inadequate to simulate effects at rare variants in large samples. Loukas Moutsianas developed a staged, iterative approach using HAPGEN2, by introducing novel variation at each step with the goals of augmenting the reference

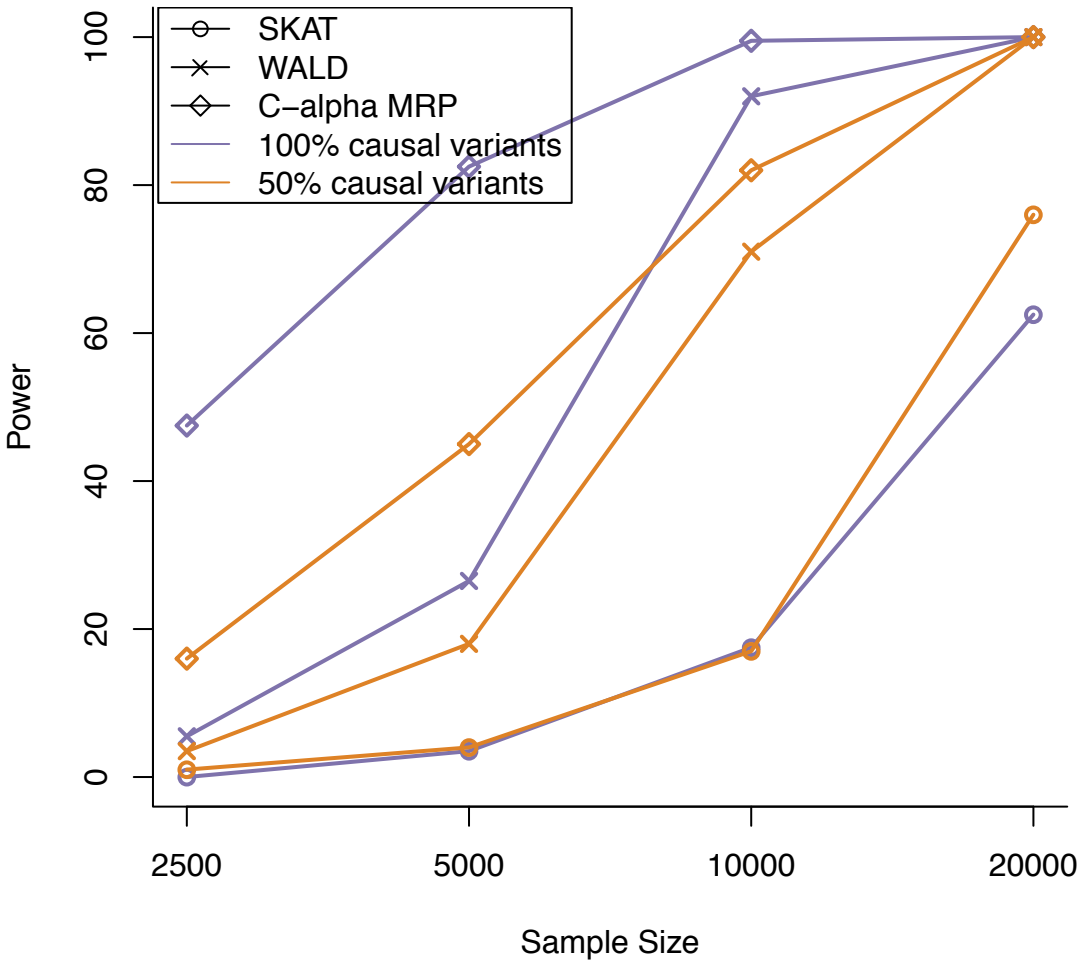


Figure 3.6: Comparison of the power to detect association in multivariate analysis with multiple quantitative trait measurements. I show that the C-alpha MRP test outperforms the multivariate Wald test and the univariate SKAT test for detecting association in simulated scenarios where 50% and 100% of variants have causal effects explaining 0.25% of the phenotypic variance across three separate traits with total correlation equal to that observed for low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG).

panel to a larger final sample size of over 12,000 samples. The simulated dataset had the following properties: i) maintained LD structure consistent with what is originally observed in the 1000G panel and ii) a full SFS consistent with that observed in empirical resequencing data. This data set is available for download from `http://mccarthy.well.ox.ac.uk/publications/2014/moutsianas_simulations/`.

The cumulative variance explained in a gene was simulated to be 1%, whenever rare variants in a gene are associated to disease. I compared power to detect association for the C-alpha MRP test to the power to detect association for the univariate SKAT test applied to the data such that the case-control comparison was carried out with a disease group with assigned effects. I applied SKAT using the linear weighted kernel, *P*-values evaluated using 'davies' method, `impute.method = fixed`, and `weights.beta = c(1,25)` implemented in the SKAT R package v0.82.

For cross-disorder analysis I found that when effects are disease-specific or shared, power to detect association at a significance level, α equal to 0.001, is higher ($80-85\%$) compared to the univariate case-control analysis (21%) (Figure 3.7). One intuitive explanation behind this is that the number of individuals observed is larger (6,000) than a simple case-control analysis (3,000) and the number of allelic copies in the disease group is larger than you expect by chance, i.e. uniformly distributed across the control and disease groups under the null. These results indicate that the C-alpha MRP test is an attractive option for gene-based testing in exome and genome sequencing studies of rare variants and multiple outcomes.

MCMC simulations. To assess whether the MCMC algorithm we proposed is properly recovering the underlying effects I simulated genotypes for 10 variants in $1,000$ individuals with a minor allele frequency of 1%. I introduced three groups (clusters) of effects across four phenotypes: i) NULL for variants labeled "simm3", "simm4", "simm5", or "simm6", no effect; ii) "CLUSTER 1" for variants labeled "simm7", "simm8" or "simm9", .5 standard deviations across all four traits; and iii) "CLUSTER 2" for variants labeled "simm0", "simm1" or "simm2", -.5 standard deviations across all four traits. I found that the clustering algorithm is able to recover the majority of introduced effects with high probability $> .8$ and appropriately recover the effect sizes introduced across the four traits for the three clusters (Figure 3.8).

3.3.5 Application of method in this thesis

In this thesis I applied the C-alpha MRP test to a targeted sequencing data set of autoimmune cases (Section 5.2), an exome sequencing study of lipid traits (Section 6.1),

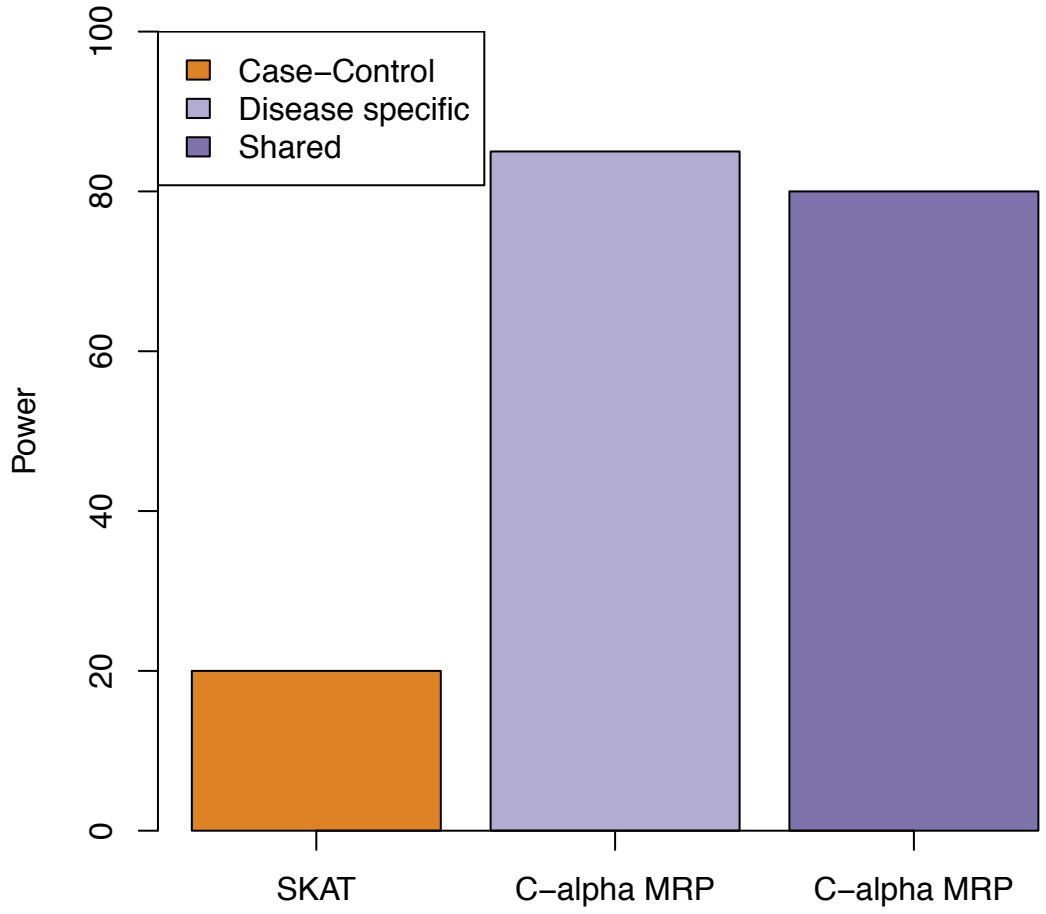


Figure 3.7: Comparison of the power to detect association ($\alpha = 0.001$) in cross-disorder analysis for three disease groups and one control group with effect introduced such that 1% of the phenotypic variance in liability scale is explained by additive genetic effects. I observe that when effects are disease-specific or shared power to detect association is higher compared to separate case-control analysis showing that pooling all the data and treating the diseases as separate case groups provide gain in power to detect association.

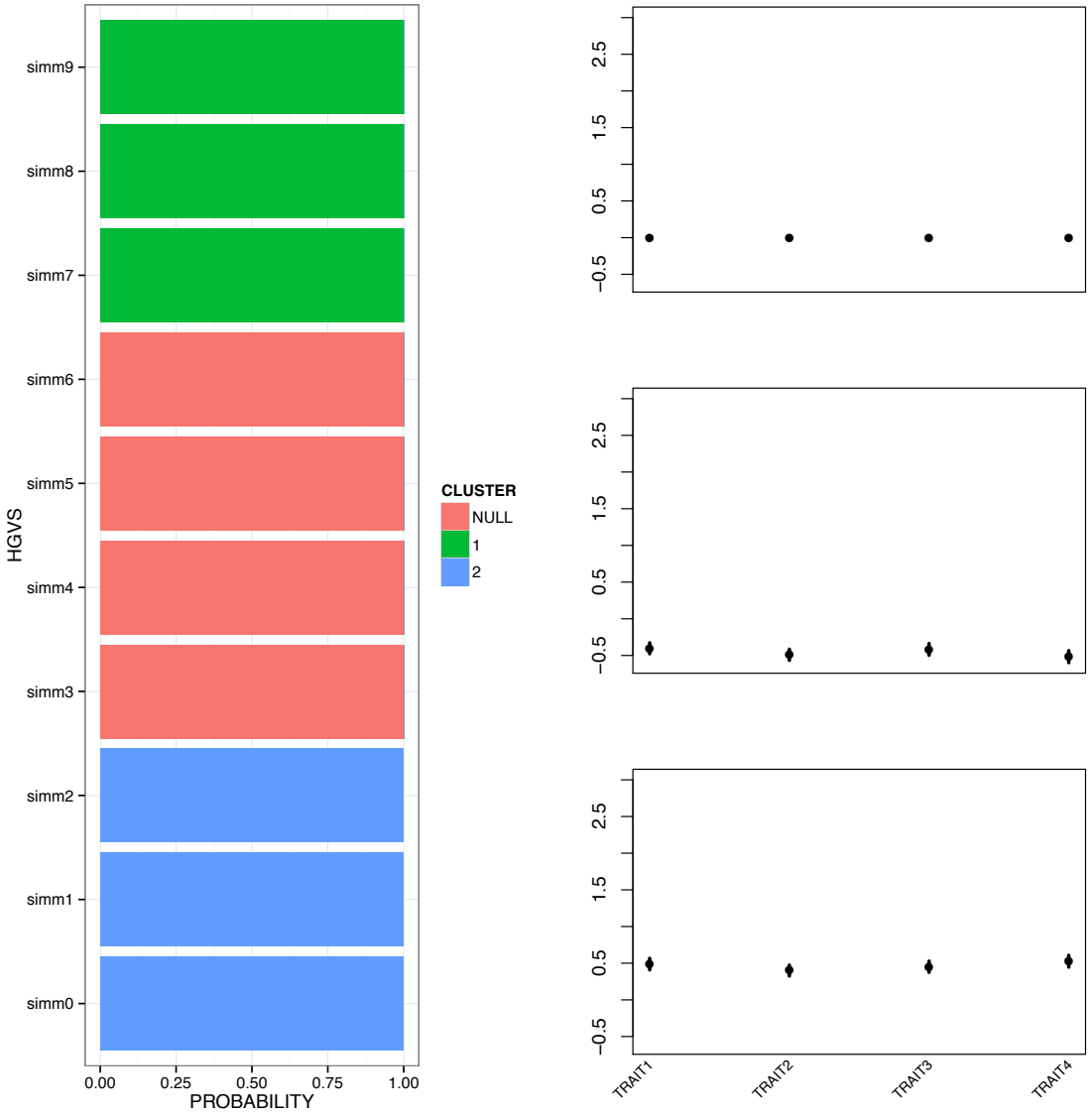


Figure 3.8: Application of MCMC algorithm to simulated data. Genotypes for 10 variants in 1,000 individuals with a minor allele frequency of 1% were simulated. Three groups (clusters) of effects across four phenotypes were introduced: i) NULL for variants labeled "simm3", "simm4", "simm5", or "simm6", no effect; ii) "CLUSTER 1" for variants labeled "simm7", "simm8" or "simm9", .5 standard deviations across all four traits; and iii) "CLUSTER 2" for variants labeled "simm0", "simm1" or "simm2", -.5 standard deviations across all four traits. On the left side of the figure a barplot of the posterior probability for the cluster membership assigned by the algorithm to each of the variants is presented. On the right side of the figure a series of three plots with the estimated effect size on the y-axis across all four traits on the x-axis is presented. Intervals between 2.5% and 97.5% quantile of the Gibbs samples of the parameter are shown.

and an exome array study of metabolic traits measured using NMR spectroscopy (Section 6.2). I employed the estimation algorithm described in this section for results identified in Section 6.1 and 6.2.

3.3.6 Limitations and next steps

The approach that we developed for assessing association between protein truncating variants and multiple phenotypes has several limitations including: 1) it does not handle relatedness; 2) in order to perform meta-analysis a method like Fisher's method or Stouffer's method is needed (Fisher, 1925; Liptak, 1958), which is not optimal as it is less powerful than other approaches that have been recently suggested for rare variant meta-analysis (Liu et al., 2014); 3) at the moment the estimation algorithm presented in this section does not estimate the number of clusters needed to fit the data, rather it must be specified by the user; 4) uncertainty estimates cannot be incorporated for the alternative, rather for the derived C-alpha MRP test a point estimate is required. Furthermore, I note that the alternative simulations do not encapsulate the universe of alternative scenarios. There are definitely scenarios where univariate approaches may be better powered to detect association compared to C-alpha MRP. It was brought to my attention that a reversible-jump Markov Chain Monte Carlo version of the estimation algorithm presented in this section would work for identifying the number of clusters to fit to the data (Green, 1995; Richardson and Green, 1997). Similarly, the statistical framework we have proposed for exploiting correlation of genetic effects can be addressed in a Bayesian modeling framework where uncertainty of the estimates of genetic effects can be incorporated by proposing a distribution of the estimate. I envisage that the limitations outlined and the proposed next steps are areas of research worthwhile to address in the near future.

3.3.7 Attributions

Together with my supervisors and Matti Pirinen I designed the C-alpha MRP test outlined in this section. Together with Matti Pirinen I participated in designing the Gibbs sampler clustering algorithm. I derived the formulation for the C-alpha MRP test - it was checked and simplified by Matti Pirinen who recognized that the test statistic could be represented as the Hadamard product of two matrices: 1) a matrix collecting the expected correlation of genetic effects, and 2) a matrix describing the Hessian of the density divided by the density. Jointly we proposed a clustering algorithm and he guided me in the derivation of the Gibbs algorithm

where the conditional distribution updates are taken from the conjugate prior page in Wikipedia `http://en.wikipedia.org/wiki/Conjugate_prior`. For a great exposition on Gibbs sampler algorithms for beginners see Chapter 7 in Hoff (2009). I implemented the C-alpha MRP test, and the clustering algorithm in the software `MAMBA`, and generated the simulations. Loukas Moutsianas provided the simulated data set I used to evaluate power for the cross-disorder analysis.

3.4 Discussion

There has been a growing interest in rare variant association studies as early results indicated that variants with strong effects remain to be found. Together, the statistical methods presented in this chapter were approaches developed for assessing protein truncating variant association across a broad range of rare variant association study designs.

The SEMCC and SEMGEM methods I presented in Sections 3.1 and 3.2 are Bayesian approaches developed for the analysis of rare variant genetic data. In particular, for assessing association between protein truncating variants and dichotomous traits or quantitative traits. One of the attractive properties of these approaches is the ability to incorporate prior belief or information related to the distribution of effect sizes one is expecting to observe under the alternative model. In principle, this would allow researchers to get a ranked list of genes that are more likely to reflect that belief compared to other frequentist approaches. For example, with the Bayesian SEMCC model presented in Section 3.1 if researchers are interested in pursuing genes with potentially large effects, and they would prefer not spend any effort in pursuing any genes with PTVs with weak or moderate effects, they could simply alter the alternative and null prior densities. Furthermore, with the Bayesian SEMCC model researchers could integrate external data sets like publicly available summary statistic data from the Exome Aggregation Consortium (ExAC)[8].

During my thesis work I became increasingly interested in the analysis of high-dimensional phenotypes. I suspected that one way we could improve my ability to detect protein truncating variant association signal was to exploit as much information as possible. In Stephens (2013), the author conducts simulation studies to demonstrate the benefits of analyzing multiple phenotypes simultaneously in genetic association studies of common variants. However, to date, very few studies have taken this approach possibly due to the difficulty in interpreting the results. Nonetheless,

[8] `http://exac.broadinstitute.org/`

I sought to develop a statistical approach for assessing association between protein truncating variants and multiple phenotypes for cross-trait analysis of either multiple quantitative traits or disease groups. I decided to develop a frequentist approach because at the time, with colleagues, I was focused on extending the binomial C-alpha test for case-control analysis of Neale et al. (2011) to the setting of single quantitative trait analysis, which was later published in Clarke et al. (2013). The C-alpha MRP test is flexible and extensible as shown in Section 3.3. An area of future development is to integrate the intuition behind the Bayesian SEMCC and SEMGEM models with that of the C-alpha MRP test into a single statistical framework. For example, in the current version of the C-alpha MRP test it is not possible to specify the magnitude of effects for which the test may be most powerful against, at least not in practice. This could, in principle, be integrated in a Bayesian framework.

While there are now a range of possible approaches for analyzing rare variant data (for many of which I was involved in their development) in the studies presented in Chapter 5 and 6, I focused on the statistical methods presented in this chapter. The development of these methods was important for being able to analyze and interpret rare variant data from these large scale array and sequencing studies.

The simulations presented in this chapter only cover a limited number of possible alternative scenarios. It is very likely that under other alternative scenarios other approaches may be preferable. I see this exercise as a possible future project where a larger number of scenarios and methods can be compared.

At the time I started my thesis development of statistical methods for rare variant analysis was an active area of research and tools were not available. I have implemented the methods in a software package called `MAMBA`. The hope is that these statistical, bioinformatic, and computational methods are of use to other researchers beyond the studies I present in this thesis.

Chapter 4

Development of statistical, bioinformatic, and computational methods for the analysis of protein truncating variants: transcriptional consequences

4.1 Introduction

In Chapter 3 I presented and described statistical and computational methods that participated in developing for assessing protein truncating variant association. In this chapter I describe statistical and computational methods for improving annotation of protein truncating variants by integrating experimental data and/or studying their functional consequences. To study the impact of these (mainly rare) variants I concentrated on RNA-sequencing data as a functional assay. The methods developed that are presented in this chapter focus on three topics: i) improving annotation of protein truncating variants by using transcript isoform quantification, ii) assessing allele-specific expression to quantify the impact of nonsense and frameshift indels on gene expression, and iii) estimating the extent and effect of splicing disruption by variants proximal to splice junctions. In Chapter 7 I present results and systematic analysis of the data from the Geuvadis and the Genotype-Tissue Expression project. Here, I focus on the methods used to analyze these data sets.

For each method I present:

- the background and rationale,
- the intuition motivating the design of the approach,

- the implementation,
- where appropriate a series of simulations to assess the method's limitations,
- a description of the application of the method in this thesis,
- a summary of the methods limitations and foreseeable next steps, and
- provide attributions.

4.2 *In silico* annotation of protein truncating variants with RNA-seq transcript quantification data

4.2.1 Background and rationale

A major challenge in the analyses of DNA sequence variants identified in whole-genome and exome sequencing studies is assessing their functional impact. In disease studies, inferences made from *in silico* prediction tools can have a crucial impact on gene discovery: for example, in the setting of rare variant association studies for complex traits, errors in variant annotation can dramatically reduce power to detect association with gene-based association tests (Zuk et al., 2014). More importantly, in the clinical application of genome sequencing variant annotation errors can dramatically impact disease diagnosis (Dewey et al., 2014).

For annotation of genomic variants discovered in DNA sequencing studies most bioinformatics tools take a reference transcript set (e.g. GENCODE, REFSEQ, or ENSEMBL) and annotate a genomic variant with the most deleterious variant prediction (McCarthy et al., 2014). This approach has the major limitation that it is dependent on software and on the exhaustive number of transcript isoforms that are available for any given gene. McCarthy et al. (2014) reported results from a comparative study where they used different transcript sets and software to annotate variants discovered in a whole genome sequencing experiment (WGS500 study). They found 44% agreement in PTV annotation when they applied the same software, but altered the reference transcript set. In addition, they found only 65% agreement in PTV annotation when they applied different software, but kept the reference transcript set constant. One of the reasons why there is such poor agreement is that different bioinformatic tools use different approaches for reporting variant annotation. For example, some choose the most deleterious variant prediction, while others may choose

the "canonical" transcript reported by the consortia providing the reference transcript set, which in some instances is defined as the longest protein-coding transcript[1]. Another reason is that transcript sets have varying number of protein-coding transcript isoforms catalogued. For example, ENSEMBL contains over 200,000 transcripts while REFSEQ contains about half that amount. The majority of these transcript isoforms are likely to either be lowly expressed, exist in very unusual conditions, or be experimental and prediction errors making them irrelevant for most practical applications in human genomic studies.

On the other extreme, in the setting of protein truncating variants (splice, nonsense single nucleotide variants (SNVs), or frameshift indels) some bioinformatic tools, e.g. VEP (McLaren et al., 2010), take a reference transcript, annotate a variant with prediction for each of the protein-coding transcripts and filter the quality of the "loss of function" prediction based on heuristic rules such as the length of the intron, size of the gene, and location of the PTV in the gene described in MacArthur et al. (2012). These heuristics, although at times useful, have not been thoroughly assessed.

For *in silico* variant effect prediction, the inclusion of transcript isoform quantification data provides a new bioinformatic framework for the annotation of genomic variants identified in DNA sequencing studies. To do this requires exome sequencing data and RNA-sequencing data. As part of the GTEx project I sought to develop a bioinformatic tool that would annotate genetic variants discovered in DNA sequencing studies with the aid of RNA-seq transcript isoform quantification from reference data sets.

4.2.2 Intuition

The main intuition behind the bioinformatic tool for DNA sequence variant annotation is that transcript isoform quantification data obtained from an RNA sequencing experiment either from a single or multiple tissues (like in GTEx) can be used to quantify the relative variant annotation support across all isoforms that were quantified. I illustrate the intuition with an example.

First, let's assume that I have identified two novel single nucleotide variant alleles in *SGCB* in a medical resequencing study of limb-girdle muscular dystrophy type 2E. Mutations in the *SGCB* gene are associated with limb-girdle muscular dystrophy type 2E, a group of related disorders characterized by muscle weakness and wasting, particularly in the shoulders, hips, and limbs (Nigro et al., 1996). In the GENCODE

[1] `http://www.ensembl.org/Help/Glossary?id=346`

version 12 reference transcript set 4 transcript isoforms are reported. Using standard bioinformatic tools for variant annotation including VEP and PLINK/SEQ I obtained a "STOP GAIN" or nonsense annotation for both SNVs since these programs report the most deleterious annotation across the 4 transcript isoforms available. However, using transcript isoform quantification data from the GTEx project it is clear that the transcript isoform ENST00000381431 is the dominant isoform and it is highly expressed across all tissues (Figure 4.1, which can be obtained in the GTEx portal at `http://gtexportal.org/home/gene/SGCB`). A common unit of measurement of mRNA abundance using RNA-seq data is the reads per kilobase of transcript per million reads mapped (RPKM) (Mortazavi et al., 2008). This unit of measurement can be used for comparing the mRNA abundance of protein-coding genes, exons, transcript isoforms, and even introns. Given the interest in a disease that affects the muscle, and a gene with prior known involvement, ideally I would like to annotate the DNA sequence variants by integrating reference isoform quantification data for skeletal muscle. I observed that ENST00000381431 is the dominant isoform expressed in skeletal muscle tissue (MSCLSK) and that the other protein-coding isoforms are lowly expressed. "SNV 1" and "SNV 2" are predicted to be a nonsense variant across two non-muscle protein-coding transcript isoforms and as a 5′-UTR or intronic variant across the remaining isoforms (Figure 4.2). Given that the empirical data supports the prediction that "SNV 2" will not lead to a truncated protein product (it is only truncated in non-muscle isoforms), and that it is likely to be "intronic", the developed bioinformatic tool should report the variant annotations as "nonsense" and "intronic" for "SNV 1" and "SNV 2", respectively.

4.2.3 Implementation

In the software program I developed, `MAMBA`, I implemented a module called `annotate` where DNA sequence variants can be annotated using the GENCODE transcript set used in the GTEx project. The tool simply makes use of the PLINK/SEQ annotation program that takes in a reference transcript set and translates *in silico* the protein-coding transcripts with the alternate allele for each of the transcript isoforms available for any overlapping gene. For each transcript isoform, the tool then loads the transcript isoform quantification data from GTEx, computes the median RPKM value across all the samples for any given tissue and stores the RPKM value into memory. Once a variant is annotated across all overlapping transcript isoforms the program either reports the annotation for the highest expressed transcript isoform in the user-specified tissue, e.g. `--tissueabbreviation MSCLSK --dominant`, or reports

the annotation for all transcript isoforms that are expressed above a user-specified RPKM threshold for a given tissue, e.g. `--tissueabbreviation MSCLSK --rpkm 0.1`. In the current implementation only 9 tissues from the pilot GTEx project are supported: 1) adipose subcutaneous (ADPSBQ); 2) artery tibial (ARTTBL); 3) heart left ventricle (HRTLV); 4) lung (LUNG); 5) muscle skeletal (MSCLSK); 6) nerve tibial (NERVET); 7) skin sun exposed (SKINS); 8) thyroid (THYROID); 9) blood (WHLBLD). If a tissue is not specified then the tool either reports the annotation for the dominant isoform calculated as the most frequent dominant isoform across all the studied tissues or reports the annotation for all transcript isoforms that are expressed above a user-specified RPKM threshold across any studied tissue. Furthermore, a user may require that the transcript isoform used for variant annotation is either the dominant isoform across a proportion threshold of the studied tissues in GTEx or that the transcript isoform is expressed higher than an RPKM level threshold across a proportion threshold of the studied tissues. This is done with the option `--prop [0-1]`.

4.2.4 Results

To demonstrate the impact of this new bioinformatic framework that I refer to as "RNA-seq informed DNA variant annotation" I used exome sequencing data and RNA sequencing data from the GTEx project. Variant genotype data and RNA-seq transcript quantification data was generated on a subset of donors as a pilot during this first phase of the GTEx project. From the variants identified in the 173 GTEx individuals with both exome-sequencing and RNA-seq data I sought to assess global properties of the predicted PTVs.

First, I used the variant annotation data generated centrally by the Genotyping Working Group of the GTEx project, a large consortium project, which will be introduced in more detail in Chapter 7. The variant annotation data was prepared using a customized version of the VEP program. The customized version of the VEP program supports annotation flags for PTVs that were proposed in MacArthur et al. (2012). For instance, PTVs that had annotation support of a PTV across all transcript isoforms for any given gene contained the "LOF=FULL" flag, whereas some PTVs that had only partial annotation support of a PTV across the transcript isoforms for any given gene contained the "LOF=PARTIAL" flag. I calculated (by dividing the number of variants with "LOF=PARTIAL" over the total number of PTVs) that only 38.4% of PTVs have PTV annotation support across all transcript isoforms for any given gene (Figure 4.3). Hence, most PTVs only had "PARTIAL" PTV support.

Second, I used the median RPKM measurements for all transcript isoforms across all tissues with at least 10 samples. I asked for all the predicted PTVs what percentage of the variants maintain a PTV annotation if I required that a fixed percentage of the dominant isoforms across all sequenced tissues support a PTV prediction. I found that the percentage of predicted PTVs with support of PTV annotation quickly decreased as I increased the threshold for the proportion of tissues with major transcript isoform support for PTV prediction: 70% of PTVs are relevant if proportion of tissues with PTV annotation for dominant transcript isoform threshold is 10%, and ∼40% of PTVs are relevant if threshold is 100% (Figure 4.3).

Third, I used the allele-specific expression (ASE) data presented in Chapter 7 to assess whether variants showing PTV support using transcript abundance analysis have different patterns of allele-specific expression compared to those that have little support of PTV annotation. I observed significant difference (P value = 0.0084, Mann-Whitney-Wilcoxon (MWW) one-sided test) in the posterior probability of no ASE across all tissues between the two groups: i) the group of variants that have PTV annotation support by transcript abundance in at least 90% of tissues ($n = 164$ with ASE data), and ii) the group group of variants that have PTV annotation support by transcript abundance in less than 10% of tissues ($n = 133$ with ASE data).

These results highlight the need for empirically derived reference transcript sets for a broad array of tissues for personal genome sequencing and disease studies.

4.2.5 Application of the approach in this thesis, limitations, and next steps

Unfortunately, the development of the bioinformatic tool and the computational framework developed in this section for annotating DNA sequence variants discovered in genome sequencing studies occured in the final stages of my thesis work. Hence, I was unable to apply the framework in either the association studies of diseases in Chapter 5 or quantititative traits in Chapter 6. Furthermore, I was developing the approach in the final stages of the research work conducted in Chapter 7 where I focus on studying the functional consequences of PTVs. I envisage that this work will be an important component in understanding the nature and consequences of PTVs.

One of the major limitations that I see from the developed approach is that it is entirely possible that multiple transcript isoforms for any given gene may be relevant to disease and the relative abundance of each in a tissue may be pertinent information. For instance, assume that in a gene there exists 3 transcript isoforms with relative abundance ratios of 50 : 30 : 20 in a tissue of interest. In practice I would like this

information to be captured when annotating DNA sequence variants. At the moment, this information is not integrated in the current implementation. In future work I envisage that we could assign a variant annotation score for all possible annotations given the transcript quantification data. Thus, in the stated example, assume that the first two transcript isoforms support the "nonsense" annotation and the third transcript isoform supports the "intronic" annotation, then the score for "nonsense" annotation will equal to 0.8 and the "intronic" annotation will equal to 0.2. All other possible annotations will have a score of 0. I see the probabilistic framework as a natural extension to the current framework I have developed. Another major issue is that in some (many) cases, even for rare diseases, the correct tissue may not be known.

4.2.6 Attributions

I conceived the idea to integrate experimental data to improve annotation of genetic variation identified in DNA sequencing experiments. I designed the computational framework, and implemented it in the software `MAMBA`. The Genotyping Working Group (GWG) of the GTEx consortium, in particular Monkol Lek and Daniel MacArthur, generated the exome sequencing callset and annotated the DNA sequence variants using a customized VEP program. Additional information about the exome sequencing callset including location of data generation and processing is described in Chapter 7. The Quantification Working Group of the GTEx consortium, in particular Michael Sammeth, generated the transcript isoform quantification data. The data was made publicly available in the GTEx portal where I downloaded it from. My supervisors provided useful comments and feedback. This work was contributed to the GTEx consortium pilot manuscript, which is currently in press.

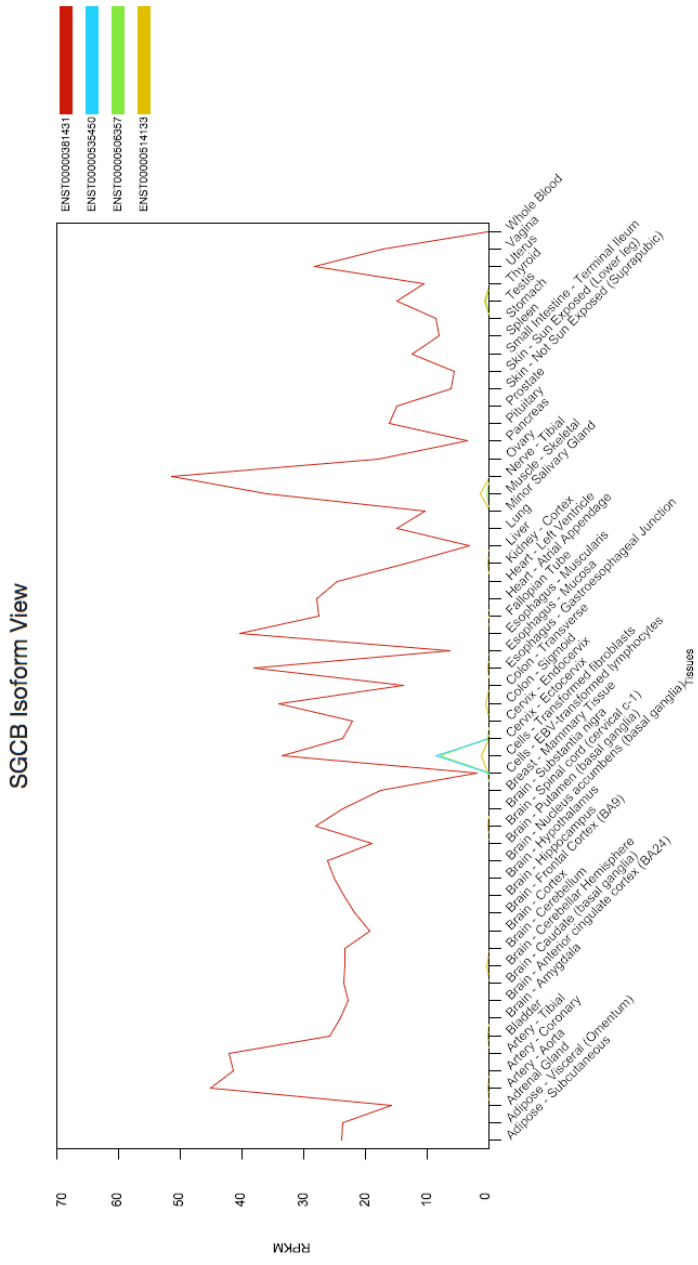


Figure 4.1: Expression levels of *SGCB* transcript isoforms across all tissues in GTEx. ENST00000381431 is the highest expressed transcript isoform across all studied tissues in GTEx.

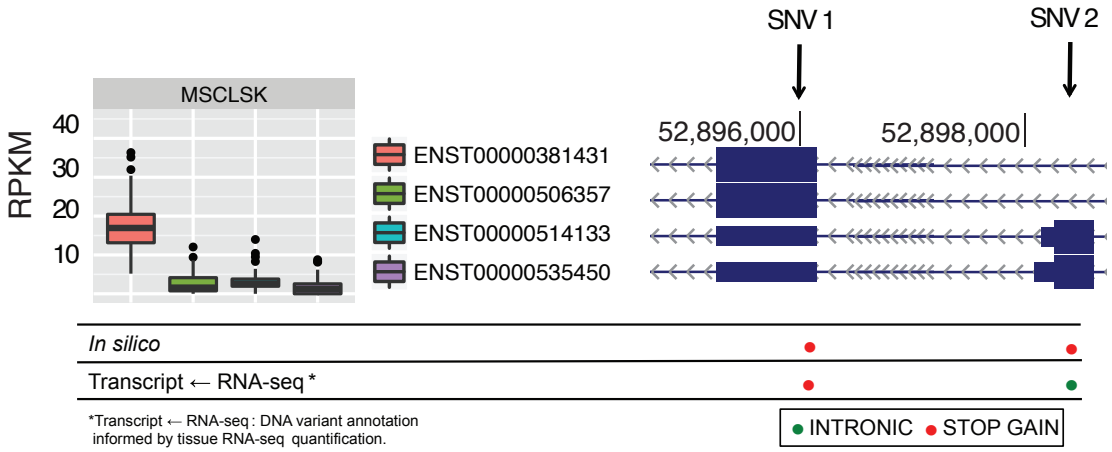


Figure 4.2: An example of how tissue-specific isoform annotation from RNA-seq can impact the interpretation of putative novel alleles in the gene *SGCB*. In the GENCODE version 12 reference transcript, a set of four transcript isoforms are reported. ENST00000381431 is the dominant isoform expressed in skeletal muscle tissue (MSCLSK) and the other protein-coding isoforms are minimally expressed. "SNV 1" is predicted to be a nonsense variant across all four protein-coding transcript isoforms, whereas SNV2 is predicted to be a nonsense variant across two protein-coding transcript isoforms and an intronic variant across the remaining isoforms. Some functional prediction programs output the most deleterious annotation, i.e., nonsense, and can lead to false-positive results for diagnostic exome sequencing screening of muscle disorder samples.

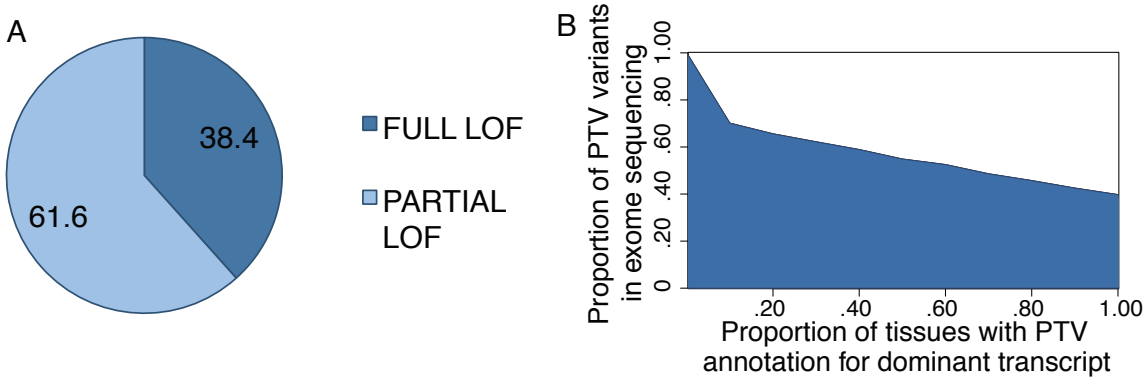


Figure 4.3: Integration of personal transcriptome data for improved annotation of predicted protein truncating genetic variants. (A) The majority of annotated PTV/LoF variants are partial PTV, meaning that only a fraction of the RNA-seq transcripts support PTV annotation. (B) For all the predicted PTVs the percentage of variants that maintain a PTV annotation if we require that a fixed percentage of the dominant isoforms across all sequenced tissues support a PTV prediction: 70% of PTVs are relevant if the threshold is 10% and, similarly, 40% of PTVs are relevant if the threshold is 100%.

4.3 Assessing allele-specific expression across multiple tissues from RNA-seq read data

4.3.1 Background and rationale

RNA-Seq is an approach to transcriptome profiling that uses short-read sequencing technologies (Mortazavi et al., 2008; Wang et al., 2009). It enables allele specific expression (ASE) studies that complement standard genotype expression studies for common variants (Lappalainen et al., 2013) and facilitates the study of the regulatory impact of rare variants (Montgomery et al., 2011). Allele specific expression analysis can be used as a proxy to interpret the functional impact of expression Quantitative Trait Loci (eQTL) variants in *c*is. For any heterozygote genotype the proportion of reads representing the coding variant in *c*is with an eQTL variant can be used to estimate the functional impact of that eQTL variant on transcription. For PTVs, the main interest of my thesis, the causal variant is the PTV. The NMD pathway is triggered when a premature truncating codon is detected and the corresponding isoforms are degraded (Lejeune and Maquat, 2005). Using allele specific expression data we can identify if, and to what extent, degradation occurs.

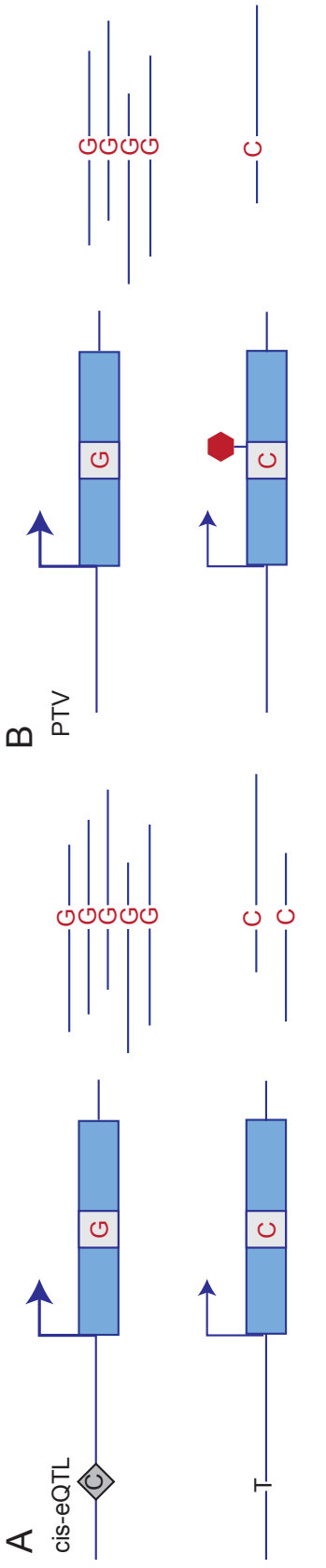


Figure 4.4: Diagram of allele specific expression analysis using short read sequencing data. (A) Representation of a coding variant serving as a proxy to interpret the functional impact of an eQTL variant in *cis*. (B) Representation of a protein truncating variant directly having a functional consequence on transcription. For both figures the number of reads representing each allele is shown.

Allele specific expression analysis is a powerful approach to interpret the functional impact of genetic variants on gene transcription. The idea is simple: Under the null model, the two gene copies are equally expressed. Under the alternative model, the two gene copies are not equally expressed. Heterozygote genotypes are used to differentiate between the two gene copies. For any heterozygote genotype, under the null model reads from RNA sequence data supporting either allele are assumed to be equally distributed. A binomial test is usually applied to read count data to assess significance of deviation from the null. However, this approach has many limitations. First, it does not take into account biases that may exist in the data, for example due to mapping biases in sequence alignments where RNA-seq reads carrying the non-reference allele of a variant loci can have lower probability to map correctly to the reference genome (Panousis et al., 2014; Skelly et al., 2011). Second, interpretation of such a test can be quite difficult because factors like the total read count (or coverage) will affect power. Third, simultaneous analysis of multiple tissues is not explicitly handled.

In this section I present the intuition and methodology for assessing allele-specific expression that resulted in more sophisticated statistical and computational models. I applied the methods presented in this section (and their extensions) to RNA sequencing data sets from the Genotype Tissue Expression project and the Geuvadis RNA sequencing project. The results are presented in Chapter 7.

4.3.2 Intuition

The main motivation for developing statistical methods for analyzing allele-specific expression patterns from RNA-seq data across multiple tissues are biological processes like nonsense-mediated decay. This phenomenon is expected to lead to strong ASE where the PTV containing allele is expressed at lower levels compared to the non-PTV containing allele.

We wanted to address the following three questions when we started developing the ASE method(s):

1. In which tissue does a heterozygous site show ASE?
2. Which tissues show similar ASE effects at the site studied?
3. What proportion of a certain class of variants (such as PTVs) show ASE in all tissues, only in some tissues, or in no tissue?

For the first question, a standard binomial test is commonly used (Lappalainen et al., 2013). For the second and third question it was not clear what is the best way to approach it and at the time we started addressing the problem there were no statistical methods available. One way to address the questions was by developing a method that would classify the protein truncating variants with ASE data conditional on the read observations across all the available tissues as either having evidence of 1) ASE effects (suggesting the transcript(s) containing the protein truncating allele were degraded by the NMD machinery), 2) no ASE effects (suggesting the transcript(s) containing the protein truncating allele escape NMD), or 3) heterogeneous ASE effects (suggesting the transcript(s) containing the protein truncating allele escaped NMD in some tissues while in other tissues they were degraded by the NMD machinery). We chose a Bayesian statistical framework where the ASE data across all variants and tissues would be treated as independent observations. In Pirinen et al. (2015) we jointly model the ASE data.

4.3.3 Implementation

The following Bayesian models were proposed:

Let y_1 and y_2 be the number of reads supporting the alternate (or the non-reference) and reference alleles, respectively. The random variable y_1, i.e. the number of reads supporting the alternate allele, will follow the binomial distribution. Let θ be the probability that a transcript carries the non-reference allele.

Under the null model, we assumed the prior density on θ to be a beta distribution (as that is the usual conjugate prior used for random variables following a binomial distribution):

$$\theta_{\text{NULL}} \sim \text{Beta}\left(a_0, b_0\right), \tag{4.1}$$

with a_0 and b_0 as the hyperparameters. In an ideal setting where technical and mapping biases are not introduced $\theta_{\text{NULL}} = \frac{1}{2}$. However, this is likely not to be the case as discussed in Skelly et al. (2011) and Panousis et al. (2014).

The likelihood is

$$P\left(y_1|\theta_{\text{NULL}}\right) = \binom{y_1 + y_2}{y_1} \theta_{\text{NULL}}^{y_1} (1 - \theta_{\text{NULL}})^{y_2}, \tag{4.2}$$

and the marginal distribution of y_1 (Wakefield, 2013) is given by

$$\int P\left(y_1|\theta_{\text{NULL}}\right) P\left(\theta_{\text{NULL}}\right) d\theta_{\text{NULL}} = \binom{y_1 + y_2}{y_1} \left[\frac{\Gamma\left(\alpha_0 + \beta_0\right)}{\Gamma\left(\alpha_0\right)\Gamma\left(\beta_0\right)} \frac{\Gamma\left(y_1 + \alpha_0\right)\Gamma\left(y_2 + \beta_0\right)}{\Gamma\left(n + \alpha_0 + \beta_0\right)}\right]. \tag{4.3}$$

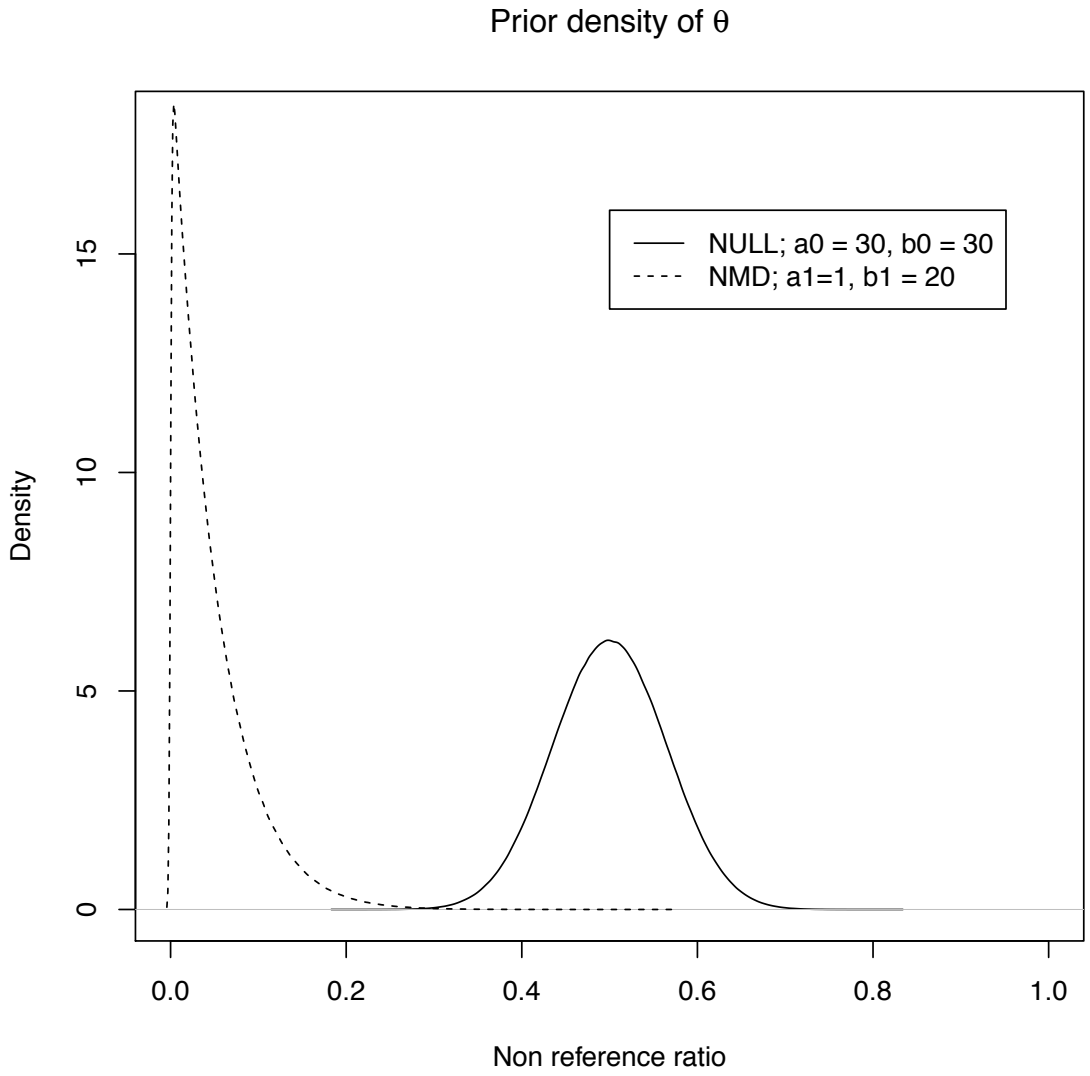


Figure 4.5: Prior density of θ for the null and alternative model. Under the null model we used a Beta $(30, 30)$ distribution. Under the alternative model we used a Beta $(1, 20)$ distribution. Without loss of generality we assume that the PTV allele is always the non-reference allele. In the scenario where the reference allele introduces a premature stop then the reference allele should be treated as the PTV allele.

Under the alternative model, i.e. nonsense mediated decay pathway gets triggered, we assumed that protein truncating variants will have a strong impact on allele specific expression readout.

We assumed that the prior on θ is a beta distribution with different hyperparameters α_1 and β_1 (we chose the hyperparameters to reflect the strength of ASE that we believed to represent the consequence of NMD),

$$\theta_{\text{NMD}} \sim \text{Beta}\left(\alpha_1, \beta_1\right). \tag{4.4}$$

and the marginal distribution of y_1 is given by

$$\int P\left(y_1|\theta_{\text{NMD}}\right) P\left(\theta_{\text{NMD}}\right) d\theta_{\text{NMD}} = \binom{y_1 + y_2}{y_1} \left[\frac{\Gamma\left(\alpha_1 + \beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)} \frac{\Gamma\left(y_1 + \alpha_1\right)\Gamma\left(y_2 + \beta_1\right)}{\Gamma\left(n + \alpha_1 + \beta_1\right)}\right]. \tag{4.5}$$

Bayes Factors were used to quantify the evidence for the alternative hypothesis as compared to the null.

$$BF_{\text{NMD}} = \frac{\int P\left(y_1|\theta_{\text{NMD}}\right) P\left(\theta_{\text{NMD}}\right) d\theta_{\text{NMD}}}{\int P\left(y_1|\theta_{\text{NULL}}\right) P\left(\theta_{\text{NULL}}\right) d\theta_{\text{NULL}}}, \tag{4.6}$$

which results in

$$BF_{\text{NMD}} = \frac{\left[\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(y_1+\alpha_1)\Gamma(y_2+\beta_1)}{\Gamma(n+\alpha_1+\beta_1)}\right]}{\left[\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(y_1+\alpha_0)\Gamma(y_2+\beta_0)}{\Gamma(n+\alpha_0+\beta_0)}\right]}. \tag{4.7}$$

Note that the statistical framework is very similar to the Bayesian SEMCC model presented in Section 3.1.

The relationship between posterior odds and Bayes Factor (Kass and Raftery, 1995) is given by

$$\text{posterior odds} = \text{prior odds} \times BF. \tag{4.8}$$

To obtain the posterior probability that a tissue shows evidence of NMD we can simply calculate it using the posterior odds as

$$P = \frac{\text{posterior odds}}{1 + \text{posterior odds}}, \tag{4.9}$$

which represented as a function of Bayes Factors and prior odds is given by

$$P = \frac{BF \times \text{prior odds}}{1 + BF \times \text{prior odds}}. \tag{4.10}$$

Then, for any tissue, the posterior probability that the tissue shows evidence of NMD is given by

$$P_{\text{NMD posterior}} = \frac{BF_{\text{NMD}} \times \text{prior odds}}{1 + BF_{\text{NMD}} \times \text{prior odds}}. \tag{4.11}$$

For definiteness, we set the prior odds to 1 as a choice reflecting symmetry as to whether or not NMD occurs. It is possible to study the sensitivity of the method to the choice of the prior by varying the effect size and the number of reads observed.

Furthermore, an updated prior may be chosen after the initial wave of RNA-seq experiments are analyzed. However, neither of these is considered in this thesis.

In the method we evaluate the probability that all tissues show evidence of NMD (state 'ALL'), all tissues do not show evidence of NMD (state 'NONE'), and tissue heterogeneity (state 'HET') first by assuming independence across tissues using the following approach:

$$P_{\text{No NMD all tissues, NONE}} = \prod_{i=1}^{k} \left(1 - P^{(i)}_{\text{NMD posterior}}\right), \tag{4.12}$$

$$P_{\text{NMD all tissues, ALL}} = \prod_{i=1}^{k} \left(P^{(i)}_{\text{NMD posterior}}\right), \text{and} \tag{4.13}$$

$$P_{\text{NMD heterogeneity, HET}} = 1 - P_{\text{NMD all tissues}} - P_{\text{No NMD all tissues}}. \tag{4.14}$$

In addition, we consider tissue specific effects (state 'TIS_SPE') where one tissue shows evidence of ASE and all others do not show evidence of ASE. The TIS_SPE state is a special case of the heterogeneous effects (state 'HET'):

$$P_{\text{NMD tissue specific, TIS\_SPE}} = \sum_{i=1}^{k} \left(P^{(i)}_{\text{NMD posterior}} \prod_{j \neq i}^{k} \left(1 - P^{(j)}_{\text{NMD posterior}}\right) \right). \tag{4.15}$$

This model is considered the "Independent Tissue Model (ITM)" in Pirinen et al. (2015) and is implemented in `MAMBA` with the option `--indep`. This model may be useful in the scenario where very deep RNA sequencing is conducted across multiple tissues from different individuals and there may be enough information contained within each tissue to assess ASE independently of each other (Kukurba et al., 2014). However, in current experiments, data is generated from multiple tissues collected from a single individual where many biological conditions may be shared (e.g. shared genome across multiple tissues).

4.3.4 Simulations

I generated a series of simulations to study the method's performance. First, I fixed the total number of reads, and I altered the total number of tissues ($t = 5, 10, 20$) and simulated four scenarios: 1) ASE effects across all tissues (Figure 4.6), 2) no ASE effects across all tissues (Figure 4.7), 3) heterogeneous effects (Figure 4.8), and 4) tissue-specific effects (Figure 4.9). Then, I altered the total number of reads ($n = 10, 50, 100, 100$) to study the method's performance as the number of reads increased.

I found that the method appropriately recovers the correct classification when the number of reads is large ($n \geq 50$, Figures 4.6-4.9). However, for simulated scenarios with homogeneous ASE effects across all tissues it appears that when the read count is small ($n \leq 20$) and the number of tissues is large ($t \geq 10$) the method incorrectly gives higher probability to the heterogeneous state (Figure 4.6 and 4.7). Furthermore, I found that when ASE effects are tissue-specific and the number of reads is small the method gives high probability to the 'NONE' state (Figure 4.9).

4.3.5 Limitations and next steps

The independent tissue model for assessing allele-specific expression across multiple tissues from RNA-seq read data has a couple of obvious limitations. First, as shown in the simulation study, the method is likely to classify a variant as having heterogeneous ASE effects when the number of reads is small and the number of tissues is large. Second, we do not model moderate ASE effects. In the literature it is assumed that the NMD machinery is efficient and it is one of the reasons that PTVs are commonly referred to as loss-of-function variants (Flannick et al., 2014; Lim et al., 2014b). However, in model organisms it is clear that degradation does not occur across all transcripts and that NMD efficiency may depend on the tissue (Linde et al., 2007a; Zetoune et al., 2008). To address these two limitations of the independent tissue model I collaborated with Matti Pirinen based at the Institute for Molecular Medicine in Finland (FIMM) to develop more sophisticated statistical models for asssessing ASE across multiple tissues. These models were mainly developed by Matti Pirinen and presented in a manuscript that may be found online on the preprint server bioRxiv[2] (Pirinen et al., 2015). I describe the models briefly in this subsection as they are the main methods I applied to the data described in Chapter 7.

Grouped Tissue Model (GTM) The Grouped Tissue Model (GTM) classifies the ASE effect in each tissue into three groups: no ASE ($\mathcal{N}$), moderate ASE ($\mathcal{M}$), and strong ASE ($\mathcal{S}$). The conjugate priors that were chosen to describe each of the three groups were:

$$\theta(\mathcal{N}) = \text{Beta}(2000, 2000), \tag{4.16}$$

$$\theta(\mathcal{M}) = \text{Beta}(36, 12), \text{and} \tag{4.17}$$

$$\theta(\mathcal{S}) = \text{Beta}(80, 1). \tag{4.18}$$

[2] http://biorxiv.org/content/early/2014/07/17/007211

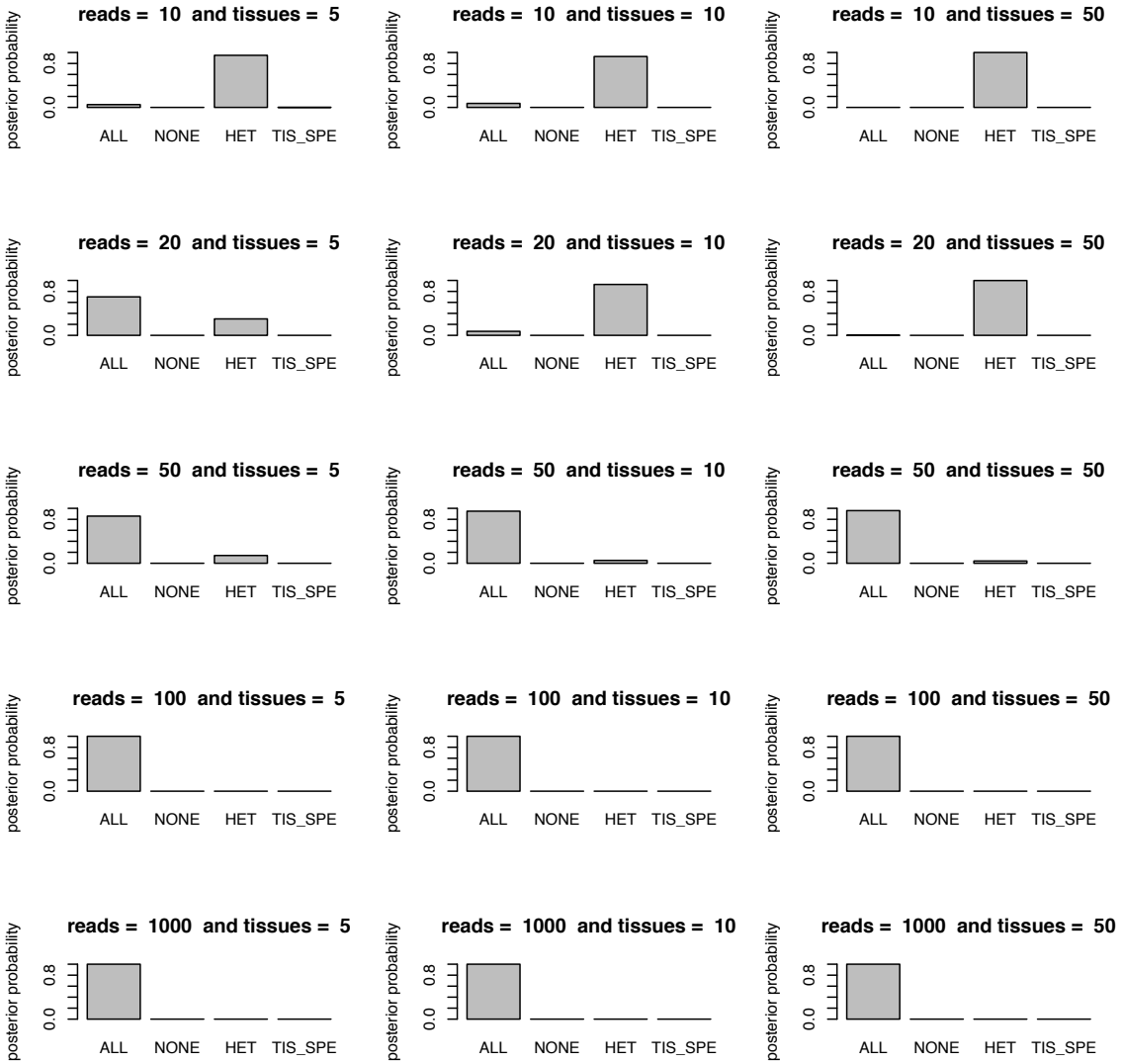


Figure 4.6: ASE simulations scenario I. Strong ASE effects across all tissues were simulated such that probability that the non-reference allele is observed is equal to 0.1. For each figure panel the posterior probability for the 'ALL' (ASE across all tissues), 'NONE' (no ASE across all tissues), 'HET' (heterogeneous ASE effects), and 'TIS_SPE' (tissue-specific ASE) states are shown. Each figure panel corresponds to a different scenario where the number of reads (reads) and/or the number of tissues (tissues) is varied.

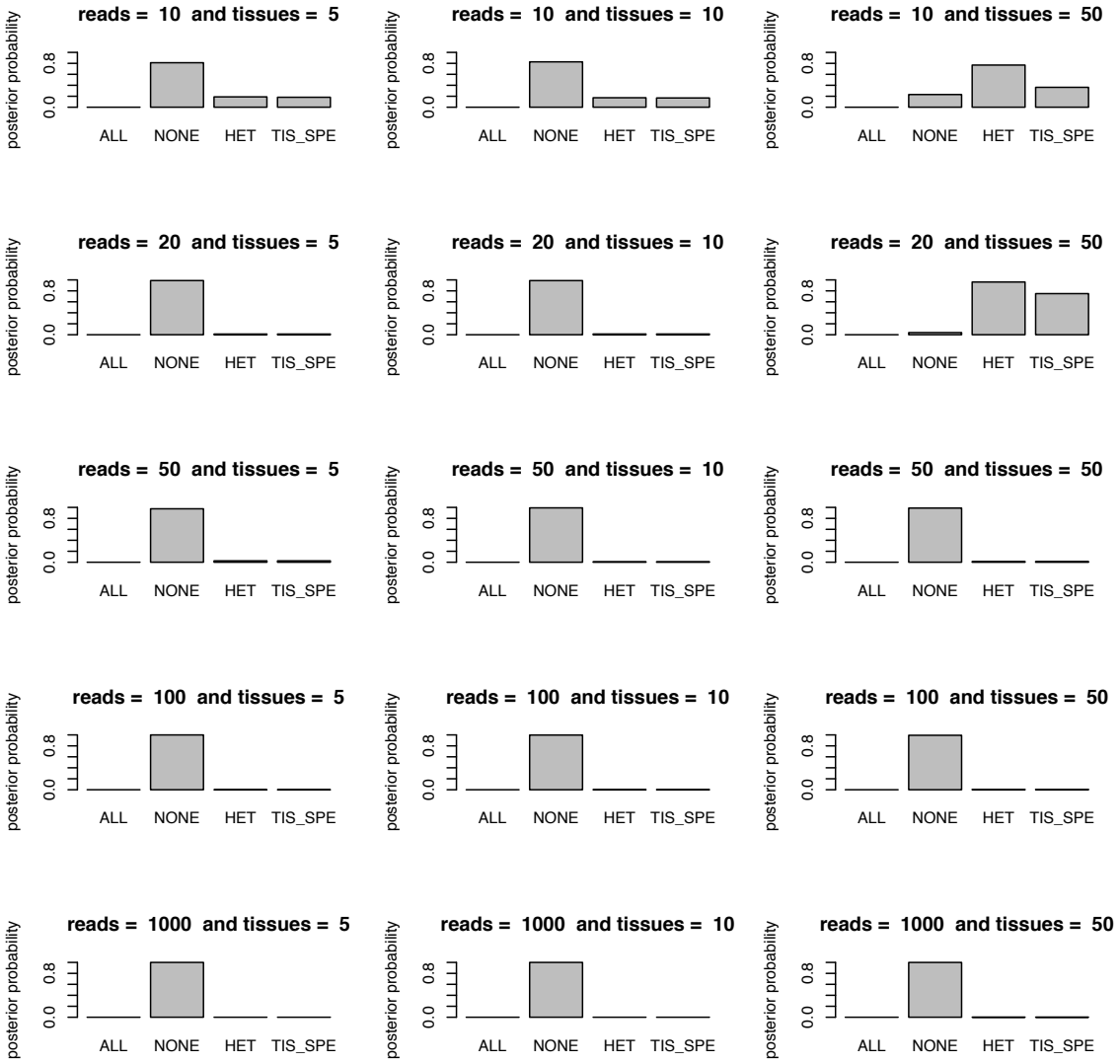


Figure 4.7: ASE simulations scenario II. No ASE effects across all tissues were simulated such that the probability that the non-reference allele is observed is equal to 0.5. For each figure panel the posterior probability for the 'ALL' (ASE across all tissues), 'NONE' (no ASE across all tissues), 'HET' (heterogeneous ASE effects), and 'TIS_SPE' (tissue-specific ASE) states are shown. Each figure panel corresponds to a different scenario where the number of reads (reads) and/or the number of tissues (tissues) is varied.

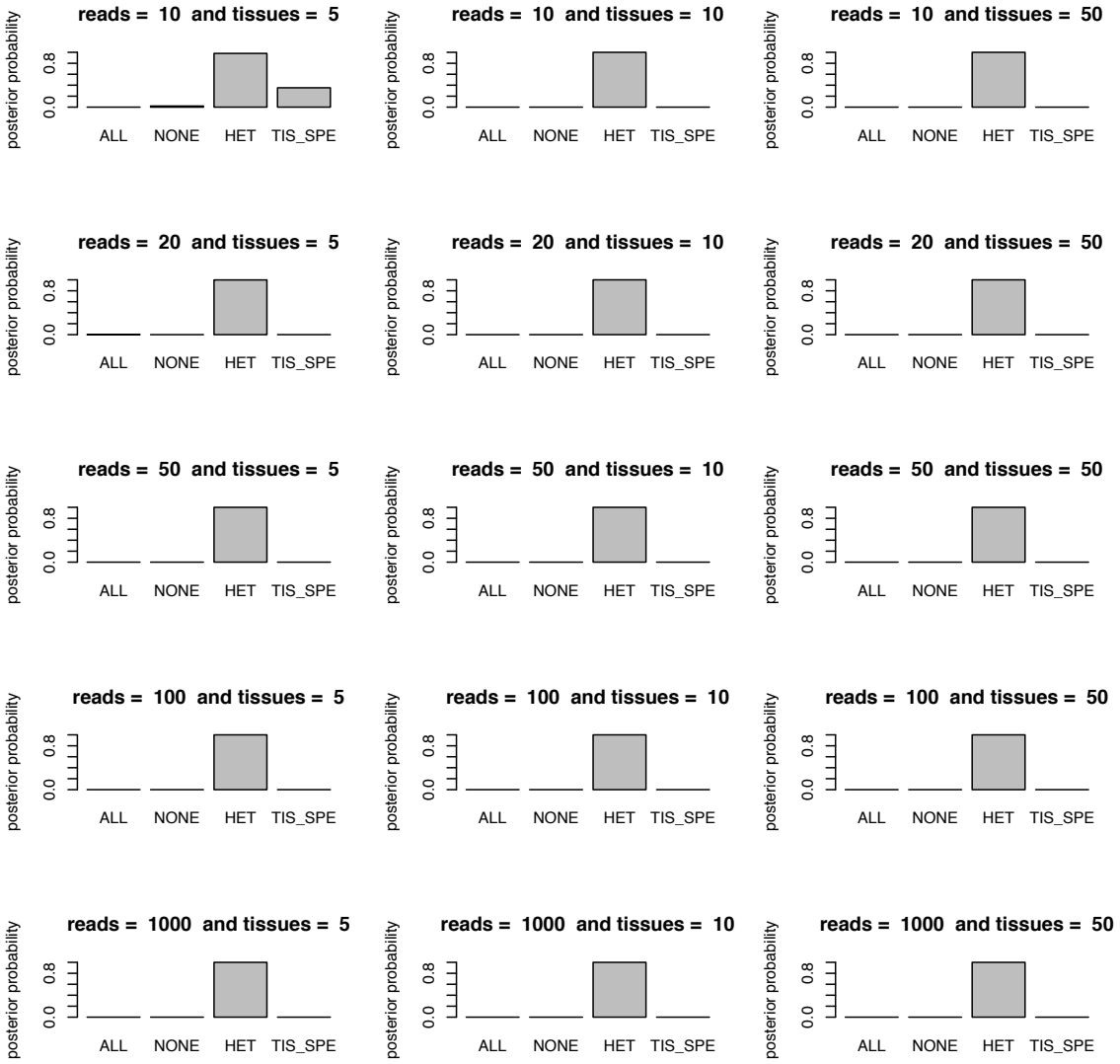


Figure 4.8: ASE simulations scenario III: heterogeneous ASE effects. Simulations were generated such that half the tissues had simulated ASE effects and the other half did not have ASE effects. For each figure panel the posterior probability for the 'ALL' (ASE across all tissues), 'NONE' (no ASE across all tissues), 'HET' (heterogeneous ASE effects), and 'TIS_SPE' (tissue-specific ASE) states are shown. Each figure panel corresponds to a different scenario where the number of reads (reads) and/or the number of tissues (tissues) is varied.

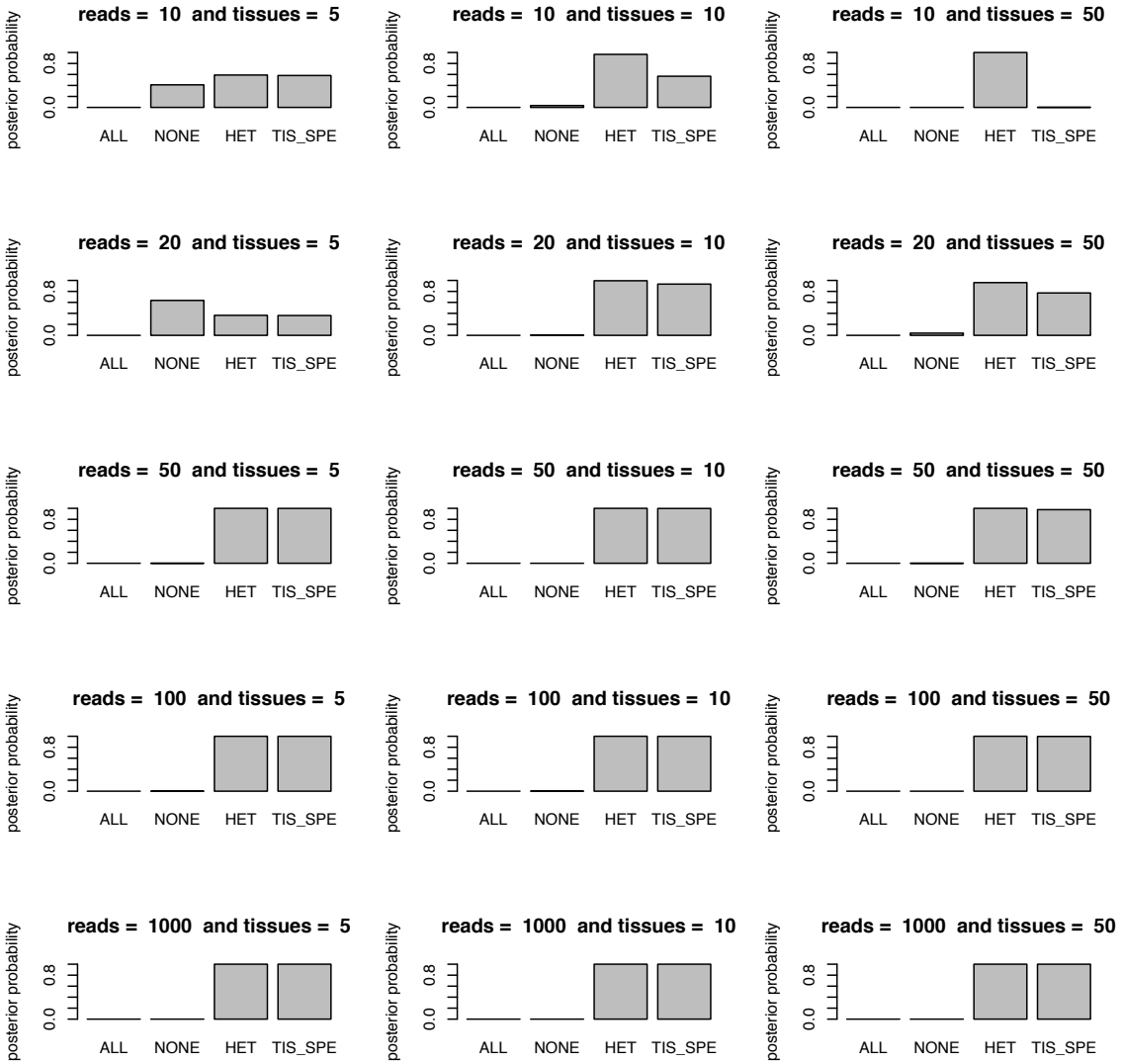


Figure 4.9: ASE simulations scenario III: tissue-specific ASE effects. Strong ASE effects were simulated for a single tissue, otherwise no ASE effects were introduced. For each figure panel the posterior probability for the 'ALL' (ASE across all tissues), 'NONE' (no ASE across all tissues), 'HET' (heterogeneous ASE effects), and 'TIS_SPE' (tissue-specific ASE) states are shown. Each figure panel corresponds to a different scenario where the number of reads (reads) and/or the number of tissues (tissues) is varied.

As in Pirinen et al. (2015), θ is represented as the probability that a read supports the reference allele. Furthermore, a mixture of two beta distributions were used to describe the prior for the group $\mathcal{M}$ and $\mathcal{S}$ allowing for the detection of ASE when either the reference or the non-reference allele was preferentially expressed. However, in the applications, given that I was interested in the NMD biological phenomenon, I chose to focus on the one-sided model where for group $\mathcal{S}$ and $\mathcal{M}$ the reference allele is preferentially expressed only. In `MAMBA` the option to run the two-sided model is `--two_sided`. In the R script the option to run the two-sided model is `two.sided =TRUE`. Next, as with the independent tissue model, one of the principal aims when we developed the GTM was to be able to characterize the configuration of the ASE effects for any given variant across multiple tissues. For any single variant the possible number of group configurations is 3^t, where t is the number of tissues. We decided to constrain the space of configurations into five ASE states that we found to be of interest and that followed the intuition of the independent tissue model:

- "NOASE" state: for any single variant application of the grouped tissue model to the data supports no ASE effects across all the tissues,
- "MODASE" state: for any single variant application of the grouped tissue model to the data supports moderate ASE effects across all the tissues,
- "SNGASE" state: for any single variant application of the grouped tissue model to the data supports strong ASE effects across all the tissues,
- "HET0" state: for any single variant application of the grouped tissue model to the data supports heterogeneous ASE effects where a mixture of no ASE effects and either moderate ASE or strong ASE effects are present, and
- "HET1" state: for any single variant application of the grouped tissue model to the data supports heterogeneous ASE effects where a mixture of moderate ASE and strong ASE effects are present.

We considered a tissue-specific sub-state of the heterogeneity states as well (as in the independent tissue model). The classification was conducted using a Gibbs sampler algorithm that generates posterior probability estimates for the variant that it falls into any of the 5 ASE states previously described. Furthermore, for each tissue it generates a posterior probability that the data supports either the $\mathcal{N}$, $\mathcal{M}$, or $\mathcal{S}$ group characterizing the ASE effect for that tissue.

Hierarchical Grouped Tissue Model (GTM*) The Hierarchical Grouped Tissue Model (GTM*) is an extension of the GTM to the case where many variants either in different genes, the same gene, or both, are analyzed simultaneously. One level of hierarchy was added to the model by introducing a vector of proportion parameters that determines the proportion of variants in each of the five states defined in the GTM. The idea is that through the use of a Gibbs sampler the parameters are learned and that information may be borrowed not only across multiple tissues, but also across multiple variants (as is likely to be important for PTVs). In `MAMBA` to analyze ASE data with GTM* or GTM the default option is `--indep=FALSE`, which is the default behavior. Otherwise, `--indep=TRUE` may be used to run the independent tissue model.

4.3.6 Application of the method in this thesis

In Chapter 7 of this thesis I applied the ASE methods described in this section to an RNA sequencing data set of multiple tissues from 173 individuals and an RNA sequencing data set from lymphoblastoid cell lines (LCLs) in 462 subjects from the 1000 Genomes Project.

4.3.7 Attributions

The methodology for assessing allele-specific expression across multiple tissues was conceived of and designed by Matti Pirinen and I. Matti Pirinen guided me in the derivation of the independent tissue model. I implemented the independent tissue model in the software `MAMBA`, and generated the simulations. I received useful feedback from my advisors on the methodology. Matti Pirinen designed GTM and GTM*, and I implemented a version in the software `MAMBA`. Tuuli Lappalainen and Emmanouil Dermitzakis provided useful feedback on the methodology and access to the RNA sequencing data set described in Chapter 7.

4.4 Assessing impact of rare variants proximal to splice junctions: the Splice Disruption Model (SDM)

4.4.1 Background and rationale

In Section 4.3 I presented methodology for assessing allele-specific expression. ASE methods are useful for studying the functional consequences of nonsense SNVs and

frameshift indels. However, at least to my knowledge, they are not useful for studying the functional impact of variants proximal to splice junctions, which are the variants that are usually predicted to disrupt splicing and believed to lead to aberrant transcript isoforms that undergo NMD degradation or create a truncated protein product. The major challenge is that variants proximal to splice junctions are usually rare (MacArthur et al., 2012) and because there will not be sufficient power to detect association for any given variant they are not amenable to standard QTL approaches. At the time I started considering the functional consequences of PTVs no method existed for the analysis of rare variants and their impact on splicing. As a result, I sought to consider the problem of assessing the extent and nature of splice disruption by rare DNA sequence variants proximal to the splice junctions.

4.4.2 Intuition

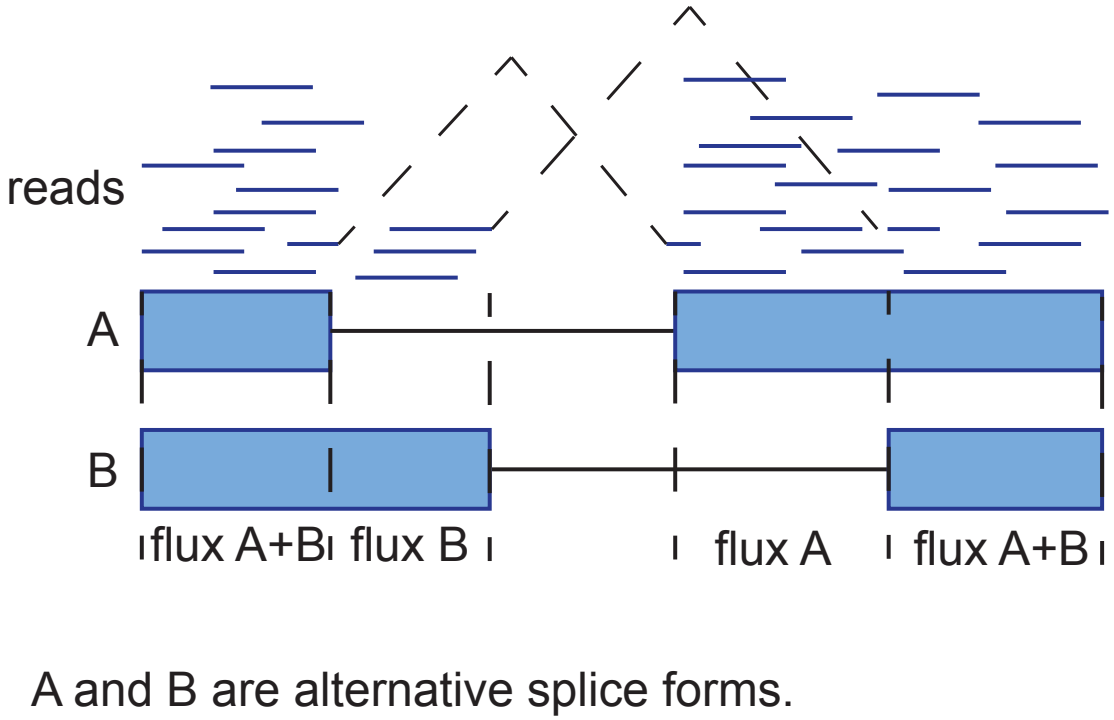


Figure 4.10: Diagram of splice junction quantification using FLUX Capacitor adapted from `http://flux.sammeth.net/capacitor.html`. Diagram of exonic structure of two splice isoforms is shown along with overlapping reads. The reads are used to deconvolute and quantify the two isoforms using flux A, flux B, and flux A+B.

RNA-seq data can be used to quantify different structures of a protein-coding gene including: 1) splice junctions, 2) exons, 3) introns, and 4) transcript isoforms.

Software tools available for quantification of expression based on these gene annotations (i.e. exons, transcripts, genes, and splice junctions) include SAMtools (Li et al., 2009), FLUXCapacitor (Montgomery et al., 2010), Cufflinks (Trapnell et al., 2010), and Scripture (Grabherr et al., 2011). In this thesis I focused on splice junction quantifications generated by FLUXCapacitor (FLUX) developed at the Centre de Regulacio Genomica (CRG) in Barcelona, Spain (Montgomery et al., 2010), which was used in the GTEx project and in the Geuvadis RNA sequencing project. The program attempts to deconvolute the reads into the separate gene bodies. For example, in Figure 4.10 multiple reads are spanning a gene model with two transcript isoforms: A and B. Some of the reads splitting across multiple exons are then used to quantify the relative abundance of different splice junctions in the gene model. In addition, the reads spanning one or multiple exons are used to quantify the relative abundance of different introns and exons in the gene model.

In a population-based RNA sequencing study like the Geuvadis and the GTEx projects multiple gene model quantification data are available across hundreds of individuals. Furthermore, genotype data across hundreds of rare variants proximal to splice junctions are available because DNA sequencing was conducted. Hence, I thought that one way to study the functional consequences of rare variants proximal to splice junctions was by borrowing as much information as possible from the splice junction quantification values (relative to the population) of individuals that share a common property. In this instance the common property is that the group of individuals contain a non-reference allele in a variant that is a genomic distance d away from a splice junction body. To estimate the impact of rare variants proximal to splice junctions we developed a statistical method referred to as the **S**plice **D**isruption **M**odel (SDM) used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population of all variants (at a distance d) proximal to a splice junction, without sub-population identity information. The idea is that some of the variants will disrupt splicing efficiency while others may not. Thus, we use the splice junction values for individuals carrying a variant allele to estimate the mixture components and estimate the extent of splice disruption (Figure 4.11).

4.4.3 Implementation

In this subsection I present the mathematical derivation of the SDM algorithm and give a description of the bioinformatic implementation.

Setup. Let k be the number of splice junctions annotated in a reference transcript model. Let n be the number of individuals in the RNA sequencing experiment. Let $y_i^{(j)}$ be the splice junction measurement of the number of reads spanning the splice junction normalized by library size (Lappalainen et al., 2013) for individual i ($i = 1, \ldots, n$) in splice junction j ($j = 1, \ldots, k$). Assume that for each splice junction j the measurements of the number of reads spanning the splice junction normalized by library size across all individuals have been transformed to a normal distribution with mean 0 and standard deviation equal to 1 by rank-based inverse normal transformations (Qiu et al., 2013), so that $\tilde{y}_i^{(j)}$ is the standardized measurement for individual i in splice junction j.

Let $v^{(d)}$ denote the collection of variants at a distance d to any of the k splice junctions in the reference transcript set model. Let m be the number of variants in $v^{(d)}$. Assume that variants a distance d from a splice junction either have no effect or the same negative effect whatever the splice site is and hence we combine information. For each d studied let $\boldsymbol{y}^{[d]}$ be the collection of normalized values $\tilde{y}_i^{(j)}$ for splice junction j and individual i if individual i has a non-reference allele at the variant in $v^{(d)}$ that is a distance d away from splice junction j. Let t be the number of standardized individual measurements in $\boldsymbol{y}^{[d]}$ so that $\boldsymbol{y}^{[d]} = \left(y_1^{[d]}, \ldots, y_t^{[d]}\right)$, and let ℓ be an index for $1, \ldots, t$.

We want to estimate, at each distance, the proportion of carriers belonging to

1. the no splice disruption group (0), or
2. the splice disruption group (1).

More formally we assumed that $\boldsymbol{y}^{[d]}$ is a collection of t conditionally independent observations from a univariate normal mixture distribution with density

$$p(\boldsymbol{y}^{[d]}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{t} \left[\pi_1 \mathcal{N}\left(y_i^{[d]}; \mu_1, \sigma_1^2\right) + \pi_0 \mathcal{N}\left(y_i^{[d]}; \mu_0, \sigma_0^2\right)\right], \tag{4.19}$$

where $\mathcal{N}(y; \mu, \sigma^2)$ denotes the density function of the univariate normal distribution with mean μ and variance σ^2; $\boldsymbol{\pi}$ are the mixture proportions which are constrained to be non-negative and sum to unity, i.e. $\pi_0 = 1 - \pi_1$ (hereinafter I will refer to π_1 only); $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the component specific parameters, where μ_0 and σ_0^2 being the specific components to the no splice disruption group are kept fixed as $\mu_0 = 0$ and $\sigma_0^2 = 1$. Let γ_ℓ be an indicator variable denoting the membership of $y_\ell^{[d]}$ in either the splice disruption group (1) or the no splice disruption group (0).

Mixture model. Then, the SDM is the following mixture model:

$$\begin{aligned}
\pi_1 &\sim \text{Beta}(1,1) \\
\gamma_\ell|\pi_1 &\sim \text{Bernoulli}(\pi_1) \\
\mu_0 &= 0 \\
\mu_1 &\sim \mathcal{N}(-1,1) \\
\sigma_0^2 &= 1 \\
\sigma_1^2 &\sim \text{IG}(1,1) \text{ (IG denotes inverse-gamma)} \\
y_\ell^{[d]}|\gamma_\ell,\mu_0,\mu_1,\sigma_0^2,\sigma_1^2 &\sim \mathcal{N}\left(y_\ell^{[d]};\mu_{\gamma_\ell},\sigma_{\gamma_\ell}^2\right). \qquad (4.20)
\end{aligned}$$

The group membership of each of the alternate allele carriers is unknown in advance. The prior for π is uniform, that is Beta $(1,1)$, because we are not favoring any possible value of π_1 over the others.

Since the measurement values for each splice junction j have been standardized jointly we model the effects of variants at a particular distance from a splice junction as either drawn from the general population distribution, i.e. the standard normal distribution $\mathcal{N}(\mu_0 = 0, \sigma_0^2 = 1)$, or from what we call a splice disruption specific normal distribution with unknown shift in mean μ_1 and unknown variance σ_1^2. The mean μ_1 by a normal distribution with mean -1 and variance 1 is meant to reflect the interest in those variants that decrease splicing efficiency. In principle, we could also have used another component to reflect variants with a putative increase in splicing efficiency. The prior for the variance parameter σ_1^2 is the inverse gamma distribution with parameters $\alpha = 1$ and $\beta = 1$. This distribution is relatively flat and covers the region where we expect the variance parameter to reside (95% probability region [0.278,39.50]) because the observations have been standardized and altogether have variance of 1.

Gibbs sampler. To fit the mixture of univariate normal distributions to the splice junction quantification data we use a Gibbs sampler. The Gibbs sampler is a Markov Chain Monte Carlo (MCMC) algorithm that constructs a dependent sequence of parameter values whose distribution approximates (when it converges) the target joint posterior distribution (Geman and Geman, 1984; Casella and George, 1992; Hoff, 2009) as described in Section 3.3.3. We used the following Gibbs sampler to analyze the SDM, with superscripts for the variables denoting their value after the corresponding iteration.

1. Initialize $\pi_1^{(0)}, \mu_1^{(0)}, (\sigma_1^2)^{(0)}$, and $\gamma_\ell^{(0)}$ for all ℓ.

2. Repeat for iter $= 1, 2, \ldots, n_{\text{burn}} + n_{\text{niter}}$

   (a) For $\ell = 1, 2, \ldots, t$, generate $\gamma_\ell^{(\text{iter})} \sim \text{Bernoulli}\left(p_\ell^{(\text{iter})}\right)$ where

$$p_\ell^{(\text{iter})} = \frac{\pi_1^{(\text{iter}-1)} \mathcal{N}\left(y_\ell^{[d]}; \mu_1^{(\text{iter}-1)}, (\sigma_1^2)^{(\text{iter}-1)}\right)}{(1 - \pi_1^{(\text{iter}-1)}) \mathcal{N}(0, 1) + \pi_1^{(\text{iter}-1)} \mathcal{N}\left(y_\ell^{[d]}; \mu_1^{(\text{iter}-1)}, (\sigma_1^2)^{(\text{iter}-1)}\right)}.$$

   (b) Generate $\pi_1^{(\text{iter})} \sim \text{Beta}\left(1 + \sum_\ell \gamma_\ell^{(\text{iter})}, 1 + t - \sum_\ell \gamma_\ell^{(\text{iter})}\right)$.

   (c) Update:

$$(\sigma_1^2)^{(\text{iter})} \sim \text{IG}\left(1 + \frac{\sum_\ell \gamma_\ell^{(\text{iter})}}{2}, 1 + \frac{1}{2} \sum_\ell \gamma_\ell^{(\text{iter})} \left(y_\ell^{[d]} - \mu_1^{(\text{iter}-1)}\right)^2\right)$$

$$\mu_1^{(\text{iter})} \sim \mathcal{N}\left(\frac{\frac{-1}{1} + \frac{\sum_\ell y_\ell^{[d]} \gamma_\ell^{(\text{iter})}}{(\sigma_1^2)^{(\text{iter})}}}{\frac{1}{1} + \frac{\sum_\ell \gamma_\ell^{(\text{iter})}}{(\sigma_1^2)^{(\text{iter})}}}, \left(\frac{1}{1} + \frac{\sum_\ell \gamma_\ell^{(\text{iter})}}{(\sigma_1^2)^{(\text{iter})}}\right)^{-1}\right).$$

As in Section 3.3.3 the Gibbs update is obtained by doing a lookup of the posterior hyperparameters using the conjugate prior table in the Wikipedia page `http://en.wikipedia.org/wiki/Conjugate_prior`.

Software implementation and MCMC diagnostics. I implemented SDM in the sofware `MAMBA` (module option `--module SDM`). By default I run the algorithm for $n_{\text{burn}} + n_{\text{niter}} = 2000$ iterations and discard the first $n_{\text{burn}} = 1000$ iterations from the analysis. Although I assigned default choices for n_{burn} and n_{niter}, before making inferences from a Gibbs sampler it is worthwhile to determine if there might be any problems with the Gibbs sampler run (Hoff, 2009). There are several recommendations to diagnose whether a Gibbs sampler is properly approximating the joint posterior distribution of the parameters (Hoff, 2009). A common practice is to examine a traceplot of the parameters that are being estimated where on the x-axis we display the iteration number and on the y-axis we display parameter value from the Gibbs sampler for the corresponding iteration number (Figure 4.12 contains examples of traceplots). In addition, it is common practice to examine the histogram of the parameter samples obtained from the Gibbs sampler. Another suggested approach is to compute Geweke's test statistic (Geweke et al., 1991) using the `R` package `coda` (Plummer et al., 2006). Geweke's approach was to test for equality of the means of the first (default choice is first 10%) and last (default choice is last 50%) part of a Markov chain after burn-in (Plummer et al., 2006). Using the `coda` package the test

statistic obtained is a standard Z-score. Hence, values less than -3 or greater than 3, for example, suggest that the two means are not equal and that a larger number of iterations should probably be applied to the Gibbs sampler. In Figures 4.12 and 4.13 I compute Geweke's statistic for parameters π_1, labelled 'gpi', and μ_1, labelled 'gmean'. In addition, I added the options `--burn` and `--niter` to change the number of iterations to discard and the number of total iterations. This can be useful when reanalyzing the data and to assess whether the application resulted in good mixing, which can be larger than the default options[3].

Assessing evidence for a second component. There are Bayesian approaches to assess the evidence for the existence of a second component (Fraley and Raftery, 2002). In principle we can compute Bayes Factors with marginal likelihoods. However, this requires integration over the parameter space, which can be difficult to compute analytically. Alternatively, reversible jump MCMC methods can be applied for Bayesian model comparison (Fraley and Raftery, 2002). Instead of Bayesian model comparison we proposed a reasonable test statistic for the second component

$$T^* = \frac{1}{n_{\text{niter}}} \sum_{t=n_{\text{burn}}+1}^{n_{\text{burn}}+n_{\text{niter}}} \pi_1^{(t)} |\mu_1^{(t)}|, \tag{4.21}$$

for each d. The intuition behind the statistic is that we want to get a sense as to whether the magnitude in shift in the mean, given by μ_1, and also whether the proportion π_1 is beyond what we would expect by chance. The statistic itself gives an average of the product between the estimated proportion of variants in the second mixture distribution and the shift in the mean for that second mixture across all the iterations in the Gibbs sampler that are used to estimate the mixture of normals.

In the implementation a standardized statistic is computed

$$T' = T^* \Bigg/ \sqrt{\frac{1}{n_{\text{niter}} - 1} \sum_{t=n_{\text{burn}}+1}^{n_{\text{burn}}+n_{\text{niter}}} \left(\pi^{(t)} |\mu_1^{(t)}| - T^* \right)^2}, \tag{4.22}$$

to calculate an empirical P value.

An empirical P value is obtained by calculating a null distribution $\left(T'_\eta\right)_{\eta=1,\ldots,\text{perms}}$ where for each null iteration η, we generate a new collection of $\boldsymbol{y}^{[d]}$ values as in the setup. However, for each j where an individual with a variant allele is identified an individual with a homozygous reference allele is randomly sampled and their standardized splice junction measurement for splice junction j is included instead. Then,

[3] `http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo`

we apply the Gibbs sampler to the new collection of $\boldsymbol{y}^{[d]}$ and calculate the test statistic using equation 4.22.

Finally, an empirical P value is obtained by calculating

$$p = \frac{\sum_{\eta=1}^{\text{perms}} I\left(T'_\eta \geq T'\right) + 1}{\text{perms} + 1}, \tag{4.23}$$

where I is an indicator function, i.e. equal to 1 if it meets the condition in $(\cdot)$, and 0 otherwise.

4.4.4 Simulations

To illustrate the performance of the method and to study its properties I conducted a series of simulation studies. I applied the SDM with $n_{\text{burn}} + n_{\text{niter}} = 2,000$ iterations and discard the first $n_{\text{burn}} = 1000$ iterations. I show two figures representing the following two mixture distributions:

1. $0.5\mathcal{N}(-1, 1) + 0.5\mathcal{N}(0, 1)$ (Figure 4.12), and
2. $0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(0, 1)$ (Figure 4.13).

For each figure I present a histogram of the estimated parameters (μ_1, π_1, and σ_1^2) after the burn-in iterations. Then, I present a plot showing the samples of the parameters π_1 and μ_1 for each iteration of the MCMC algorithm (including the burn-in stage). In the simulation study I varied the number of standardized splice junction measurement values (ncopies $= 20, 50, 100, 200, 2000$) to illustrate the changes in performance with larger data sets.

The results of the simulation study show that the method achieves good mixing behavior when the effects are large ($\mu_1 \geq -1$) even in situations where the number of measurements available for analysis is small (≤ 100, Figures 4.12 and 4.13).

4.4.5 Application of the method in this thesis

In this thesis I applied the SDM algorithm to rare variant and RNA sequencing data from the Geuvadis and GTEx projects. The results are presented in Chapter 7.

4.4.6 Limitations and next steps

A major limitation of the SDM algorithm presented in this section is that the number of mixture components is fixed ($k = 2$). However, it is possible that variants proximal to splice junctions may have varying effects (some may actually increase splicing efficiency!). Hence, in future work I would like to be able to estimate the number of mixture components and assess the existence of these components in a Bayesian framework.

Another limitation of the approach taken here is that the measurement value is univariate. In the GTEx study design an individual may have multiple values because multiple tissues may have been collected. At the moment this information is not exploited, and I envisage this to be a worthwhile next step to pursue in future research for the analysis of rare variants.

4.4.7 Attributions

I conceived and designed the idea to exploit as much information as possible for variants proximal to splice junctions by first normalizing with respect to the general population and then estimating the mixture component. Matti Pirinen recommended Matthew Stephens' thesis, which presents Bayesian methods for finite mixtures of normal distributions, and guided me in the derivation of the mixture model and its implementation. My supervisors provided useful feedback and comments on the approach.

4.5 Discussion

In order to properly understand the functional consequences of predicted protein truncating variants experimental data should be considered. Together, the statistical, bioinformatic, and computational methods presented in this chapter were approaches developed to address the challenge of improving annotation of PTVs.

The RNA-seq informed DNA variant annotation bioinformatic framework I presented in Section 4.2 allows researchers to integrate RNA-seq transcript isoform quantification data for tissue(s) of interest and exclude *in silico* variant annotations that are not corroborated by empirical data. In the GTEx exome sequencing study I showed that over 60% of PTVs have partial PTV annotation support across all the transcript isoforms and that as much as 50% of the PTV annotation data may be

irrelevant. This may have severe consequences when conducting rare variant association studies that first begin with a permissive criteria for identifying a variant as a PTV. More importantly, this framework can help guide variant annotation in clinical sequencing studies as shown with the *SGCB* example.

The ASE model presented in Section 4.3 was developed to study the functional consequences of nonsense SNVs and frameshift indels across multiple tissues, and the SDM model presented in Section 4.4 was developed to study the functional consequences of variants proximal to splice junctions. The main motivation was that by studying the functional consequences of PTVs we would be able to improve annotation of PTVs.

The methods presented in this chapter are implemented in the software package `MAMBA`, which can be downloaded from `http://www.well.ox.ac.uk/~rivas/mamba/`. Data visualization for the output of the ASE and SDM models is currently being supported by the GTEx portal `http://gtexportal.org`.

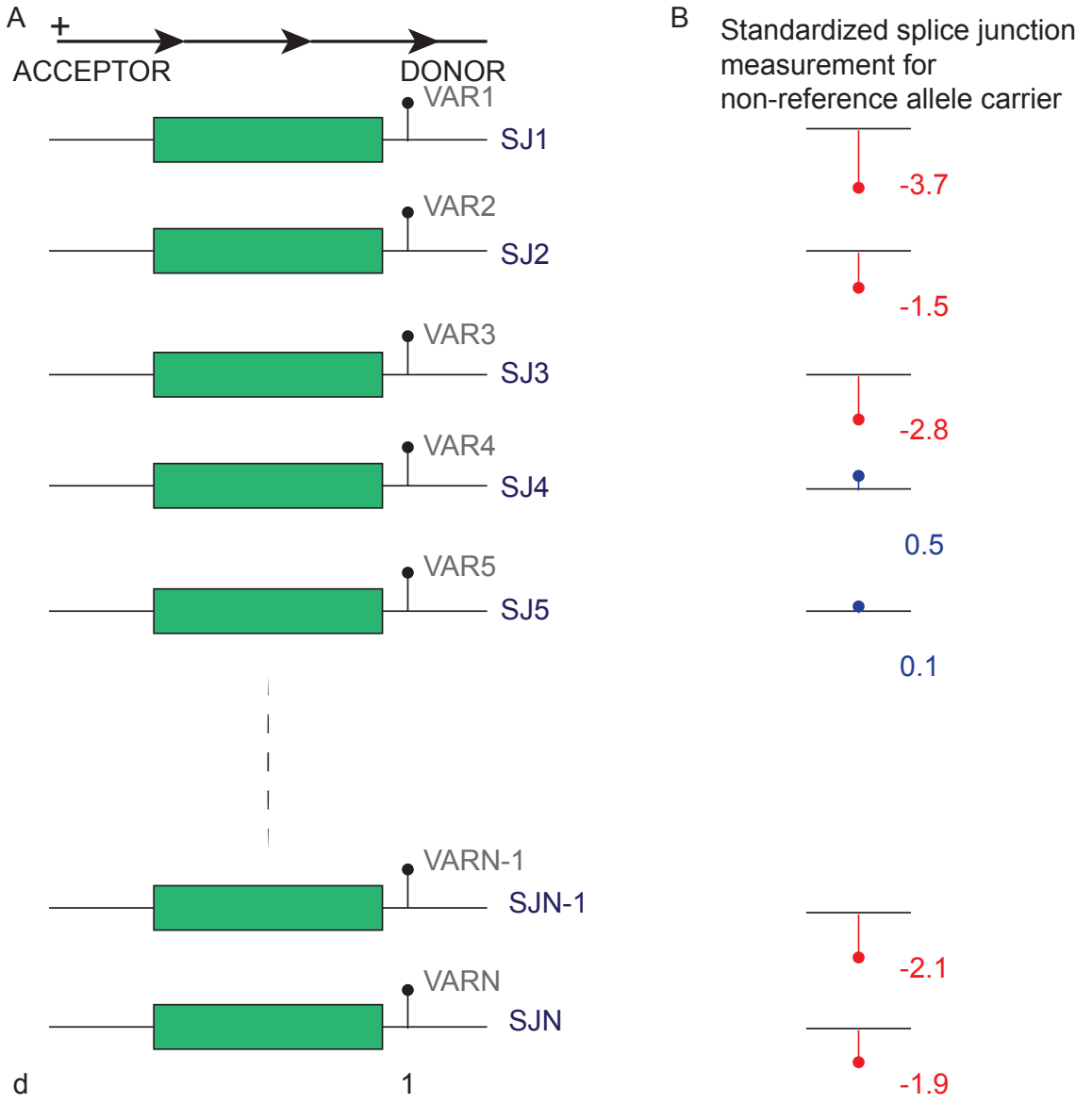


Figure 4.11: Diagram of signal sought using the SDM algorithm. (A) N variants are identified that are at a distance 1 base pair away from a donor splice junction in the reference transcript set model. In this toy diagram for every splice junction in the reference transcript set model there exists a variant 1 base pair away from the donor splice site. (B) The measurement for the individual is the number of reads spanning a reference transcript set splice junction normalized by the library size. Additional procedures to adjust for technical confounding variables in the data may be used (Stegle et al., 2012). In the diagram the splice junction measurement value, after rank-based inverse normal transformation has been applied, for the individual with a non-reference allele at each 'VAR' is shown. In the diagram variants at a distance $d = 1$ from the donor splice junction are shown. These standardized measurements are then used to estimate the proportion of variants that disrupt splicing for $d = 1$, and to estimate the magnitude of the effect for those variants that are classified as likely disrupting splicing.

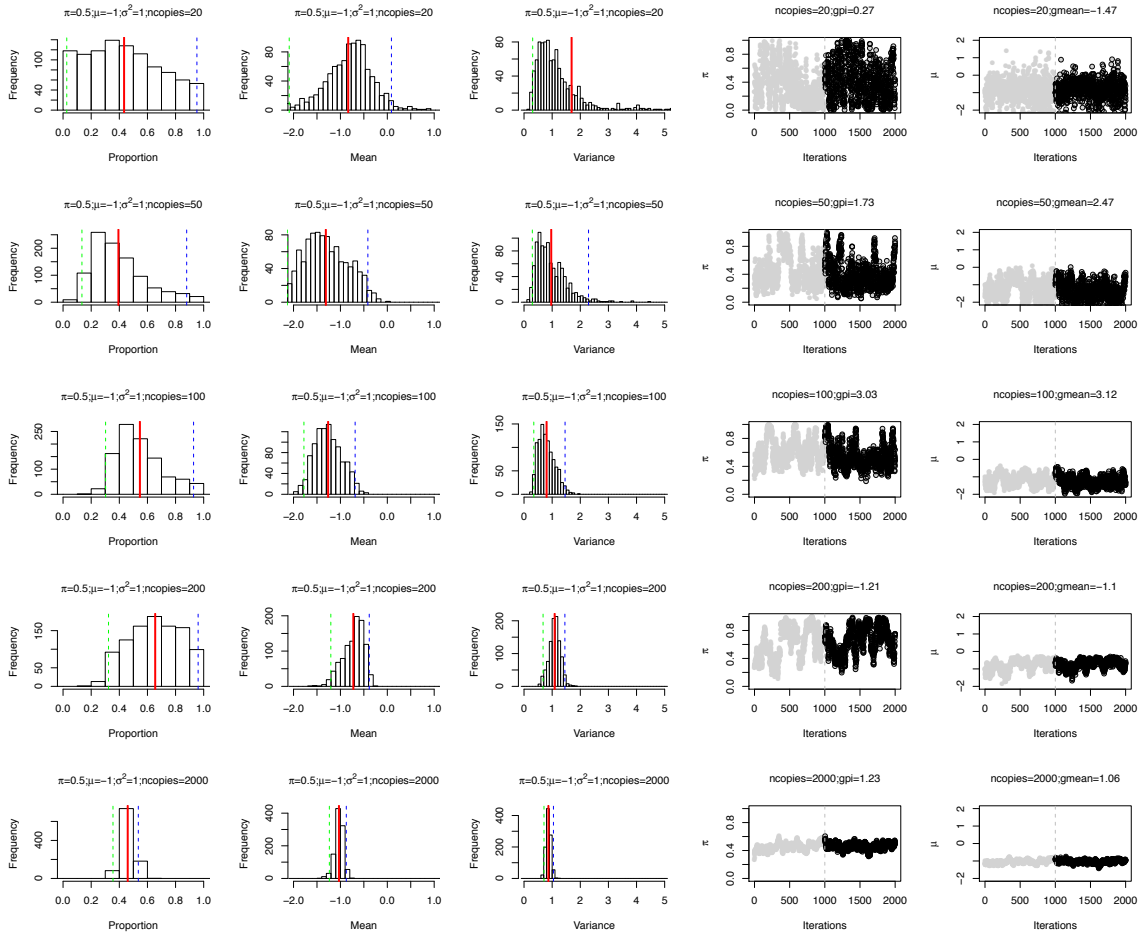


Figure 4.12: SDM simulations scenario I: $0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(0,1)$. In each row results for different number of standardized splice junction measurement values (ncopies $= 20, 50, 100, 200, 2000$) are shown. In the first column histograms of the sample values of the estimated parameter π_1, i.e. the proportion of measurement values coming from the mixture component with shift in the distribution μ_1 are displayed. In the second column histograms of the sample values of the estimated parameter μ_1 are shown. In the third column histograms of the sample values of the estimated parameter σ_1^2 are shown. Dotted lines represent quantiles of the samples of the parameter in the sampler (average value: red solid line, 2.5 % quantile of Gibbs samples of the parameter [2.5]: green, 97.5% quantile of Gibbs samples of the parameter [97.5]: blue). In the fourth column traceplots of the sample values at each iteration (including the burn-in stage [light-gray]) for parameter π_1 are shown. In the fifth column traceplots of the sample values at each iteration (including the burn-in stage [light-gray]) for parameter μ_1 are shown. To examine convergence of the Gibbs sampler Geweke's statistic was computed for parameters π_1, labelled as 'gpi', and μ_1^2, labelled as 'gmean', which can be interpreted as a standardized Z-score.

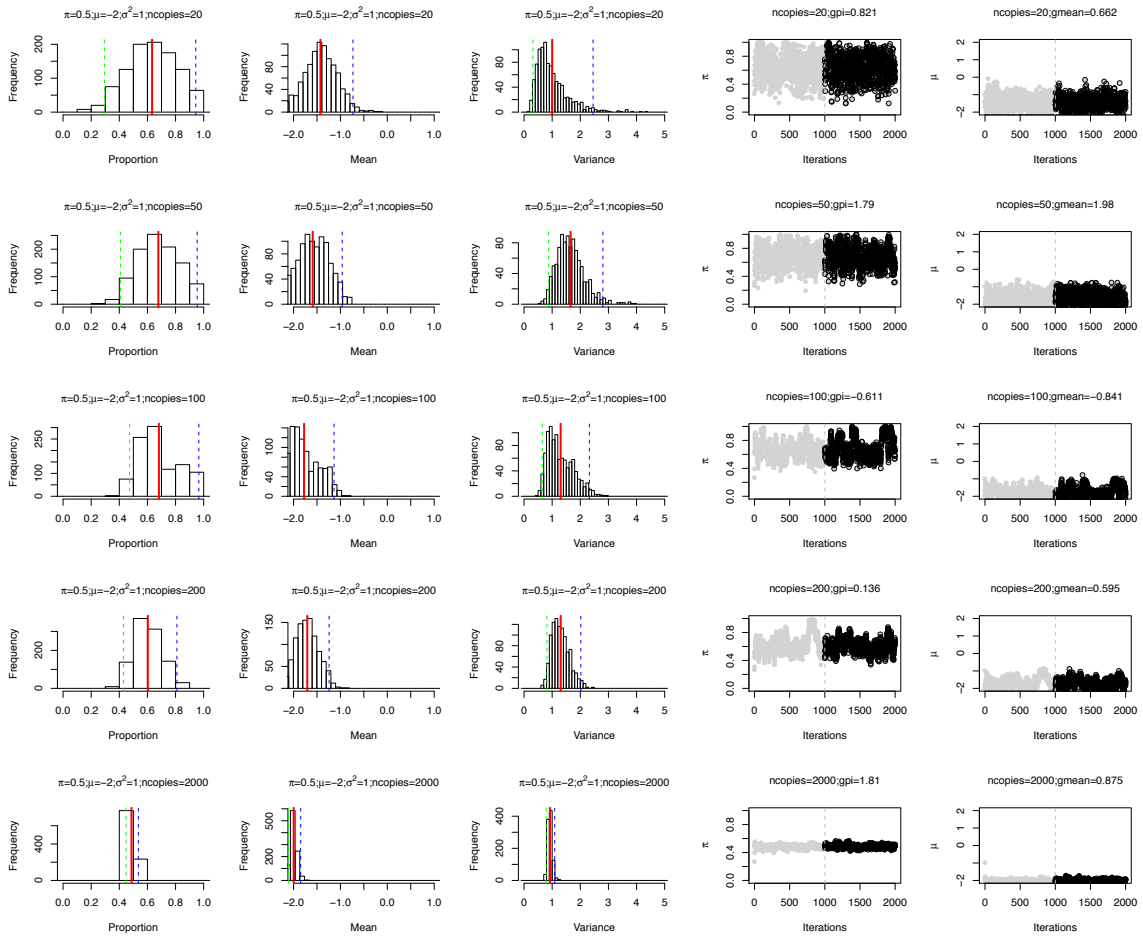


Figure 4.13: SDM simulations scenario II: $0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(0,1)$. In each row results for different number of standardized splice junction measurement values (ncopies $= 20, 50, 100, 200, 2000$) are shown. In the first column histograms of the sample values of the estimated parameter π_1, i.e. the proportion of measurement values coming from the mixture component with shift in the distribution μ_1 are displayed. In the second column histograms of the sample values of the estimated parameter μ_1 are shown. In the third column histograms of the sample values of the estimated parameter σ_1^2 are shown. Dotted lines represent quantiles of the samples of the parameter in the sampler (average value: red solid line, 2.5 % quantile of Gibbs samples of the parameter [2.5]: green, 97.5% quantile of Gibbs samples of the parameter [97.5]: blue). In the fourth column traceplots of the sample values at each iteration (including the burn-in stage [light-gray]) for parameter π_1 are shown. In the fifth column traceplots of the sample values at each iteration (including the burn-in stage [light-gray]) for parameter μ_1 are shown. To examine convergence of the Gibbs sampler Geweke's statistic was computed for parameters π_1, labelled as 'gpi', and μ_1^2, labelled as 'gmean', which can be interpreted as a standardized Z-score.

Chapter 5

Application of methods to the study of protein truncating variants and their relevance to medical traits: disease studies

Fewer than 10 PTVs have been associated to a common disease (Ogura et al., 2001; Cohen et al., 2006; Nejentsev et al., 2009; Rivas et al., 2011; Jørgensen et al., 2014; Myocardial Infarction Genetics Consortium, 2014; Moltke et al., 2014; Lim et al., 2014b; Flannick et al., 2014). However, results from a recent study (Flannick et al., 2013) suggests that PTVs found to associate with monogenic disease are probably subject to overestimation of penetrance and likely contributing to disease predisposition of the common form of the trait, suggesting that the distinction between what is monogenic and common is blurred. In this chapter I present results from three studies where I, and colleagues (when conducted as part of a consortium), attempted to identify additional PTVs associated to common diseases. First, I present results from a collaborative study to address the contribution of PTVs in 750 genes involved in the "DNA repair" pathway to breast cancer predisposition where I participated in analyzing a pooled sequencing data set of $1,150$ breast cancer patients. I applied methodology that was available at the time of the study and update the analysis with new approaches and data sets described in Chapter 3. Second, I present analysis of PTVs in 25 genes from a published targeted sequencing data set that were chosen because they were proximal to common variants identified to contribute disease predisposition for at least two of the six studied autoimmune diseases in genome-wide association studies. I studied association of PTVs to any of the six autoimmune diseases in this data using the C-alpha MRP approach presented in Chapter 3. Lastly, I present association analysis of PTVs to type 2 diabetes risk. I applied methodology

in Chapter 3 to a multi-ethnic T2D case-control exome sequencing data set from the Genetics of Type 2 Diabetes (GoT2D) and the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) projects.

For each of the studies in this chapter, I first give a brief overview of the main study where I describe the samples collected, the data generated, the data QC, results that were generated, and provide explicit attributions. Afterwards, I describe my role in the study providing additional details that were not explicitly described in either publications describing the study or manuscripts that are in preparation presenting the primary findings. Finally, I present results from my analysis of the data.

I hope that these findings, and the approach taken, will contribute to our current understanding of PTVs and the role they play in common diseases.

5.1 Mosaic protein truncating variants in *PPM1D* contribute to breast and ovarian cancer predisposition

5.1.1 Overview of the study

The breast cancer pooled sequencing study presented in this section was conducted as part of the Wellcome Trust Case Control Consortium (WTCCC) effort led by Nazneen Rahman (Institute of Cancer Research) to investigate the role of PTVs in breast cancer predisposition. As part of the analysis group of the consortium, I contributed to the analysis of the data and the development of associated software.

To set the scene, I summarize the design of the study, the materials and methods, and main findings of the study in this subsection.

In the results subsection I will briefly describe the major findings and provide details of my role in identifying the *PPM1D* PTV signal. Then, I will present additional data that corroborates the finding and I present updated analysis with the Bayesian similar effects model for case control (SEMCC) data (presented in Section 3.1) and the ExAC consortium reference data set. The results from this study were published in Nature (Ruark et al., 2012).

Study design. To assess the contribution of PTVs to breast cancer predisposition the consortium conducted two phases of sequencing: 1) discovery, and 2) follow-up and replication.

Discovery phase. Phase I of the study consisted of pooled sequencing of 7,789 targets totaling 1.7 megabases (Mb) of genomic sequence in lymphocyte DNA from 1,150 women with breast cancer. My main involvement in this project was in the first phase.

Targets were chosen by colleagues at the Institute of Cancer Research to overlap exonic sequence of genes annotated as "DNA repair" in the gene ontology (GO) database or genes identified to be an interacting partner of *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *CHEK2*, and *PALB2* with high confidence (≥ 0.9) in the STRING database (Ruark et al., 2012). These genes were chosen because they likely participated in the DNA repair pathway in humans, which has been shown to play a critical role in breast cancer predisposition (Rahman et al., 2006).

Forty-eight pools of DNA consisting of 24 individuals per pool (200 nanogram (ng) of DNA per individual) were created. Target enrichment was performed using Agilent SureSelect by the high-throughput Genomics team at the Oxford Genomics Centre in the Wellcome Trust Centre for Human Genetics. Sequencing was performed by the high-throughput Genomics team at the Oxford Genomics Centre and MRC hub in the Wellcome Trust Centre for Human Genetics in Oxford using a Illumina HiSeq 2000 sequencing instrument with one lane of sequencing dedicated per pool generating $2 \times 100-$bp reads. An average coverage $119 - 231$x was achieved per sample.

Variant calling was performed with `Syzygy version 1.2.2-1.2.4` by Peter Humburg (a post-doctoral fellow based at the Wellcome Trust Centre for Human Genetics).

Variants were annotated by Peter Humburg using a customized script written at the Wellcome Trust Centre for Human Genetics using ENSEMBL version 65 reference transcript set. Single nucleotide nonsense variants, coding frameshift indels, and essential splice disrupting variants were considered as PTVs in this experiment.

Genes were ranked and prioritized in an initial phase of case only sequencing (1,150 breast cancer samples).

Follow-up and replication phase. Phase II of the study was primarily conducted by colleagues at the Institute of Cancer Research.

First, *PPM1D* was chosen for follow-up and replication targeted sequencing in additional breast and ovarian cancer cases, and population based controls. *PPM1D* gene Sanger sequencing was performed in 2,456 cases and 1,347 population based controls. PCR primers were designed using exon-primer from the UCSC genome browser (Ruark et al., 2012). PCR reactions were performed using the Qiagen multiplex

PCR kit, and amplicons were unidirectionally sequenced using the BigDye terminator cycle sequencing kit on an ABI3730 automated sequencer (ABI PerkinElmer). The sequencing traces were analyzed by two analysts using both automated software (Mutation Surveyor, SoftGenetics) and manual visual inspection (Ruark et al., 2012). In total, 10 PTVs were identified in the cases and 0 in the controls.

Then, since all of the PTVs detected occured in the final exon of *PPM1D*, replication targeted sequencing of the final exon in *PPM1D* was conducted in $5,325$ cases and $4,514$ controls using the same instrumentation and protocol described for the follow-up stage. In the replication stage 15 PTVs were observed in cases and 1 PTV was observed in controls for a significant combined evidence of association ($P = 1.12 \times 10^{-5}$, 25 copies in $7,781$ cases versus 1 copy in $5,861$ population controls).

After careful examination of the Sanger sequencing traces, the analysts hypothesized that the detected PTVs in *PPM1D* were mosaic in blood. Thus, indexed libraries using the Nextera technology from Illumina were generated of PCR products generated from lymphocyte DNA amplified targeting the final exon of *PPM1D* in 20 individuals by the high-throughput Genomics team at the Oxford Genomics Centre in the Wellcome Trust Centre for Human Genetics. An Illumina MiSEQ desktop sequencer was used to sequence the libraries generating median coverage greater than 500X across the final exon of *PPM1D* (average median coverage was $3,384$X)(Ruark et al., 2012).

5.1.2 Overview of my role

The analysis of the breast cancer pooled sequencing data set was done collaboratively, involving multiple researchers across both the Institute of Cancer Research and the Wellcome Trust Centre for Human Genetics. As a member of the team I was involved in many parts of the analysis, from making sense of the raw high throughput sequencing data generated to comparing against reference data sets. Here, I focus on the aspects of the study where I had direct involvement that highlight my contribution, which may not be described in detail in the final publication (Ruark et al., 2012).

Variant calling. Processing large amounts of sequencing data requires the right software and appropriate computational speed that will allow the consortium to conduct the study. In this study I optimized the variant calling software, Syzygy, which I wrote in 2010 prior to beginning my thesis work. I optimized Syzygy to deal with targets larger than 1Mb, and modified the software to perform indel calling in the

pooled sample sequencing study design. Given the focus on PTVs this was critical for properly conducting the study.

Comparison against reference data sets. In this study I, with advice and supervision from my supervisor (Peter Donnelly) and Peter Humburg, compared the composite PTV allelic counts in the prioritized genes to reference data sets due to the lack of controls in the first phase of the experiment.

Subsequent analysis. After the study was published, I reanalyzed the data with additional reference data sets to examine whether any additional genes had evidence of association. In the next subsection I describe the analysis as examples of application of the methods that are described in Chapter 3.

5.1.3 Results

Variant calling. Identifying variants from DNA sequencing data can be challenging, and errors can severely impact the investigators' ability to properly execute an experiment (DePristo et al., 2011). A sequencing experiment where multiple individual DNAs are pooled together, commonly referred to as a pooled sequencing experiment, can present many analytical and computational challenges. A primary analytical challenge is separating signal from error. The average error rate for any genomic position in a sequencing experiment can be as high as 5% (Rivas et al., 2011). Thus, careful attention is required to account for proper quality score calibration, filtering bases with low base and mapping quality, and modeling systematic error processes (DePristo et al., 2011; Rivas et al., 2011). This is especially true for pooled sequencing data where, in a pool of 50 individuals, for example, a polymorphic DNA sequence variant can be observed in about 1 in 100 reads.

The variant calling software used in the breast cancer sequencing study, Syzygy, was originally designed for the analysis of small targets ($<$ 60 kilobases [kb]) and for the detection of SNVs. Indel detection was not supported and larger targets would require multiple weeks to complete. Given the aims of the breast cancer pooled sequencing study: 1) to detect variation in a pooled sequencing experiment of 48 pools of DNA consisting of 24 individuals per pool in a total genomic target of 1.7 Mb; and 2) to prioritize genes for PTV signal, which would required detection of indels; I set out to accomplish two variant calling aims in this project: 1) to improve computational performance; and 2) to support detection of indels in pooled sequencing data sets, which may account for half of all PTVs (MacArthur et al., 2012).

To address the first variant calling aim, i.e. to improve computational performance, I implemented, with help from Andrew Rimmer (a researcher and senior software developer at the Wellcome Trust Centre for Human Genetics), some of the modules in Syzygy with Cython, an optimising static compiler for the Python programming language (Behnel et al., 2011). Then, I applied computational shortcuts including reducing the number of iterations required in the Expectation Maximization (EM) algorithm required for detecting strand bias, which is indicative of systematic sequencing errors. Finally, and most importantly, I implemented changes that would enable parallel and distributed computing. To enable parallel computing I incorporated the multiprocessing package[1] to support spawning processes across multiple processors. To enable distributed computing I implemented a "scatter/gather" option that would allow the genomic targets to be processed independently in chunks across multiple nodes in a cluster and combined only in modules where the combined information would be required.

When Peter Humburg conducted the variant calling, reads and bases were filtered using the default parameters in Syzygy, i.e. reads with mapping quality 0 were removed, and bases with a call quality less than 22 were filtered.

After the optimizations were implemented, and variant calling was conducted, we identified $15,658$ variants including:

- $9,450$ coding SNVs and 211 coding indels,
- 176 nonsense SNVs,
- 159 frameshift indels, and
- 123 essential splice site disrupting SNVs.

The transition to transversion ratio (Ti/Tv) for coding SNVs was 2.52 suggesting a good quality callset.

In this study there was experimental data to assess performance. Syzygy successfully identified the positive control SNVs and indels with high sensitivity: 92.3% for SNVs and 94.4% for rare indels, which consisted of: 24 out of 26 SNVs, 14 out of 14 insertions, 30 out of 32 deletions, and 7 out of 8 complex indels. 132 PTVs were selected without considering Syzygy filter for validation to assess specificity of variant calling. 103 of 132 PTVs (78%) were identified as true positives. Syzygy includes score assignments to exclude variants that have poor quality: 0, indicating

[1] `https://pypi.python.org/pypi/multiprocessing/`

high quality, which is defined using the strand logarithm of odds score (SLOD ≤ 0, defined by the difference in support for a variant by one strand only compared to support for a variant by both strands (Rivas et al., 2011)) and the information about the number variants proximal to the candidate variant (a high quality variant has zero variants within 5 base pairs); 1, indicating very poor quality (SLOD ≥ 5); and 2 or 3, indicating moderate quality (2 reflects that the variant is within 5 base pairs of another variant, and 3 reflects that the variant has an SLOD less than 5, but greater than 0).

I provide a breakdown of the validation data highlighting that these filters were useful for prioritizing real PTVs:

- 86 of the 91 PTVs (95%) with high quality flag successfully validated,
- 13 of the 16 PTVs (81%) with moderate quality flag validated, and
- 4 of the 25 PTVs (16%) with poor quality flag validated.

Prioritizing genes for follow-up and replication. The goal of this study was to identify breast cancer predisposition genes. The primary approach to mapping genes was to focus on PTVs, which was largely motivated by prior successes in identifying breast cancer predisposition genes (Rahman et al., 2006). Due to the lack of controls in Phase I of the study, the consortium had to decide on a strategy to prioritize gene(s) for follow-up and replication sequencing where formal case-control tests of association could be performed.

After variant annotation was applied by Peter Humburg and genes were ranked by the number of PTVs in the data set I compared the estimated PTV allelic counts to reference data sets. We decided to focus on genes with at least four PTVs, and because of the lack of controls in our initial screen I compared the composite PTV allelic counts in the prioritized genes to reference data sets (combined PTV allelic counts from an exome sequencing data set of 6500 individuals generated as part of the Exome Sequencing Project (ESP) and obtained from `http://evs.gs.washington.edu/EVS/`, and another exome sequencing data set of 1000 individuals from the 1000 Genomes Project; Table 5.1). I observed 0 PTV copies in *PPM1D*, the top-ranked gene by collaborators, in the reference data sets (Table 5.1).

Mosaic nature of *PPM1D* PTVs. After careful visual inspection of the Sanger sequencing traces it was postulated that the *PPM1D* PTVs were mosaic (or somatic). This was suspected because the alternate allele traces were at lower height compared to the reference allele traces. An Illumina MiSEQ desktop sequencer was used to sequence the final exon of 20 individual *PPM1D* PTV carriers generating an average median coverage of 3,384X per individual across the final exon (Ruark et al., 2012). In Syzygy I implemented a module, `ReadBAM`, that would provide coverage summary for the reference and all possible alternate reads. I, with Peter Humburg, confirmed and estimated the extent of mosaicism after analyzing deep targeted *PPM1D* sequencing data generated from lymphoblastoid DNA of individuals identified to carry a *PPM1D* PTV in the final exon (Figure 5.1).

Subsequent analysis. To obtain updated estimates of the carrier frequency of the final exon *PPM1D* PTVs in the general population I examined the Exome Aggregation Consortium data set[2] (a collection of over 62,000 exomes with summary statistics made publicly available). In the ExAC data set 19 PTVs are catalogued. The total PTV allelic count is 25 resulting in a *PPM1D* PTV carrier frequency of 4 in 10,000 individuals (Figure 5.2). The majority of the PTVs in the ExAC consortium data set were also identified to be in the final exon (only 5 PTVs did not occur in the final exon). I obtained permission from collaborators (A. Palotie) to investigate in a Finnish subset of these individuals whether the observed *PPM1D* PTVs were mosaic. For all possible PTVs detected, even those that failed filters, coverage statistics of the reference allele and the alternate allele were obtained. To compute the mosaic (somatic) frequency I computed the ratio of alternate reads to total number of reads across 13 *PPM1D* PTV carriers. Then, I computed a 95% confidence interval using the normal approximation interval[3]. The alternate allele ratio, i.e. number of reads supporting the alternate read vs the total number of reads observed at the variant site, for 12 of the 13 PTV carriers was less than .5 (P= 0.0017, binomial test), which suggests the presence of mosaic PTVs.

When I performed the comparison of combined PTV allelic counts in the breast cancer pooled sequencing data set I had not developed any of the statistical methods described in Chapter 3. In addition, summary statistics for only 6,500 exomes was available at the time (2012). Recently, the ExAC consortium released version 0.3 of the ExAC data set (January 2015). This data set includes summary statistic

[2] `http://exac.broadinstitute.org`

[3] `http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval`

for variation identified from over 60,000 exomes. This presented an opportunity to reanalyze the breast cancer pooled sequencing data set and compared it with the ExAC data set using the Bayesian SEMCC method. I applied the Bayesian SEMCC approach described in Section 3.1, which handles summary count data.

First, I estimated the composite PTV allelic counts from the pooled resequencing data set across all the sequenced genes. Then, I used a composite PTV allelic count table for the ExAC data set that was prepared by Konrad Karczewski (a postdoctoral fellow at the Broad Institute). I excluded from the analysis 9 genes that contained positive controls in the experiment: *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *CHEK2*, *PALB2*, *RAD50*, *RAD51D*, and *TP53*. I applied the Bayesian SEMCC approach implemented in `MAMBA` with default priors. I considered the following scales of association: Bayes Factor (BF) > 1 as weak; BF > 10 as suggestive; and BF > 100 as substantial (Table 5.2). I considered 36 genes with multiple PTVs (> 1) in the pooled resequencing data set for further scrutiny and analysis.

RAD52, *POLD3*, and *TDG* had a Bayes Factor greater than 1000. These signals were driven by a variety of error modes including: 1) error in variant frequency estimate in the pooled sequencing experiment; 2) additional annotation filters applied in the ExAC data set; and 3) a false positive indel call. For *TDG*, 58 copies of a PTV predicted to disrupt one of the acceptor splice sites (c.793−1G>T) was erroneously estimated to be present across all the pools. For *RAD52*, two low frequency nonsense variants (p.Y415X and p.S346X) were not annotated as high-confidence LoF variants in ExAC. Hence, there was a clear overestimation of the composite allele frequency in the pooled sequencing data set compared to the reference data set. For *POLD3*, a low frequency frameshift indel (MAF= 2%, p.Arg300GlyfsX5) was identified in the pooled sequencing experiment. This variant has a 19 base pair homopolymer run and it is not present in ExAC, suggesting it is likely a false positive indel call.

PPM1D and *SPO11* had a Bayes Factor greater than 100. As expected, I found evidence of association (BF = 106.5) between PTVs in *PPM1D* and breast cancer by comparing the combined PTV allelic counts from the phase I pooled sequencing data set (5 PTV alleles, carrier frequency = 0.4%) and the combined PTV allelic counts from the ExAC reference data set (25 PTV alleles, carrier frequency = 0.05%). I examined the association in *SPO11* (BF = 124) driven by two protein truncating variants: 1) c.844+1G>C (1 PTV allele) and 2) p.Q56X (4 PTV alleles). Interestingly, p.Q56X is present in the ExAC reference data set and observed in six individuals in non-Finnish Europeans. The exon where the nonsense variant has over 30X coverage in greater than 90% of the individual in ExAC suggesting that high

quality information is provided for this particular site. *SPO11* is a protein-coding gene involved in meiotic recombination and in the creation of double stranded DNA breaks (Keeney et al., 1997). In mice, disruption of *Spo11* leads to severe gonadal abnormalities in both males and females due to defective meiosis. Heterozygotes are fully fertile, but homozygous males and females are not as indicated in the Jackson Laboratory database for stock number 019117 (Baudat et al., 2000).

Gene	Number of PTVs in Breast Cancer Sequencing	Total PTV Allelic Counts (frequency)	Initial PTV Allelic Count screen (frequency)	ExAC PTV Allelic Counts (frequency)	Bayes Factor (SEM) Breast Cancer v ExAC
BRCA2	16	63 (5.5%)	167 (2.2%)	48 (0.08%)	NA
BRCA1	14	16 (1.4%)	8 (0.1%)	106 (0.17%)	NA
PALB2	9	22 (1.9%)	2 (0.03%)	17 (0.03%)	NA
BRIP1	6	7 (0.6%)	11 (0.15%)	106 (0.17%)	NA
CHEK2	6	17 (1.5%)	2 (0.03%)	4444 (7.1%)	NA
CCDC63	5	9 (0.8%)	57 (0.76%)	394 (0.63%)	<1
CNTLN	5	12 (1.0%)	5740 (likely false positives)	320 (0.51%)	<1
PPM1D	5	5 (0.4%)	0 (0.0%)	25 (0.04%)	106.5
PRKDC	5	51 (4.4%)	1820 (24%)	NA	NA
MLKL	4	7 (0.6%)	19 (0.25%)	71 (0.11%)	32.7
CEP164	4	79 (6.9%)	49 (0.65%)	NA	NA
FANCA	4	5 (0.4%)	591 (7.9%)	441 (0.71%)	<1
FANCM	4	13 (1.1%)	71 (0.9%)	372 (0.6%)	<1
IGHMBP2	4	4 (0.35%)	0 (0.0%)	72 (0.12%)	<1
PARP4	4	4 (0.35%)	0 (0.0%)	209 (0.33%)	<1
TOP3A	4	4 (0.35%)	6 (0.08%)	106 (0.17%)	<1
TP53	4	6 (0.5%)	0 (0.0%)	1 (0.002%)	NA

Table 5.1: DNA repair genes with ≥ 4 PTVs in the breast cancer pooled sequencing experiment. We ranked the list of genes (column 1) by the number of PTVs (column 2) observed in the case only pooled sequencing experiment ($1,150$ breast cancer cases [69 ovarian cancer]). I used a publicly available data set with about $7,500$ individuals from the Exome Sequencing Project and the 1000 Genomes Project. We prioritized *PPM1D* for follow-up and replication sequencing. After publication (Ruark et al., 2012), I used the published pooled sequencing data set and compared the composite PTV allelic counts in these genes to pre-computed composite PTV allelic counts from the ExAC consortium (column 5, $\sim 62,500$ samples, personal communications with Konrad Karczewski). I computed an updated Bayes Factor using SEMCC (column 6), described in Section 3.1. For the nine genes with positive control samples I did not calculate a Bayes Factor as they were not random samples. In column 3, labelled "Total PTV Allelic Counts (frequency)", the estimated combined PTV allelic counts in the breast cancer pooled sequencing data set is shown. In parenthesis, the estimated PTV carrier frequency, i.e. the proportion of individuals in the pooled sequencing data set carrying a PTV in the corresponding gene is shown. In column 4, labelled "Initial PTV Allelic Count screen (frequency)", the estimated combined PTV allelic counts in the $7,500$ reference exome sequencing data set is shown.

Gene	Number of PTVs in breast cancer pooled sequencing data set	Total PTV allelic counts (frequency)	Number of PTVs in ExAC	ExAC PTV allelic counts (frequency)	Bayes Factor (SEMCC) breast cancer data set compared to ExAC
SPO11	2	5 (0.4%)	14	24 (0.038%)	124.0
PPM1D	5	5 (0.4%)	16	25 (0.04%)	106.5
NR1H2	2	3 (0.3%)	5	5 (0.008%)	61.6
MLKL	4	7 (0.6%)	19	71 (0.11%)	32.7
BTBD16	3	5 (0.4%)	25	42 (0.07%)	14.9
WRN	3	6 (0.5%)	41	66 (0.1%)	10.6
BCCIP	3	4 (0.35%)	12	47 (0.08%)	2.3
MSH4	3	4 (0.35%)	27	51 (0.08%)	1.8
ERCC3	3	5 (0.4%)	25	81 (0.13%)	1.3
POLK	3	6 (0.5%)	18	110 (0.18%)	1.2

Table 5.2: Comparison of composite PTV allelic counts in the breast cancer pooled sequencing data with the ExAC reference data set (genes with Bayes Factor > 1 and multiple (> 1) PTVs shown). After publication (Ruark et al., 2012), I used the published pooled sequencing data set and compare the composite PTV allelic counts for any gene with multiple PTVs to pre-computed composite PTV allelic counts from the ExAC consortium (column 5, $\sim 62,500$ samples, personal communications with Konrad Karczewski). I computed an updated Bayes Factor using SEMCC (column 6), described in Section 3.1. For the nine genes with positive control samples I did not calculate a Bayes Factor as they were not random samples. I excluded three genes with Bayes Factor > 1000 as it was clear that these results were driven by various error modes described in the text of this section. In column 3, labelled "Total PTV Allelic Counts (frequency)", the estimated combined PTV allelic counts in the breast cancer pooled sequencing data set is shown. In parenthesis, the estimated PTV carrier frequency, i.e. the proportion of individuals in the pooled sequencing data set carrying a PTV in the corresponding gene is shown. In column 2 the number of PTVs identified in the breast cancer pooled sequencing data set is shown. In column 4 the number of PTVs identified in the ExAC reference data set is shown.

5.1.4 Conclusion

In this study, primarily conducted as a collaborative effort by an analysis group in the consortium, we sought to identify genes in the DNA repair pathway that contribute to breast cancer predisposition by focusing on PTVs discovered in a breast cancer pooled sequencing experiment. We prioritized PTV enrichment in breast cancer cases by focusing on the total number of PTVs in a case only sequencing study design. We were able to prioritize a gene, *PPM1D*, based only on the total number of PTVs observed in the data set and by filtering out genes that contained a substantial number of PTVs

in reference data sets or genes that were already well established. After follow-up and replication sequencing it was clear that PTVs in *PPM1D* were associated to breast and ovarian cancer predisposition. Curiously, the PTVs were located in the final exon of the gene and were identified to be mosaic in blood with a somatic frequency ranging from $10-25\%$ in breast cancer patients. The enrichment of PTVs in the final exon suggested that risk was conferred by PTV containing transcripts that were not subject to nonsense-mediated decay (Maquat, 2004). Thus, this raised the possibility that the functional consequence of the resulting gene product was that of a "gain of function" rather than "loss of function", which is usually assumed for gene products resulting from the presence of PTVs (MacArthur et al., 2012). Follow-up functional work by our colleagues showed that the *PPM1D* PTVs result in enhanced suppression of p53 in response to ionizing radiation exposure suggesting that PTVs in the final exon of *PPM1D* have a "gain of function" effect (Ruark et al., 2012).

Following the publication of this study, I investigated the carrier frequency of final exon *PPM1D* PTVs in the general population by integrating the ExAC consortium data set. This resulted in a carrier frequency estimate of about 4 in 10,000 individuals with strong support from detailed inspection of the alternate read ratio that these were indeed mosaic variants. Interestingly, the PTV carrier frequency in the ExAC consortium data set changes from 0.04% in the general population to 0.1% depending on whether PTVs that were filtered by the variant caller are included. Given the enrichment of PTVs in the final exon in the filtered set and the observation that these PTVs are indeed mosaic it is likely that these events are true positive variants. The finding that PTVs in the final exon of *PPM1D* contribute to breast and ovarian cancer predisposition has been replicated (Akbari et al., 2013). In addition, I assessed association between PTVs in genes sequenced in the pooled sequencing experiment to breast cancer predisposition by using composite PTV allele counts from the ExAC consortium data set. I excluded 9 genes that included individuals with positive control variants and focused on genes with more than one PTV. This exercise resulted in a list of three genes with very strong evidence of association (BF > 1000). Closer examination of both the pooled sequencing data set and the ExAC reference data set showed that these signals were driven by various error modes highlighting that critical inspection is necessary when integrating such diverse data sets. Nonetheless, two genes had suggestive evidence of association (BF > 100) including *PPM1D* (as expected from the published result) and *SPO11*. A lookup of PTVs in *SPO11* in an independent germline cancer study with a mixture of cancer types shows a PTV carrier frequency of 6 in $5,929$ cancer patients (0.1%) and 10 in $12,875$ controls (0.078%;

P value = 0.60; OR = 1.3 95% confidence interval 0.39 to 3.96). Ongoing large-scale germline cancer sequencing projects will hopefully be able to address unequivocally the relevance of *SPO11* to breast cancer.

Surprisingly, in the published study the mosaic PTVs observed in blood were absent in sequencing data generated from the tumor samples (Ruark et al., 2012). In a recent study (Zhang et al., 2014), mosaic final exon *PPM1D* PTVs were identified in brainstem gliomas in about 37.5% of the tumor samples and were infrequent in non-brainstem gliomas. This raises a couple of interesting observations: i) final exon *PPM1D* PTVs are observed in cancers other than breast and ovarian cancer, and ii) they may be absent (breast cancer) or present (brainstem gliomas) from the tumor DNA. Further research will be required to investigate the carrier frequency of mosaic *PPM1D* PTVs in blood across multiple cancer types and contrast it to the degree of mosaicism observed in the tumor samples.

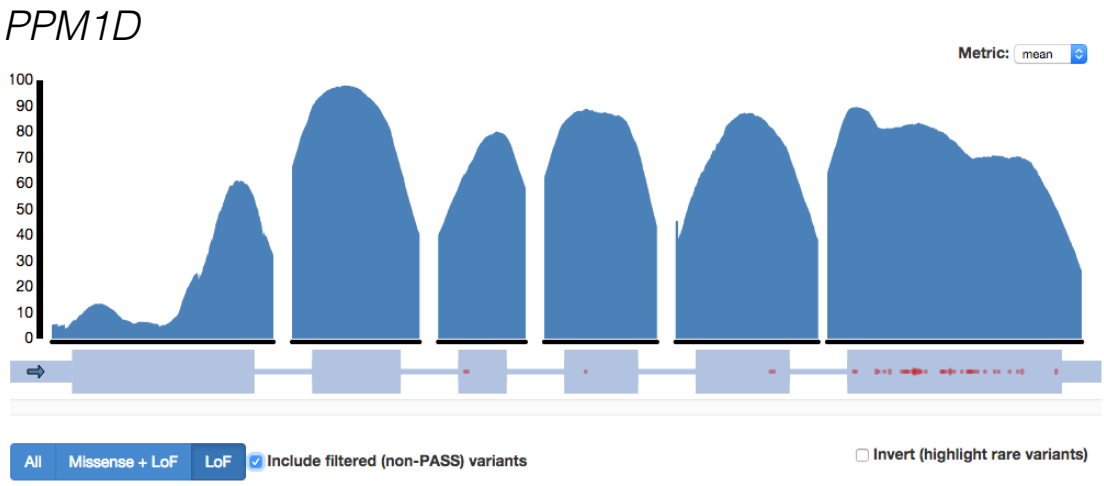


Figure 5.2: Plot of *PPM1D* PTVs in the ExAC data set. The majority of *PPM1D* PTVs discovered in the ExAC data set are in the final exon. The plot generated in the ExAC browser[4] shows the canonical transcript (chosen by ENSEMBL) of *PPM1D*, a summary of coverage on the y-axis, and highlighted dots are *PTVs* in the data set.

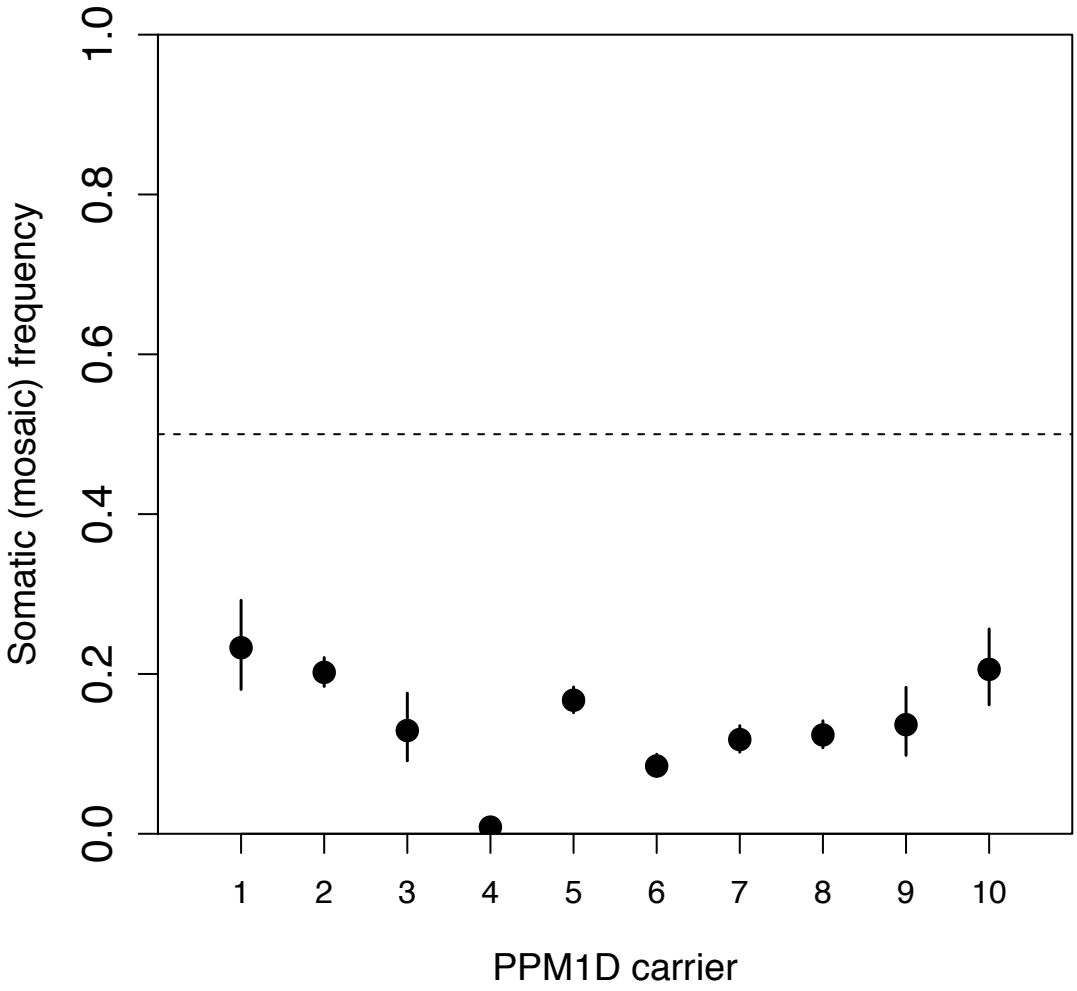


Figure 5.1: Somatic (mosaic) frequency estimated from MiSEQ sequencing data of *PPM1D* PTVs identified through full gene Sanger sequencing. We performed deep MiSEQ amplicon sequencing of the final exon in the 10 individuals identified to carry a mosaic *PPM1D* PTV in Phase 2 of the case-control *PPM1D* full gene Sanger sequencing experiment. The estimate of the mosaic frequency in blood (y-axis) and the 95% confidence interval are shown.

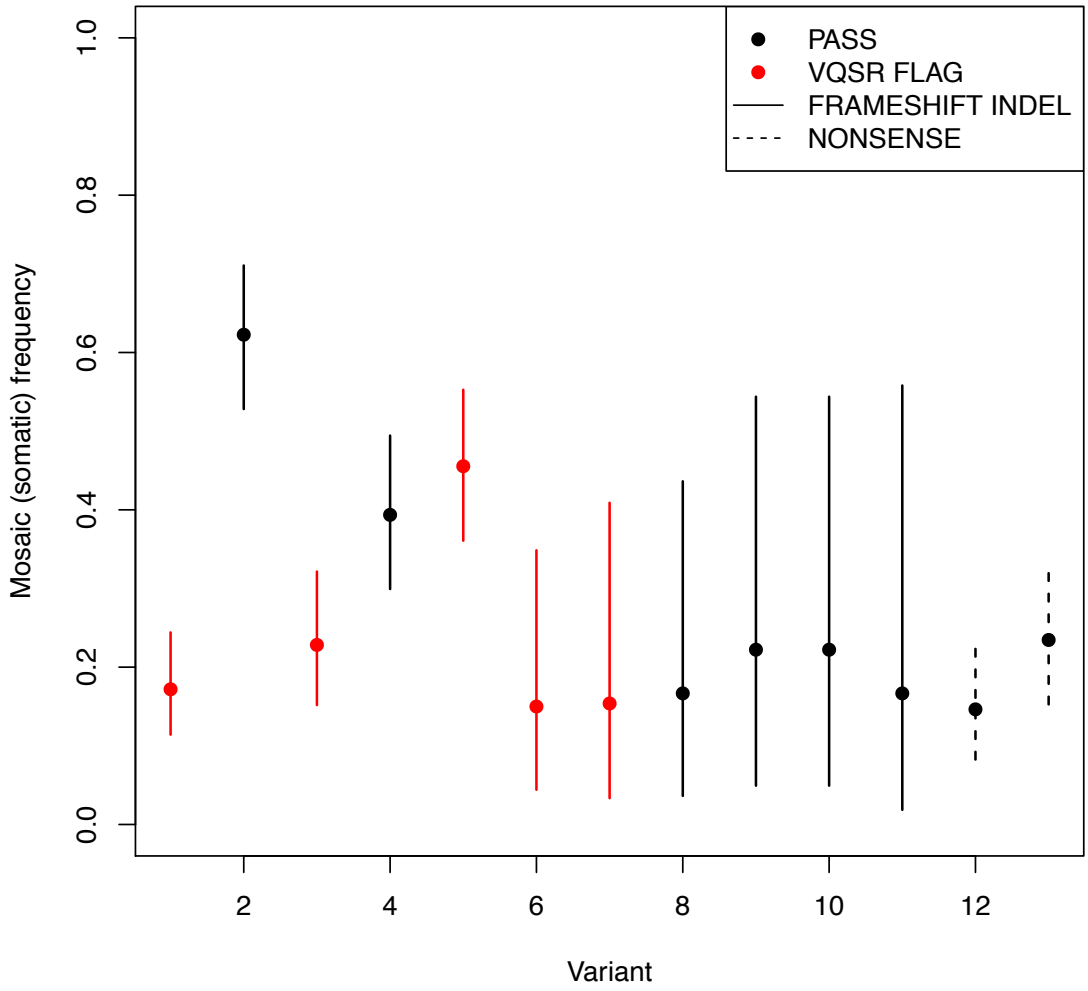


Figure 5.3: Identification of mosaic PPM1D PTVs in population unselected controls. In an exome sequencing data set of $\sim 12,000$ individuals I observe 13 PTVs (8 PASS filters) giving a population carrier frequency of about 1 in a 1000 with mosaic somatic frequency in whole blood < 0.5 (with a single exception). The estimate of the mosaic frequency in blood (y-axis) and the 95% confidence interval are shown.

5.2 Cross-disorder analysis of six autoimmune diseases: Protein truncating variants in *TNFAIP3* show suggestive evidence of association to type 1 diabetes

5.2.1 Background

In Section 3.3 I presented the C-alpha MRP test for rare variant association testing of multiple phenotypes, which exploits correlation of genetic effects across variants and multiple phenotypes. I applied it to the publicly available data set from Hunt et al. (2013) to evaluate performance and to examine whether I could identify association signals that were missed. In Hunt et al. (2013) they generated a targeted sequencing data set of 25 genes (identified with common variant association to autoimmune diseases) in 41,911 UK residents of white European origin, comprising 24,892 subjects with one of six autoimmune disease phenotypes:

1. autoimmune thyroid disease (AITD),
2. coeliac disease (CeD),
3. Crohn's disease (CD),
4. psoriasis (PD),
5. multiple sclerosis (MS),
6. type 1 diabetes (T1D),

and 17,019 controls.

According to Hunt et al. (2013) all 25 genes selected for targeted sequencing were risk loci for at least two of the six autoimmune diseases studied, had a known immune system function, and were from loci with only a single strong candidate immune gene. 34 libraries (each of $1,536$ barcoded samples) were generated and sequenced one per lane on Illumina HiSeq sequencers achieving coverage greater than 20X for the majority ($> 99\%$) of the targeted amplicons (Hunt et al., 2013).

In the original study, the investigators annotated the variants using ANNOVAR and the GENCODE v14 reference transcript set. Rare functional variants were identified based on nonsense, frameshift indel, missense SNV, or splice predictions. Protein truncating variants were identified based on nonsense, frameshift indel, or essential

splice site disrupting predictions. The investigators applied 5 gene based association tests including: i) the binomial C-alpha test, ii) the burden test, iii) conditional gene-based burden test, iv) count of case-unique rare alelles (UNIQ) test, and v) count of control-unique rare alleles (UNIQ) test, using the PLINK/SEQ toolkit. They applied it across the 25 genes and across 7 phenotype definitions, separately, where the seventh phenotype definition consisted of all autoimmune diseases combined. The distribution of association statistics for all five pooled gene tests across the seven phenotypes tested were consistent with the global null of no association (Hunt et al., 2013) with no genes showing a P value < 0.001. The authors did not focus explicitly on protein truncating variants, they grouped missense and protein truncating variants, which could have led to null results. In addition, the authors did not exploit the entire data set. For example, for a disease specific effects a simple strategy would be to use all the disease samples and the control samples as a combined control set. However, this does add an additional multiple testing burden as all six phenotypes would need to be treated in the same manner.

I contacted the principal investigator of the study, David van Heel, who kindly provided the genotype and phenotype data set. My approach, unlike theirs, was to take advantage of the group definitions to improve power to detect rare variant association, in particular with a focus on protein truncating variants.

For my analysis I performed annotation of all the DNA sequence variants discovered using PLINK/SEQ and the Gencode reference transcript set in the locus database (locdb) available for download from the PLINK/SEQ website[5]. Protein truncating variants were identified based on worst annotation equal to "nonsense", "frameshift", or "esplice".

5.2.2 Results

I restricted the analysis to protein truncating variants. I applied the C-alpha MRP test for cross-disorder rare variant association analysis. I chose a sub-matrix $\mathbf{R}_{\text{var}}$ such that for any given gene the correlation of genetic effects between a pair of protein truncating variants and a single trait is equal to 1, the assumption being that PTVs in a gene have similar effects on a phenotype (for a more nuanced treatment of PTV annotation and *in silico* predictions of the functional consequences see the description of the SEMGEM models in Section 3.2). In the absence of information about correlation of genetic effects between diseases I assumed that they were independent,

[5] `http://atgu.mgh.harvard.edu/plinkseq/`

i.e. I used the identity matrix for sub-matrix $\mathbf{R}_{\text{phen}}$ as recommended in Section 3.3. The matrix $\mathbf{U}$ is the Kronecker product of these two sub-matrices. I applied the test to 21 genes with at least one PTV (Table 5.3).

For the 21 protein-coding genes I found nominal evidence of association, $P < 0.05$, for three genes: *PTPN2*, *IL12A*, and *TNFAIP3*. In Hunt et al. (2013) across the 7 phenotypes and 5 gene based tests applied to rare variants in *PTPN2*, *IL12A*, and *TNFAIP3* exactly 0, 1, and 1 test and phenotype combinations achieved a P value < 0.05, respectively.

I found suggestive evidence of association between protein truncating variants (C-alpha MRP P value$= 8.2 \times 10^{-5}$, minimum univariate burden test P value $= 3.6 \times 10^{-3}$) in *TNFAIP3* and autoimmune diseases, mainly driven by PTV singleton observations ($n = 5$) in individuals with type 1 diabetes, ($n = 1$) psoriasis, or ($n = 1$) multiple sclerosis (Figure 5.4, Figures 5.6a - 5.9 for IGV snapshots of variants provided by David van Heel). The PTV annotation for all 7 PTVs was supported by transcript ENST00000237289, the dominant isoform across all studied tissues in the GTEx project (Figure 5.5).

There are two main reasons that I may have detected this signal and it was not found in the published analysis: i) I restricted the analysis to PTVs, and ii) in the setting of disease-specific effects across multiple groups the C-alpha MRP test may gain power to detect association compared to univariate approaches as shown in Section 3.3. Another benefit of using the C-alpha MRP test is that it is designed to work with shared controls, which is an issue for univariate tests because shared controls means that P values are not independent across diseases. One possible way that the signal in *TNFAIP3* may have been detected by Hunt et al. (2013) would have been by treating each disease group as a case group and treating the remaining case groups and control group as one super control group. However, this probably would have added an additional 6 phenotypes to test across 5 gene based tests, which would have increased the multiple testing burden.

I sought to estimate the combined allele frequency of protein truncating variants at *TNFAIP3* in other data sets. I focused on the ExAC consortium data set. There are pros and cons to using this data set. One concern that may arise from integrating the ExAC data set is that there is phenotypic, technical, and population heterogeneity in the samples included. Thus, extreme care must be taken whenever inferences are made after integrating this data set in an analysis. For example, some of the contributing projects include genotype data from type 2 diabetes, schizophrenia, myocardial infarction, or inflammatory bowel disease cases. This may weaken the ability

to claim evidence of association because it is possible that some effects may be shared with one of the aforementioned diseases. Furthermore, it is possible that some exons and/or genes are not well captured and absence of any PTV in the data set may just well be indicative of missing data. However, the ExAC data set is the largest reference exome data set publicly available. The sheer number of individuals in this data set, $63,000$, is a useful starting point for understanding the nature and frequency of PTVs in a gene.

In the exome variant server from the ExAC consortium with data from over $60,000$ individuals I observed three PTV copies for a combined frequency of approximately 10^{-5}. I asked Matti Pirinen (a postdoctoral fellow at University of Oxford at the time and currently a postdoctoral fellow at the Institute for Molecular Medicine in Finland) to apply code he had developed to perform model comparisons to the data set that I generated. We assumed four orders of magnitudes for the PTV carrier frequency: i) 0.001, ii) 10^{-4}, iii) 10^{-5}, and iv) 10^{-6} and assessed the posterior probability of each of the models conditional on the observed data for each of the four groups I was interested in: i) type 1 diabetes group, ii) other autoimmune disease group, iii) controls, and iv) controls + ExAC. For the T1D group the model assuming a PTV carrier frequency of 1 in 1000 had a posterior probability > 0.99 over all other computed models. In the controls and controls+ExAC group the models assuming a PTV carrier frequency less than 10^{-4} had combined posterior probability equal to 1 (Figure 5.4). The ExAC data set supports the observation that PTVs in *TNFAIP3* are extremely rare and that the carrier frequency in type 1 diabetes cases is unusual.

The Tumor necrosis factor, alpha-induced protein 3, commonly referred to as A20 protein, is a protein that in humans is encoded by the *TNFAIP3* gene. The protein encoded by this gene is a negative regulator of tumor necrosis factor (TNF)-induced signaling pathways leading to apoptosis, stress response and inflammation (Lademann et al., 2001). The observations from this study support the mouse model results that A20 inhibition will have physiological consequences related to beta-cell function and may provide, if replicated, the first line of genetic evidence in humans that PTVs in *TNFAIP3* confer strong - as it appears to be an order of magnitude more frequent in the type 1 diabetes group compared to all controls or reference data set - risk to type 1 diabetes (Liuwantara et al., 2006).

In addition to applying the C-alpha MRP testing framework to each gene I wanted to apply the test to the set of genes sequenced by treating the list of genes as a gene set. As described in Section 3.3 the C-alpha MRP testing framework allows a simple extension to the analysis of gene-sets, in particular, the analysis of PTVs across

multiple genes. In that setting I assume among a group of PTVs in a gene that the correlation of genetic effects is 1 across a trait and make an assumption that the expected correlation of genetic effects between a pair of variants in separate genes is 0, as the correlation of genetic effects between a pair of genes and a trait is unclear. Intuitively, this results in an approach where I collapse the PTVs in a gene and treat each gene entry in a test of dispersion. I restricted the analysis to genes with significant GWAS association, P value < 0.0001, according to Supplementary Table 1 in Hunt et al. (2013). I applied the C-alpha MRP test to the gene set for each of the six autoimmune disease groups separately, and found suggestive evidence of association between protein truncating variants across the gene set and type 1 diabetes (P value $= 0.0017$), and nominal evidence of association between protein truncating variants across the gene set and multiple sclerosis (P value $= 0.045$). An interesting exercise, after detecting association, is identifying which variants or genes are driving the aggregate association signal. I applied the EM algorithm described in Neale et al. (2011) to the type 1 diabetes gene set, and found that 4 of the 8 genes implicated to T1D through GWAS studies have posterior probability $> .75$ of being associated to disease, including evidence of PTVs at *UBASH3A* conferring protection where I observed 15 PTV alleles in the 17,019 controls (0.088%) and 1 PTV allele in the $6,494$ type 1 diabetes cases (0.015%, Figure 5.10). The thirteen PTVs are in separate haplotype blocks and unfortunately we do not have the GWAS risk variant rs876498 in the targeted sequencing data set to be able to study the haplotype structure with the thirteen variants. *UBASH3A* encodes the ubiquitin associated and SH3 domain containing A protein, which belongs to the T-cell ubiquitin ligan family involved in regulating T-cell signaling (Tsygankov, 2009). Homozygous null mice are viable and healthy with no abnormalities detected according to the Jackson Laboratory database for MGI:3051962.

If PTVs in *UBASH3A* cause protection, then one might expect the GWAS risk variant (rs876498) to increase *UBASH3A* expression (most likely in pancreas tissue or beta cells). I checked for association between rs876498 and expression of *UBASH3A* in pancreas ($n = 58$) using the GTEx portal (`http://gtexportal.org/home/eqtls/calc?tissueName=Pancreas&geneId=UBASH3A&snpId=rs876498`). There is weak (P= 0.04) evidence of association with direction of effect consistent with increased expression. In an updated unpublished data set the signal is abrogated (personal communications with Martijn van de Bunt). Interestingly, in the GEUVADIS RNA-sequencing data set (Lappalainen et al., 2013), rs876498 is a significant exon QTL for coding exon 19 in *UBASH3A* ($P = 1.62 \times 10^{-9}$; personal communications with Martijn van de

Bunt), also with direction of effect consistent with increased expression. The evidence for association between PTVs in *UBASH3A* and type 1 diabetes directly was $P = 0.087$ using the univariate SKAT test (Wu et al., 2011).

5.2.3 Conclusion

In this study I used the C-alpha MRP approach to test for association between PTVs in a highly curated list of 25 genes and six autoimmune disease phenotypes. I identified new signal of PTV association in *TNFAIP3* largely stemming from 5 PTV singleton observations in type 1 diabetes cases. I used a large 62k reference data set to confirm that the PTV composite allele frequency in the general population is low $(10^{-4} - 10^{-5})$ and conclude that PTVs in *TNFAIP3* confer moderate to strong risk to type 1 diabetes.

The findings from this study emphasize the value that new innovative approaches may have in extracting exciting new associations that were essentially missed in the original study. I expect that large-scale published rare variant case-control data sets will continue to be reanalyzed in order to identify novel associations. This will likely be true when high throughput functional assays can be integrated or when *in silico* predictions of the pathogenic nature of DNA sequence variants improve.

Gene	Controls (17019)	T1D (6494)	MS (3834)	PD (2096)	CD (3271)	CeD (6587)	AITD (2610)	ExAC (frequency)	P
TNFAIP3	0	5	1	1	0	0	0	2 (0.003%)	8.2×10^{-5}
IL12A	3	0	4	0	0	1	0	2 (0.003%)	0.011
PTPN2	2	5	0	0	0	2	2	3 (0.005%)	0.03
ZFP36L1	0	3	0	0	0	1	1	18 (0.02%)	0.069
IL21	0	0	0	0	0	1	1	1 (0.002%)	0.19
PTPRK	2	0	2	1	1	0	0	56 (0.09%)	0.22
IL10	0	0	0	0	1	1	0	7 (0.01%)	0.25
UBASH3A	15	1	1	2	2	6	0	37 (0.06%)	0.25
TNFRSF14	1	1	0	1	0	0	0	19 (0.03%)	0.28
THEMIS	7	1	1	0	2	0	0	38 (0.06%)	0.38
CTLA4	0	1	0	0	0	0	0	2 (0.003%)	0.42
BACH2	0	1	0	0	0	0	0	4 (0.007%)	0.45
STAT4	1	0	0	1	0	0	0	8 (0.01%)	0.45
ZMIZ1	1	0	1	0	0	0	0	5 (0.008%)	0.55
IL18RAP	4	1	1	1	0	3	0	23 (0.04%)	0.68
IL18R1	5	1	0	0	2	2	1	24 (0.04%)	0.73
TAGAP	5	1	1	0	0	0	1	15 (0.025%)	0.78
NCF2	17	5	5	3	5	9	2	30 (0.05%)	0.79
ICOS	1	0	0	0	0	1	0	6 (0.01%)	0.86
CD226	1	1	1	1	0	1	0	5 (0.008%)	0.89
RGS1	2	0	0	1	0	0	0	6 (0.01%)	0.97

Table 5.3: Results from cross-disorder analysis of PTVs and six autoimmune diseases. In this table we show: the combined allele count of PTVs per gene in each case group, the combined allele counts of PTVs in the ExAC browser, and the C-alpha MRP P value.

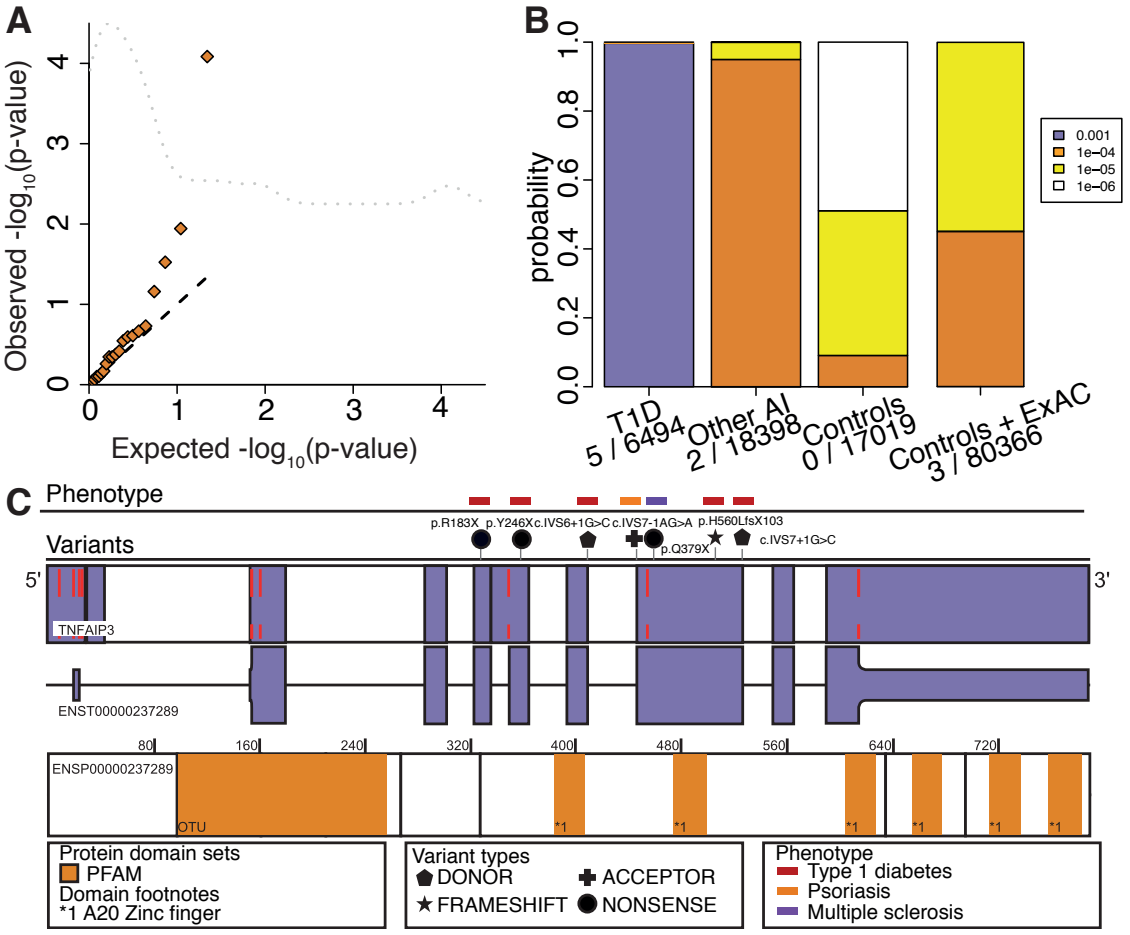


Figure 5.4: *TNFAIP3* PTV association. A. Q-Q plots for cross-disorder rare variant association analyses of protein truncating variants and autoimmune diseases. Application of the C-alpha MRP test identifies evidence of association of protein truncating variants at *TNFAIP3* and autoimmune disease phenotypes. B. Model comparison of the frequencies of protein truncating variants in different groups (T1D, Other autoimmune diseases, controls, and controls + ExAC consortium data). This quantifies the evidence that T1D has different frequencies from Other AI. C. Protein and transcript diagram showing the genomic location of protein truncating variants identified in *TNFAIP3*.

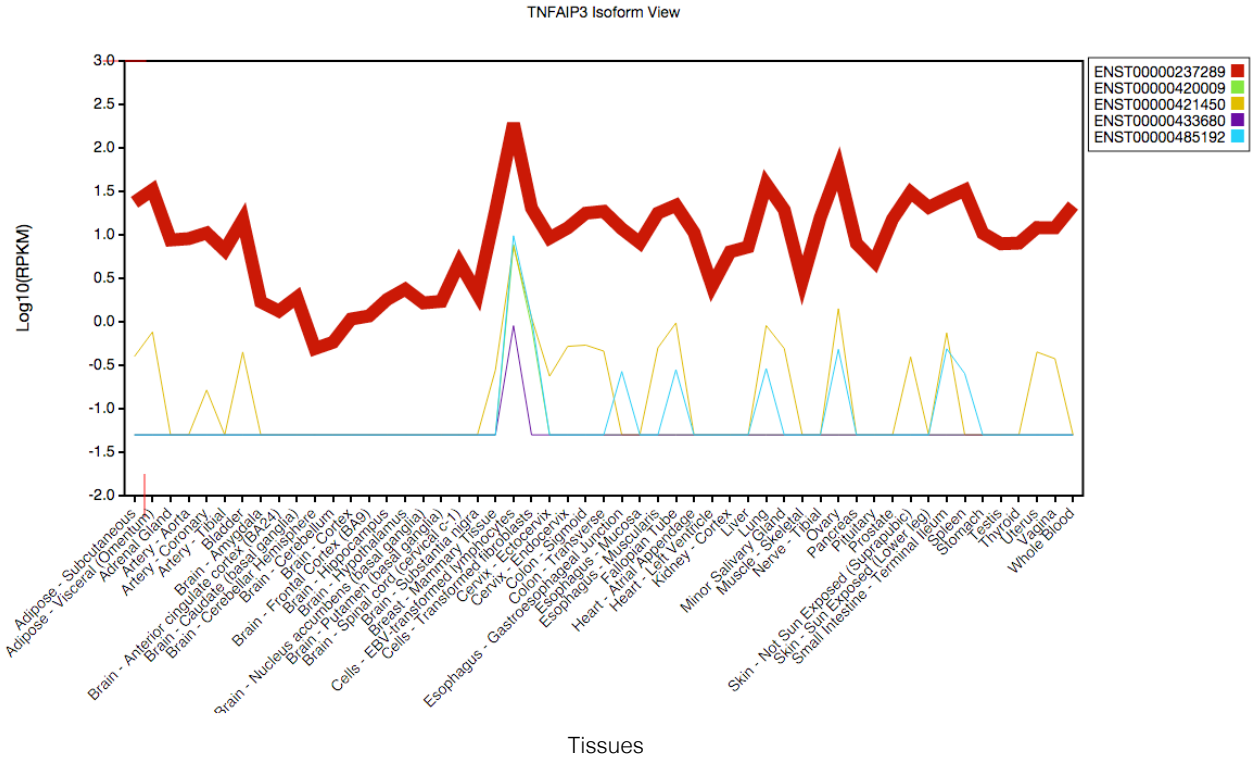
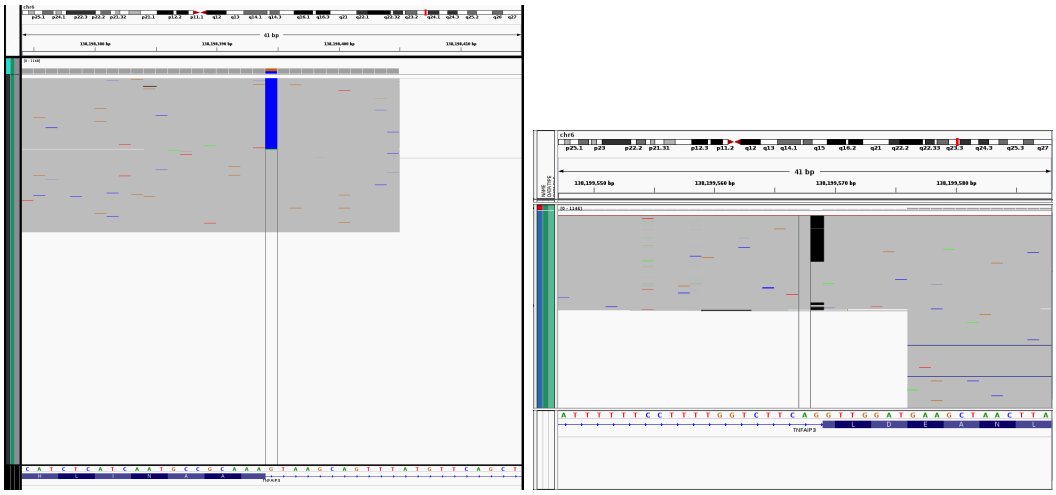


Figure 5.5: *TNFAIP3* transcript isoform expression across the GTEx tissues. Protein-coding transcript ENST00000237289 is dominantly expressed across all tissues studied in the GTEx project. Figure available from the GTEx portal in the 'Gene expression for TNFAIP3' tab with options Plot=Isoform, Scale=Log, Sort=ABC, Range=Absolute (`http://gtexportal.org/home/gene/TNFAIP3`).

(a) IGV snapshot of p.R183X in *TNFAIP3*. (b) IGV snapshot of p.Y246X in *TNFAIP3*.

Figure 5.6: IGV snapshots I.



(a) IGV snapshot of c.IVS6 + 1G>C in *TNFAIP3*. (b) IGV snapshot of c.IVS7 − 1AG>A in *TNFAIP3*.

Figure 5.7: IGV snapshots II.

(a) IGV snapshot of p.Q379X in *TNFAIP3*. (b) IGV snapshot of p.H560LfsX103 in *TNFAIP3*.

Figure 5.8: IGV snapshots III.

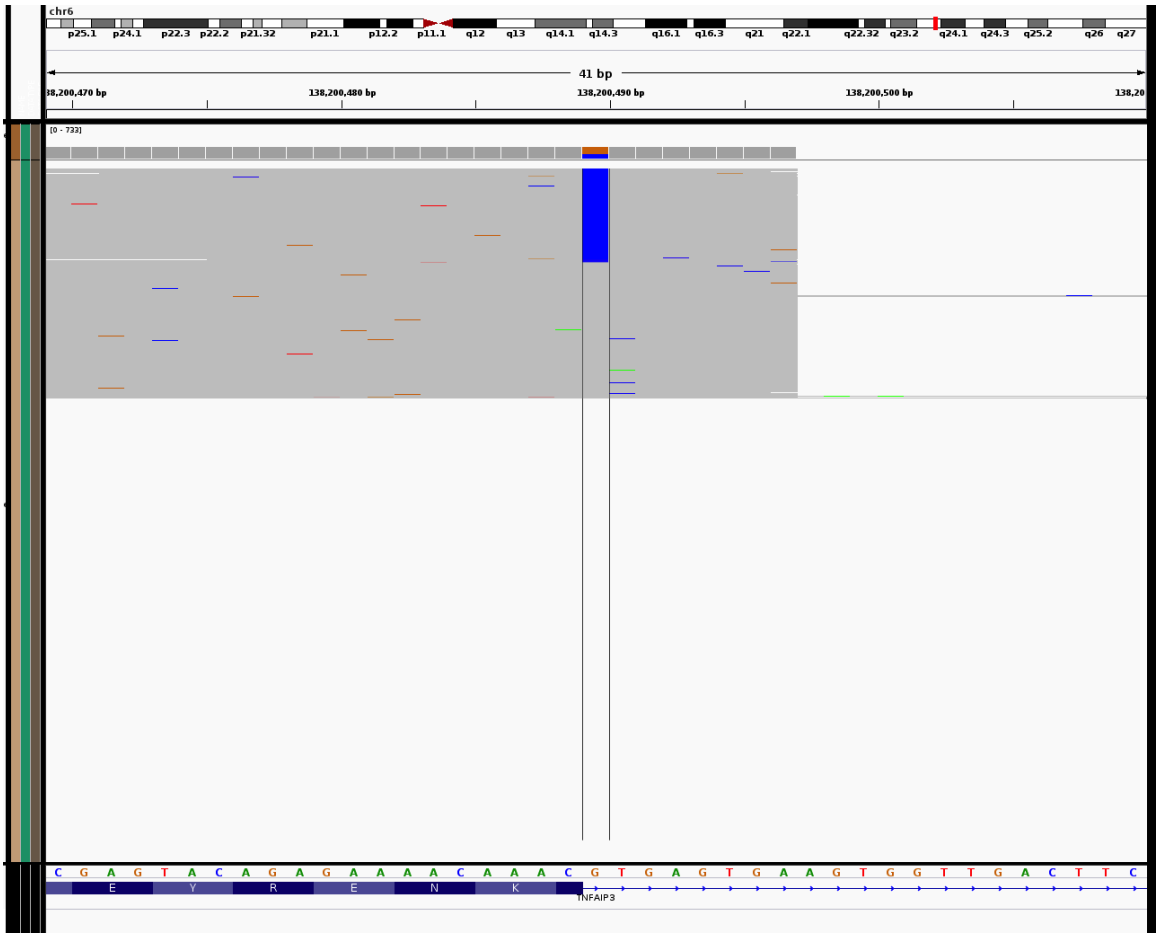


Figure 5.9: IGV snapshot of c.IVS7 + 1G>C in *TNFAIP3*.

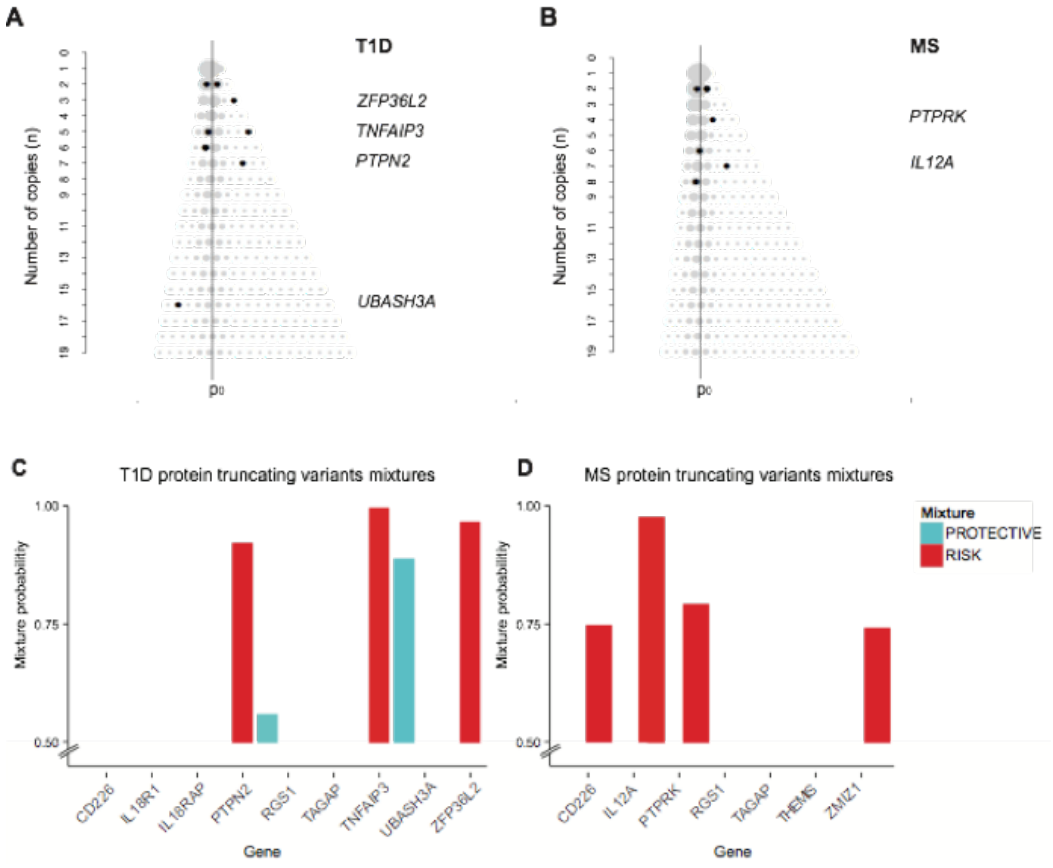


Figure 5.10: Extension of C-alpha MRP test to the analysis of gene-sets and application to autoimmune targeted sequencing data set. Analysis for each disease was restricted to genes with significant GWAS association signal, P value $< .0001$, according to Supplementary Table 1 in Hunt et al. (2013) for each disease. Nominally significant ($P < 0.05$) association was identified between protein truncating variants across the genes to A) type 1 diabetes (P value $= .0017$) and to B) multiple sclerosis (P value $= .045$). A-B) The dispersion plot shows an entry for gene used to evaluate the statistical evidence (black dot) and a representation of the probability of the observation under the null (gray dot). For example, in figure panel A *ZFP36L2* is highlighted, which has the sum of protein truncating alleles, y_1, and the total number of truncating alleles, n, equal to 3. C-D) Bar plots of the probability that the gene belongs in the protective (blue) or risk (red) mixture component according to the EM algorithm presented in Neale et al. (2011) are shown.

5.3 Multi-ethnic exome-wide association study of protein truncating variants to type 2 diabetes

5.3.1 Overview of the study

In this section I present results from my analysis of a type 2 diabetes exome sequencing study. The primary analysis of the type 2 diabetes exome sequencing study was conducted as part of a consortium effort to investigate the role of coding variation in type 2 diabetes susceptibility. As part of the analysis group of the consortium I contributed to the analysis of the data by applying software and methodology presented in Chapter 3, and by applying additional publicly available software.

To set the scene, I summarize the design of the study and present the materials and methods.

In the following subsections I will provide an overview of my role and present results focused on addressing the challenges and opportunities inherent in exome sequencing study design by: i) applying methods developed in Chapter 3, ii) applying methodology used in a published exome sequencing study of schizophrenia, and iii) examining the effect sizes required to detect association at significance levels that take into account the multiple testing burden observed in these studies.

Study design. To evaluate the contribution of coding variation to T2D risk the consortium generated exome sequencing data at the Broad Institute for 12,940 individuals (6,505 T2D cases and 6,436 controls) from five major ancestry groups:

- 4,541 individuals of European ancestry (2,359 T2D cases),
- 2,074 African Americans (1,018 cases),
- 2,217 individuals of South Asian ancestry (1,094 cases),
- 2,165 individuals of East Asian ancestry (1,012 cases), and
- 1,943 Hispanics (1,021 cases).

Sequence reads were processed and aligned by the production group at the Broad Institute with Picard[6]. Variants and genotypes were called by the production group with GATK (DePristo et al., 2011). All variant QC of the exome sequencing data set was conducted by collaborators at the Broad Institute. Extensive QC was applied

[6] `http://picard.sourceforge.net`

to the data resulting in a data set consisting of 1.78 million coding variants. A high quality call set was generated and made available in a Variant Call Format (VCF) file. The VCF file was made available through the Broad ftp network. For my analysis, and all analysis conducted in this study, genotypes were considered as "missing" if the genotype quality (GQ) was less than 20 representing a predicted error rate greater than 1%.

Variant annotation was performed by three analysts in the consortium (Kyle Gaulton [Oxford], Pablo Cingolani [McGill], Jason Flannick [Broad]). They decided to merge the variant annotations generated by three separate variant annotation programs: CHAoS v0.6.3, SnpEff v3.1, and VEP v2.7. Variants with "nonsense", "frameshift", and "essential splice site" predictions were considered PTVs. Overall, 67,789 PTVs were identified with an average of 235 PTVs per individual. For my analysis I used the PTV definition agreed to by the consortium.

5.3.2 Overview of my role

The analysis of the type 2 diabetes exome sequencing data set was done collaboratively, involving multiple researchers across multiple institutions including: 1) Broad Institute, 2) University of Michigan, and 3) University of Oxford. As a member of the team I was involved in many parts of the analysis. Here, I focus on the aspects of the study where I had direct involvement that highlight my contribution, particularly emphasizing the application of methods presented in Chapter 3 for the analysis of PTVs.

Individual gene level analysis of PTVs. In this study, I applied the Bayesian SEMCC models, described in Section 3.1, across all protein-coding genes with at least a single PTV in the final exome sequencing data release to assess association between PTVs and type 2 diabetes.

PTV analysis of gene sets. In Purcell et al. (2014) they found a polygenic burden of rare protein truncating variants in schizophrenia. In this study, I suggested to the consortium that we should take a similar approach. I applied the statistics/-matrix/permutation procedure described in Purcell et al. (2014) to assess association between protein truncating variants in 18 manually curated (manually curated by Pablo Cingolani and Robert Sladek [McGill]) gene sets of Mendelian diseases and pathways for type 2 diabetes and T2D status (Appendix Table 1).

Power study. Prior to conducting a genetic association study power calculations are usually performed to assess the range of effects that may be detected in the proposed study design. In principle, power studies should be done before conducting a study (Thomas, 1997). In practice, retrospective power analysis, although controversial, may be able to highlight the limitations of a study design after the data has been observed. In this study, I conducted power calculations to assess the range of effects that could be detected at the level of single-marker analysis and gene-based analysis. I implemented a module, `StudyDesign`, in the software `MAMBA` to perform a broad range of power calculations with different gene-based tests. To calculate the power to detect association using single marker association I used the Genetic Power Calculator (GPC) available from `http://pngu.mgh.harvard.edu/~purcell/gpc/` (Purcell et al., 2003).

5.3.3 Results

Individual gene level analysis of PTVs. I assessed association between PTVs and type 2 diabetes by aggregating allelic effects for each protein-coding gene. I employed the Bayesian SEMCC models tailored specifically to gene-level analysis of rare protein truncating variants and described in Section 3.1. In this analysis I did not address ancestry, thus it is possible that stratification may drive some of the suggestive association results described below.

I characterize results at the following scales of association: 1) $\log_{10}(\text{BF}) > 2$ as suggestive; 2) $\log_{10}(\text{BF}) > 3$ as substantial; and 3) $\log_{10}(\text{BF}) > 4$ as strong. In combined analysis of all $12,940$ individuals no gene (of the $14,415$ containing at least one PTV) had a $\log_{10}(\text{BF}) > 4$. Three genes had suggestive signal of overall association with $\log_{10}(\text{BF}) > 2$ including *HIST1H2AD*, *OSGIN2*, and *C6orf48* (Table 5.4).

The association signal at *HIST1H2AD* (Bayes Factor = 886.93) was driven by 8 protein truncating variants (p.K130X, p.S123TfsX8, p.H124PfsX7, p.E122XfsX0, p.R82X, p.K75QfsX23, p.Y58X, and p.Q25X) with 16 PTV allelic copies observed in type 2 diabetes cases and 0 PTV allelic counts in controls.

The association signal at *OSGIN2* (Bayes Factor = 229.77) was driven by 3 protein truncating variants (p.H493RfsX8, p.F495LfsX7, and p.R498KfsX9) with 42 PTV allelic copies observed in type 2 diabetes cases and 86 PTV allelic counts in controls.

The association signal at *C6orf48* (Bayes Factor = 174.96) was driven by 6 protein truncating variants (c.IVS2 + 1GGTGA>G, c.IVS3 − 2A>G, p.Q41IfsX30, p.Q41LfsX30, and p.Y54X) with 14 PTV allelic copies observed in type 2 diabetes cases and 1 PTV allelic count in controls.

Biologically driven gene set analysis of PTVs. Another strategy for improving power to detect association signal in exome sequence data that has been successfully employed in exome sequencing studies of schizophrenia involves attempting to aggregate association signal across sets of genes with high prior knowledge of their relationship to disease etiology (Purcell et al., 2014). I used a manually curated collection of gene sets of Mendelian diseases and pathways for T2D. I used the SMP (statistics/-matrix/permutation) geneset enrichment procedure implemented in the PLINK/Seq package (`http://atgu.mgh.harvard.edu/plinkseq/`), which is basically a burden test across the aggregate allelic counts of PTVs in cases compared to controls. At the time this analysis was performed the C-alpha MRP approach had not been developed, which would serve as an alternative approach to the SMP procedure. The SMP procedure uses gene-based association statistics, and forms sums of these statistics over all genes in a set, the significance of which is evaluated by permutation. To account for different ancestries in the study design I applied stratified permutations (only label-swapping phenotype within ancestry). I considered the relative enrichment statistic, SSET/SEXOME, with significance evaluated empirically (10,000 replicates) based on the null distribution of this ratio. The reported effect sizes from the gene-set enrichment estimate analysis are estimates of the unconditional odds ratio that do not take exome-wide differences in case/control rates into account (Purcell et al., 2014). As a consortium, we selected 18 "premium" sets of genes that reflect the current knowledge of pathways ($N = 15$) involved in type 2 diabetes and the three sets of genes involved in monogenic form of diabetes above: 'Monogenic All' ($N = 81$); 'Monogenic Primary' ($N = 28$); and 'Monogenic OMIM' ($N = 13$). I restricted these analyses to singleton and ultra-rare (MAF $< 0.1\%$) PTVs. Using the SMP approach, we find association ($P < 0.05$) in the more inclusive monogenic gene set ("Monogenic All", 81 genes: $P = 0.006$, odds ratio [OR] $= 1.35$ for singletons; $P = 0.04$, OR $= 1.12$ for ultra-rare alleles) and the more restrictive monogenic gene set ("Monogenic OMIM", 13 genes: $P = 0.0088$, OR $= 2.39$ for singletons; $P = 0.02$, OR $= 1.82$ for ultra-rare alleles). We also detected a separate burden signal for increased T2D risk attributable to singleton alleles within the mTOR pathway ($P = 0.012$, OR $= 3.6$). The drivers of the mTOR pathway signal include three case-only PTV singletons in both *MNK1* and *MTOR* (Figure 5.11).

Quantifying power to detect association at exome-wide significance. These results highlight the challenges that exist in exome sequencing studies. To quantify

the effects that would be required to detect exome-wide signals of association I employed power calculations for the current exome sequencing study design. A Bonferroni correction for 100,000 tests was used corresponding to a conservative exome-wide significance threshold of $P < 5 \times 10^{-7}$. I made use of the genetic power calculator (`http://pngu.mgh.harvard.edu/~purcell/gpc/`) to estimate power to detect association of T2D (assuming 8% prevalence) with coding variants over a range of realistic genetic models assuming: (i) a fixed-effect across all ethnicities (6,500 cases and 6,500 controls); and (ii) an effect specific to one ancestry group (1,000 cases and 1,000 controls). I find that a genotype relative risk greater than 4.5 would be required to detect significant association at a minor allele frequency equal to 0.1%, and even for samples sizes like that of the combined exome array and exome sequencing study design ($> 80,000$ individuals) variant effects with genotype relative risk greater than 2 would be required, Figure 5.12.

For gene-based tests, I considered a Bonferroni correction for 20,000 genes, corresponding to $P < 2.5 \times 10^{-6}$. I used a simulated haplotype dataset from the SKAT package (`http://cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf`) and estimated the power of SKAT-O (Lee et al., 2012) to detect association of variants within a gene at this threshold as a function of the (phenotypic) variance in the liability scale explained by additive genetic effects and the percentage of variants that were causal (50% and 100%). As for single variant power calculations, I considered: (i) a fixed-effect across all ethnicities (6,500 cases and 6,500 controls); and (ii) an effect specific to one ancestry group (1,000 cases and 1,000 controls). I found that for the most optimistic scenario, where 100% of the variants contribute to signal and that effects were shared across all ancestry group, 80% power to detect association is achieved for effects that in aggregate explain greater than 1% of the phenotypic variance.

Gene	Number of PTVs	Total PTV Allelic Counts in T2D	Total PTV Allelic Counts in Controls	Bayes Factor (SEMCC)
HIST1H2AD	8	13	0	886.93
OSGIN2	3	42	86	229.77
C6orf48	6	14	1	174.96
TMCO2	8	25	6	80.02
MPL	8	2	15	73.73
ABHD4	3	15	2	60.04
ARHGEF19	7	0	8	45.10
TMPRSS9	12	7	25	42.81
CYB5RL	6	2	14	42.79
PRSS23	3	8	0	40.87
HNMT	4	8	0	40.86
CPT1C	7	8	0	40.86

Table 5.4: Top 10 PTV gene association results from the type 2 diabetes multi-ethnic exome sequencing study. Interestingly, *PRSS23* encodes a trypsinogen, which is a precursor form of the pancreatic enzyme trypsin. Activation of inappropriate trypsinogen in the pancreas can result in pancreatitis (Whitcomb et al., 1996).

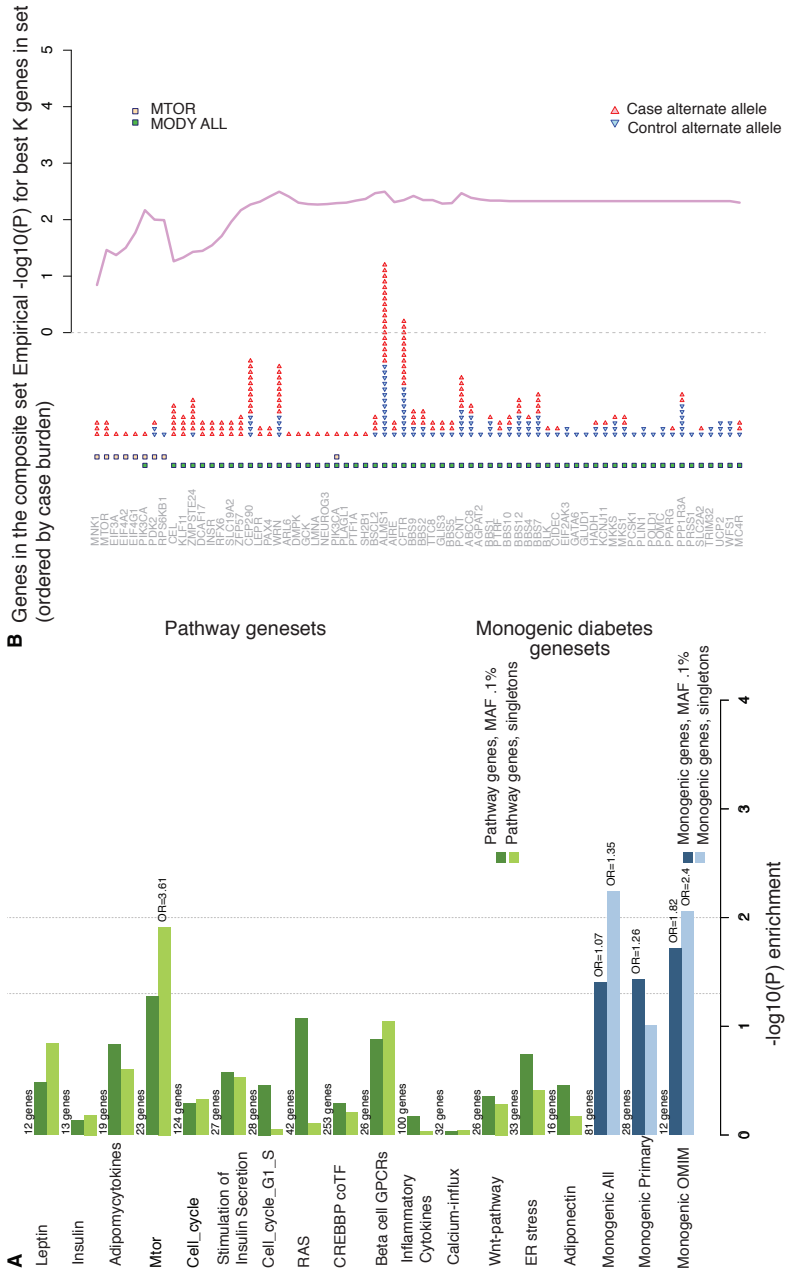


Figure 5.11: Accumulation of ultra-rare protein truncating alleles amongst genes contributing to pre-specified "biologically-driven" gene-sets. The analysis was designed specifically to determine whether the burden of rare, deleterious, alleles was differentially distributed amongst gene-set members in cases vs controls (using the "statistics/matrix/permutation" [SMP] utility in PLINK/SEQ, application of a burden test). We restricted these analyses to singleton and ultra-rare (MAF $< 0.1\%$) PTVs. A. Using this approach, we find association in the more inclusive monogenic gene set ("Monogenic All", 81 genes: $P = 0.006$, odds ratio [OR] $= 1.35$ for singletons; $P = 0.04$, OR $= 1.12$ for ultra-rare alleles) and the more restrictive monogenic gene set ("Monogenic OMIM", 13 genes: $P = 0.0088$, OR $= 2.39$ for singletons; $P = 0.02$, OR $= 1.82$ for ultra-rare alleles). We also detected a separate burden signal for increased T2D risk attributable to singleton alleles within the mTOR pathway ($P = 0.012$, OR $= 3.6$). B. The drivers of the mTOR pathway signal include three case-only PTV singletons in both *MNK1* and *MTOR*.

5.3.4 Conclusion

In this study I used the PTV Bayesian model for case-control analysis to assess association between PTVs and type 2 diabetes. In addition, I used the SMP procedure presented in Purcell et al. (2014) to assess association between PTVs in manually curated gene sets (with prior relationships to T2D biology) and type 2 diabetes. Evaluation of power to detect association indicates that with $12,940$ individuals we have $> 80\%$ power only for aggregate effects accounting for $> 1\%$ of the phenotypic variance in the liability scale. The findings from this study illustrate that larger sample sizes will be required to detect signal, that there exists evidence of a polygenic signature of PTVs in genes with prior relationships to T2D biology, and that for even sets of genes causal for monogenic and syndromic forms of type 2 diabetes heterozygous PTVs, on average, only have modest effects (OR < 3) on disease risk.

5.4 Discussion

In this chapter I have presented results from the analysis of protein truncating variants in three studies of common diseases. I applied the methods presented in Chapter 3 for assessing protein truncating variant association.

The results shed light on the benefits of using the proposed methods. For example, in the breast cancer targeted sequencing study the results highlight proof of principle of identifying established signals and comparing with new reference data sets, and highlight some of the bioinformatic and computational challenges inherent in data analysis of large-scale sequencing studies. Furthermore, using publicly available data from a published targeted sequencing study of autoimmune diseases I found (by applying the C-alpha MRP test) suggestive evidence of PTV association in *TNFAIP3* and type 1 diabetes risk. In addition, in the type 2 diabetes exome sequencing study I showed the challenges presented by large-scale exome sequencing study by conducting a study of power to detect association at exome-wide significance. However, I also found evidence of a polygenic signature of PTVs in sets of genes with prior relationship to T2D.

These promising results indicate that there are signals to be discovered in rare variant association studies and that investigators should carefully consider the statistical and bioinformatic approaches applied to the data sets.

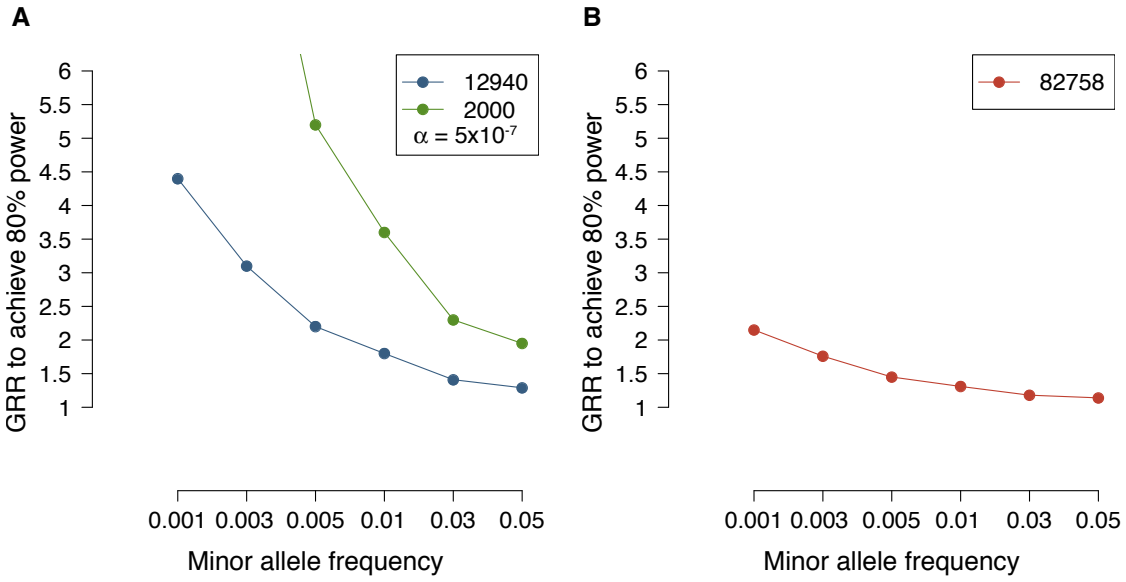


Figure 5.12: Power to detect single variant association. The Genetic Power Calculator (GPC) was used to illustrate the nature of effect sizes required to achieve 80% power to detect association, using single-marker analysis, at a significance level $\alpha = 5 \times 10^{-7}$. A) The genotype relative risk (GRR) required to achieve $> 80\%$ power to detect single variant association for two sample size scenarios: i) ancestry-specific variant in 2,000 samples and ii) variant observed with shared frequency across all ancestry groups in 12,940 samples in the exome sequencing study only, and B) the genotype relative risk (GRR) required to achieve $> 80\%$ power to detect single variant association at varying minor allele frequencies in the combined exome array and sequencing data set in the exome focused T2D study (effective sample size $82,758$).

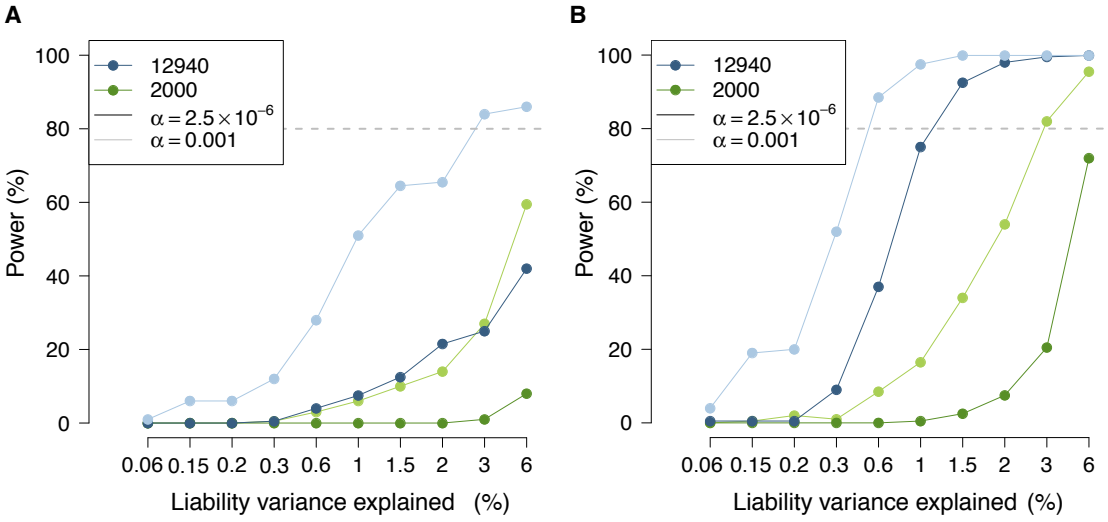


Figure 5.13: Power to detect gene-based association. The power calculator implemented in MAMBA was used to illustrate the nature of cumulative liability scale variance explained by PTVs required to detect exome-wide significance threshold of $\alpha = 2.5 \times 10^{-6}$. SKAT-O was used for illustrative purposes only. A) Power for gene based test of association with SKAT-O at different levels of variance explained on the liability scale with 50% of the variants contributing to disease risk while the remaining 50% have no effect on disease risk in $2,000$ (ancestry-specific effects; green) and $12,940$ (ancestry-shared effects; blue) samples. B) Power for gene based test of association with SKAT-O at different levels of variance explained on the liability scale with 100% of the variants contributing to disease risk in $2,000$ (ancestry-specific effects; green) and $12,940$ (ancestry-shared effects; blue) samples. From the simulation studies it is clear that under the optimistic model where effects are shared across all ethnicities (blue line) and (b) all variants contribute power is $> 75\%$ for $\alpha = 2.5 \times 10^{-6}$. However, as shown in A) and/or B) power drops rapidly if we relax either criterion, i.e. reduce the sample size to 2000, or allow in 50% neutral variants.

Chapter 6

Application of methods to the study of protein truncating variants and their relevance to medical traits: quantitative traits

Quantitative trait measurements can be key for making biological and clinical insights from genetic studies. For instance, the identification of protein truncating variants in *PCSK9* conferring protection to coronary heart disease was facilitated by the well established relationship that high LDL causes heart disease (Manninen et al., 1992; Sorlie et al., 1999; Cohen et al., 2005, 2006). To date, it remains unclear whether genetic associations of other quantitative traits will yield similar insights. Nonetheless, a recently published study finds, with the help of common variant associations and mendelian randomization, that triglycerides is causal for coronary heart disease (Do et al., 2013).

In this chapter, I use the Bayesian SEMGEM method from Section 3.2 and the C-alpha MRP test from Section 3.3 in two studies. In the first study, I assess the association of PTVs to low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC) by applying the Bayesian SEMGEM method and the C-alpha MRP test in Chapter 3 to the genotype and phenotype data from the T2D multi-ethnic exome sequencing study presented in Section 5.3. In the second study, I assess the association of PTVs in the Illumina HumanExome BeadChip array by applying the Bayesian SEMGEM method and the C-alpha MRP test in Chapter 3 to metabolic measurements generated with nuclear magnetic resonance (NMR) spectroscopy in $4,097$ individuals from the Oxford Biobank study.

For each of the studies in this chapter, I first give a brief overview of the main study where I describe the samples collected, the data generated, and the QC steps taken to clean up the data, which may have been primarily conducted by other members in the consortium. Then, I describe my role in the study. Finally, I present results from my analysis of the data.

I expect that the approach taken in this chapter will highlight the utility of the methods for analyzing protein truncating variants and quantitative traits presented in Chapter 3.

6.1 Exome sequencing analysis of protein truncating variants and the lipid profile in a multi-ethnic type 2 diabetes case-control study

6.1.1 Overview of the study

Plasma concentrations of total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and triglycerides are important heritable risk factors for cardiovascular disease in individuals with and without type 2 diabetes. In patients with type 2 diabetes in particular, the usual abnormalities include high levels triglycerides and low levels of high-density lipoprotein-cholesterol (Chehade et al., 2013), and may contribute to the increased risk of cardiovascular disease in T2D (Laakso, 1999). Increasingly large-scale genome-wide meta-analyses, mostly in individuals of European descent, continue to identify association of common noncoding single-nucleotide variants with small effects on plasma lipid levels (Teslovich et al., 2010; Do et al., 2013; Global Lipids Genetics Consortium, 2013). Recently, genetic studies have gravitated towards understanding the contribution of low-frequency and rare genetic variants to trait variation (Lange et al., 2014; Peloso et al., 2014). In rare variant association studies of lipid traits, in particular, the study of protein truncating variants has yielded extremely valuable insights into the role they may play in disease (Cohen et al., 2006; Jørgensen et al., 2014; Myocardial Infarction Genetics Consortium, 2014). The most recent success confirmed that targeting *NPC1L1* for inhibition to treat coronary heart disease is supported by genetic data (Myocardial Infarction Genetics Consortium, 2014). Dissecting the role of genetic variants, more importantly PTVs, to lipid trait variation can be crucial for identifying novel drug targets that could be of utility for the management of dyslipidemia, and subsequently to its comorbidities. The T2D-GENES and GoT2D consortia sought to investigate the role of

coding variants in lipid trait variation. The primary analysis of the exome sequencing study was conducted as part of a consortium effort. As part of the analysis group of the consortium, I contributed to the analysis of the data by applying software and methodology that I presented in Chapter 3.

To set the scene, I summarize the design of the study and present materials and methods.

Phenotype and genotype data. To assess the role of protein truncating variants in lipid levels across different ancestries, the T2D-GENES and GoT2D consortia made available lipid phenotype data for 11,284 individuals with at least one of the four blood lipid levels, and exome sequence data generated as part of the T2D study described in Section 5.3 to a subgroup focused on analyzing quantitative traits. The samples consisted of 5,502 type 2 diabetes patients and 5,782 controls ascertained from multiple balanced type 2 diabetes case control studies from five major ancestry groups (1,022 African-Americans, 2,160 East-Asians, 4,115 Europeans, 1,770 Hispanics, and 2,217 South-Asians).

The sequencing data, as explained in Section 5.3, had been generated at the Broad Institute and called using GATK (DePristo et al., 2011). All variant QC of the exome sequencing data set was conducted by the Broad Institute and explained in detail in Section 5.3. I obtained the VCF file that I used for the analyses presented in this Section from the Broad ftp where genotype data passing QC was shared.

Phenotype QC and normalization. Phenotype QC and normalization was conducted by Xueling Sim (University of Michigan) and Pierre Fontanillas (Broad Institute) who were members of the GoT2D and T2D-GENES consortia. LDL-C was calculated using Friedewald formula, LDL-C = TC − HDL-C − (TG/5), for those with TG < 400mg/dl (Friedewald et al., 1972; Warnick et al., 1990). To account for effects of lipid-lowering medication they estimated untreated lipid levels (LDL-C, TC) in individuals reported to be on lipid-lowering medication using the proposed procedure in Peloso et al. (2014). The principle behind the approach is that it has been demonstrated to perform well in accounting for treatment effects in studies of quantitative traits (Tobin et al., 2005). In individuals on lipid medication, they first divided TC by 0.8 (as it is estimated that statins at average dose reduce TC by 20%), and this adjusted TC was used in the calculation of LDL-C. In studies where measured LDL-C was available, they divided LDL-C by 0.7 (as it is estimated that statins at average dose reduce LDL-C by 30%) for individuals on lipid medication.

No adjustment was made for HDL-C and TG. For cases and controls, separately, they used the residuals for the four lipid traits adjusted for age, gender, and the first ten principal components. For the four traits, study- and disease-specific inverse-rank normalized residuals were calculated. Overall, T2D cases had higher TG (increase of $33-48\%$ compared to controls), lower HDL-C (reduction of $6-16\%$), and a higher percentage of them were on lipid-lowering medication, consistent with characteristics of diabetic dyslipidemia and higher likelihood of using medication to achieve a lower LDL-C threshold in T2D patients (Chehade et al., 2013). After adjusting for medication, LDL-C and TC were higher in African-American, East-Asian, and Hispanic individuals with type 2 diabetes. They observed similar patterns when restricted to individuals not on any lipid-lowering medication. African-Americans had the lowest TG and highest HDL-C, as previously observed (Sorlie et al., 1999).

Variant annotation. Variant annotation was described in Section 5.3.

6.1.2 Overview of my role

The analysis of the lipid trait data set combined with exome sequencing data was done collaboratively, involving multiple researchers. As a member of the analysis group I was involved in many parts of the analysis. However, in this section, I focus on the application of the methods presented in Chapter 3 for the analysis of PTVs.

Individual gene level analysis of PTVs. In this study, I applied the Bayesian SEMGEM models, described in Section 3.2, across all protein-coding genes with at least a single PTV to assess association between PTVs and the four studied lipid traits. In addition, I applied the C-alpha MRP test, described in Section 3.3, across all protein-coding genes with at least a single PTV to assess association between PTVs and the joint lipid profile.

PTV analysis of gene sets. Motivated by the findings in Purcell et al. (2014), where a polygenic burden of PTVs in schizophrenia was discovered, I sought to assess whether any evidence of a polygenic PTV association signature to the four studied lipid traits existed in the data set. I applied the C-alpha MRP test to assess association between PTVs in a manually curated gene set, consisting of 18 genes involved in monogenic forms of lipid disorders (dyslipidemias; Table 6.3), and the four lipid traits.

Analysis of T2D patients with *APOC3* PTVs. Multiple recent publications (Pollin et al., 2008; Hofker, 2010; Tachmazidou et al., 2013; Jørgensen et al., 2014), including one that I co-authored (Rivas et al., 2013), have shown that PTVs in *APOC3* are associated with reduced triglyceride levels in population controls. However, to date, nothing is known about the impact it has on triglyceride levels in T2D patients. In this study, I conducted the analysis of *APOC3* PTVs in T2D patients.

6.1.3 Results

Individual gene level analysis of PTVs. I assessed association between PTVs and the four lipid traits in the standard lipid profile by employing the three Bayesian models described in Chapter 3 (SEM, GEM, and GEM-NMD). I characterize results at the following scales of association: 1) $\log_{10}(\text{BF}) > 2$ as suggestive; 2) $\log_{10}(\text{BF}) > 3$ as substantial; and 3) $\log_{10}(\text{BF}) > 4$ as strong. I found suggestive ($\log_{10}(\text{BF}) > 2$) evidence of association for 26 genes to at least one trait, six genes showed substantial evidence of association ($\log_{10}(\text{BF}) > 3$; *APOB*, *CETP*, *PCSK9*, *LPL*, *CABLES2*, and *APOC3*) with three showing strong evidence of association ($\log_{10}(\text{BF}) > 4$; *APOB*, *LPL*, and *APOC3*)(Table 6.1).

PTVs in *APOB* were associated with LDL-C and TC with trait levels (-2.1 standard deviation [s.d.] below the population mean) consistent with a cardioprotective profile and with a hypobetalipoproteinemia clinical feature (OMIM #615558, Pulai et al. (1998)). Interestingly, the GEM-NMD model, which incorporates predictions of nonsense-mediated decay to perform the grouping of PTVs, had stronger evidence of association ($\log_{10}(\text{BF}) = 10.2$) than the models that do not NMD predictions to group variants including the SEM model ($\log_{10}(\text{BF}) = 8.1$), and the GEM model ($\log_{10}(\text{BF}) = 10.1$). Upon further examination, I found that an individual carrier of a frameshift indel (p.E4441SfsX29, *APOB*) predicted to escape nonsense-mediated decay had very high levels of LDL-C and TC (2.03 and 1.91 s.d. above the population mean, respectively). It is possible that this variant results in a truncated form of apoB with defective LDLR binding affinity, which have been documented to associate with obesity, high cholesterol levels, and increased risk of coronary artery disease (Corsini et al., 1989) potentially illustrating the drastic consequences that the positional context of PTVs may have on human health.

PTVs in *CETP* were associated with HDL-C ($\log_{10}(\text{BF}) = 3.1$) with trait levels (1.0 s.d. above the population mean) consistent with the clinical features of hyperalphalipoproteinemia (OMIM #143470,Glueck et al. (1975)), which has been found, at

least in rats, dogs, and pigs, to associate with protection to atherosclerosis (Russell and Proctor, 2006).

In *PCSK9* I found evidence of PTV association consistent with findings from Cohen et al. (2005), where LDL-C and TC levels were lower than the population mean (-0.90 and -0.76, respectively).

Association signal in *LPL* was mainly driven by a common nonsense variant, p.S447X, which was associated with lower triglyceride (0.16 s.d. below the population mean) and higher HDL-C (0.13 s.d. above the population mean) levels.

PTVs in *CABLES2* were associated with HDL-C levels (2.1 s.d. above the population mean). Only four PTV copies were observed for this gene raising the concern that this may be a spurious association and that further follow-up and replication data will be required.

In *APOC3* I found strong evidence of association between 5 PTVs and triglyceride levels (1.5 s.d. below the population mean), and suggestive evidence of association to HDL-C levels (0.78 s.d. above the population mean). This result further corroborates findings from a published study showing association between *APOC3* PTVs and triglyceride levels in the Oxford Biobank cohort (Rivas et al., 2013).

I assessed association between PTVs and the joint lipid profile (the four lipid traits) by employing the C-alpha MRP test described in Section 3.3. I assumed that the expected correlation of genetic effects between protein truncating variants all in the same gene and any single trait to be 1. Hence, $\mathbf{R}_{\text{var}}$ was a matrix of ones. Furthermore, I incorporated estimates of the pairwise correlation of genetic effects for the four lipid traits from a reference database. The reference database was provided by Matti Pirinen and generated by applying a bivariate linear mixed model (biLMM) (Pirinen, 2015) to a common variant genotype and metabolomics data set described in Kettunen et al. (2012), which included LDL-C, HDL-C, TG, and TC as phenotypes. Hence, $\mathbf{R}_{\text{phen}}$ was a matrix of point estimates of the expected correlation of genetic effects between a single variant and a pair of traits.

After applying the C-alpha MRP test to $13,453$ genes separately, only one gene (*APOC3*) reached exome-wide significance ($\alpha = 2.5 \times 10^{-6}$) and eight genes (*LPL*, *C4orf50*, *CETP*, *HIST1H4F*, *PPAN*, *APOB*, and *STOML1*) showed suggestive evidence of association ($P \leq 1 \times 10^{-4}$) (Table 6.2). Four of the eight genes highlighted (*APOB, CETP, LPL, APOC3*) had evidence of association in the single trait analysis. *HIST1H4F* is 147.5kb upstream of a common effect association (rs1800562) identified from a large-scale meta-analysis study of lipid levels (Global Lipids Genetics Consortium, 2013).

Enrichment of PTV association in genes involved in monogenic dyslipidemias. Deep understanding of the impact of PTVs in genes with known clinical relevance is critical for proper interpretation of novel variation in clinical sequencing studies. I generated a list of 18 genes identified in Hegele (2001) to be involved in monogenic forms of lipid disorders (dyslipidemias, Table 6.3). I loaded the gene set to the PLINK/SEQ locdb database as a locus-set using the command: `pseq proj locset-load --group gencode --name dyslipidemia --file dyslipidemia.txt` . To assess association between protein truncating variants in the dyslipidemia gene set and the joint lipid profile I applied the C-alpha MRP test extension for gene sets described in Subsection 3.3.3. I assumed that the expected correlation of genetic effects between protein truncating variants and any single trait to be 1 if and only if the protein truncating variants belonged to the same gene, otherwise I assumed the expected correlation of genetic effects to be 0. Hence, $\mathbf{R}_{\text{var}}$ was a block matrix of ones where each block corresponded to a gene and off-diagonal blocks represented the expected correlation of genetic effects between PTVs across two different genes (assumed to be 0). $\mathbf{R}_{\text{phen}}$ was a matrix of point estimates of the expected correlation of genetic effects between a single variant and a pair of traits. The expected correlation estimates were obtained using the reference database described in the previous paragraph. Furthermore, I focused the analysis on three different scenarios (motivated by (Purcell et al., 2014)) by restricting the frequencies of PTVs: i) singletons, ii) ultra-rare (singletons to at most 10 copies), and iii) all PTVs (no allelic count cutoff).

I observed an aggregate PTV association signal to the lipid profile $P < 1 \times 10^{-5}$ after applying $100,000$ permutations across the gene set when all PTV were included (Figure 6.1). When I restricted the analysis to singletons and ultra-rare PTVs I found nominal evidence of association ($P < 0.01$, Figure 6.1).

Five of the eighteen genes included in the gene set had already been identified to associate with one of the four lipid traits (either $\log_{10}(\text{BF}) > 2$ or C-alpha MRP P value $< 1 \times 10^{-4}$) in single gene analysis. It was unclear if any signal existed in the gene set beyond these five genes. Therefore, I excluded *PCSK9*, *APOB*, *CETP*, *LDLR*, and *LPL* from the list of 18 genes and generated a new gene set that I referred to as the "DYSLIPIDEMIA_ExSigRemoved" gene set. I applied the C-alpha MRP test using the same choice for $\mathbf{R}_{\text{var}}$ and $\mathbf{R}_{\text{phen}}$ as indicated with the more inclusive gene set. For ultra-rare PTV alleles, I observed nominal evidence of association ($P = 0.0069$) between the "DYSLIPIDEMIA_ExSigRemoved" gene set and the joint lipid profile (Figure 6.1).

To better understand which variants were driving the signal of association and to learn their impact across the four lipid traits in the standard lipid profile I applied the MCMC clustering algorithm described in Subsection 3.3.3. I fixed the number of clusters to 3 (I examined additional numbers of clusters including 2 and 4, but there appears to be enough information for 3 clusters and not enough information to learn 4 clusters - as two of the four clusters converged to the null cluster of zero effects). I used the options `--burn 5000` and `--niter 10000` in `MAMBA` to reflect the total number of iterations $(10,000)$, and the number of initial iterations to discard.

After I applied the MCMC clustering algorithm, I found that the two alternative cluster components estimated correspond to either i) strong PTV effects lowering total cholesterol and LDL-C levels (1.14 [2.5% and 97.5% samples quantiles of the parameters are 0.65 and 1.63, respectively] and 1.01 [0.46 to 1.53] standard deviations below the population mean, respectively) or ii) strong PTV effects increasing total cholesterol and triglyceride levels (2.54 [1.5 to 3.59] and 1.80 [0.85 to 2.80] standard deviations above the population mean, respectively) (Figure 6.2). Interestingly, one of the learned clusters is consistent with a cardioprotective profile thus raising the possibility that some of the variants/genes, just like *PCSK9*, may be informative drug targets for inhibition. Further research into the sensitivity of the method on sample size, priors, and number of clusters will be required. Nonetheless, this exercise highlights the potential in characterizing drivers of association.

Type 2 diabetes patients carrying *APOC3* PTVs have normal triglyceride levels. It appears that loss of APOC3 function confers protection to coronary heart disease (The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, 2014). However, Cohen et al. (2014) contest that because PTVs in *APOC3* result in a 17% reduction in LDL-C the cardioprotective effect may be acting through LDL-C and not directly through triglyceride. Moreover, Cohen et al. (2014) claim that PCSK9 inhibition may be a better strategy for therapeutic intervention due to the magnitude of cardioprotective effect conferred by PTVs in *PCSK9*. Thus, the extent to which the *APOC3* discovery can provide major clinical benefits remains unclear. However, it is clear that dyslipidemia, in particular hypertriglyceridemia, is a major complication of insulin resistance and type 2 diabetes due to VLDL overproduction (Lewis et al., 2002). The fact that half of the samples in the exome sequencing study design had type 2 diabetes gave me the chance to examine whether or not individuals with T2D carrying a PTV in *APOC3* have lower levels of triglycerides compared to diabetes patients that do not

have a PTV in *APOC3*. I identified nine PTV carriers across three PTVs (p.R19X, c.IVS2 + 1G>A, and p.T94LfsX6) with low triglyceride levels (1.74 s.d. below the mean of T2D cases; $\log_{10}$(BF) of association = 5.17) equivalent to a 62% mean reduction in triglyceride levels compared to T2D cases without loss of function variants in *APOC3*. All 9 individuals had reduced triglyceride levels compared to the mean measurement in T2D for their corresponding population (Figure 6.3):

- 1 East Asian PTV carrier with a 76.4% reduction in TG level compared to the mean TG level in T2D cases in the East Asian sample collection,
- 2 South Asian PTV carriers with a 59.8% reduction in TG level compared to the mean TG in South Asian T2D cases,
- 4 European PTV carriers with a 53.1% reduction in TG level compared to the mean TG in European T2D cases, and
- 2 African American PTV carriers with a 73.5% reduction in TG level compared to the mean TG in African American T2D cases.

In controls, I calculated the mean reduction in triglyceride levels to be 50.7% for PTV carriers compared to non-carriers without type 2 diabetes.

To put it in perspective I compared the magnitude of effects observed in T2D cases for *APOC3* PTVs to effects estimated from common variants associated to triglyceride levels in genome-wide association studies, and to the estimated effects for the same group of variants in the T2D controls (Figure 6.4) and show that the strength of the effects observed for *APOC3* PTVs is almost a standard deviation stronger compared to the common variant associations and that the effect in T2D is consistent (possibly stronger) with the effect observed in controls. The observation that type 2 diabetes patients carrying PTVs in *APOC3* have normal triglyceride levels supports the hypothesis that APOC3 inhibition might be a valuable strategy in the management of diabetic hypertriglyceridemia in particular.

Gene	Number of PTVs	Number of Carriers	LDL-C Bayes Factor (Mean, Model)	TG Bayes Factor (Mean, Model)	HDL-C Bayes Factor (Mean, Model)	TC Bayes Factor (Mean, Model)
PCSK9	8	24	7033 (-0.90,GEM)	0.07 (-0.06,SEM)	0.45 (0.39,SEM)	582.3 (-0.76,GEM)
TBX19	6	16	148.5 (0.95,SEM)	0.08 (-0.04,SEM)	0.10 (-0.12,SEM)	3.6 (0.65,SEM)
APOB	9	9	>100,000 (-2.1,GEM-NMD)	8.1 (-0.94,SEM)	0.12 (0.09,SEM)	>100,000 (-2.1,GEM-NMD)
IMPG2	7	8	127.3 (1.3,SEM)	0.14 (0.14,SEM)	0.13 (-0.11,SEM)	34.0 (1.15,SEM)
MEPE	6	23	513.4 (-0.68,GEM-NMD)	0.10 (0.17,SEM)	0.37 (-0.37,SEM)	10.2 (-0.45,GEM-NMD)
NKD2	2	3	146.7 (2.1,SEM)	0.30 (0.45,SEM)	5.1 (-0.42,GEM-NMD)	100.2 (2.0,SEM)
TRPC4	2	2	141.4 (-2.6,SEM)	2.0 (0.95,GEM)	1.2 (0.13,GEM)	2.3 (-0.91,GEM)
MYH3	12	26	156.4 (-0.71,GEM)	0.07 (-0.06,SEM)	1.8 (-0.39,GEM-NMD)	10.9 (-0.56,GEM)
ATP11B	5	5	0.19 (0.16,SEM)	191.6 (-1.7,SEM)	0.54 (0.69,SEM)	0.18 (-0.18,SEM)
ATP10D	13	17	0.26 (-0.35,SEM)	515.4 (-0.56,GEM)	8.8 (0.68,SEM)	0.07 (-0.05,SEM)
HIST1H2BM	4	10	0.71 (-0.59,SEM)	343.6 (-1.2,SEM)	0.11 (-0.02,SEM)	8.5 (-0.91,SEM)
LPL	4	1825	0.0074 (0.005,SEM)	>100,000 (-0.16,SEM)	>100,000 (0.13,SEM)	0.35 (-0.03,GEM)
APOC3	5	19	0.08 (-0.004,SEM)	>100,000 (-1.5,SEM)	52.4 (0.78,GEM)	0.11 (-0.18,SEM)
KRT2	3	3	0.25 (-0.29,SEM)	733.7 (-2.3,SEM)	0.31 (0.15,GEM)	0.68 (-0.83,SEM)
TXNDC16	6	6	0.30 (-0.25,GEM)	154.8 (-1.5,SEM)	0.52 (0.54,GEM)	0.20 (-0.25,GEM)
GLT8D1	7	20	0.65 (-0.46,SEM)	0.31 (0.27,GEM)	108.9 (-0.32,GEM)	0.78 (-0.27,GEM)
PRMT10	7	7	1.5 (-0.85,SEM)	0.20 (0.13,GEM)	100.3 (-1.3,SEM)	13.7 (-0.58,GEM)
SLC26A5	8	11	0.42 (-0.51,GEM)	1.7 (0.55,GEM)	129.3 (-1.1,SEM)	1.5 (-0.33,GEM)
KIF26A	3	4	0.60 (0.55,GEM)	0.40 (0.55,GEM)	156.6 (-1.8,SEM)	0.29 (0.42,GEM)
CETP	11	18	0.11 (-0.18,SEM)	0.23 (0.35,SEM)	1195.8 (1.0,SEM)	0.19 (0.30,SEM)
CABLES2	4	4	4.9 (-0.84,GEM)	1.00 (-0.73,GEM)	1575.0 (2.1,SEM)	0.40 (-0.003,GEM)
RFX5	4	7	65.2 (0.36,GEM)	0.36 (0.52,SEM)	0.20 (0.33,SEM)	134.2 (0.60,GEM)
FGFR1OP	2	4	33.0 (-1.6,SEM)	21.0 (-1.5,SEM)	0.22 (0.29,SEM)	140.3 (-1.8,SEM)
POLR1C	2	3	1.2 (0.26,GEM-NMD)	0.27 (0.07,SEM)	4.8 (0.97,GEM-NMD)	166.9 (-0.12,GEM-NMD)
TPMT	3	8	54.1 (-0.63,GEM)	1.1 (-0.71,SEM)	0.13 (0.075,SEM)	135.7 (-0.68,GEM)
RTDR1	2	2	63.6 (2.4,SEM)	2.7 (2.02,SEM)	0.51 (0.61,GEM)	133.5 (2.6,SEM)

Table 6.1: PTV gene based association results for univariate Bayesian models applied to LDL-C, TG, HDL-C, and TC levels in the multi-ethnic exome sequencing study. Results with Bayes Factor (BF) > 100 to at least one of the lipid traits are displayed (BF > 1000: dark purple, BF > 100: light purple). Mean quantitative trait level of PTV carriers is shown in parenthesis. Bayesian model (SEM,GEM, or GEM-NMD) with maximum Bayes Factor is shown.

Gene	Number of PTVs	Number of PTV alleles	P
APOC3	5	23	2.0×10^{-6}
LPL	4	2516	4.6×10^{-6}
C4orf50	9	332	1.7×10^{-5}
CETP	9	15	5.1×10^{-5}
HIST1H4F	3	25	4.1×10^{-5}
PPAN	2	13	4×10^{-5}
APOB	15	15	1×10^{-4}
STOML1	9	9	1×10^{-4}

Table 6.2: PTV gene based association results for C-alpha MRP test applied to LDL-C, TG, HDL-C, and TC levels in the multi-ethnic exome sequencing study. Results for genes with P value $\leq 1 \times 10^{-4}$ are shown.

Gene	Direction of effect and related trait	Position	Condition
ABCA1	Low HDL-C	9q31.1	Tangier disease
ABCG5	High LDL-C	2p21	Sitosterolemia
ABCG8	High LDL-C	2p21	Sitosterolemia
APOA1	Low HDL-C	11q23	ApoA-I deficiency
APOA5	High VLDL, high chylomicrons	11q23	ApoA-V deficiency
APOB	Low LDL-C, High LDL-C	2p24	Familial hypobetalipoproteinemia, Familial defective ApoB-100
APOC2	High chylomicrons	19q13	Familial ApoC-II deficiency
APOE	High VLDL, high chylomicrons	19q13	Familial dysbetalipoproteinemia
CETP	High HDL-C	16q13	Cholesteryl ester transfer protein deficiency
LCAT	Low HDL-C	16q22	Lecithin-cholesterol acyltransferase deficiency (fish-eye disease)
LDLR	High LDL-C	19p13	Familial hypercholesterolemia
LDLRAP1	High LDL-C	1p36	Autosomal recessive hypercholesterolemia
LIPC	High VLDL remnants	15q22	Familial hepatic lipase deficiency
LMF1	High triglycerides	16p13	Combined lipase deficiency
LPL	High chylomicrons	8p21	Lipoprotein lipase deficiency
MTTP	Low LDL-C	4q24	Abetalipoproteinemia
PCSK9	Low LDL-C, High LDL-C	1p32	PCSK9 deficiency Autosomal-dominant hypercholesterolemia
SAR1B	Low chylomicrons	5q31.1	Chylomicron retention disease

Table 6.3: Table of monogenic dyslipidemia genes with the name of condition in OMIM (Rahalkar and Hegele, 2008; Hegele, 2001).

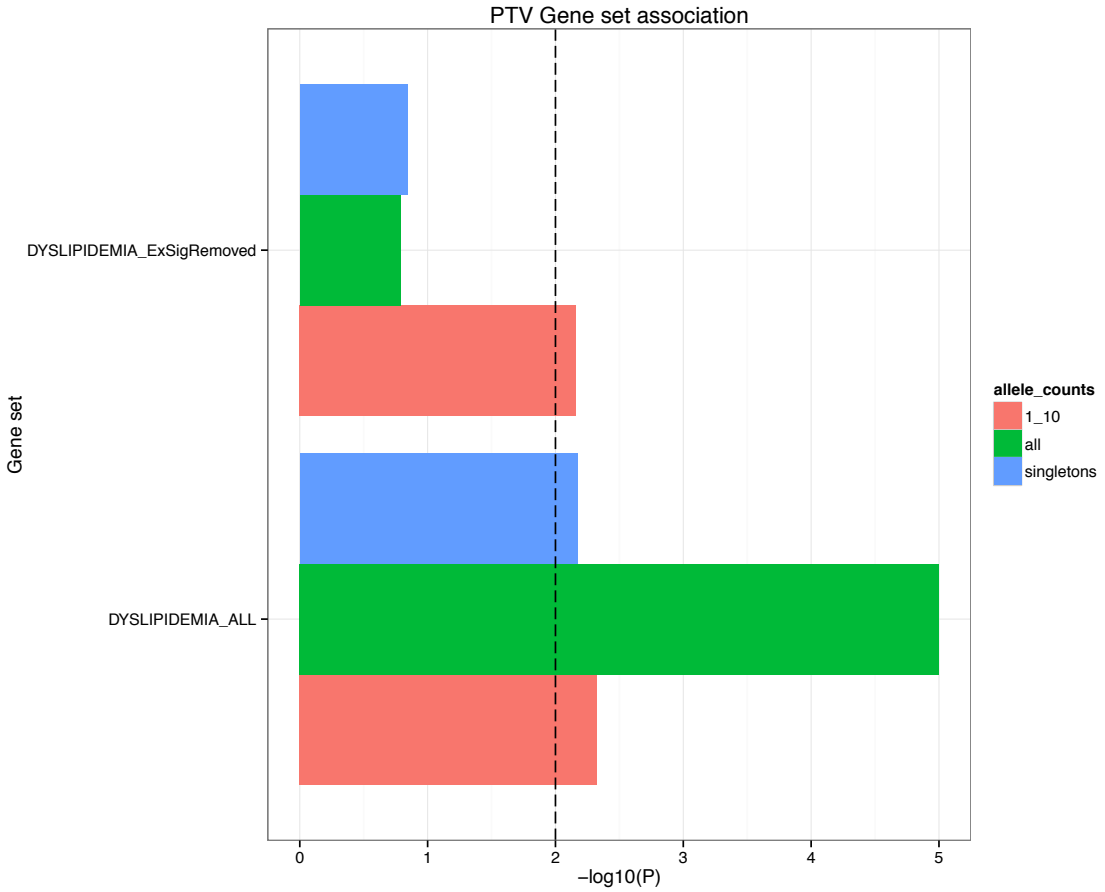


Figure 6.1: C-alpha MRP applied to the dyslipidemia gene sets. Shown are results, $-\log_{10}(P \text{ value})$ of enrichment of association, for PTV singletons (minor allele count = 1; blue), singletons to variants observed with less than or equal to 10 copies (red), and all variants without allele count thresholds (green). Nominal significant enrichment of association in the dyslipidemia gene set ($n = 18$, P value $< .01$) for every possible allele count threshold chosen was observed. The strongest evidence of association (P value $< 1 \times 10^{-5}$) was observed when all PTVs were included. Nominal evidence of association for the gene set "DYSLIPIDEMIA_ExSigRemoved" ($n = 13$; $P = 0.0069$) including PTVs at an allele count threshold of 10.

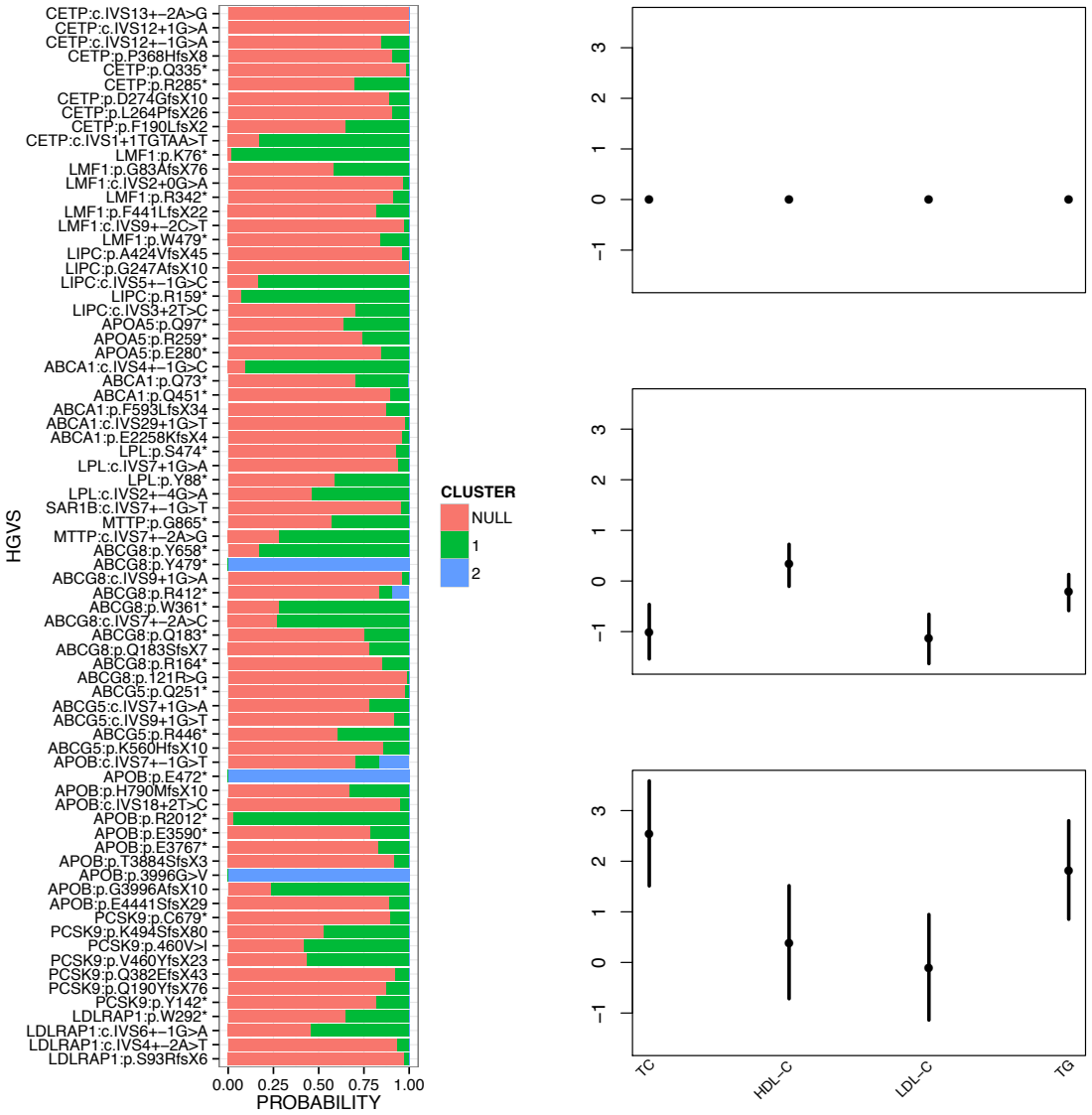


Figure 6.2: MCMC clustering algorithm described in Chapter 3 applied to PTVs in the monogenic dyslipidemia gene set and the standard lipid profile. Shown are the PTVs in the gene set and a horizontal barplot of the posterior probability that the variant belongs to one of the three clusters. On the right the estimated effect sizes per cluster for each of the four traits with 2.5% and 97.5% quantiles of the Gibbs samples shown. The plot at the top corresponds to the null cluster.

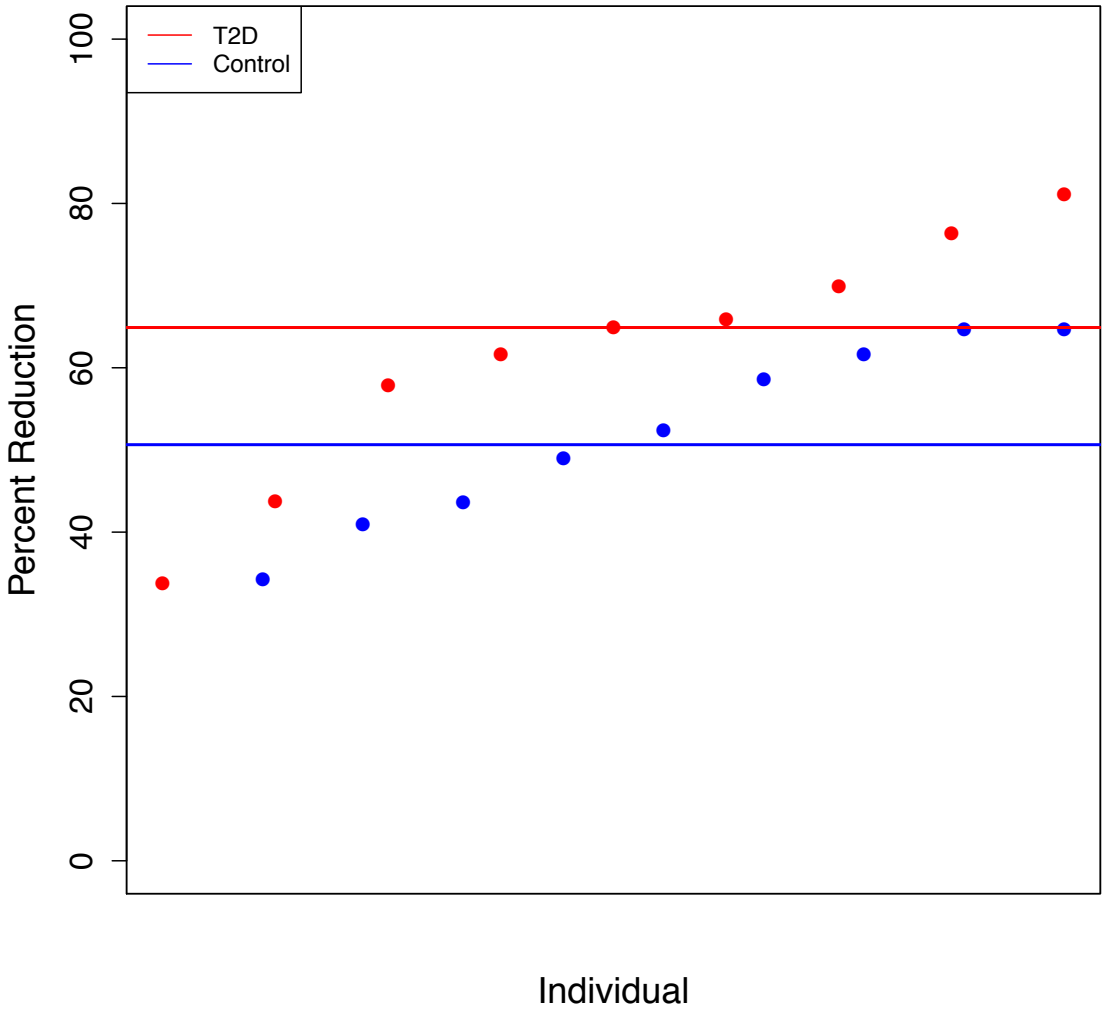


Figure 6.3: Percent triglyceride level reduction compared to disease or control population. A mean 62% reduction in triglyceride levels for T2D patients carrying a PTV in *APOC3* compared to T2D patients without a PTV in *APOC3* was estimated. A mean 50.6% reduction in triglyceride levels for controls carrying a PTV compared to the average population control was estimated. The individuals are plotted by rank in percent triglyceride level reduction.

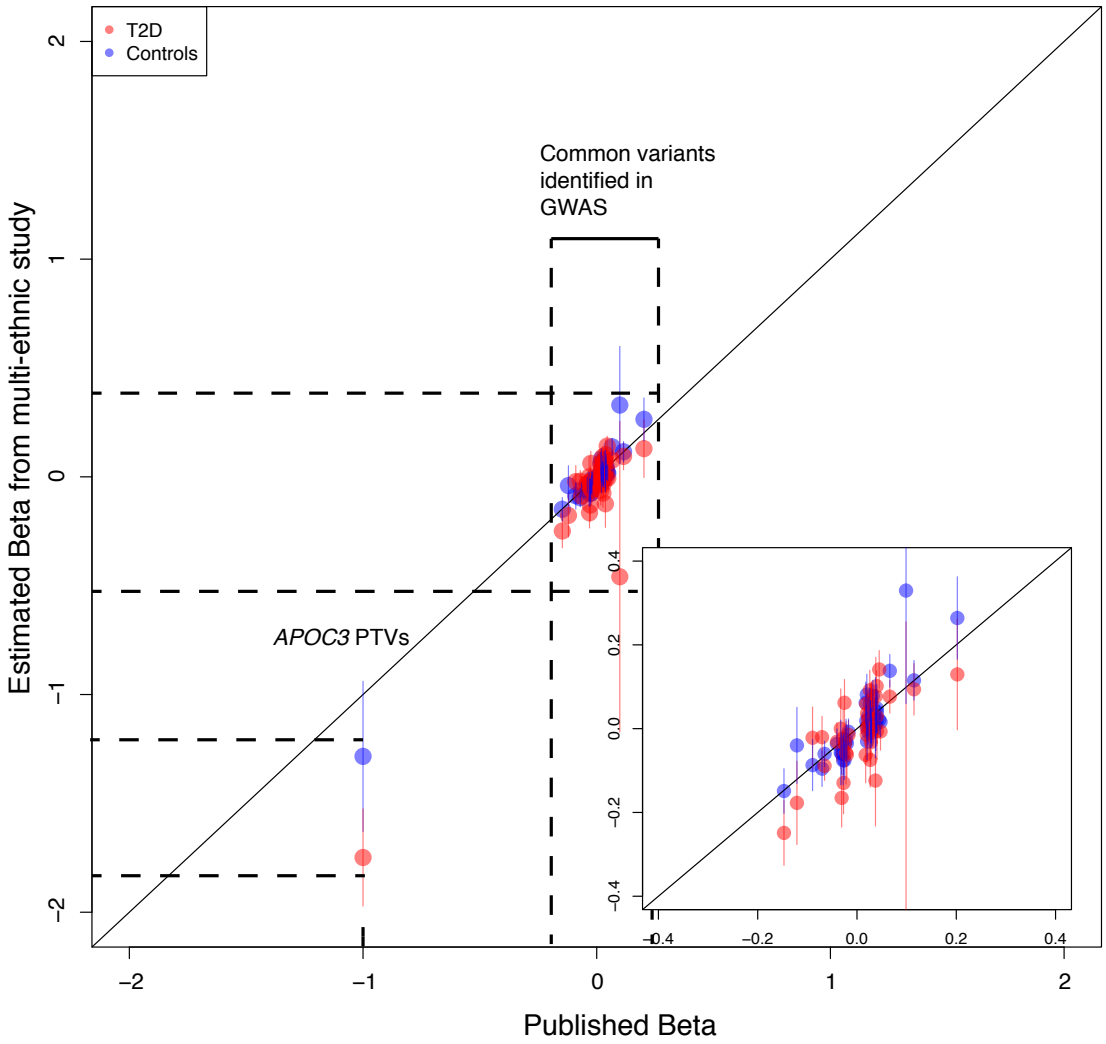


Figure 6.4: Comparison of effect sizes reported in publications for variants associated to triglyceride levels to effect sizes estimated from the multi-ethnic exome sequencing study in T2D cases and controls. On the x-axis the point estimate of the effect size, β, reported in published studies (Global Lipids Genetics Consortium, 2013; Tachmazidou et al., 2013; Timpson et al., 2014) is shown. Shown are the point estimates of the effect size, β, in the multi-ethnic exome sequencing study and its corresponding standard error.

6.1.4 Conclusion

Exome sequencing was conducted in greater than 11,000 multi-ethnic individuals in the GoT2D and T2D-GENES projects. I assessed association between PTVs in protein coding genes and the four main lipid traits measured (LDL-C, HDL-C, TG, and TC). I found three genes with strong evidence of association ($\log_{10}(\text{BF}) > 4$) that had already been identified to contain PTVs associated to one of the four lipid traits - demonstrating the challenges of gene discovery with rare variants, in particular PTVs. In addition, I found enrichment of PTV association to the standard lipid profile in genes catalogued to be causal for a lipid disorder, and evidence that T2D patients carrying an *APOC3* PTV have low triglyceride levels. It was brought to my attention that there may be more than 18 genes implicated in monogenic dyslipidemias. I have found, from additional literature review, that the genes *SLC10A2* and *LIPA* should be incorporated in the analysis of the monogenic dyslipidemia gene set. These findings show the value of exome sequencing data and the focused analysis of PTVs. Currently, ISIS pharmaceuticals is conducting Phase III trials for ISIS-APOCIII$_{\text{Rx}}$, an antisense drug designed to reduce apoC-III protein production (Isis Pharmaceuticals, 2014). ISIS-APOCIII$_{\text{Rx}}$ may not be the drug of choice for decreasing coronary heart disease risk in the general population, however, as supported by these findings, it is entirely plausible that it could be repositioned (Ashburn and Thor, 2004; Sanseau et al., 2012) to treat hypertriglyceridemia in T2D patients.

6.2 Exome-wide association study of protein truncating variants and blood metabolite measurements in the Oxford Biobank Study

6.2.1 Overview of the study

Metabolomics, the study of chemical processes involving metabolites, is currently undergoing a major revolution due to improvements in the instrumentation used for measuring metabolite levels in serum, plasma, urine, and tissue extracts, which has also led to a dramatic drop in costs (Beckonert et al., 2007). Furthermore, the integration of metabolite level measurements with human genetic data is a promising new avenue of research. The hope is that we can begin to integrate these data sets to understand the genetic architecture of metabolites and dissect their relationship to diseases, in particular those with a cardiometabolic component. Nuclear magnetic resonance (NMR) and mass spectroscopy are two research techniques that are currently being used to measure metabolite levels in human genetic association studies (Kettunen et al., 2012; Shin et al., 2014). The most recent study, for example, reported genome-wide significant associations at 145 metabolic loci across 400 metabolites in human blood (Shin et al., 2014). However, none of the published studies has focused on protein truncating variants. This is largely a consequence of two things: i) our catalogue of PTVs was incomplete in the era of common variant association studies, and ii) high throughput sequencing technologies have only recently come to market making rare variant association studies possible. The Oxford Biobank Exome Array Metabolomics study sought to investigate the contribution of coding variants to metabolite trait variation. The primary analysis of the exome array and the metabolomics data set was conducted by one additional analyst (as well as myself), Karina Banasik, a postdoc with Mark McCarthy (Oxford). As part of the analysis group of this study, I contributed to the analysis of the data by applying the Bayesian SEMGEM models presented in Section 3.2 and the C-alpha MRP test presented in Section 3.3 to assess for association between PTVs and the 123 metabolite traits both in a univariate and multivariate fashion across genes and gene sets (Figure 6.5).

Samples. The $4,097$ subjects that passed all QC steps in this study were collected as part of the Oxford Biobank, a collection of 30-50 year old healthy men and women living in Oxford recruited from primary care who underwent a detailed examination at a screening visit, donated DNA, and gave informed consent to be re-approached. The Primary Investigator of the Oxford Biobank collection is Fredrik Karpe.

Genotype data set. Exome array genotyping data was generated at Oxford University, and variant and individual quality control was performed by analysts in Mark McCarthy's lab. I obtained the PED file after QC, and a list of markers to exclude. I used this genotyping data set for my analysis.

Phenotype generation and QC. Metabolite measurements for 123 traits were generated using NMR spectroscopy from serum samples by Mika Ala-Korpela (Oulu), Appendix Table 2. The phenotype data was QC'ed for missingness data and outliers (zero or NA values) by Karina Banasik, Matthew Neville, and Fredrik Karpe (Oxford). Two phenotypes, glycerol and lactate, were replaced with measurements made by the Oxford Biobank group due to poor correlation with the measured variables at Oxford and large number of subjects with zero values. All 123 trait measurements were rank normalized across $4,097$ individuals using an in-house script to transform the traits to standard normal measurements after regressing out batch effects, age, gender, and first 10 principal components (Goh and Yap, 2009). The data harmonization was conducted by Karina Banasik (Oxford).

Variant annotation. Variant annotation of the exome-array was conducted as part of the T2D-GENES and GoT2D exome project by Pablo Cingolani (McGill) using the same annotation programs applied to the exome sequencing data set described in Section 5.3.

Estimating genetic correlation of metabolites. The pairwise correlations between the 123 NMR metabolite traits was generated by Matti Pirinen (FIMM) after applying a bivariate linear mixed model (biLMM) (Pirinen, 2015) that decomposes the total correlation into genetic and environmental parts to a data set that is described by Kettunen et al. (2012). I used the estimates of the expected correlation of genetic effects to apply the C-alpha MRP test.

6.2.2 Overview of my role

The analysis of the metabolomics data set combined with the exome array genotyping data set was done collaboratively, involving two analysts. As a member of the analysis group I was involved in many parts of the analysis. However, in this section, I focus on the application of the Bayesian SEMGEM model described in Section 3.2 and the C-alpha MRP test described in Section 3.3 to assess association between PTVs and the 123 metabolite traits.

Univariate and multivariate gene level analysis of PTVs. In this study, I applied the Bayesian SEMGEM models, described in Section 3.2, across 2,829 protein-coding genes with at least a single polymorphic PTV in the exome array data set for all 123 metabolite traits. I constructed an expected correlation of genetic effects matrix and applied the agglomerative hierarchical clustering algorithm described in Inouye et al. (2012) in order to identify networks of metabolites that are likely to share genetic effects. I applied the multivariate rare variant association test I presented in Section 3.3 to the networks of metabolites that are likely to share genetic effects and incorporated the pairwise estimates of expected genetic correlation of effects in testing.

Multivariate gene set analysis. To highlight the value of the method I sought to apply the C-alpha MRP test to sets of genes with genes considered to be involved in metabolism, in particular lipid metabolism. I applied the C-alpha MRP test to a manually curated list of gene sets (Johannes Kettunen [University of Oulu, Finland]) to assess association between PTVs in each gene set and the networks of metabolites.

Cluster	Number of Traits	Trait Labels
1	16	XL.HDL.P_c,XL.HDL.L_c,XL.HDL.PL_c,XL.HDL.FC_c,L.HDL.P_c, L.HDL.L_c,L.HDL.PL_c, L.HDL.C_c,L.HDL.CE_c, L.HDL.FC_c,HDL.C_c,HDL.D_c,DBinFA_c,BIStoDB_c,BIStoFA_c,FALen_c
2	13	XS.VLDL.PL_c,IDL.P_c,IDL.L_c, IDL.PL_c,L.LDL.P_c,L.LDL.L_c,L.LDL.PL_c,L.LDL.C_c, L.LDL.CE_c,M.LDL.P_c,M.LDL.PL_c,M.LDL.CE_c,LDL.C_c
3	13	IDL.C_c,IDL.FC_c,L.LDL.FC_c,XL.HDL.TG_c,Serum.C_c,EstC_c,FreeC_c,FAw3_c, otPUFA_c,DHA_c,TotPG_c,PC_c,SM_c
4	11	XXL.VLDL.P_c,XL.VLDL.P_c,XL.VLDL.PL_c, L.VLDL.P_c,L.VLDL.PL_c,L.VLDL.C_c, L.VLDL.CE_c,L.VLDL.FC_c,M.VLDL.P_c,M.VLDL.L_c,Gp_c
5	11	XXL.VLDL.L_c,XXL.VLDL.PL_c,XXL.VLDL.TG_c,XL.VLDL.L_c,XL.VLDL.TG_c, L.VLDL.L_c,L.VLDL.TG_c,M.VLDL.TG_c,VLDL.D_c,Ile_c,Leu_c
6	11	XS.VLDL.P_c,XS.VLDL.L_c,IDL.TG_c,M.LDL.L_c,M.LDL.C_c,S.LDL.P_c, S.LDL.L_c,S.LDL.C_c,ApoB_c,FAw6_c,LA_c
7	9	M.VLDL.PL_c,M.VLDL.C_c,M.VLDL.FC_c,S.VLDL.P_c,S.VLDL.TG_c, S.HDL.TG_c,Serum.TG_c,FAw79S_c,MUFA_c
8	9	XL.HDL.C_c,XL.HDL.CE_c,M.HDL.P_c,M.HDL.L_c,M.HDL.PL_c,M.HDL.C_c, M.HDL.CE_c,M.HDL.FC_c,ApoA1_c
9	8	Glycerol,S.HDL.P_c,S.HDL.L_c,Ala_c,Lactate,Phe_c,Pyr_c,Tyr_c
10	8	LDL.D_c,bOHBut_c,Ace_c,Cit_c,Crea_c,Gln_c,Gly_c,Urea_c
11	7	M.VLDL.CE_c,S.VLDL.L_c,S.VLDL.PL_c,S.VLDL.C_c, S.VLDL.FC_c,XS.VLDL.TG_c,TotFA_c
12	7	Alb_c,AcAce_c,Glc_c,His_c,Val_c,CH2inFA_c,CH2toDB_c

Table 6.4: Grouping of metabolites into 12 metabolic clusters.

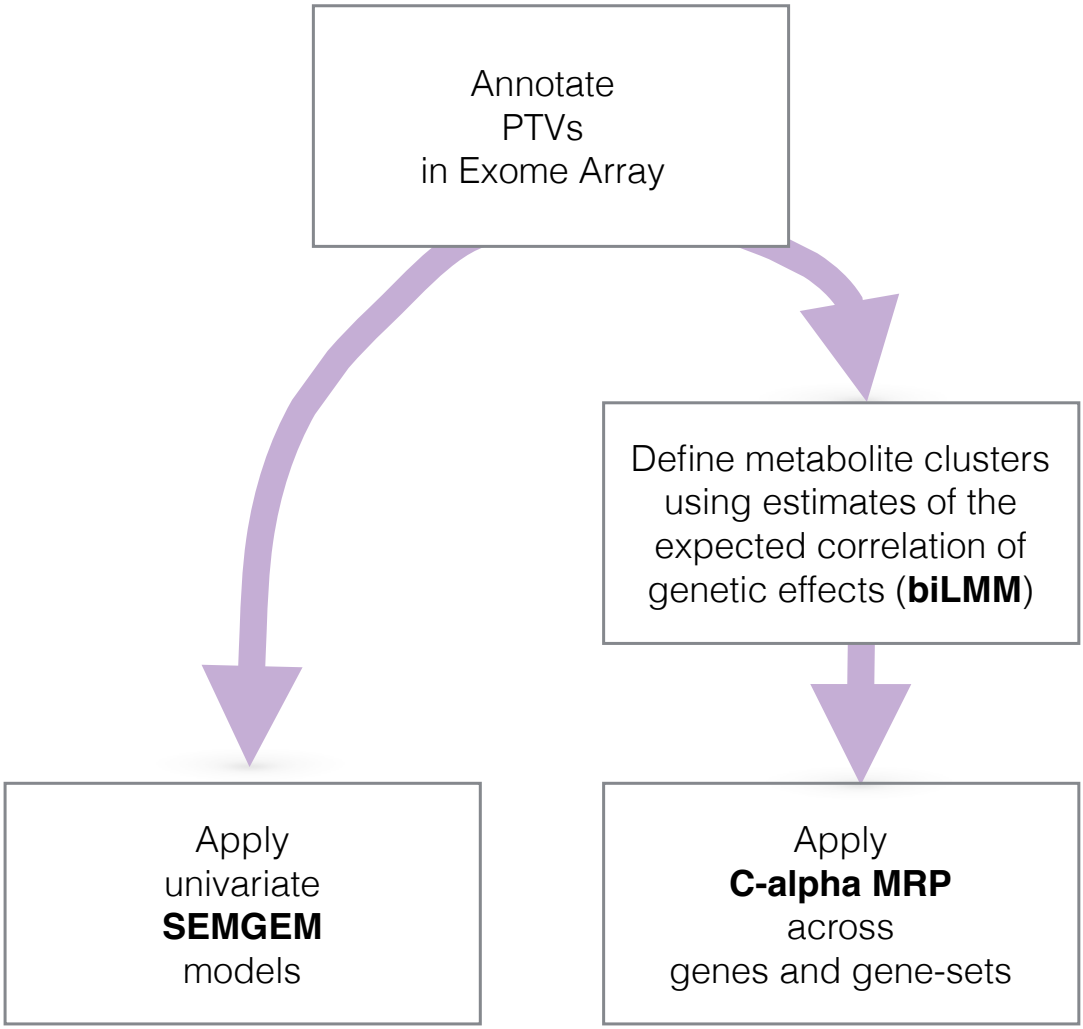


Figure 6.5: Diagram of PTV analysis in the Oxford Biobank study and NMR metabolite measurements. First, I integrated the PTV annotation for the exome array defined by the GoT2D/DIAGRAM study (Mahajan et al., 2015). To study the impact of PTVs on the metabolite levels I applied the SEMGEM models to the 123 measurements. I defined 12 metabolic clusters using the approach presented in Inouye et al. (2012) (dynamic tree cutting algorithm) to estimates of the expected correlation of genetic effects from a Finnish population study obtained using a bivariate linear mixed model (biLMM) algorithm (Pirinen, 2015). Then, I applied the C-alpha MRP test using these estimates of the expected correlation of genetic effects across the 12 metabolic clusters.

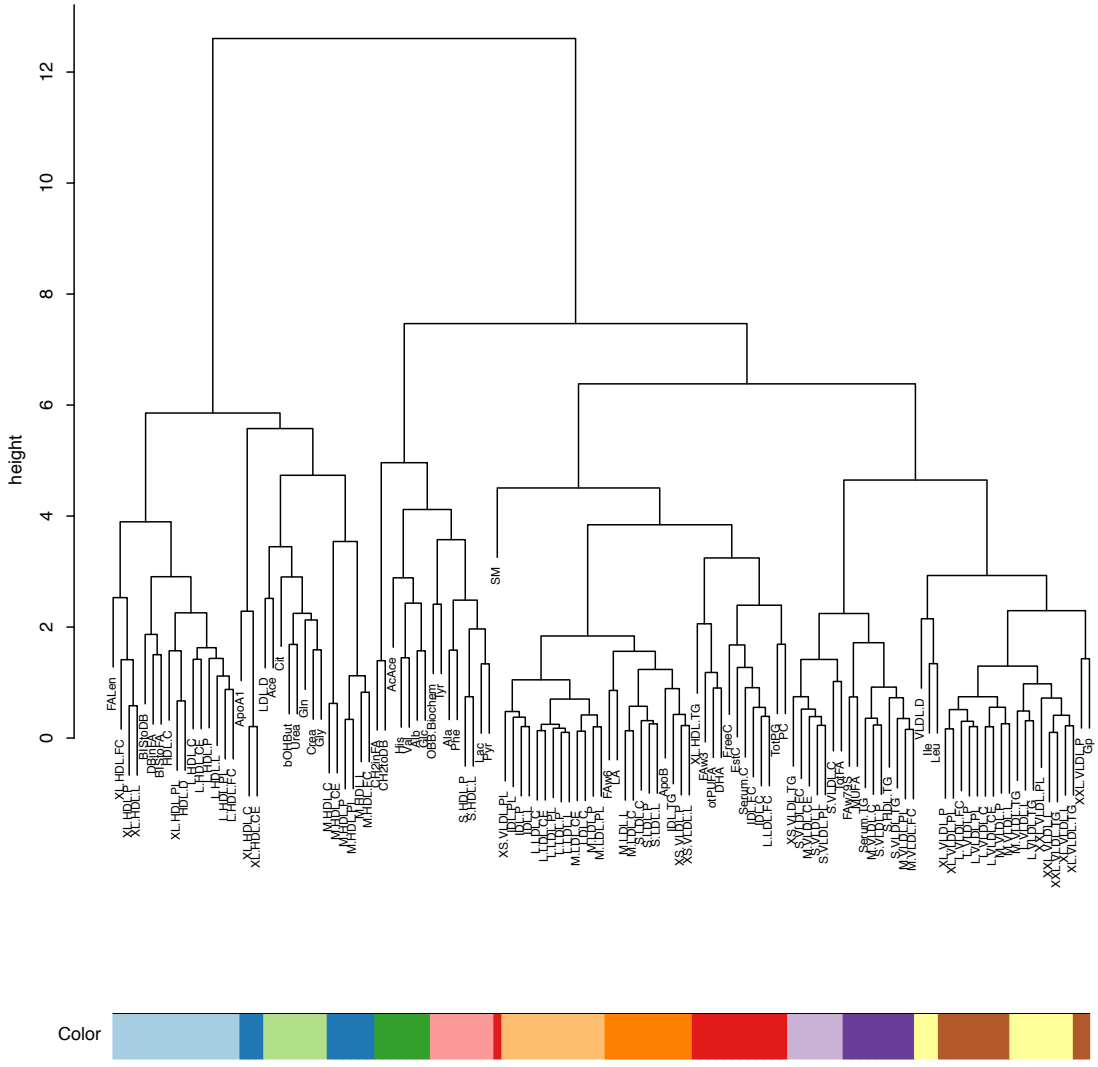


Figure 6.6: Grouping metabolomic traits for multivariate association analyses. Hierarchical clustering of 12 metabolic networks using the approach presented in Inouye et al. (2012).

6.2.3 Results

Individual gene level analysis of PTVs. I assessed asssociation between PTVs and the 123 metabolite traits by employing the three Bayesian models described in Chapter 3 (SEM, GEM, GEM-NMD) using the software `MAMBA` across 2,829 genes with informative PTVs (minor allele count ≥ 1). I found 40 genes at a $\log_{10}(\text{BF}) \geq 2$ and 3 genes at a $\log_{10}(\text{BF}) \geq 3$ (Figure 6.7). The three genes include *AHSA2* (activator of heat shock 90kDa protein ATPase homolog 2; number of PTVs = 1, composite PTV allelic count = 2) with association to the amino acid Isoleucine, *PARP3* (Poly [ADP-ribose] polymerase 3; number of PTVs = 2, composite PTV allelic count = 9) with association to total phosphoglyceride - the main component of biological membranes, and *COQ10A* (Coenzyme Q10 Homolog A; number of PTVs = 1, composite PTV allelic count = 466) with association to the amino acid Glutamine.

Given the challenges in analyzing > 100 traits simultaneously, including reduction of power and computational complexity, I decided to apply the same methodology described in Inouye et al. (2012) for grouping traits for multivariate association analysis of common variants. It is unclear whether this approach is optimal and further research into the analysis of large (> 100) phenotypes is necessary. The main difference in the approach to grouping of the traits from that presented in Inouye et al. (2012) is that I focused on the genetic component of the pairwise correlations expecting that the gain in power to detect association will come from the expected correlation of genetic effects (as that is what the C-alpha MRP test attempts to exploit). I applied the distance function in *R*, *dist*, to the genetic correlation matrix to obtain a distance matrix using the Euclidean distance metric (Mardia and Kent, 1979; Borg and Groenen, 1997). I performed hierarchical clustering analysis on the distance matrix and applied the dynamic tree cutting algorithm used in Inouye et al. (2012). I proceeded to downstream analysis with `deep Split = 4`, `minClusterSize = 5`, corresponding to 12 metabolic networks (Figure 6.6 and Table 6.4).

I applied the C-alpha MRP approach and tested for association across the 2,829 genes to 12 clusters of metabolite traits and integrated the estimates of the expected correlation of genetic effects. I identified two genes, *LPL* and *CT62*, at a significance level $P < 0.0001$ ($P = 1.31 \times 10^{-5}$ and $P = 7.8 \times 10^{-5}$, respectively; Figure 6.8). The *LPL* (lipoprotein lipase) association is driven by a single variant, the common nonsense variant p.S474X, which has been well studied (Rip et al., 2006). I examined the *CT62* (cancer/testis antigen 62; number of PTVs = 2, composite PTV allelic count = 5) association by applying the MCMC estimation algorithm described in Chapter 3. I find that the association is mainly driven by a single

nonsense variant, p.C69X, with effects mainly concentrated across five phenotypes: "L.VLDL.TG_c" (triglycerides in large VLDL, 2.14 [0.91 - 3.35] s.d. above the population mean),"M.VLDL.TG_c" (triglyicerides in medium VLDL, 1.47 [0.17 - 2.73] s.d. above the population mean),"VLDL.D_c" (mean diameter for VLDL particles, 1.03 [-0.31 - 2.38] s.d. above the population mean),"Ile_c" (isoleucine amino acid, 1.59 [0.11 - 3.04] s.d. above the population mean), and "Leu_c" (leucine amino acid, 1.35 [-0.25 - 2.95] s.d. above the population mean) (Figure 6.9).

Enrichment of PTV association in genes involved in lipoprotein metabolism, lipid metabolic process, and lipid digestion and transport. Finally, I applied the C-alpha MRP test by integrating estimates of the expected correlation of genetic effects and testing for association between four gene sets selected (gene sets were selected by Johannes Kettunen, personal communications) based on prior curation of protein-coding genes involved in lipoprotein metabolism, lipid metabolic process, and lipid digestion and transport from the Molecular Signature Database v4.0 `http://www.broadinstitute.org/gsea/msigdb/index.jsp` and found nominal ($P < 0.01$) evidence of association across 7 of the 12 tested metabolic clusters with minimum P value $< 1 \times 10^{-5}$ for metabolic cluster 11. Restricting the analysis to rare (minor allele frequency $< 1\%$) PTVs did not yield any significant results (Figure 6.10).

6.2.4 Conclusion

In this study I applied the Bayesian models and the C-alpha MRP test presented in Chapter 3 to assess association between protein truncating variants and 123 metabolite traits measured using NMR spectroscopy in $4,091$ individuals from the Oxford Biobank study. I used PTV content in the Illumina human ExomeArray data and after applying the Bayesian models identified 40 genes at a $\log_{10}(\text{BF}) \geq 2$ across 97 associations, and 3 genes at a $\log_{10}(\text{BF}) \geq 3$. I generated twelve clusters of metabolite traits using estimates of the expected correlation of genetic effects and applied the C-alpha MRP test across the $2,829$ genes to the 12 clusters. I identified two genes, *LPL* and *CT62*, at a significance level $P < 0.0001$. I examined the *CT62* (cancer/testis antigen 62; number of PTVs = 2, composite PTV allelic count = 5) association and find that the association is mainly driven by a single nonsense variant, p.C69X, with effects mainly concentrated across five metabolite traits. Furthermore, I applied the C-alpha MRP test to a gene set chosen based on its biological relationship to the metabolites that are measured by the NMR spectroscopy platform. I found evidence

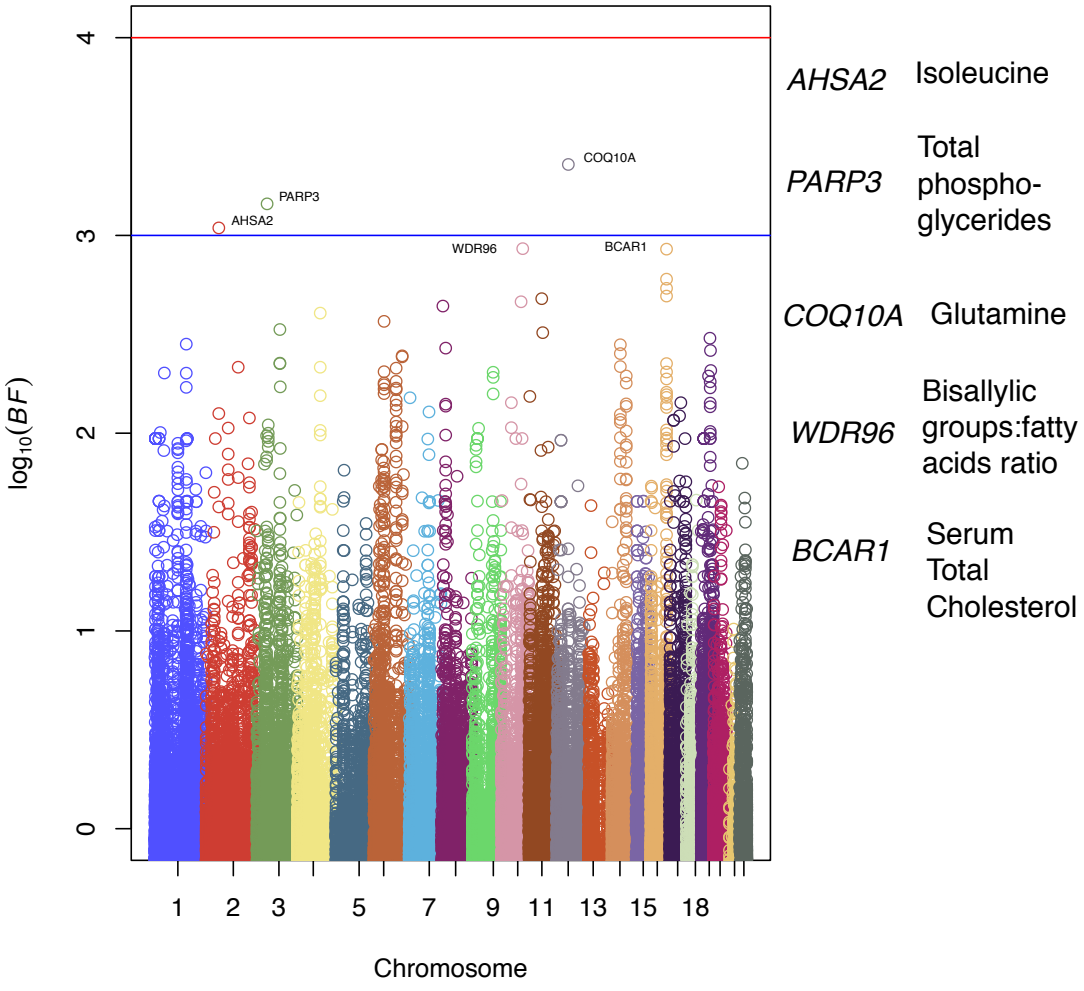


Figure 6.7: Manhattan plot of protein truncating variant association to blood metabolite measurements in the Oxford Biobank study using the Bayesian SEMGEM models. I find 40 genes at a $\log_{10}(\text{BF}) \geq 2$ and 3 genes at a $\log_{10}(\text{BF}) \geq 3$. The three genes include *AHSA2* (activator of heat shock 90kDa protein ATPase homolog 2) with association to the amino acid Isoleucine, *PARP3* (Poly [ADP-ribose] polymerase 3) with association to total phosphoglyceride, and *COQ10A* (Coenzyme Q10 Homolog A) with association to the amino acid Glutamine.

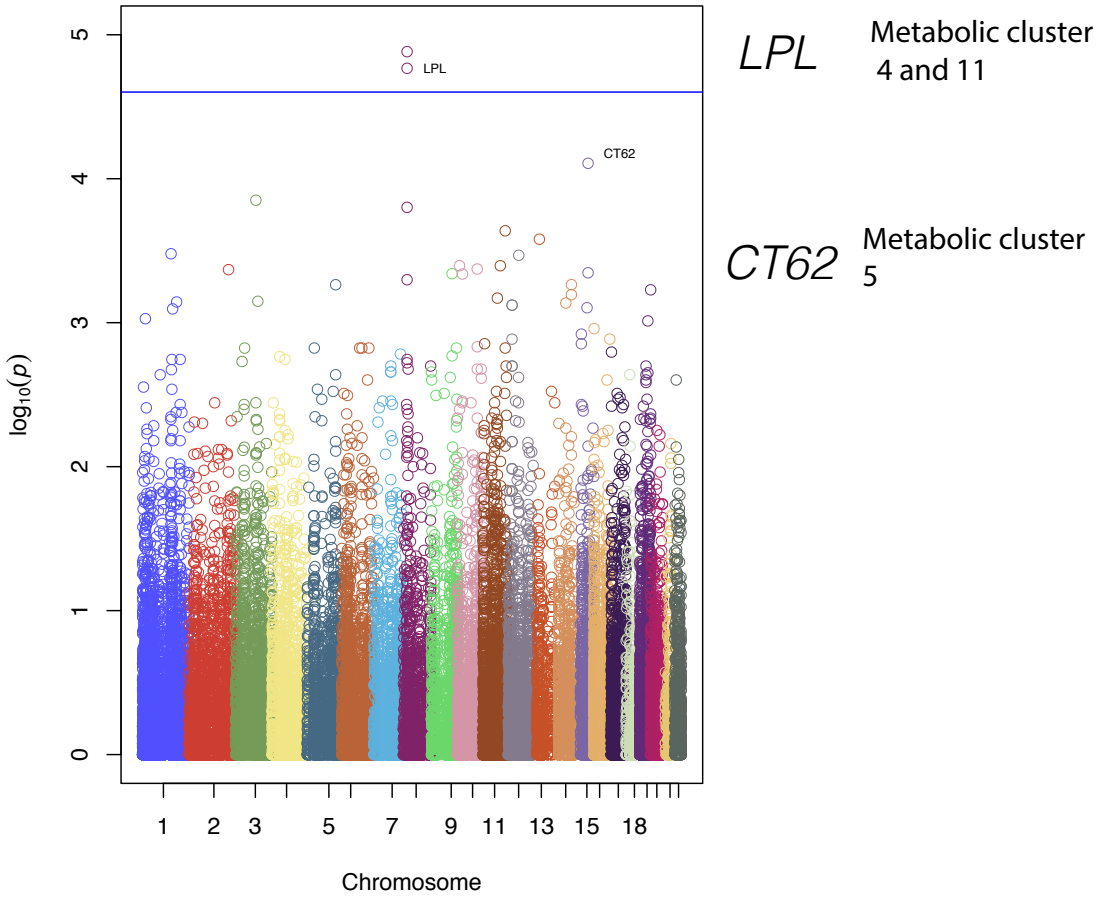


Figure 6.8: Manhattan plot of protein truncating variant association to blood metabolite measurements in the Oxford Biobank study using the C-alpha MRP test. All P values across the 12 metabolic clusters are shown. The single PTV in *LPL*, p.S447X, is associated with metabolic cluster 11 (P value $= 1.31 \times 10^{-5}$) and 4 (P value $= 1.71 \times 10^{-5}$). PTVs in *CT62* are found to be associated with metabolic cluster 5 (P value $= 7.81 \times 10^{-5}$).

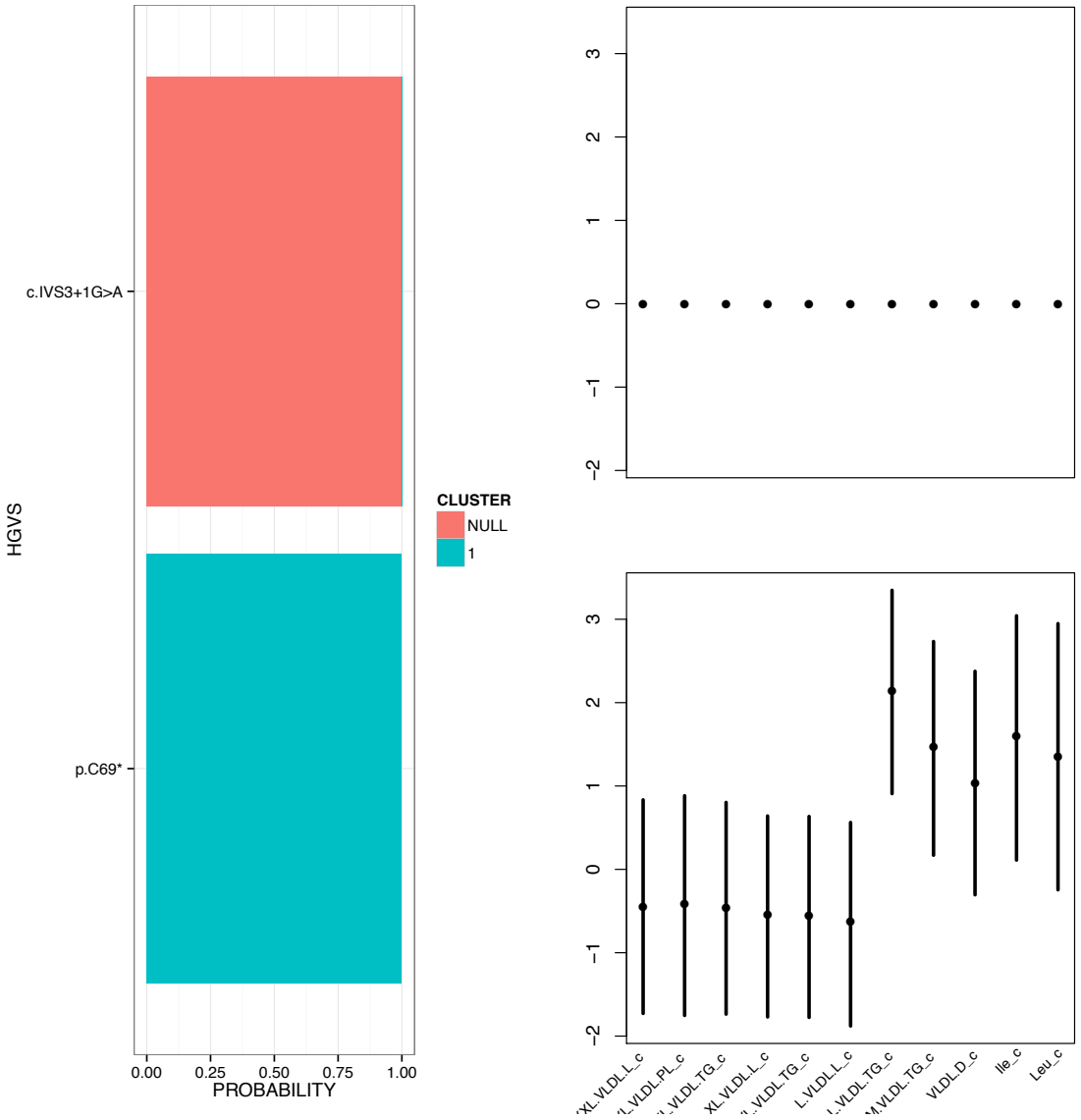


Figure 6.9: MCMC estimation algorithm applied to PTV data in exome array for *CT62* and metabolic cluster 5. PTV, c.IVS3+1G>A, a predicted splice disrupting variant, belongs to cluster 1 with probability equal to 1. Cluster 1 represents the null cluster, i.e. no effect across all traits. PTV, p.C69X, a nonsense variant, belongs to cluster 2 with probability equal to 1. Cluster 2 represents an alternative cluster with effect sizes denoted by the plot on the bottom right of this figure. It is clear that the effects are concentrated on 5 traits and careful examination shows that the effect is mainly on triglycerides in large VLDL ("L.VLDL.TG_c", 2.14 [0.91 - 3.35] s.d. above the population mean).

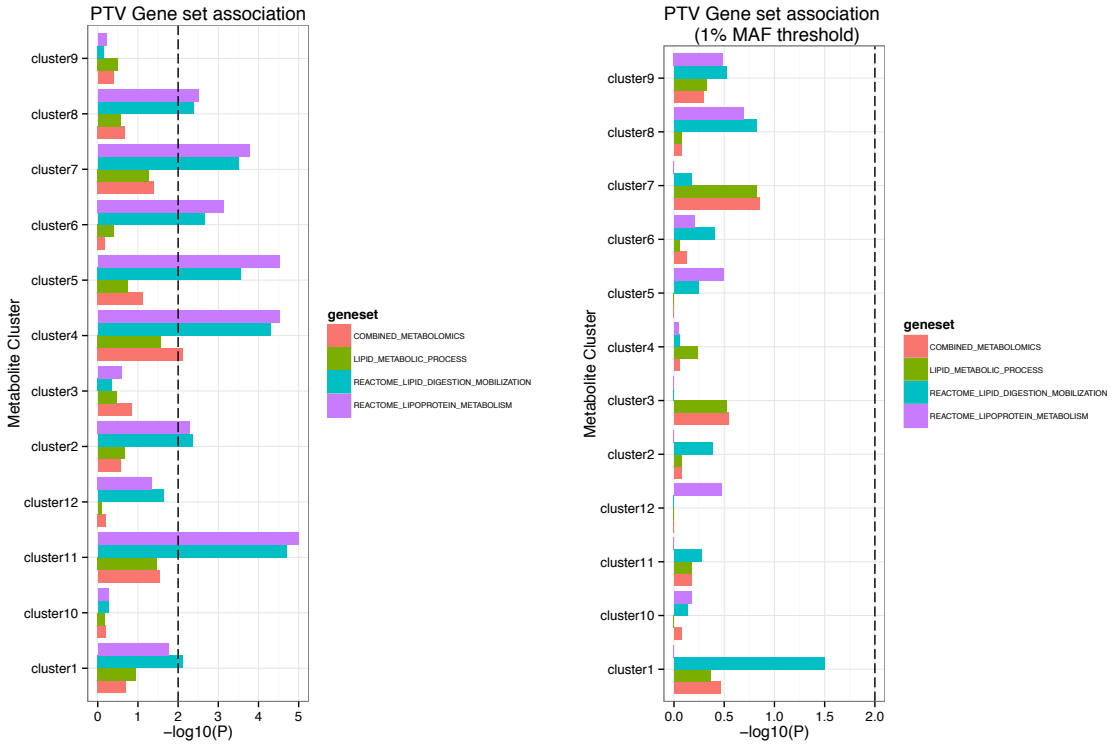


Figure 6.10: PTV association results for biological gene sets selected to be relevant to NMR metabolite levels. The analysis was designed specifically to determine whether association between PTVs across the three gene sets chosen from the Molecular Signature database (and a fourth combining all the genes into a single gene set) and the 12 metabolic clusters exist. Analyses was restricted to all PTVs and PTVs with a minor allele frequency $< 1\%$. Using this approach, evidence of association is found for the REACTOME_LIPID_DIGESTION_MOBILIZATION and the REACTOME_LIPOPROTEIN_METABOLISM gene sets across 7 of the 12 tested metabolic clusters with a P value < 0.01. No evidence of association was detected when the analysis was restricted to rare (MAF $< 1\%$) PTVs.

of association across the gene set suggesting that a signature possibly of many variants or genes may exist. A major limitation of the Illumina Human ExomeArray is that there is limited PTV content. Only 10% of protein-coding genes have at least one informative PTV making gene discovery and assessment of the contribution of PTVs to either disease predisposition or quantitative trait variation difficult. I hope the approach taken here and that the initial findings will be useful for future exome array, and more importantly, exome and genome sequencing studies of metabolite measurements.

6.3 Discussion

In this chapter I have presented results from the application of the SEMGEM method presented in Section 3.2 and the C-alpha MRP test presented in Section 3.3 to the analysis of protein truncating variants in two studies of quantitative traits.

I found new PTV signals in genes with prior established roles in lipid biology and metabolism. However, the new variant signals show variants with strong effects. A promising result is the discovered impact of APOC3 PTVs on triglyceride levels even in the setting of type 2 diabetes. This result stresses the importance of taking into consideration the impact of PTVs under different conditions and evaluating their impact across multiple phenotypes.

A limiting factor in the use of the exome array is that it is enriched for more common PTVs and these may be less likely to have functional consequences, which can limit the ability to detect signal. It is not clear whether the new gene signals I found in the exome sequencing study presented in Section 6.1 or in the metabolomics exome array study presented in Section 6.2 are real. The next steps will include to replicate the signals, and also to participate in larger and better powered gene discovery efforts.

Chapter 7

Improving annotation of protein truncating variants by studying their functional consequences

7.1 Background

Protein truncating variants (PTVs) are typically expected to have large effects on gene function. These variants are highly enriched for severe disease-causing mutations (Holbrook et al., 2004; Stenson et al., 2014), but in other instances may actually be protective against disease (Cohen et al., 2006, 2005; Nejentsev et al., 2009; Rivas et al., 2011; Sullivan et al., 2012; Plenge et al., 2013; Flannick et al., 2014; Kaiser, 2014). However, PTVs are abundant in the genomes of healthy population controls (MacArthur et al., 2012), demonstrating that they do not necessarily have drastic phenotypic consequences. In addition, while PTVs are often described as loss-of-function variants, in most cases their precise impact on gene function has not been molecularly characterized, and in other cases such variants have been shown to have gain-of-function effects like the *PPM1D* example presented in Section 5.1 (for more examples see Holbrook et al. (2004)). Predicting the clinical impact and improving annotation of PTVs will require detailed characterization of their downstream functional consequences.

One potential source of information about the functional impact of genetic variation is the transcriptome (the collection of expressed RNA transcripts), which informs interpretation of the effects of variants on gene expression levels and transcript structure. Several studies have highlighted the value of transcriptome data for mapping regulatory variants (Montgomery et al., 2010; Lappalainen et al., 2013; Ardlie et al., 2015), and targeted analysis of mode of inheritance (Rip et al., 2006; Linde et al.,

2007b; Ruark et al., 2012) and the relevant tissue-type (Bateman et al., 2003) of disease-causing mutations. However, to date there has been no systematic characterization of the impact of PTVs on the human transcriptome.

In this chapter I present results from a focused analysis of the functional consequences of PTVs by using a multi-tissue RNA sequencing data set from the Genotype Tissue Expression (GTEx) project and a population based RNA sequencing data set from the Geuvadis project. The results of this study are published in *Science* in Rivas et al. (2015).

A comprehensive catalog of predicted PTVs and their transcriptomic impact in human population samples of two studies was generated: 462 individuals with DNA and RNA sequencing from lymphoblastoid cell lines in the Geuvadis study (Lappalainen et al., 2013; AC't Hoen et al., 2013), and 173 individuals with combined DNA and mRNA sequencing from a total of 1,634 samples from multiple tissues in the Genotype-Tissue Expression study (Lonsdale et al., 2013). Methodology presented in Chapter 4 and further developed in Pirinen et al. (2015) was applied to the data sets. The integrated data sets were used to characterize the degradation of premature stop-containing transcripts via nonsense-mediated decay, the impact of large deletions on gene expression, and the splicing effects of variants proximal to exon-intron junctions (Figure 7.1). Finally, resources for exploring the transcriptome impact of PTVs in these reference data sets were provided to the community.

As in Chapter 5 and 6 I present an overview of the study, overview of my role in the study, and the results from my analysis of the data sets.

7.2 Overview of the study

This study was conducted with members of the GTEx and the Geuvadis projects. The GTEx project is a program led by the National Institutes of Health (NIH) Common Fund in the United States with the primary aim to establish a resource database and associated tissue bank in which to study the relationship between genetic variation and gene expression in human tissues (Lonsdale et al., 2013). The group of investigators involved in the project are part of the GTEx consortium. The Geuvadis project is a European project whose aim is to bring together the knowledge and resources on medical genome sequencing at a European level and allow researchers to develop and test new hypotheses on the genetic basis of disease (`http://www.geuvadis.org/web/geuvadis/RNAseq-project`). The RNA sequencing arm of the project was led by Manolis Dermitzakis (University of Geneva) and

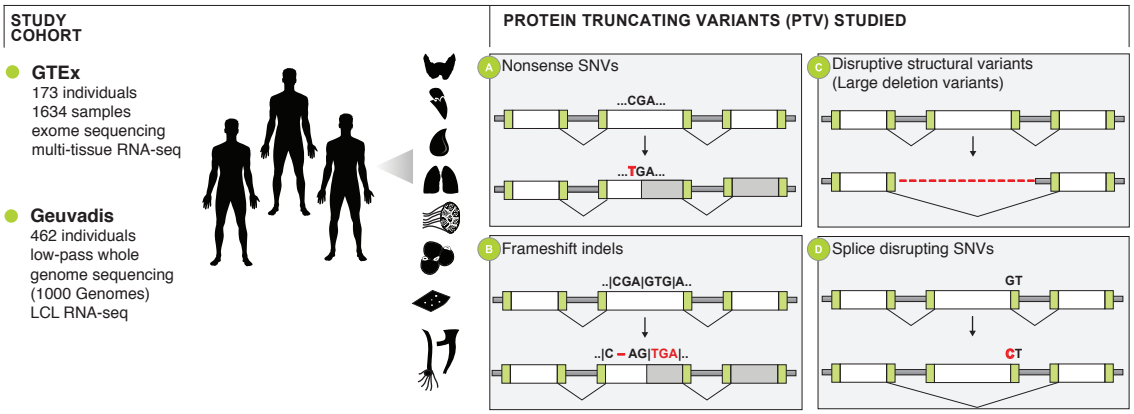


Figure 7.1: An integrated DNA and RNA sequencing data set was prepared by combining the pilot phase of the GTEx project of 173 individuals with up to 30 tissues per individual (total = 1634 samples) and the Geuvadis project of lymphoblastoid cell line (LCL) DNA and RNA sequencing in 462 individuals. The nine main tissues in the GTEx dataset are thyroid, heart ventricle, whole blood, lung, skeletal muscle, subcutaneous adipose, sun exposed skin, and tibial nerve and artery. From these data, I analyzed the effect of predicted protein-truncating genetic variants on the human transcriptome, including: a) nonsense SNVs; b) frameshift indels; c) large deletion CNVs; and d) splice-disrupting SNVs.

Tuuli Lappalainen (a postdoctoral fellow at the University of Geneva at the start of the project). I was an analyst in both the GTEx and the Geuvadis consortia whose primary responsibility was to deliver on the analysis of PTVs working with my supervisors and others including Tuuli Lappalainen (University of Geneva/Columbia/New York Genome Center) and Daniel MacArthur (MGH/Broad). An overview of my role is outlined in Section 7.3.

Genotype data in Geuvadis. Genotype data in the Geuvadis project was obtained from the Phase 1 release of the 1000 Genomes project (1KG) for 462 individuals (genotypes for 41 individuals were imputed) available from `http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/genotypes/` (The 1000 Genomes Consortium, 2012; Lappalainen et al., 2013). Organization of the CNV genotype data set and the CNV validation data set for the 462 individuals in the Geuvadis study was conducted by Donald Conrad (Washington University in St. Louis). The CNV calls for GEUVADIS samples were extracted from the official Phase 1 data release of the 1000 Genomes project. A deletion was considered to be validated if it passed at least one of the three validation experiments (described in detail in the supplementary material of the Phase 1 paper of the 1000 Genomes project (The 1000 Genomes Consortium, 2012)) and it did not fail PCR or CGH validation. 1,425 PTV deletion events were

manually curated, which produced a short, conservative list of 145 CNV calls with extremely high confidence for a total of 59 unique PTV deletions. The definition of large deletions was restricted to just those that remove the entire CDS of a gene, due to technical concerns about the interplay between structural variation and mapping accuracy, producing a list of 33 large deletions before manual curation and a list of 25 large deletions following manual curation.

Genotype data in GTEx. The Genotyping Working Group of the GTEx consortium led by Daniel MacArthur and Monkol Lek was responsible for generating an exome sequencing callset. Whole exome sequencing (WES) on blood DNA samples from 180 GTEx pilot phase donors was performed by the GTEx consortium at the Broad Institute's Genomics Platform. The BWA aligner (`http://bio-bwa.sourceforge.net`) was used to map reads to the human genome build 37 (hg19) (Li and Durbin, 2009). GATKs UnifiedGenotyper package was then used for SNV calling jointly across all 180 samples (McKenna et al., 2010; DePristo et al., 2011). GATKs HaplotypeCaller (v2.8) was used for indel calling across all 180 samples (Auwera et al., 2013). Default filters were applied to SNV and indel calls using GATK's Variant Quality Score Recalibration (VQSR) approach, which resulted in a VCF file for downstream analyses (Danecek et al., 2011).

Large deletion variant calls and genotypes were generated for this study. Donald Conrad was responsible for generating a call set that combined the exome sequencing data and the genotype array data set. Menachem Fromer (Mt. Sinai School of Medicine) was responsible for generating the exome sequencing CNV callset. All GTEx pilot samples were run on two Illumina Human microarray platforms: the Illumina Omni 5M and the Illumina Exome array. Following QC, high quality genotypes were obtained for 30 common PTV deletions. The XHMM software was used to detect copy number variation in whole-exome sequencing data from the GTEx project in 180 individuals (Fromer et al., 2012).

Variant Annotation. Variants were identified in several categories expected to represent the majority of true PTVs, with predicted annotation based on Gencode v12 gene models and additional predefined PTV annotation fields as described in MacArthur et al. (2012) (Table 7.1). 13,182 candidate PTVs were identified using whole-genome sequencing data from 421 individuals analyzed as part of Phase 1 of the 1000 Genomes Project (The 1000 Genomes Consortium, 2012) and included in the Geuvadis RNA sequencing project, as well as 4,680 candidate PTVs using exome

PTV Flags	Description
ANC_ALLELE	PTV is the ancestral allele
NON_CAN_EXON	Exon is surrounded by non-canonical splice site (i.e. not AG/GT)
END_TRUNC	PTV removes less than 5% of remaining protein
SINGLE_EXON	Transcript only has one coding exon
SMALL_INTRON	Splice site mutation within intron smaller than 15 bp
NON_CAN_SPLICE	Splice site is non-canonical OR other splice site within same intron is non-canonical
EXON_INTRON_UNDEF	Unable to determine exon/intron boundaries surrounding variant

Table 7.1: PTV annotation flags used in the annotation pipeline.

sequencing data from 173 individuals in the GTEx project, for a combined total of 16,286 unique candidate variants (Table 7.2). Large deletion variant and genotype calls were required to be supported by independent experimental data and manual curation.

RNA sequencing. In the Geuvadis project, RNA sequencing of the Epstein-Barr virus (EBV) growth-transformed lymphoblastoid cell lines was performed in multiple European sequencing facilities on the Illumina HiSeq2000 platform with 75 base pair paired-end sequencing using the Illumina TruSeq library construction protocol. This is a non-strand specific polyA+ selected library. Further details on cell line processing, RNA extraction, RNA sequencing, and raw data processing are described in Lappalainen et al. (2013); AC't Hoen et al. (2013).

In the GTEx project, RNA sequencing of the tissues collected in the pilot phase of the project was performed using the Illumina TruSeq library construction protocol. The sequencing produced 75 base pair paired-end reads. Further detail on the samples, read alignment, post-processing, tissue and sample collection is described in the main GTEx analysis manuscript (Ardlie et al., 2015). RNA sequencing of the GTEx samples was performed by the Broad Institute Genomics platform.

mRNA quantifications. From the aligned RNA-seq read data from both GTEx and Geuvadis, several features of transcripts were quantified based on the GENCODE v12 gene annotation: genes, exons, introns, and splice junctions. In the GTEx project

Variant type	**GTEx**		**Geuvadis**	
	total (HC)	avg.; homozygous (HC)	total (HC)	avg.; homozygous (HC)
nonsense	1345 (1020)	57.09; 11.56 (29.71; 4.75)	5987 (4682)	71.67; 12.57 (36.55; 3.81)
splice	845 (661)	58.64; 17.17 (29.12; 6.64)	6113 (3252)	125.28; 28.11 (29.39; 4.87)
frameshift	2324 (1746)	107.19; 14.74 (68.10; 6.61)	1023 (606)	16.94; 0.15 (29.73; 0.31)
large deletions	70 (8)	NA	59 (25)	NA
total	4584 (3435)	NA	13182 (8565)	NA

Table 7.2: Number of PTVs discovered in the GTEx exome sequencing data set and in the Geuvadis/1KG Phase 1 whole-genome data set. Total numbers of PTVs and average number of PTVs per individual; average number of homozygous PTVs per individual are shown for each PTV class and data set (in parenthesis data shown for variants with HC flags only, i.e. those that do not have any of the filters described in Table 7.1). For the Geuvadis data set I report the numbers for the 421 individuals with genome sequence data. For large deletions I only report the total number of PTV deletion with manual curation in the study.

these measurements were generated for the nine tissues with the highest number of subjects:

1. Adipose Subcutaneous (ADPSBQ, $n = 94$);
2. Artery Tibial (ARTTBL, $n = 112$);
3. Heart Left Ventricle (HRTLV, $n = 83$);
4. Lung (LUNG, $n = 119$);
5. Muscle Skeletal (MSCLSK, $n = 138$);
6. Nerve Tibial (NERVET, $n = 88$);
7. Skin Sun Exposed (SKINS, $n = 96$);
8. Thyroid (THYROID, $n = 105$); and

9. Blood (WHLBLD, $n = 155$).

Gene and exon quantifications are described in Ardlie et al. (2015); Lappalainen et al. (2013). The two studies used very similar methods for generating exon and gene quantifications. The data files of these quantifications were made available in two separate web links

- Geuvadis: `http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/`, and
- GTEx: `http://www.broadinstitute.org/gtex/datasets`.

The mRNA quantification data generation and processing was performed using FLUX[1] and led by Michael Sammeth (a postdoctoral fellow working under the supervision of Roderic Guigo at the Centre for Genomic Regulation [CRG] in Barcelona, Spain).

mRNA quantifications were normalized using methods presented in Lappalainen et al. (2013) for both the GTEx and Geuvadis data sets so that in the analysis data from the two projects could be combined. All read count quantifications were first normalized by sequencing depth by dividing them by the total number of mapped reads per sample. The per sample exon, intron, and junction normalized quantifications were scaled by multiplying by the median number of mapped reads in the project (85M in GTEx). The data was split to tissues and all subsequent analysis was done separately for each of the tissues with > 80 samples. Elements (genes/exons/junctions etc) were filtered to keep only those with > 0 expression in $> 50\%$ of the individuals, except for introns where this filtering step was omitted (since the analysis focused on searching for rare *inclusions* of introns). Technical variation was removed using PEER, a software that includes a collection of Bayesian approaches to infer hidden determinants and their effects from gene expression profiles using factor analysis (Stegle et al., 2012; Lappalainen et al., 2013). PEER was applied to 20,000 quantification units (e.g. exons or genes) with 10 factors ($K = 10$). Covariates from this analysis were regressed out from all the quantifications, and the mean was added to the residuals. These quantifications across all individuals were then transformed to a normal distribution with mean 0 and standard deviation equal to 1 by rank-based inverse normal transformations (Qiu et al., 2013) for the splice disruption analysis.

[1] `http://sammeth.net/confluence/display/FLUX/Home`

Transcriptional properties of PTV-containing transcripts. The data set of median gene expression values across individuals per tissue was prepared by Konrad Karczewski (a postdoctoral fellow at the Broad Institute). The data set of median splice junction values across individuals per tissue was prepared by Michael Sammeth (CRG).

Allele specific expression (ASE) data set. Allele-specific expression analysis was based on allelic counts in the RNA-seq reads of heterozygous sites genotyped from DNA, within each individual. For the GTEx data set only heterozygous genotypes were included with genotype quality ≥ 60 (Phred-scale) for SNVs and ≥ 95 for indels. For the Geuvadis data set the maximum genotype quality in the VCF is 50. To ensure that we used confident heterozygous genotypes we focused on genotypes with genotype quality equal to 50. Additionally, sites that are susceptible to allelic mapping bias were excluded: 1) sites with 50bp mappability < 1 based on the UCSC mapability track, implying that the 50bp flanking region of the site is non-unique in the genome, and 2) simulated RNA-seq reads overlapping the site show $> 5\%$ difference in the mapping of reads that carry the reference or alternate allele (Panousis et al., 2014). Uniquely mapping reads (mapping quality > 150), NM ≥ 6, and sites with base quality > 10 were the only sites used. Furthermore, we only used sites in our ASE analysis with at least 8 RNA-seq reads in the heterozygote individual. A customized localized personal mapping pipeline was developed to account for biases that may be inherent in mapping indels for ASE analysis. The alignment was performed to local reference sequences that have been modified to contain both the reference and alternative alleles.

Indel ASE data set. In this study, the Genotyping Working Group of the GTEx project applied a new haplotype-based genotype calling algorithm implemented within GATK v2.8 (Auwera et al., 2013) to obtain substantially improved variant and genotype calls of indels, and Tuuli Lappalainen developed a new ASE approach designed to accurately align and count RNA-seq reads over both reference and non-reference indel alleles for these variants.

mmPCR experimental validation. In this study, the ASE validation experiment for the consortia was led by Stephen Montgomery and involved lab personnel and additional analysts including Emily Tsang, Kimberly Kukurba, Kevin Smith, and Rui Zhang, all from Stanford University. Allelic counts from RNA-seq read data

were obtained for a total of 1,814 PTVs (with $\geq$ 8 RNA-seq reads), and the allelic ratios obtained from RNA-seq data were validated using mmPCR-seq data (Zhang et al., 2009) from 682 variants in 121 tissue samples from 9 GTEx individuals showing general fidelity for nonsense SNVs (Pearson correlation 0.79, $P = 7.3 \times 10^{-14}$).

Analysis of common large gene deletion polymorphisms. The analysis of common large gene deletion polymorphisms was conducted by Donald Conrad (Washington University in St. Louis). The aim of this analysis was to compare and contrast with the results obtained from the analysis of rare large gene deletions.

Analysis of common variants proximal to splice junctions. The analysis of common variants proximal to splice junctions was conducted by Pedro Ferreira (University of Geneva). The aim of this analysis was to compare and contrast with the results obtained from the analysis of rare variants proximal to splice junctions.

7.3 Overview of my role

I was an analyst in both the GTEx and the Geuvadis consortia. My primary responsibility was to deliver on the analysis of PTVs working with my supervisors and others including Tuuli Lappalainen (University of Geneva/Columbia/New York Genome Center) and Daniel MacArthur (MGH/Broad). For the work presented in this Chapter I co-led the design of the PTV analysis with Tuuli Lappalainen and Daniel MacArthur, I co-led the development of methodology for analysis of the data with Matti Pirinen (who kindly spent time with me once a week discussing approaches for analyzing the data and providing continuous feedback), and I analyzed the data set presented in the next section, unless otherwise noted.

7.4 Results

7.4.1 Transcriptional properties of PTV-containing transcripts

Each human carries at least 100 PTVs, but the majority of these variants in an individual appear to be common in the population and reside in genes that are likely tolerant of dramatic disruption (MacArthur et al., 2012). One possible mechanism underlying such tolerance is that PTVs often do not affect all transcripts of a gene (MacArthur et al., 2012), and tissue-specific expression of the different transcripts may affect the penetrance of the PTV (Ardlie et al., 2015).

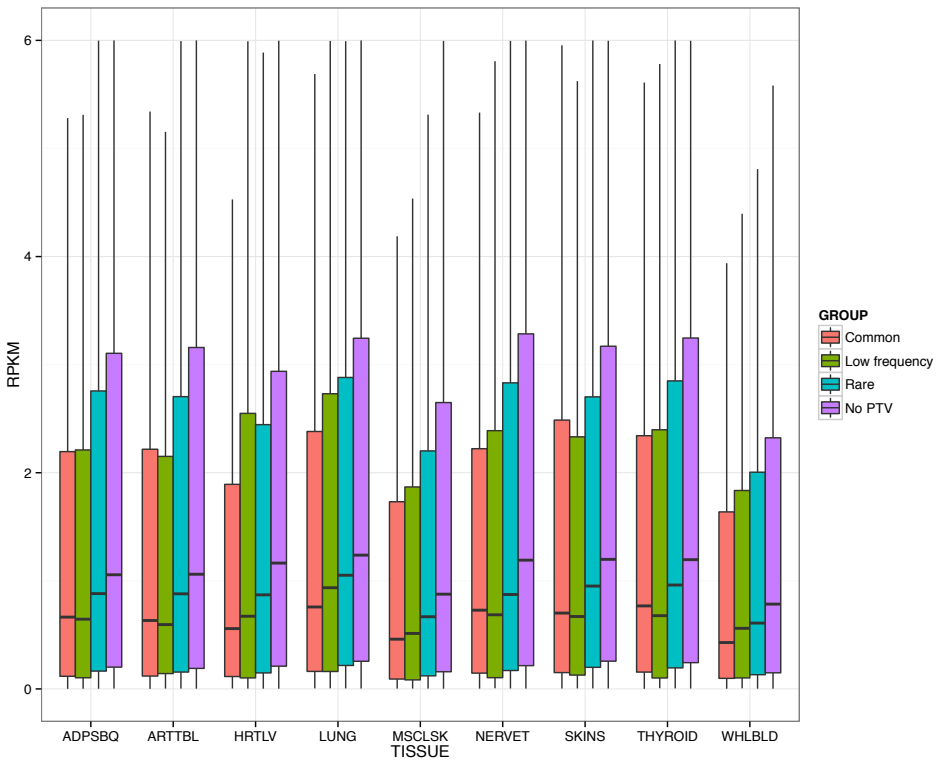


Figure 7.2: Transcriptional properties of PTV containing transcripts: tissue-wide expression profile for PTV containing genes. Comparison of the distribution of median gene expression values for PTV containing genes across tissues and categories. I found evidence that genes containing common PTVs are consistently lower expressed across tissues compared to genes that do not contain PTVs in these data sets (P value per tissue $< 2 \times 10^{-16}$, MWW test) and the same is observed also for genes containing low-frequency or rare PTVs (P value per tissue 6.1×10^{-7} to 2.4×10^{-11} and $< 2 \times 10^{-16}$, MWW test).

In order to characterize the transcriptional properties of PTV-containing transcripts I partitioned the PTVs into categories by minor allele frequency (MAF): a) common (MAF $\geq .05$; 1,607 genes), b) low frequency ($.01 <$ MAF $< .05$; 864 genes), and c) rare (MAF $\leq .01$; 5,096 genes). I compared these sets to the set of protein-coding genes where no PTVs were observed (13,372 genes).

Common and rare variants comparisons: gene expression. I compared the distribution of median gene expression values for PTV containing genes across tissues using the Mann-Whitney-Wilcoxon (MWW) one-sided (`wilcox.test` in R, `alternative ="less"`) test (Figure 7.2). I used the same grouping of genes to assess how the different groups of genes are expressed across all tissues. I compared the proportion of genes with median gene expression value above a Reads Per Kilobase per Mil-

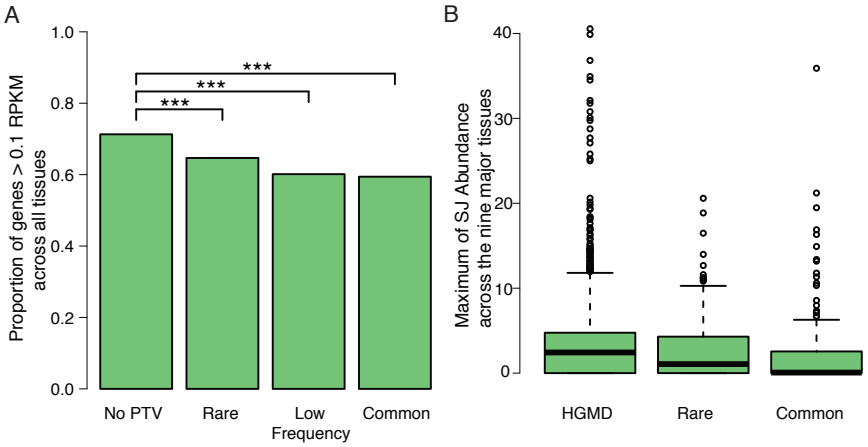


Figure 7.3: Transcriptional properties of PTV containing transcripts: ubiquitous expression and splice junction usage. a) Proportion of ubiquitously expressed genes, defined as median gene expression level (calculated across individuals) > 0.1 RPKM for all nine tissues in GTEx, in protein-coding genes where no PTVs, and genes with PTVs of different frequency: common (MAF $\geq$ .05), low frequency (.01 < MAF < .05), and c) rare (MAF $\leq$ .01). All three categories of PTV-containing genes were less likely to be ubiquitously expressed compared to the set of protein-coding genes with no PTVs ($P < 1.9 \times 10^{-10}$, two-proportion z-test). b) Maximum of the median splice junction abundance across the nine tissues in junctions containing splice-disrupting variants that are common, rare, and present in the Human Genome Mutation Database (HGMD) across the nine tissues. The junction abundance is measured from individuals not carrying the variants. Junctions with common splice-disrupting variants are less often used compared to those containing rare splice-disrupting variants (maximum SJ comparison $P = 0.0015$, MWW test) and reported disease-causing variants in HGMD (SJ comparison $P < 3.3 \times 10^{-12}$, MWW test).

lion mapped reads (RPKM; a common method of quantifying gene expression from RNA sequencing data (Mortazavi et al., 2008)) cutoff for "rare", "low frequency", "common", and "no PTV" protein-coding genes (Figure 7.3). All three categories of PTV-containing genes were less likely to be ubiquitously expressed compared to the set of protein-coding genes with no PTVs ($P < 1.9 \times 10^{-10}$, two-proportion z-test).

Common and rare variants comparisons: splice junction usage. Genes and splice junctions that contain common PTVs or splice variants were expressed at lower levels and in a more tissue-specific manner compared to those containing rare splice-disrupting variants (maximum splice junction value [SJ] comparison $P = 0.0015$, MWW test) and reported disease-causing variants in HGMD (SJ comparison $P < 3.3 \times 10^{-12}$, MWW test) (Figures 7.3).

This confirmed previous work suggesting that PTVs reaching high frequency in the population are able to do so because the affected genes have less critical organismal

functions (MacArthur et al., 2012), and for the first time showing that the same observation holds true at the level of commonly disrupted splice junctions.

7.4.2 Insights into nonsense-mediated decay

Background and methods used for ASE analysis. Variants that introduce premature stop codons may trigger nonsense-mediated decay (NMD), as noted earlier in Section 4.3 this is a cellular mechanism that degrades transcripts with premature stop-codons to prevent their translation into truncated proteins, and thus often protects the individual from their detrimental effects (Hentze and Kulozik, 1999). However, the presence or absence of NMD can have major clinical consequences: PTVs that trigger NMD in dosage-sensitive genes may cause disease via haploinsufficiency, whereas those escaping NMD may create a truncated protein that leads to dominant negative or gain-of-function effects (Holbrook et al., 2004). Thus, accurate inference of NMD is critical for variant interpretation, but its prediction from sequence data alone remains challenging: in previous studies (MacArthur et al., 2012; Lappalainen et al., 2013) fewer than 50% of variants predicted to trigger NMD showed transcriptional evidence of RNA degradation.

Allele-specific expression (ASE) analysis from RNA sequencing data is a powerful approach for detecting NMD in individuals heterozygous for PTVs (MacArthur et al., 2012; Lappalainen et al., 2013; Kukurba et al., 2014). This approach allows direct comparison of transcript levels between the PTV and the non-PTV allele within the same individual, with degradation of PTV containing transcripts manifesting as lower levels of that allele. I also assessed the transcriptome impact of frameshift indels, which cause as many as 50% of all premature stop codons but have been neglected by previous RNA-seq studies of NMD (MacArthur et al., 2012; Lappalainen et al., 2013) due to the substantial technical challenges raised by mapping bias (Montgomery et al., 2013; Skelly et al., 2011; Stevenson et al., 2013). This yielded the first genome-wide estimates of ASE over indel sites.

I analyzed these data of allelic counts from RNA-seq by applying statistical methods described in Pirinen et al. (2015) to all nonsense variants with ASE data (minimum read depth = 8). In this Chapter, when I apply the model to the GTEx data, I focus on the Hierarchical Grouped Tissue model (GTM*) described in Section 4.3, that allows many variants and tissues to be analyzed simultaneously. One of the requirements of the current GTM* implementation is that it requires each variant included in the analysis to have at least two tissues of data available. On the other

hand, when I apply it to single tissue data (Geuvadis), I estimated grouped probabilities with the `gtm` implementation in the software `MAMBA` that is also described in Section 4.3. Briefly, these models allowed the following questions to be addressed:

1. For any particular variant, does the variant have allelic-imbalance across any of the tissues available?

2. For any particular variant, which tissues (when multiple tissues are available) show ASE effects?

3. Across all variants, what proportion of variants show ASE effects in all tissues, only in some tissues, or in no tissues?

For each individual and tissue pair I obtained a posterior probability that the variant belongs to the no ASE group ($\mathcal{N}$), moderate ASE group ($\mathcal{M}$), and the strong ASE group ($\mathcal{S}$). When I analyzed the variants simultaneously, for each variant, I obtained posterior probability estimates that the variant belongs to one of the five states defined earlier in Section 4.3.5: (N=NOASE, M=MODASE, S=SNGASE, H0=HET0 and H1=HET1), where HET0 is the heterogeneous state with at least one tissue showing no ASE, and HET1 is the heterogeneous state with all tissues showing some ASE (some moderate, some strong). In addition, using the Hierarchical model (GTM*) I obtained estimates of the proportion of variants in each of the five states. I ran the ASE models with `nburn=30` and `niter=100`, which showed stable estimates across five MCMC chains (Figure 7.4). The implication of the results will be discussed later in this subsection.

Common and rare variants comparisons: patterns of ASE effects. To compare the allelic expression patterns for common (MAF $\geq .05$) and rare (MAF $\leq .01$) PTVs I used estimates of the proportion of variants reported using the ASE module in the software `MAMBA`. I combined the estimates from the multi-tissue model in GTEx as MODASE (moderate effects across all tissues), SNGASE (strong ASE across all tissues), HET0 (mixture of NOASE and/or MODASE, SNGASE) and HET1 (mixture of MODASE and SNGASE) to compare the two categories of variants (common and rare). In addition, I calculated the 2.5% and 97.5% quantiles of the Gibbs samples from the hierarchical model implementation applied to the data. For Geuvadis, a 95% frequentist confidence interval (CI) for the proportion estimates was obtained using the normal approximation interval since the current implementation of the Grouped Tissue Model only supports reporting of quantiles for data with at least two tissues.

I computed a two-sided P value for a two-proportion z-test pooled for $H_0 : p_1 = p_2$ where p_1 is proportion of common variants showing no ASE and p_2 is the corresponding proportion of common variants using the `prop.test` function with default parameters (R Development Core Team, 2005; Newcombe, 1998; Wilson, 1927). I observed a significantly higher proportion of strong or moderate allelic imbalance in rare and singleton nonsense SNVs compared to common nonsense variants (54.3%, 55.4%, and 35.7%, respectively), suggesting that rare PTVs are more likely to trigger NMD (Figure 7.5).

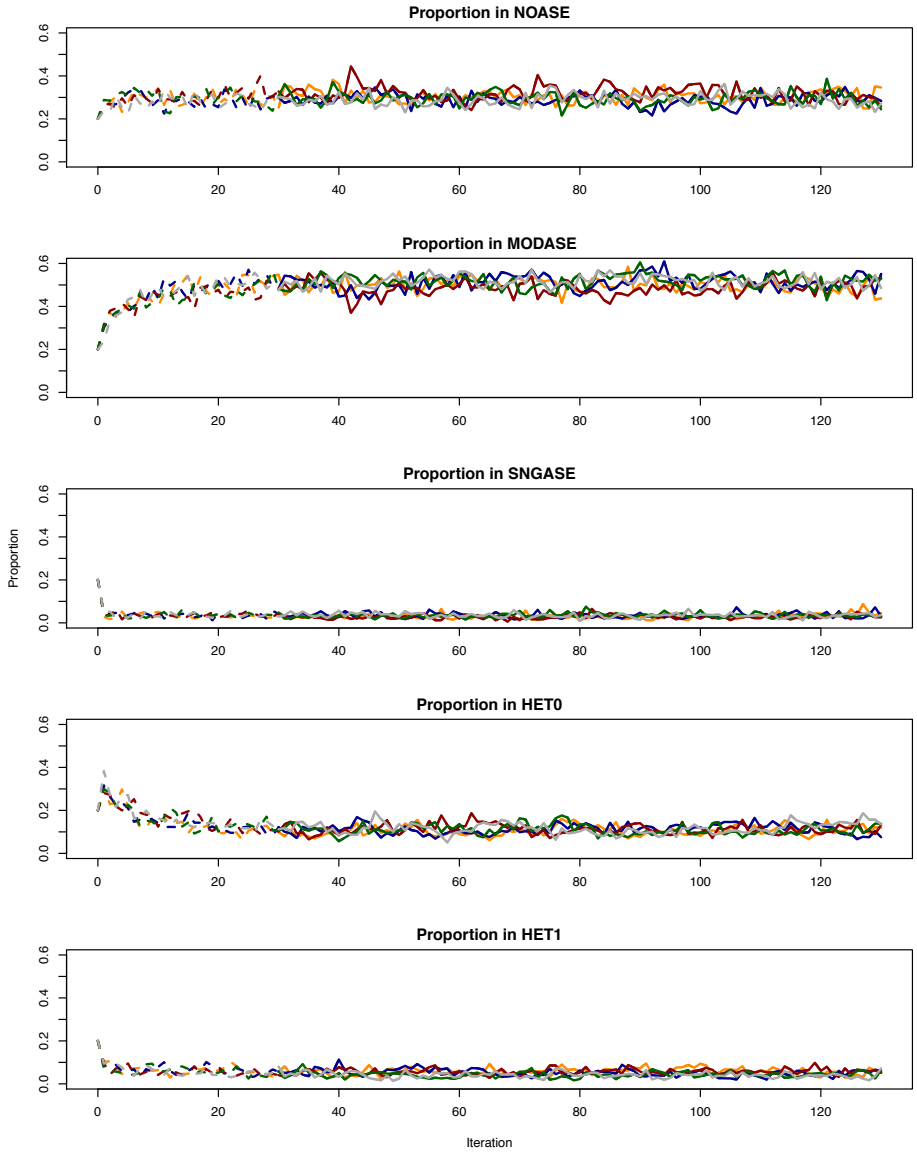


Figure 7.4: When applying MCMC algorithms to a data set it is customary to show the performance of the algorithm across all stages of the experiment (including the burn-in). The GTM* algorithm generates stable proportion estimates for all of the five states studied: NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE across all tissues), HET0 (mixture of NOASE and/or MODASE, SNGASE) and HET1 (mixture of MODASE and SNGASE). For each state I demonstrate the proportion estimate during the burn-in stage of the experiment (30 iterations, dashed lines) and the state of the experiment used to obtain the global estimates (100 iterations, solid lines) reported in this chapter. Five different chains with different random number seeds are highlighted to show that the estimates are stable.

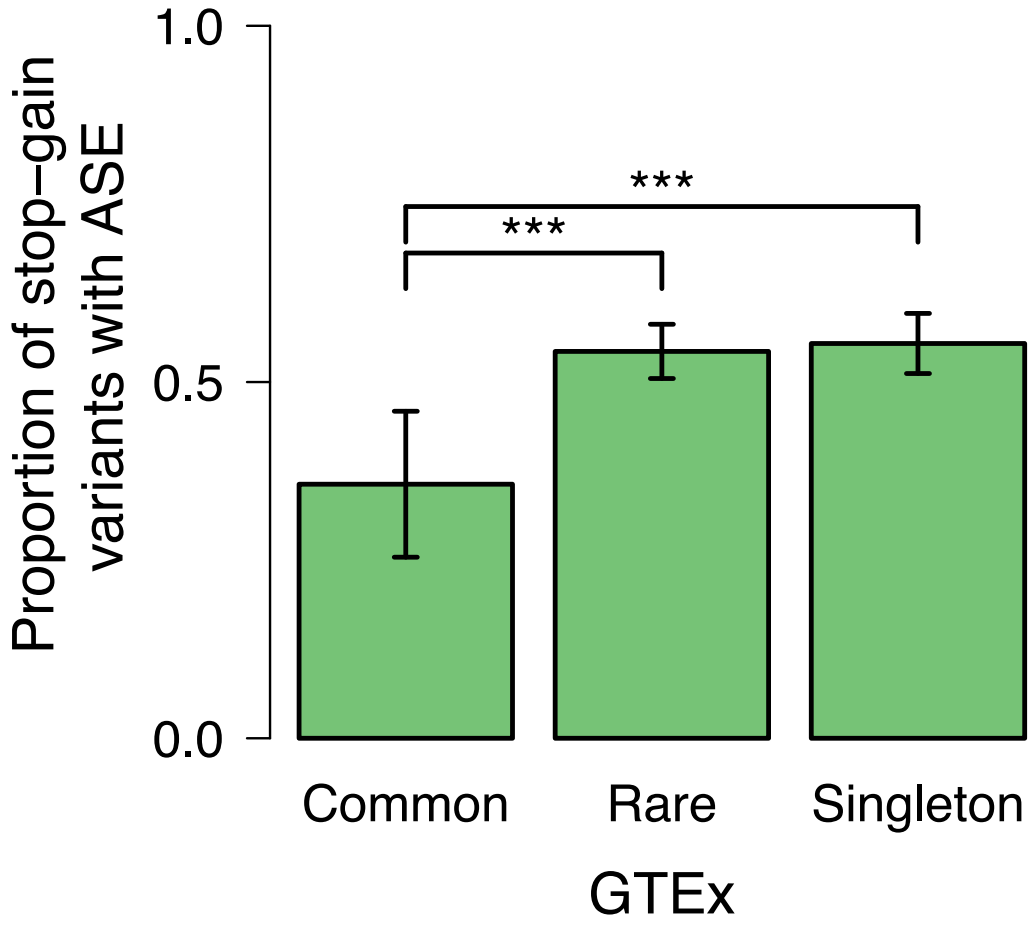


Figure 7.5: Proportion of nonsense variants with allele-specific expression effects in the GTEx data set in different frequency classes: common (MAF ≥ 0.05; $n = 84$), rare (MAF ≤ 0.01; $n = 657$), and singleton ($n = 532$) variants. I observed a significantly higher proportion of allelic imbalance in rare and singleton nonsense variants (54.3%, 2.5% and 97.5% quantiles 50.5% to 58.1%; and 55.4%, $51.2 - 59.6\%$; respectively) compared to common nonsense variants (35.7%, $25.4 - 45.9\%$).

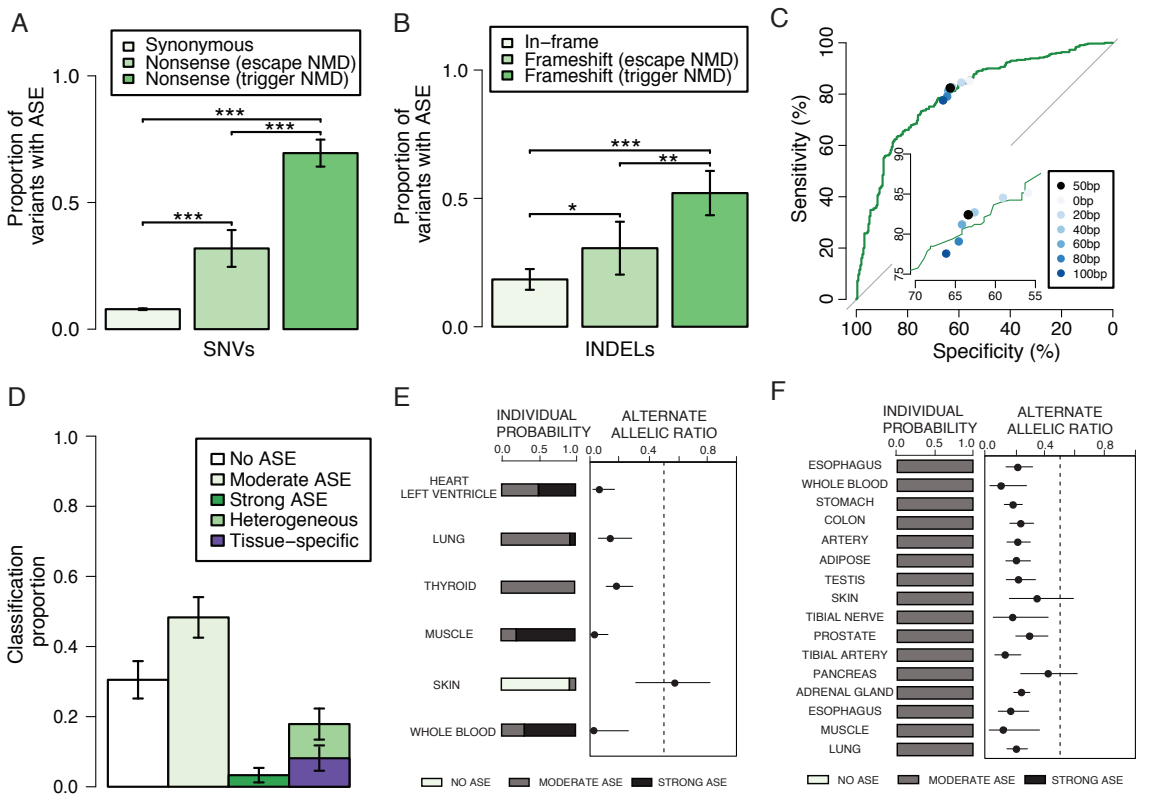


Figure 7.6: Characterizing allele-specific expression patterns of PTVs. A. Proportion of rare SNVs with ASE for synonymous variants (n = 25,233) and nonsense variants predicted to escape (n = 158) or trigger (n = 287) NMD. B. Proportion of rare indels with ASE for inframe (n = 355) and frameshift indel variants predicted to escape (n = 77) or trigger (n = 129) NMD. Due to different quality filters, the proportions are not directly comparable to those in panel A. C. ROC curve for predicting NMD with binary classification defined as no ASE (= escape) and moderate, strong, or heterogeneous ASE (= trigger). The filled circles show the specificity and sensitivity for NMD prediction with alternative simple distance rules (inset). D. Multi-tissue ASE classification for rare nonsense variants predicted to trigger NMD (n = 287). E. Example of ASE data across 6 tissues for a heterozygous carrier of the nonsense variant rs149244943 in gene *PHKB* classified as having heterogeneous ASE effects across the seven tissues. We confirmed that this effect is not driven by a common tissue-specific eQTL. F. Example of ASE data across 16 tissues for a heterozygous carrier of the nonsense variant rs119455955, a disease mutation for recessive late-infantile neuronal ceroid lipofuscinosis in gene *TPP1* (tripeptidyl peptidase I), classified as having moderate ASE across all tissues.

Assessing the accuracy of NMD predictions with ASE data. We sought to empirically evaluate the accuracy of NMD predictions that are based on sequence data alone. The most commonly used method, the 50 base-pair (bp) rule proposed by Nagy and Maquat (Nagy and Maquat, 1998) states that only termination codons located more than 50-55 nucleotides upstream of the 3'- most exon-exon junction

(measured after splicing) trigger mRNA degradation. To analyze NMD, I analyzed allelic counts of the following categories of variants:

- SNVs:
    1. synonymous variants (silent);
    2. nonsense variants predicted to escape NMD;
    3. nonsense variants predicted to trigger NMD,
- indels:
    1. in-frame indels;
    2. frameshift indel variants predicted to escape NMD;
    3. frameshift indel variants predicted to trigger NMD.

I applied GTM* using options `two.sided=FALSE` (given my interest in transcript degradation due to premature stop codon) and `indep=FALSE` in the software `MAMBA` described in Section 4.3.5. Indeed, rare nonsense SNVs predicted to trigger NMD according to the 50bp rule had a substantially larger proportion of ASE than those predicted to escape it (69.5% vs 31.9% respectively), both significantly higher than synonymous variants (7.9%, $P < 0.001$ across all comparisons, two-proportion z-test, Figure 7.6). The same is observed in rare indels, where frameshift indels predicted to trigger NMD according to the 50bp rule had more ASE (52.1%) than those predicted to escape NMD (30.6%) followed by in-frame indels (18.4%, Figure 7.6).

Improving NMD predictions. We asked whether we could improve NMD prediction beyond the 50bp rule by modeling sequence and other genomic features such as conservation and expression (ENCODE Project Consortium, 2012; Kircher et al., 2014). I collected 38 features (Table 7.3) and used the GTEx ASE outcome as a training data set with binary ASE classification of no ASE (escape; posterior probability $> .8$) or some form of ASE (trigger; MODASE, SNGASE, HET0, HET1, with sum of the posterior probability $> .8$) for all nonsense SNVs. I partitioned the data set into a training and a test set using 80% of the data to train the model and 20% to test. I sought to apply a machine learning algorithm to the data that would be able to predict NMD outcome from these 38 features. I chose random forest (Breiman, 2001) because of its good performance in empirical comparisons of supervised learning algorithms (Caruana and Niculescu-Mizil, 2006) and in a machine learning competition

using 121 different data sets (Fernández-Delgado et al., 2014), and its applicability with the `caret R` package (Kuhn, 2008). I applied a random forest algorithm using the `caret` package to generate a predictive model for no ASE versus strong/moderate ASE for all nonsense SNVs. I used the GEUVADIS ASE outcome as an independent test data set with binary ASE classification of no ASE (escape; posterior probability $> .8$) or moderate/strong ASE (trigger; with sum of the posterior probabilitiy $> .8$) and find that the predictive model predicts NMD better than the 50bp rule, with an Area Under the Curve (AUC) = 80.3% (95% CI $76.8 - 83.9\%$) compared to 50bp rule AUC = 72.9% ($69.3 - 76.5\%$) (Figures 7.6, 7.7, 7.8). The features were ranked by variable importance measures (Breiman, 2002) referred to as "MeanDecreaseAccuracy" and "MeanDecreaseGini". This is automatically generated by the `R` package `randomforestr` using the command `varImpPlot`. Interestingly, one of the top ranked features was the distance to the donor site supporting the hypothesis that pre-mRNA splicing is linked to NMD in humans (Casadio et al., 2015). Furthermore, the number of downstream exons was ranked above the 50bp rule indicating that the the absolute number is an important factor.

Additionally, I tested alternative distance rules analogous to the 50bp rule. I found that applying a 100bp threshold increases specificity but reduces sensitivity, while a 0bp threshold had the inverse effect; of all thresholds tested, the 50bp rule has the highest predictive value, consistent with the original estimates from Nagy and Maquat (Nagy and Maquat, 1998). These results provided the first large-scale quantitative estimates of the value of NMD predictions in humans, illustrating that the 50bp rule, despite being generated from a limited data set over 15 years ago, remains a valuable heuristic. Nonetheless, the more comprehensive model improved prediction of this important functional outcome from variant data alone, and it is possible that the predictive features highlighted will provide a framework for better understanding of molecular mechanisms of NMD (Figure 7.8).

No.	**Predictor**	**Description**
1	X50bp	50bp rule
2	startdist	distance to start codon
3	stopdist	distance to stop codon
4	utr3distend	distance to 3'-UTR end
5	utr5diststart	distance to 5'-UTR start
6	utr3diststart	distance to 3'-UTR start
7	utr5distend	distance to 5'-UTR end
8	utr3size	size of 3'-UTR
9	utr5size	size of 5'-UTR
10	trnscaffected	Variant is annotated as a PTV in all or some transcripts
11	nexons	number of exons
12	ntrnsc	number of alternative isoforms
13	donordist	distance to donor splice site
14	acceptordist	distance to acceptor splice site
15	onecodingexon	indicator variable (= 1 if gene contains only one coding exon)
16	af	allele frequency
17	GC	Percent GC in a window of +/- 75bp
18	CpG	Percent CpG in a window of +/- 75bp
19	priPhCons	Primate PhastCons conservation score (excl. humans)
20	mamPhCons	Mammalian PhastCons conservation score (excl. humans)
21	verPhCons	Vertebrate PhastCons conservation score (excl. humans)
22	priPhyloP	Primate PhyloP score (excl. humans)
23	mamPhyloP	Mammalian PhyloP (excl. humans)
24	verPhyloP	Vertebrate PhyloP (excl. humans)
25	GerpN	Neutral evolution score defined by GERP++
26	GerpS	Rejected Substitution score defined by GERP++
27	EncExp	Maximum ENCODE expression value
28	EncH3K27Ac	Maximum ENCODE H3K27 acetylation level
29	EncH3K4Me1	Maximum ENCODE H3K4 methylation level
30	EncH3K4Me3	Maximum ENCODE H3K4 trimethylation level
31	EncNucleo	Maximum of ENCODE Nucleosome position track score
32	minDistTSS	Distance to closest Transcribed Sequence Start (TSS)
33	minDistTSE	Distance to closest Transcribed Sequence End (TSE)
34	relcDNApos	Relative position in transcript
35	relCDSpos	Relative position in coding sequence
36	relProtPos	Relative position in protein codon
37	lofflag	LOF flag proposed in MacArthur et al. 2012
38	downstreamexons	Number of exons downstream of the PTV

Table 7.3: List of 38 predictors used for modeling NMD.

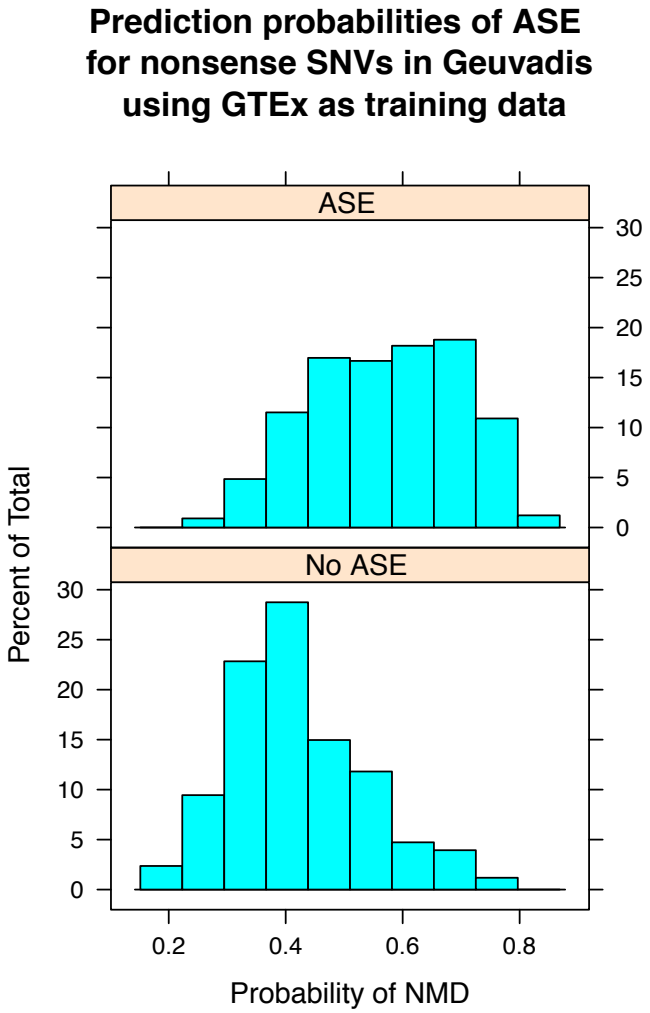


Figure 7.7: Insights into nonsense-mediated decay: modeling NMD with ASE outcome. When applying machine learning algorithms to a test data set that is independent from the training data set it is common practice to examine the prediction accuracy. To predict the outcome of the independent test data set (Geuvadis) I used the `predict.train` function (`caret` package) using the option `type = "prob"` to compute class probabilities. In this plot a histogram of the predicted probabilities of NMD assigned to nonsense SNVs in the Geuvadis data set is shown. For those variants that have no ASE the model predicted for 22.8% of variants to have probability $> .5$ of having ASE signal indicative of NMD. Conversely, it predicted for 77.2% of those variants to have probability $\leq .5$ of having ASE signal. Similarly, for those variants that have some ASE the model predicted for 68.8% of variants to have probability $\geq .5$ of having ASE signal indicative of NMD. Conversely, it predicted for 31.2% of those variants that have some ASE to have probability $< .5$ of having ASE signal.

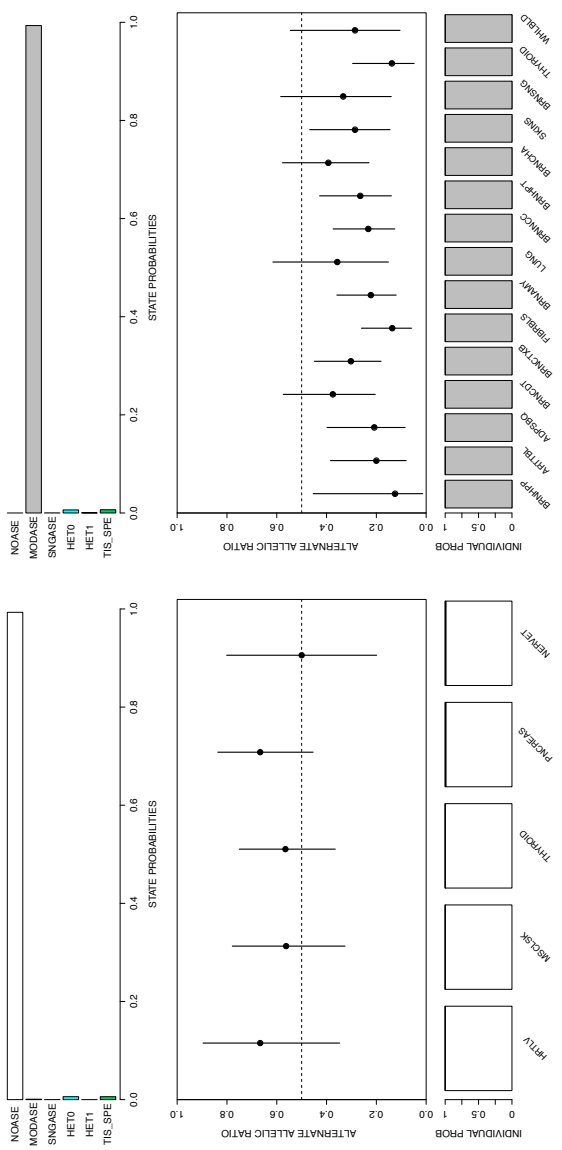
(a) Feature (variable) importance plot: mean decrease in accuracy

(b) Feature (variable) importance plot: mean decrease in gini

Figure 7.8: Insights into nonsense-mediated decay: feature (variable) importance plots for the random forest algorithm. For each feature in the training exercise the feature importance plot tells the user how important that feature is in classifying the data. The top 30 features are represented on the graphs and an estimate of their importance is given on the x-axis. The most important features are at the top. In the training exercise I have an 80:20 percent split, i.e. treating 80% of the GTEx data set as the training set and 20% as the test set (commonly referred to as the out of bag observations). This is separate from the independent test data set (Geuvadis) that is used to evaluate the accuracy of the predictions. a) By contrasting the out of bag predictions with the known outcomes, an estimate of the prediction error rate is obtained. For each feature, the mean decrease in accuracy is computed by comparing the prediction error rate to the case when the values of the feature are permuted. b) A higher mean decrease in Gini means that the feature plays a greater role in partitioning the data into the defined classes (some ASE or no ASE). This is automatically generated by the `R` package `randomforest` (Liaw and Wiener, 2002) using the command `varImpPlot`. For a description and a list of the 38 sequence and genomic features used see Table 7.3.

Variation in NMD across tissues. The GTEx study design, with RNA sequencing data for numerous tissues of the same individual, presents an unprecedented opportunity to study variation in NMD across tissues. To this end, I applied the Bayesian hierarchical model (Pirinen et al., 2015) to rare nonsense variants predicted to trigger NMD according to the 50bp rule and with ASE data from at least two tissues. I estimated that 30.5% of these nonsense variants have no ASE in any tissue, 48.3% have moderate ASE across all tissues, and 3.3% have strong ASE across all tissues (Figures 7.9, 7.10, 7.11 for examples). Finally, I estimated that 17.9% of these nonsense variants have heterogeneous effects across tissues, and 8.1% of ASE effects are specific to a single tissue (Figure 7.6). These results confirmed and quantified how NMD typically does not result in complete degradation of the affected transcripts.

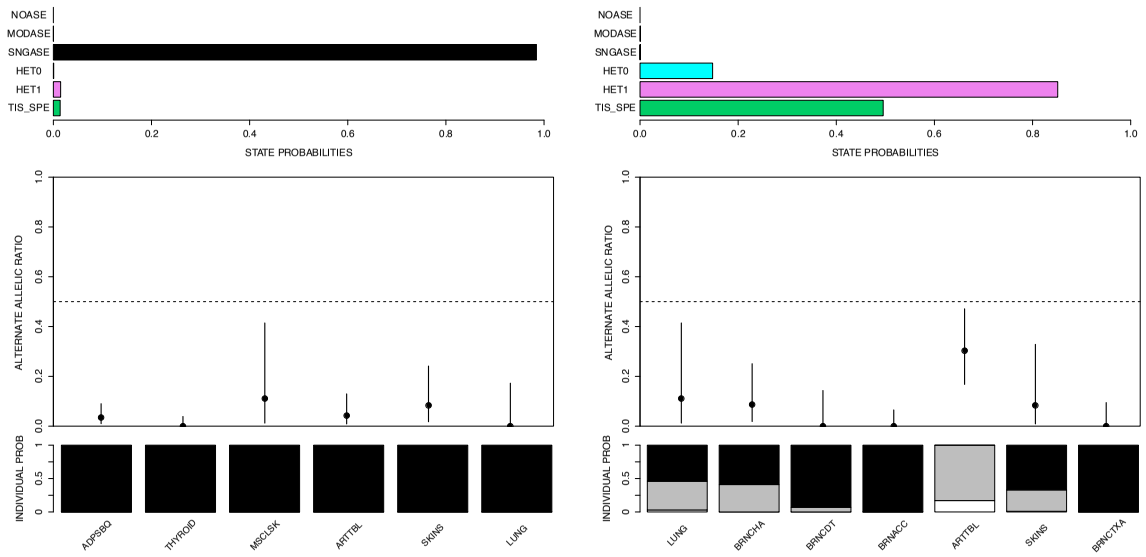
Example of ASE signals. I illustrated the multi-tissue analysis by considering three nonsense SNVs. The first nonsense SNV considered, rs149244943 in *PHKB*, is classified as having heterogeneous ASE effects across the six tissues of the heterozygous carrier. The second nonsense SNV considered, rs119455955, a recessive disease causing mutation (recessive late-infantile neuronal ceroid lipofuscinosis) in gene *TPP1* (Sleat et al., 1997), is classified as having moderate ASE across all 16 collected tissues (Figure 7.6). The third variant considered, rs328, is a nonsense mutation (p.S447X) in lipoprotein lipase *LPL* that is associated with lower triglyceride levels (Saxena et al., 2007). Across all studied tissues the data showed that this variant does not have an ASE effect (posterior probability = 1, Figure 7.12), and that transcripts with the mutation are retained. This finding suggests that a truncated protein product may be translated consistent with reports of a truncated protein with a gain-of-function mutation (Rip et al., 2006) and additionally suggested that such proteins are likely present across all tissues.



(a) NO ASE

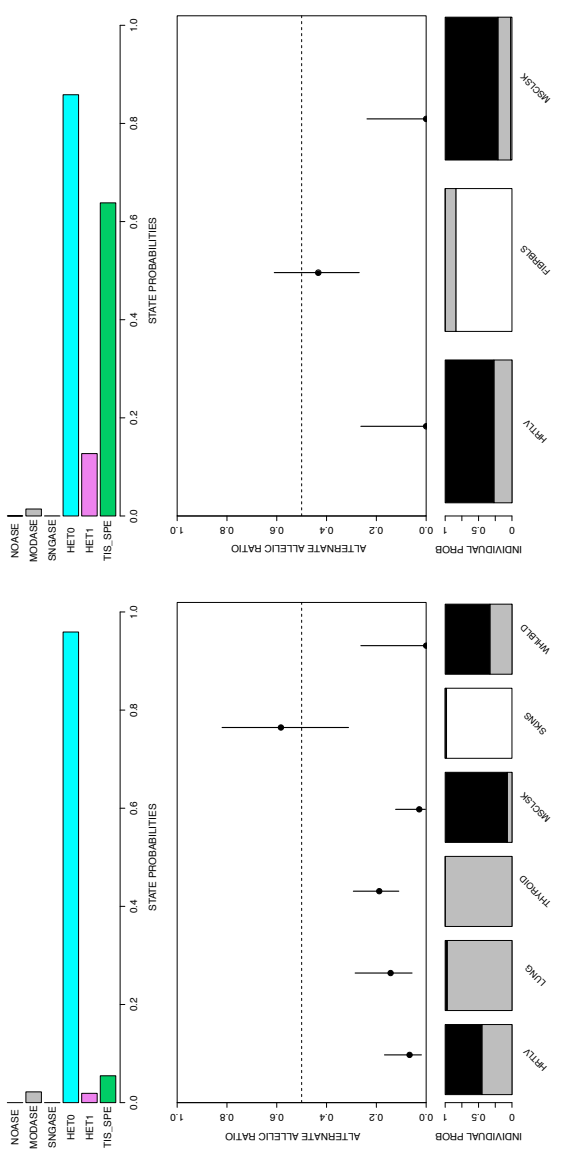
(b) Moderate ASE across all tissues

Figure 7.9: ASE classification examples: no ASE and moderate ASE across all tissues. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.S88X (rs41296182) in the gene *TRIM45* (tripartite motif containing 4), classified as having no ASE effects across all studied tissues (posterior probability for the NOASE state = 0.99). b) An example of a PTV, p.Q59X (rs121908176) in the gene *BBS2* (Bardet-Biedl syndrome 2), classified as having moderate ASE effects across all studied tissues (posterior probability for the MODASE state = 0.99).



(a) Strong ASE across all tissues (b) Mixture of moderate and strong ASE

Figure 7.10: ASE classification examples: strong ASE across all tissues and mixture of moderate and strong ASE. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.R388X (snp_2_201476115) in the gene *AOX1* (aldehyde oxidase 1), classified as having strong ASE effects across all studied tissues (posterior probability for the SNGASE state = 0.98). b) An example of a PTV, p.E318X (snp_6_86256830) in the gene *SNX14* (sorting nexin 14), classified as having a mixture of moderate and strong ASE effects (posterior probability for the HET1 state = 0.87).



(a) Mixture of no ASE and ASE

(b) Tissue-specific ASE

Figure 7.11: ASE classification examples: mixture of no ASE and ASE effect and tissue-specific ASE. The plots shown represent several properties of the data: (top) posterior probabilities of the variant classification states across all tissues including NOASE (no ASE effects across all tissues), MODASE (moderate ASE effects across all tissues), SNGASE (strong ASE effects across all tissues), HET0 (heterogeneous effects with some tissues having no ASE and others some form of ASE), HET1 (heterogeneous effects with some tissues having moderate ASE effects others strong ASE effects), TIS_SPE (tissue specific effect); (center) shows the alternate allelic ratio (maximum likelihood estimate and 95% confidence interval) per tissue; (bottom) INDIVIDUAL PROB representing the individual tissue posterior probability of no ASE (white), moderate ASE (gray), or strong ASE (black). a) An example of a PTV, p.Q776X (rs149244943) in the gene *PHKB* (phosphorylase kinase, beta), classified as having a mixture of no ASE and ASE effects (posterior probability for the HET0 state = 0.96). b) An example of a PTV, p.Q66X (snp_14_7817080) in the gene *ALKBH1* (alkB, alkylation repair homolog 1), classified as having a tissue-specific ASE effect (posterior probability for the TIS_SPE state = 0.64, which is a sub-state of HET0 as described in Section 4.3).

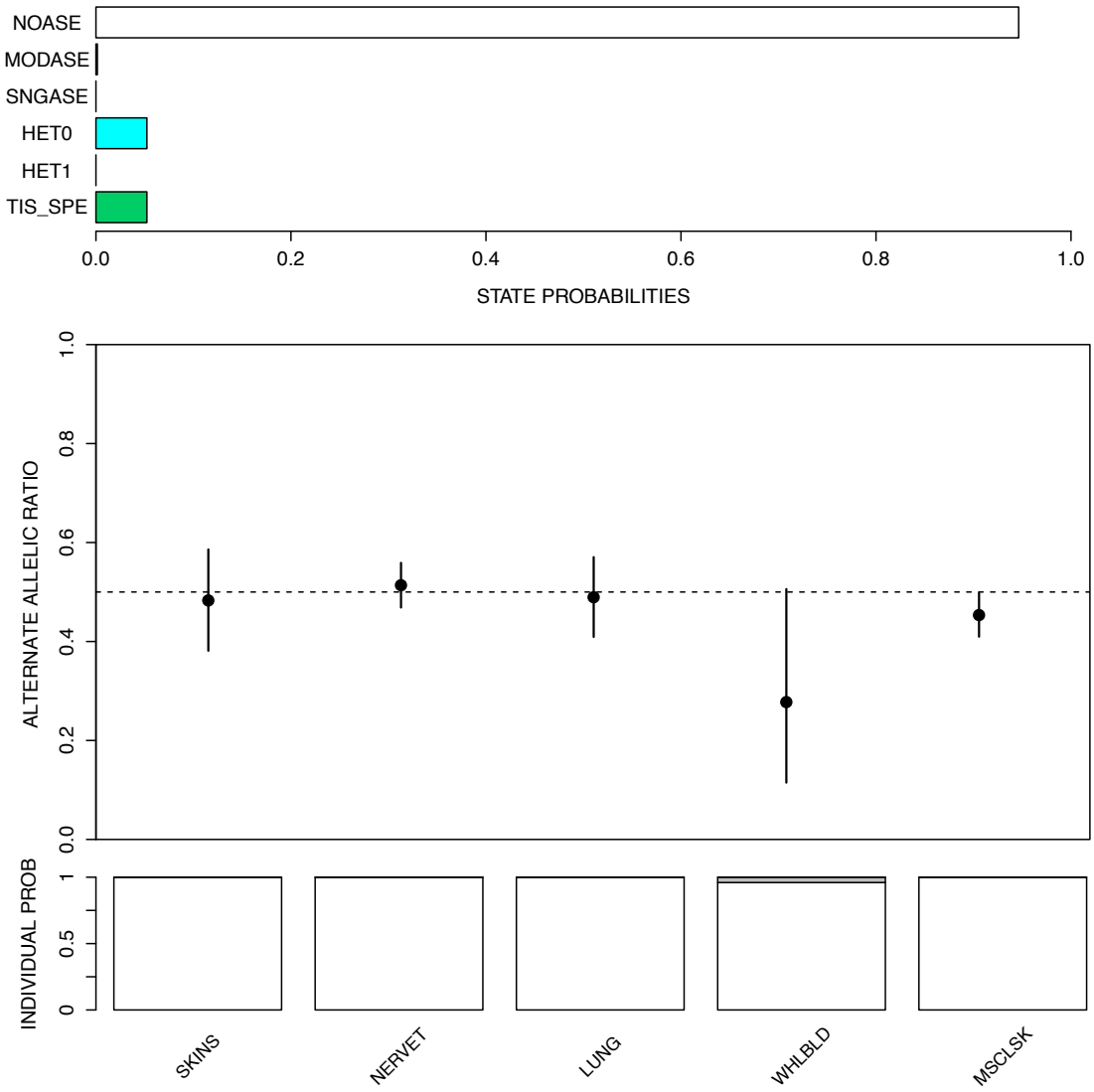


Figure 7.12: ASE data for p.S474X (rs328) in the gene *LPL* (lipoprotein lipase). The variant is classified as having no ASE across all tissues (posterior probability for the NOASE state > 0.99) in the RNA-seq data set, supporting the observation that transcripts with the mutation are retained. This is consistent with reports of a truncated protein with a gain-of-function mutation and suggests that such proteins are likely present across all tissues.

7.4.3 Dosage compensation for heterozygous PTVs

Background and motivation. Large structural deletions that partially or completely remove genes are confidently expected to cause complete loss of function of the affected genes. Thus, such deletions provide an opportunity to examine the possibility that heterozygous carriers of loss-of-function variants might exhibit compensatory

up-regulation of the functional allele. This phenomenon has been reported to occur for at least 80% of deleted genes in a recent study of *Drosophila melanogaster* (Zhou et al., 2011) and reported to correlate with gene expression level (McAnally and Yampolsky, 2010), but no systematic study of this phenomenon has yet been performed in humans. If present, such a buffering of the variant effect could be one mechanism contributing to tolerance of PTVs and could explain the widespread haplosufficiency of human genes (Huang et al., 2010; Henrichsen et al., 2009). In addition to the clinical ramifications, it would point to widespread cellular feedback mechanisms for homeostatic maintenance of transcript levels. In this study we sought to examine whether we could identify signatures of dosage compensation.

A key challenge in the detection of compensation is genotyping error, which is known to be a major challenge for deletion calls from sequencing data (The 1000 Genomes Consortium, 2012; Craddock et al., 2010; White et al., 1994) and is expected to produce a signal identical to dosage compensation because the "heterozygous" individual actually has two functional copies. In the study we focused only on biallelic whole-gene deletion variants with strong experimental support and manual curation.

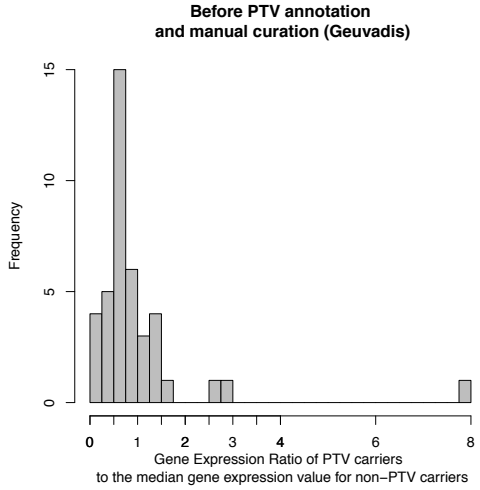
Assessing dosage compensation by analyzing common and rare large gene deletions. To put the comparison of common and rare large deletions into perspective I briefly summarize the finding from the common deletion analysis described in Section 7.2. Few examples of common whole-gene deletion polymorphisms were analyzed in the study. In the GTEx data set reliable genotypes were obtained for a common deletion of the gene *UGT2B17*. In the Geuvadis data set reliable genotypes were obtained for common deletions of the genes *DDT*, *GSTT2*, *FAM106A*, *LGALS9C* and *OR2T10*. For 5/6 (83%) of these genes, the additive model relating gene expression to gene copy number provided a better fit than a dominant model, thus providing no evidence for dosage compensation.

I sought for signal of dosage compensation in rare deletions in the GTEx and Geuvadis data, analyzing whether the expression levels of heterozygous deletion carriers tend to be half those of the population average. While the raw data show a strong signal of dosage compensation, this signal is largely removed by LOF annotation filtering and manual curation of CNV genotypes, suggesting a very strong impact of genotyping and annotation error that may have confounded previous analyses of this phenomenon (Figures 7.13, 7.14). Another potential source of error is somatic mutation. In the DNA data of one GTEx individual a large (20Mb) mosaic deletion

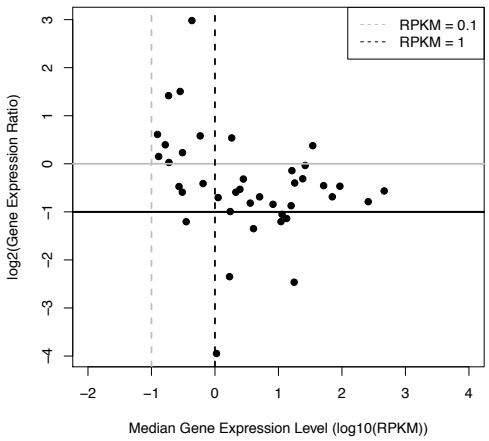
of the long arm of chromosome 20 was identified. This large deletion had been associated with myeloid malignancies (White et al., 1994). However, after I analyzed the multi-tissue RNA-seq data it was evident that the deletion was found only in the individual's blood (where DNA was extracted from), and in other tissues the normal expression of genes spanned by this variant was apparently not due to compensation but by the cells not carrying the somatic deletion (Figure 7.15). After manual curation, individuals heterozygous for a total of 11 rare PTV deletions (25 genes) had consistently decreased expression of the respective gene compared to the median of the population (P= 1.37×10^{-5}, one-sided binomial test).

Assessing dosage compensation by analyzing nonsense SNVs with strong ASE signals. To assess dosage compensation of a gene by upregulation of the non-PTV containing copy we focused on nonsense SNVs with strong ASE signals. It is possible that dosage compensation may exist in a scenario where the copy of the gene containing a PTV is degraded by NMD, and upregulation of the PTV-containing copy of the gene occurs. For the latter scenario we are unable to address whether gene dosage compensation exists with the data we have generated.

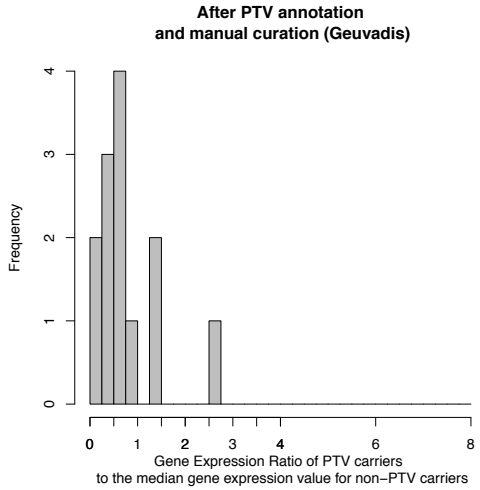
I analyzed 53 nonsense PTVs with strong ASE signals in the Geuvadis and the GTEx data sets and found that the individual carriers of the PTVs had consistently decreased expression of the respective gene compared to the median of the population ($P = 2.90 \times 10^{-9}$, one-sided binomial test; Figures 7.16, 7.17). These results suggest that full dosage compensation by upregulation of the non-PTV containing copy of gene is rare for human genes, and highlight the need for extreme care in accounting for genotyping error when performing surveys of this phenomenon. These findings suggest that the widespread haplosufficiency among human genes (Huang et al., 2010) may be driven by cellular tolerance to reduced gene expression levels, rather than direct compensatory upregulation of affected genes.



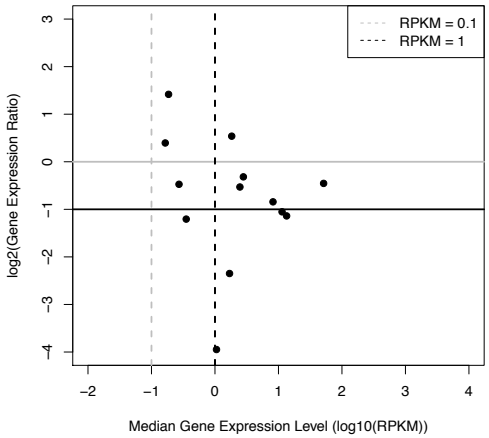
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

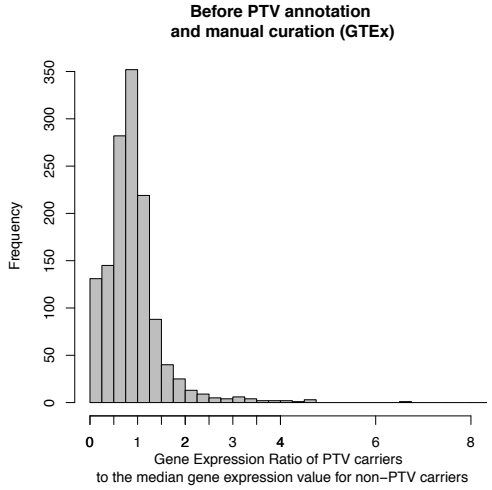


(c) Histogram of gene expression ratios

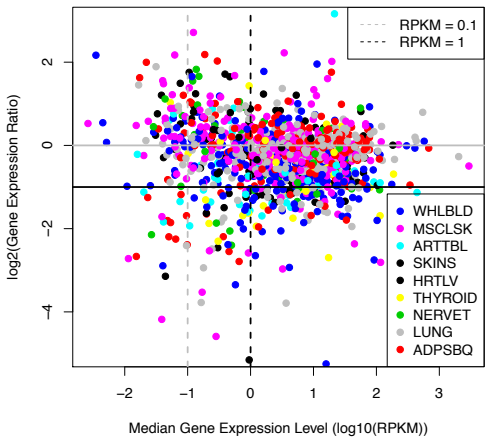


(d) Scatter plot of the median gene expression value and gene expression ratio

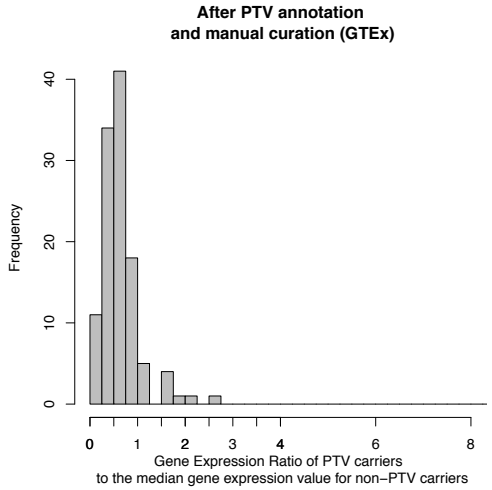
Figure 7.13: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the deleted gene(s) in PTV carriers to the median gene expression value of non-PTV carriers. a, c) Histograms of the gene expression ratios in the Geuvadis data set before ($n = 27$ CNV deletions) and after ($n = 8$ CNV deletions) PTV annotation and manual curation filtering, respectively. b,d) Scatter plots of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after PTV annotation and manual curation filtering, respectively.



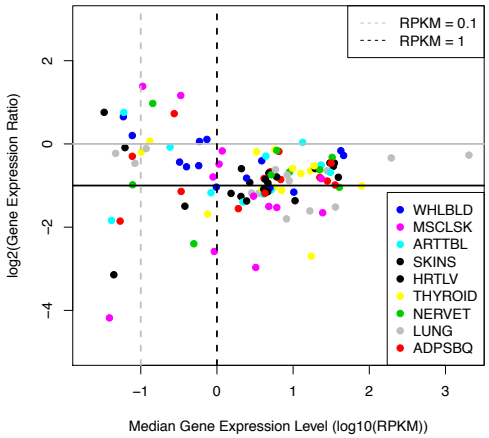
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

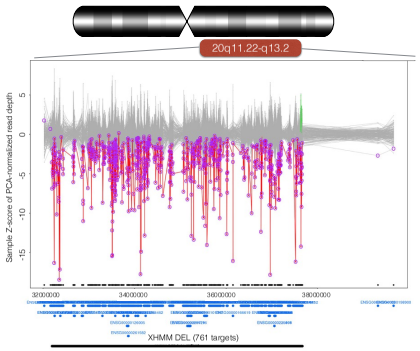


(c) Histogram of gene expression ratios

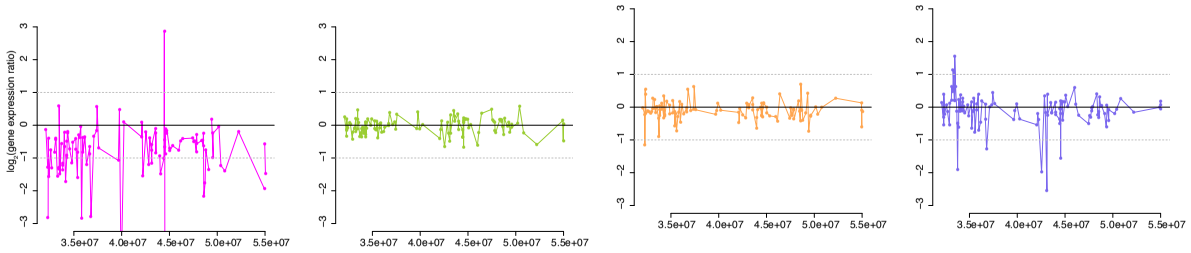


(d) Scatter plot of the median gene expression value and gene expression ratio

Figure 7.14: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the deleted gene(s) in PTV carriers to the median gene expression value of non-PTV carriers. a,c) Histograms of the gene expression ratios in the GTEx data set before ($n = 65$ CNV deletions) and after ($n = 3$ CNV deletions) PTV annotation and manual curation filtering, respectively. b,d) Scatter plots of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after PTV annotation and manual curation filtering, respectively.

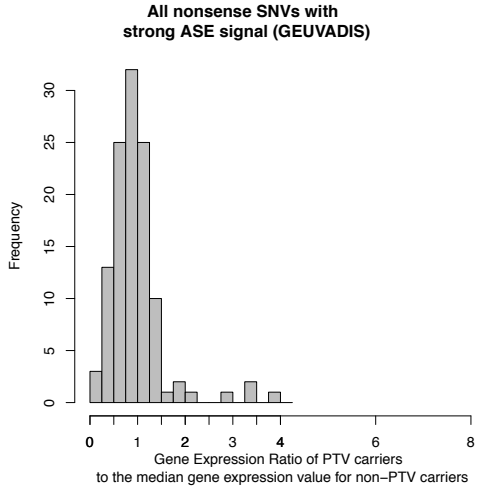


(a) XHMM 20Mb deletion

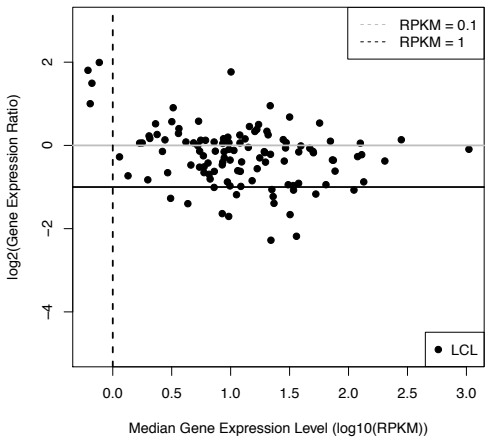


(b) Blood (WHLBLD) (c) Lung (LUNG) (d) Adipose Subcutaneous (ADPSBQ) (e) Muscle Skeletal (MSCLSK)

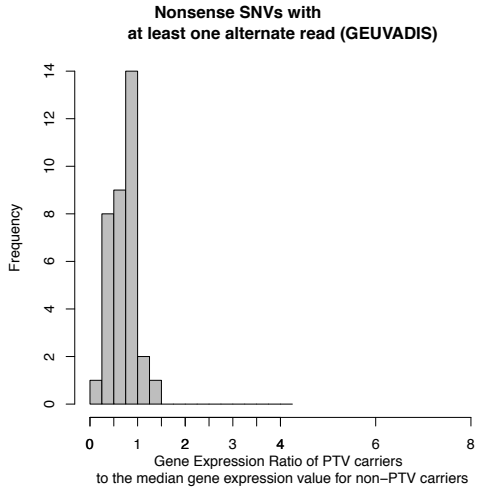
Figure 7.15: Insights into dosage compensation: impact of somatic mutations. a) A large deletion event in a single individual that spanned over 20 megabases (Mb). In red the sample Z-score of PCA normalized read depth for the sample containing the 20Mb event is shown (gray for all other samples). The large deletion was validated with the OMNI 2.5 array data set (data not shown). Somatic mosaicism in blood was inferred by the frequency of heterozygous genotypes and confirmed by the patterns of $\log_2$ gene expression ratios (individual carrier gene expression value/median non-PTV expression values, y-axis) across the 20Mb region in b) whole blood (mosaic deletion is present), c) lung (mosaic deletion is absent), d) subcutaneous adipose (mosaic deletion is absent), and e) skeletal muscle (mosaic deletion is absent). When studying dosage compensation the patterns of no difference in gene expression across all the genes in the 20Mb deletion in lung, adipose, and muscle could be mistaken as evidence of gene dosage compensation.



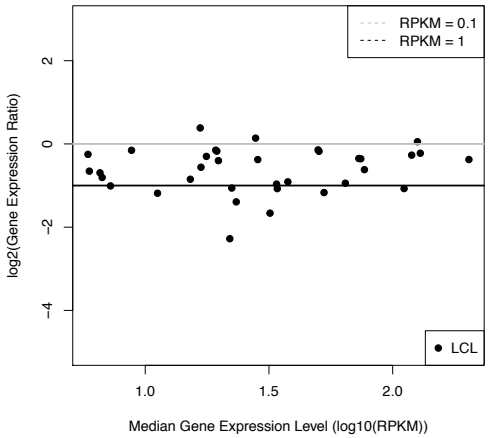
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio

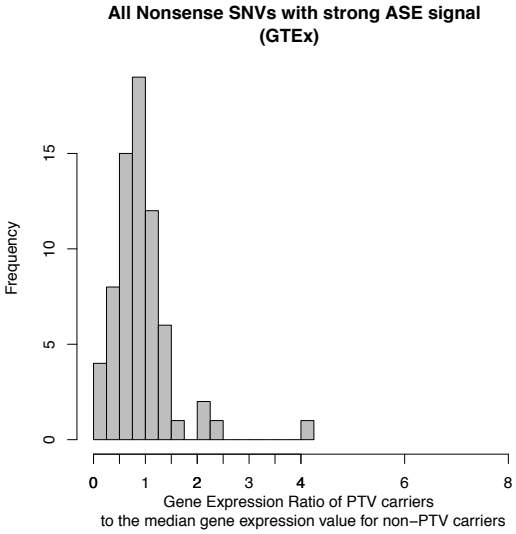


(c) Histogram of gene expression ratios

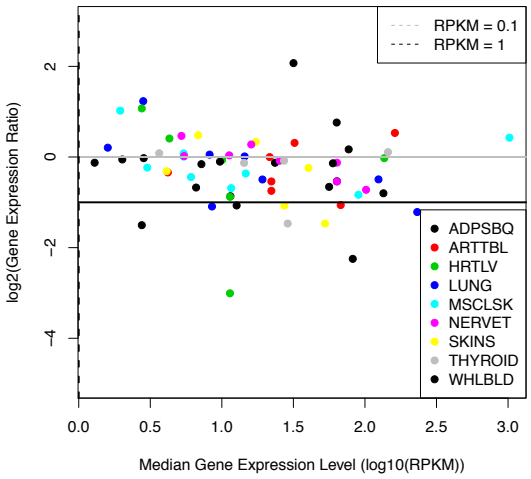


(d) Scatter plot of the median gene expression value and gene expression ratio

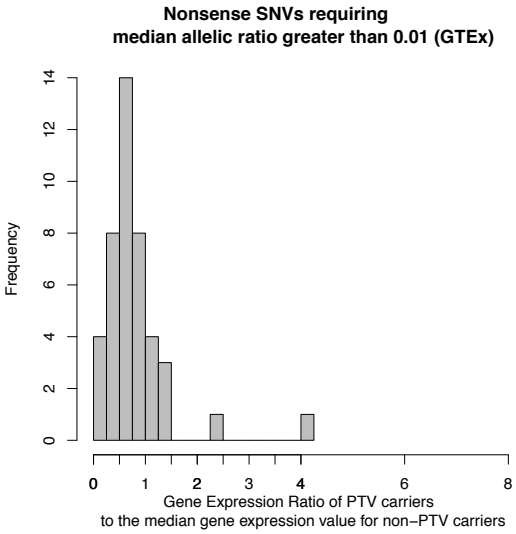
Figure 7.16: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the PTV carrier to the median gene expression value of non-PTV carriers. a,c) Histograms of the gene expression ratios in the GEUVADIS data set before ($n = 116$ nonsense SNVs) and after ($n = 35$ nonsense SNVs) requiring at least one alternate read in the ASE data set, respectively. b,d) Scatter plots of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after filtering, respectively.



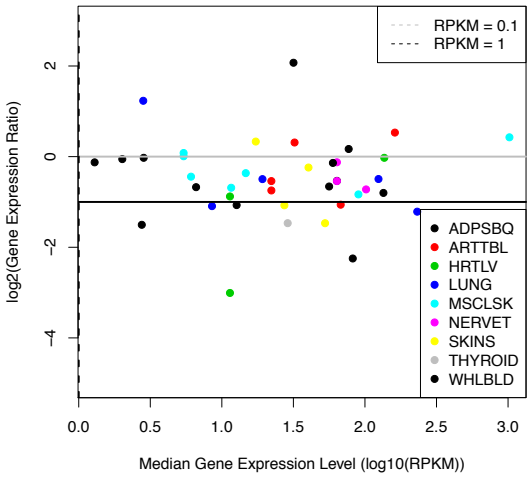
(a) Histogram of gene expression ratios



(b) Scatter plot of the median gene expression value and gene expression ratio



(c) Histogram of gene expression ratios



(d) Scatter plot of the median gene expression value and gene expression ratio

Figure 7.17: Insights into dosage compensation: examining gene expression ratios measured by comparing the gene expression value of the PTV carrier to the median gene expression value of non-PTV carriers. a,c) Histograms of the gene expression ratios in the GTEx data set before ($n = 25$ nonsense SNVs) and after ($n = 18$ nonsense SNVs) requiring at least one alternate read in the ASE data set, respectively. b,d) Scatter plots of the median gene expression across all individuals (x-axis) and the gene expression ratio (y-axis) before and after filtering, respectively.

7.4.4 Transcriptional impact of variants proximal to splice junctions

Background and motivation. A survey of reported disease-causing mutations from the Human Gene Mutation Database (HGMD) v2012.1 (Stenson et al., 2014) suggests that over 10% of disease causing mutations impact splicing, and splicing-QTL studies have catalogued thousands of loci with common variants associated with splicing variation (Lappalainen et al., 2013; Ardlie et al., 2015; Battle et al., 2014). However, widely used splicing variant annotation tools typically focus only on the two bases at either end of a spliced intron, regularly referred to as the "essential splice sites", despite the fact that other more distant sites are known to play critical roles in this process (Burge et al., 1999; Faustino and Cooper, 2003). In addition to difficulty in predicting whether splicing changes occur in general, also the type of change, such as exon skipping or elongation, is usually unknown from genetic data alone. Splicing changes always lead to major changes in protein structure either via in-frame changes in exon structure or by introducing a premature stop codon. This data set provided a valuable opportunity to assess the impact of genetic variation in the region around splice sites on transcript splicing.

Impact of rare variants proximal to splice junctions. To interpret the impact of rare variants (MAF ≤ 0.01), the majority of variants around splice junctions, on transcript structure I standardized the population distribution of each splice-junction quantification per tissue studied, and grouped variants according to their distance from their respective donor and acceptor sites. I then applied the Splice Disruption Model described in Chapter 4 to analyze if the individuals carrying variants in these positions differ from the rest of the population in splice junction quantifications, and in the expression levels of the proximal exon and intron (Figure 7.18).

In the Geuvadis data set, the largest sample size, I found that up to 79% of variants in the four essential splice site loci show evidence of splice disruption in the RNA-seq data (P value < 0.01; Figure 7.18). I also found evidence of significant splice disruption from variants outside these regions, especially at position 1-5bp of intronic donor splice sites and 1bp into the adjacent exon, but also more distantly, including the -24 position from the acceptor site with 9% of variants with strong effects (1.4 standard deviations below the mean); this might reflect the peak location of the branch-point position required for pre-mRNA splicing (Corvelo et al., 2010).

The patterns of splice disruption observed in the RNA-seq data are consistent with other estimates of functional effects: sites with enrichment of splice-disrupting

variation in the data set also display increased median Genomic Evolutionary Rate Profiling (GERP, Cooper et al. (2005)) scores (Figure 7.18), depletion of common variants in a large exome sequencing data set from over 4,500 Swedish individuals (Figure 7.18, (Purcell et al., 2014)), and a markedly higher prevalence of disease causing mutations (Figure 7.18). I provided our posterior probability estimates for sites with significant alternative distributions ($P < 0.05$) as a resource for future studies (Figure 7.19 for an example, and Figure 7.20 for a splice disrupting variant identified in HGMD in a non-canonical site), which is now hosted on the GTEx portal `gtexportal.org`. For an example of a visualization please see `http://kataviz.github.io/html/ptv.html` and `http://kataviz.github.io/html/ase.html`.

To put it in perspective I present the results from the common variant analysis described in Section 7.2. In order to analyze whether common variants in the 50bp region have a similar pattern of splice effects, common variants were mapped associating to exon inclusion levels measured as percentage spliced in, or psiQTLs. Analyzing the nine GTEx tissues with > 80 samples separately, an average of 313 exons per tissue in the population having a psiQTL at FDR of 5% were discovered, and 144 exons with a psiQTL in a joint analysis across tissues. The top variants of the psiQTLs are uniformly distributed across the 50bp window, without a similar enrichment at sites proximal to splice junctions as rare variants. The results demonstrated that analyzing common variants is unlikely to capture the properties of rare variants with drastic splicing effects.

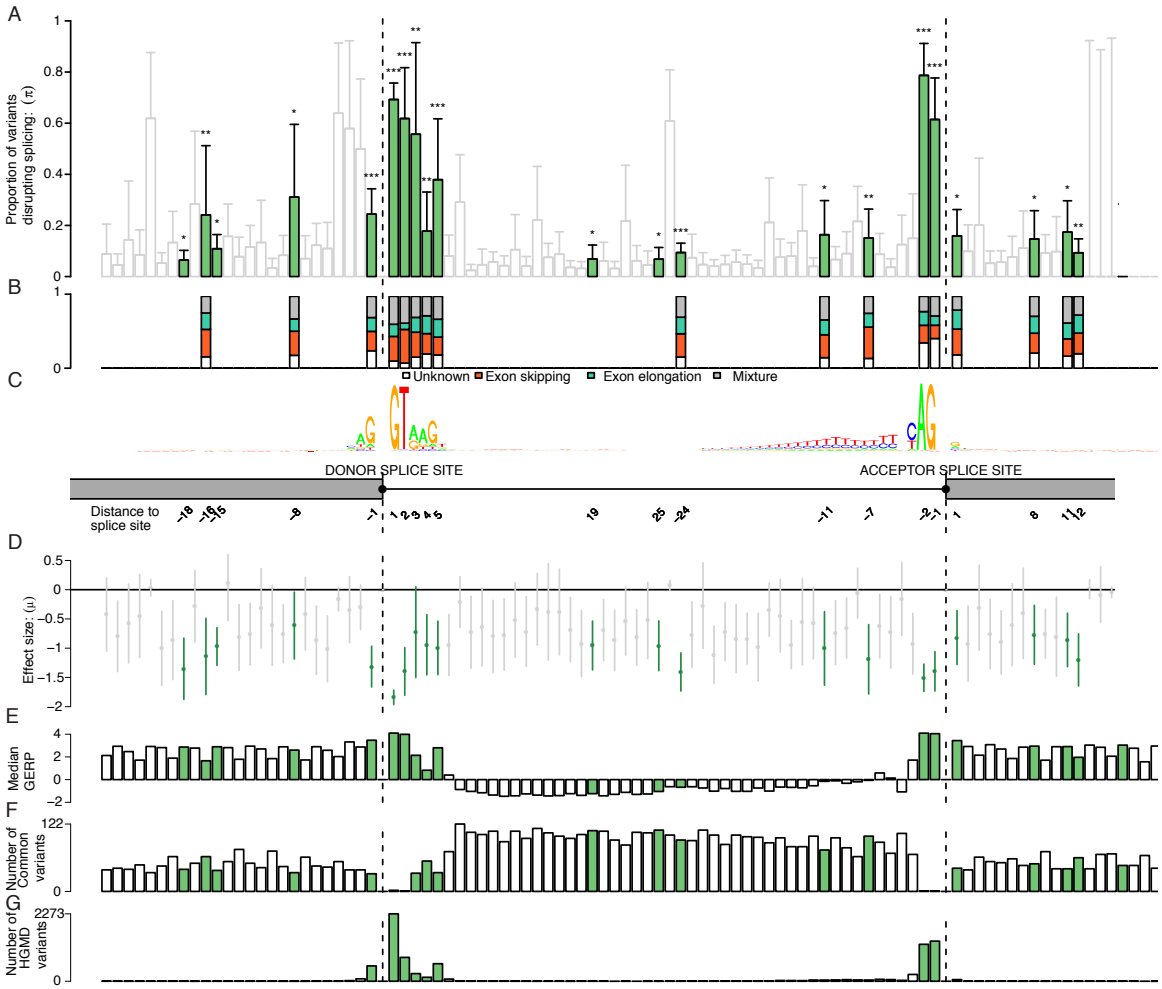
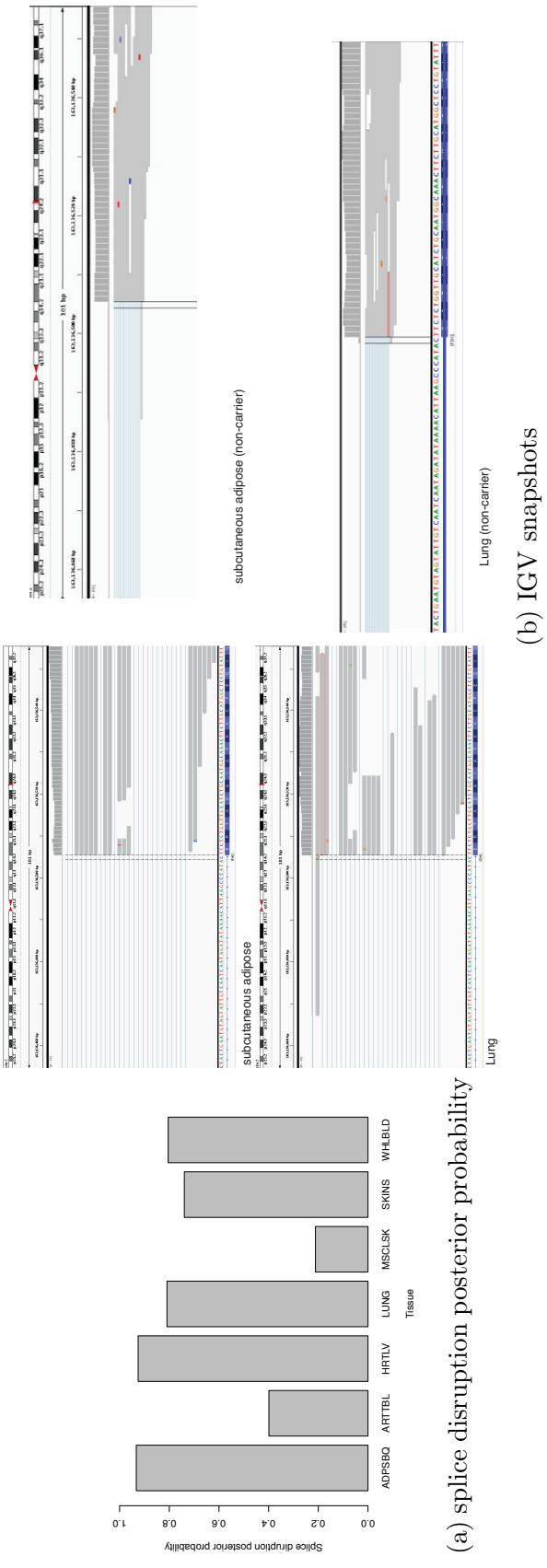


Figure 7.18: A. Proportion of variants disrupting splicing at each distance ± 1-25bp from donor and acceptor site, ($^*0.01 \leq P < 0.05$, $^{**}0.001 \leq P < 0.01$, $^{***}P < 0.001$; green for any distance with $P < 0.05$; SDM P value evaluated using the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). B. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). C. Diagram of donor and acceptor splice junctions and sequence logo of represented sequences. D. Effect size estimates (in standard deviations from the population distribution) of the variants on splice junction quantification value. E. Median Genomic Evolutionary Rate Profiling (GERP) measure of all variants. F. Distribution of common variants identified in an independent exome sequencing study of 4,500 Swedish individuals. G. Distribution of reported disease-causing variants in HGMD.



(a) splice disruption posterior probability

(b) IGV snapshots

Figure 7.19: Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, c.IVS8+1G>C (rs35337543), in the gene *IFIH1* (interferon induced with helicase C domain 1). a) The splice disruption posterior probability across seven tissues (abbreviations described in Section 7.2) after applying the splice disruption model to the GTEx data set. b) IGV snapshots of RNA-seq data from two tissues (subcutaneous adipose and lung) in the c.IVS8+1G>C carrier (top and bottom left). The snapshots show splice disruption in approximately half the reads (light blue lines). IGV snapshot of RNA-seq data from the same two tissues (subcutaneous adipose and lung) in an individual that does not carry the c.IVS8+1G>C PTV variant are shown (top and bottom right).

(a) effect size of alternative mixture distribution

(b) splice disruption posterior probability

Figure 7.20: Transcriptional impact of variants proximal to splice junctions: a splice disrupting variant, rs116928232, in a non-canonical splice site, in the gene *LIPA* (lipase A). a) The alternative shift in μ (the estimated effect size for variants predicted to disrupt splicing described in Section 4.4) for the splice disruption group, and b) the posterior probability that the variant belongs to the splice disruption group across ten tissues in the GTEx data set and Geuvadis data set (LCL). The variant rs116928232 is found in the HGMD data base and annotated as a causal mutation for an autosomal recessive disorder Cholesteryl ester storage disease (CESD, OMIM:27800) associated with reducted activity and genetic defects of lysosomal acid lipase. In (Aslanidis et al., 1996), a compound heterozygote CESD patient is identified. The patient has a nonsense mutation in the paternal allele while the maternal allele contains the A allele in rs116928232. This data is consistent with the observation that the A allele in rs116928232 disrupts splicing in the patient.

7.5 Conclusions

Understanding the functional consequences of PTVs is critical for improving variant annotation from DNA sequence variants. In this chapter, I have demonstrated the value of transcriptome data for the interpretation of protein-truncating variants, provided the first comprehensive characterization of the effects seen in normal populations, and presented new approaches for improving the annotation of PTVs.

The results presented in this chapter showed differences between the genomic distribution and functional effect of common and rare PTVs. The multi-tissue transcriptome data from the GTEx project provided an unprecedented opportunity to quantify PTV effects across human tissues, and I provided evidence that NMD has heterogeneous effects across tissues in around 18% of PTVs. I showed that while computational predictions of nonsense-mediated decay are valuable, they are far from perfect. Many of the splice-disrupting effects of variants fell in a broad window around exonic splice sites emphasizing the limitations of current variant annotation tools, which mainly focus solely on the "essential" splice sites found at the ends of spliced introns. The results emphasize the need for further research into the underlying molecular mechanisms, as well as the development of improved algorithms for predicting splice disruption and NMD.

A key question for disease studies is whether a PTV truly disrupts cellular function and potentially leads to disease, or whether the PTV effect is tolerated via different mechanisms. Contrary to previous reports in model organisms (Zhou et al., 2011; McAnally and Yampolsky, 2010) I found limited evidence for dosage compensation maintaining normal expression levels of genes affected by heterozygous LoF variants.

Chapter 8

Summary and Future Work

8.1 Summary

GWAS have successfully identified thousands of loci associated to a large number of diseases. Nonetheless, identifying the culprit gene and/or the culprit variant has presented some additional challenges (Maller et al., 2012). PTVs provide a great opportunity for exploring the effects of inhibiting genes in these regions and exome-wide because they are likely to have a loss of function impact, and because they are likely to be rare they suffer less from LD compared to their common variant counterparts that may have been used to localize the association signal (Rivas et al., 2011; MacArthur et al., 2012; Gibson, 2012). To date, many types of drugs including small molecules and protein antibodies serve to inhibit gene function, and anchoring disease association knowledge gained from PTVs may help validate these therapeutic targets (Plenge et al., 2013). The study of PTVs serves as an important complement to GWAS for understanding disease predisposition and potentially for understanding how to treat disease (e.g. *PCSK9* and heart disease).

The motivation for the project behind this thesis is to identify the medical relevance of protein truncating variants in relation to common diseases and biologically relevant quantitative traits, to understand their functional consequences, and to develop methods to detect these associations and characterize their impact.

When I first set out with this goal, large scale population sequencing data was scarce. It was unclear to the community how to approach rare variant association studies. One of my DPhil advisors (Peter Donnelly) was keen to understand the relevance of PTVs across multiple complex traits. My interest in PTVs aligned with his, and it was based on the belief that biological and clinical insights may be gained from their study as I had already found a single PTV with very strong effect in

CARD9 to inflammatory bowel disease prior to the start of my DPhil. Both of my DPhil advisors thought it would be a worthwhile pursuit.

At the start of my DPhil few methods existed for the analysis of rare variants in groups. These included the CMC approach by Li and Leal (2008), the frequency weighted burden test by Madsen and Browning (2009), the variable threshold test (aim was to choose optimal frequency cutoff for rare variant analysis) of Price and Sunyaev (Price et al., 2010), and the binomial C-alpha test of Neale and Rivas, which I co-developed with Benjamin Neale and colleagues. Ben Neale had the intuition of testing for dispersion and I worked on the formulation of a test statistic and bioinformatics development. None of the methods aforementioned really appreciated the need for variant annotation and a focused analysis on PTVs. My appreciation of PTVs largely came from the discovery of the strongly protective *CARD9* splice disrupting variant, c.IVS11+1G>C, which to date is the strongest effect documented to Crohn's disease and ulcerative colitis (Rivas et al., 2011). In addition, germline cancer genetics, as a field, has a remarkable history in understanding the medical consequences of these variants (Rahman et al., 2006). Over time, data was being generated and what became clear to me from the first set of sequencing studies that I participated in was that a naive approach to analyzing these variants, for instance ranking by the number of PTVs in a gene, was not going to work well in exome scale sequencing studies because the number of PTVs observed in any given gene largely depended on the number of samples sequenced (data that motivated this intuition is not presented). The approaches presented in Chapter 3 were developed with my DPhil advors and with Matti Pirinen to address this and as I continued analyzing additional data sets I realized that additional methods development would be required. The presentation order of the methods represent that progression.

First, we set out to develop a method for the analysis of rare protein truncating variants in case-control data sets. Taking the advice from my DPhil advisor (Peter Donnelly) we chose a Bayesian framework because Bayes Factors are likely to be a natural way to rank genes with case-control data. Alternatively, a frequentist approach could have been chosen, but this, as I've learned, has the limitations that sorting by P values, especially for small number of events, is largely dependent on the number of observations available. This approach, which I have referred to in different ways throughout the course of my thesis, including BayesLOF or simply the similar effects model, has over time presented its advantages and limitations. First, the PTV SEM approach mainly works with rare PTVs. Common PTVs, relatedness, or any other covariate for that matter was not addressed when it was developed. Second,

just like frequentist approaches, a substantial number of PTV observations would be required to achieve signal. An attractive property of the approach is that it can easily handle summary level data, like PTV counts. As a result, statistical evaluation of evidence could be continuously updated with new reference data sets like the summary level data made publicly available by the ExAC consortium of $62,500$ exomes from different a variety of disease sequencing studies.

Second, my DPhil advisor (Peter Donnelly) had the insight to develop an approach to assess association between protein truncating variants and quantitative traits. He suggested the method, and with Matti Pirinen, who provided the analytical derivation of the similar effects model, I extended it to grouped effects model to incorporate information about nonsense-mediated decay and I implemented all versions of the approach in the software `MAMBA`. Again, the Bayesian framework was chosen because it had attractive properties for ranking genes with quantitative trait data and we could easily incorporate prior beliefs of the type of effects we were hoping to find from the empirical data sets, i.e. variants with strong effects. For the study of PTVs, I believed, it would be extremely relevant as these variants may have opposite effects within the same gene, which can be predicted with *in silico* predictions of their functional consequences. Hence, I developed a grouped effects model (GEM), and the more relevant grouped effects model using nonsense-mediated decay (GEM-NMD) predictions. I studied the properties of the approach through simulations and showed comparable performance to other published methods. However, as in the case of the Bayesian case control data, this method does have limitations. First, we do not consider relatedness, an issue that may arise in population cohort or clinical collections. Second, traits are required to be transformed to a standard normal distribution - a requirement documented to impact power of statistical tests (Qiu et al., 2013).

Third, it was clear to me that for analyzing rare protein truncating variants we will need to exploit as much information as possible to gain insights into their medical relevance. One way of doing this is to aggregate signal not only aross variants, but also across multiple phenotypes. I set out to develop an approach that would allow me to test for association between protein truncating variants and either multiple disorders or multiple quantitative traits. The intuition, at the time, was that by integrating multiple disease groups we may get a better handle as to whether effects are disease specific or shared and I wanted to develop an association test that would handle these scenarios. Other methods that attempt to exploit this type of information, but in the common variant association framework include the Bayesian unified statistical framework of Stephens (2013). In addition, I wanted to develop a test that allowed

me to exploit the genetic relationship of effects that we have learned from common variant association studies as my belief is that most of these signals will likely follow these patterns. Luckily, at the time that I was developing this method Matti Pirinen was coming up with computationally tractable ways of decomposing the pairwise correlations between traits into the genetic and environmental components. Of course, a penalty may be paid when effects do not follow this assumption, but even in those scenarios we may have power to detect association and it is likely that these events occur infrequetly. I chose a frequentist framework, specifically the dispersion testing framework of Zelterman and Chen (1988). The reason, at the time, and in hindsight probably the better decision, is that it afforded a computationally tractable solution to what seemed to me a very complicated problem and I had been working on the generalized version of the binomial C-alpha test in univariate space with colleagues at Oxford (Clarke et al., 2013). In Stephens (2013) the author discusses one of the challenges with the Bayesian framework including enumerating all possible combination of effects and setting an upper bound (10) on the number of phenotypes that could be tested. Taking up the challenge was a daunting task at the time. The statistical test (an application of Section 4 in Zelterman and Chen (1988)) is computationally tractable, handles large number of phenotypes (> 100), and is flexible allowing the analysis of gene sets, pathways, networks, and can conceivably integrate empirical functional data where estimates of the correlation of genetic effects can be obtained. One of the limitations of the test is that it only incorporates point estimates of the correlation of genetic effects, something that could easily be handled in a Bayesian framework because they can better capture the distribution of the correlation of genetic effects. Second, as it is implemented, more naive approaches for performing meta analysis are supported like Fisher's method and Z-score weighing. Third, it does not handle relatedness, and it requires permutations (which is probably a good property given the poor performance of almost all rare variant approaches when P values are estimated using asymptotic theory). The test that we developed was designed to exploit expected correlation of genetic effects. However, even after detecting association I wanted to be able to estimate the effects in an unbiased fashion. One framework, that was brought to my attention by Matti Pirinen, was to perform clustering. The idea is that we can cluster variant effects in a multidimensional space. Jointly with Matti Pirinen we developed and designed an algorithm that would do this. I applied it to various data sets, and I made software available.

The results presented in Chapter 5 and 6 where I apply the methods to targeted sequencing, exome sequencing, and array based data sets of diseases and quantitative

traits are encouraging. In this thesis when I applied the SEMCC approach to the exome-sequencing data sets I considered the following three thresholds of association: 1) $\log_{10}(\text{BF}) > 2$ as suggestive; 2) $\log_{10}(\text{BF}) > 3$ as substantial; and 4) $\log_{10}(\text{BF}) > 4$ as strong. Whenever I applied the SEMCC approach to targeted sequencing data sets I considered the number of genes analyzed. Furthermore, whenever I applied the frequentist tests (e.g. C-alpha MRP) I considered $\alpha = 2.5 \times 10^{-6}$ to be exome-wide significant. However, I have also reported results for lower thresholds of significance.

I contributed to the analysis of the raw high throughput sequencing data generated for the pooled breast cancer sequencing project and compared PTV allelic counts to reference data sets, which possibly aided Nazneen Rahman (Institute of Cancer Research) in prioritizing *PPM1D* as a worthwhile gene to pursue for follow-up and replication activities. In addition, I updated the results with the Bayesian SEMCC method and new reference data sets and found suggestive signals of PTV association in the data in *SPO11*. After developing the C-alpha MRP test in Chapter 3, I was looking for available sequence data that would be suitable for further evaluation. Interestingly, after applying it to a published targeted sequencing data set of six autoimmune diseases I found evidence of a new association that was missed by the published analysis. Third, I applied the approach to an exome sequencing data set focused on type 2 diabetes. I did not find any evidence of association with the developed approaches, but managed to apply another previously published approach that hinted at some evidence of association in gene sets with prior relationship to diabetes. Fourth, I applied the univariate and multivariate approaches to quantitative trait data including the standard lipid profile identifying new associations of PTVs and across prioritized gene sets, and suggestive evidence of association in data collected for 123 metabolite traits using NMR spectroscopy. I was easily able to conduct application of the methods and tools to very large data sets and obtain sensible results. From the associations reported in this thesis the finding that I am most interested in pursuing is the association of PTVs in *APOC3* to triglyceride levels in type 2 diabetes patients, which appears to exert a stronger effect compared to controls. Multiple lines of evidence including the epidemiological observation that a complication of type 2 diabetes is hypertriglyceridemia (Chehade et al., 2013), and recent results from Mendelian randomization studies showing that triglyceride levels are causal for heart disease (Do et al., 2013), suggests that this is an attractive association to pursue for understanding how to reduce heart disease risk in type 2 diabetes patients.

A second major focus of my thesis was to understand the functional consequences of PTVs. I contributed to the development of new methodology for the interpretation of protein truncating variants. This required three separate approaches presented in Chapter 4: i) understanding how variant annotation works and how it could be improved, which led to me proposing integrating RNA-seq transcript isoform quantification in variant annotation software; ii) assessing allele specific expression to estimate the magnitude of effect and sharing across multiple tissues, which was later extended by my colleague Matti Pirinen in a framework that would borrow information across multiple variants and tissues (Pirinen et al., 2015); and iii) assessing the impact of rare variants proximal to splice junctions on splicing, which required a framework separate from the standard expression quantitative trait loci (eQTL) frameworks. I successfully applied methods ii) and iii) to RNA-seq data sets from the Geuvadis and the GTEx projects and was able to characterize the functional consequences of PTVs across the genome and provide a resource to researchers that could be used for downstream follow-up of their favorite PTV. Ideally, variant annotation including that of PTVs would be based on transcript for the relevant tissue. However, in most cases this information is not known. In practice we can use the information on transcript quantifications to derive better priors of which transcripts and non-regulatory annotations are important. For instance, one can envision using these priors/weights in an aggregate testing framework whereby the grouping is done per transcript, the transcript is assigned a weight, and then combined with all other transcripts in a gene to provide a "gene" measure of association. Furthermore, it has become clear that accurately predicting the deleteriousness of a variant is critical for empowering rare variant association studies (Thormaehlen et al., 2015). More empirical assessment of which variants are likely to be truly deleterious would improve our ability to detect signal of genes that harbour rare variants contributing to disease predisposition. Ultimately, we would need to combine information on the transcript to prioritize these genes combined with high-throughput functional screens evaluating deleteriousness of newly identified alleles.

8.2 Future Work

I have now come to the end of this thesis. At the time of writing various approaches and analyses could not be implemented, and I list them here hoping that it could be extended in the near future.

Bayesian multivariate linear mixed models for rare variant association studies. In Chapter 3 I introduced the C-alpha MRP test, a general framework for rare variant association studies across multiple phenotypes (applicable to multiple diseases or quantitative traits). It is now clear to me that the methods developed in this thesis have limitations that could easily be addressed in another statistical framework. Linear mixed models are attractive in that they can be developed to take into account factors like relatedness, and that uncertainty in the distribution of effect sizes can be considered. Furthermore, a Bayesian framework could better capture the distribution of the correlation of genetic effects.

Evaluating the optimal number of clusters k to fit to the data. In Chapter 3 I introduced an algorithm for clustering variant effects and estimating the alternative mixture components. Unfortunately, I was not able to develop a principled algorithm that could learn the optimal number of clusters k to fit the data. In Neale et al. (2011) a Gradient Diagnostics algorithm was introduced that makes this choice. The reversible-jump Markov chain Monte Carlo was brought to my attention as a possible solution[1].

General framework for jointly assessing the functional consequenes of regulatory and protein truncating variants: from DNA to protein to downstream causal effects. In Chapter 4 I introduced approaches for *in silico* annotation of protein truncating variants using RNA-seq transcript isoform quantification. It is clear to me that this is extremely important for rare variant association testing and inferring the consequences of newly discovered variants. It should be incorporated in new association approaches and in standard annotation pipelines. Furthermore, I introduced a simple framework for assessing the allele-specific expression effects of nonsense and frameshift indel variants and the splice disruption potential of variants proximal to splice junctions. During my thesis work I only regarded the transcriptome (the collection of expressed mRNAs). An area where this could be extended is the joint assessment of the functional consequences of rare regulatory and protein truncating variants across different levels of molecular biology hierarchy, e.g. DNA, RNA, protein, cell-signaling.

[1] `http://en.wikipedia.org/wiki/Reversible-jump_Markov_chain_Monte_Carlo`

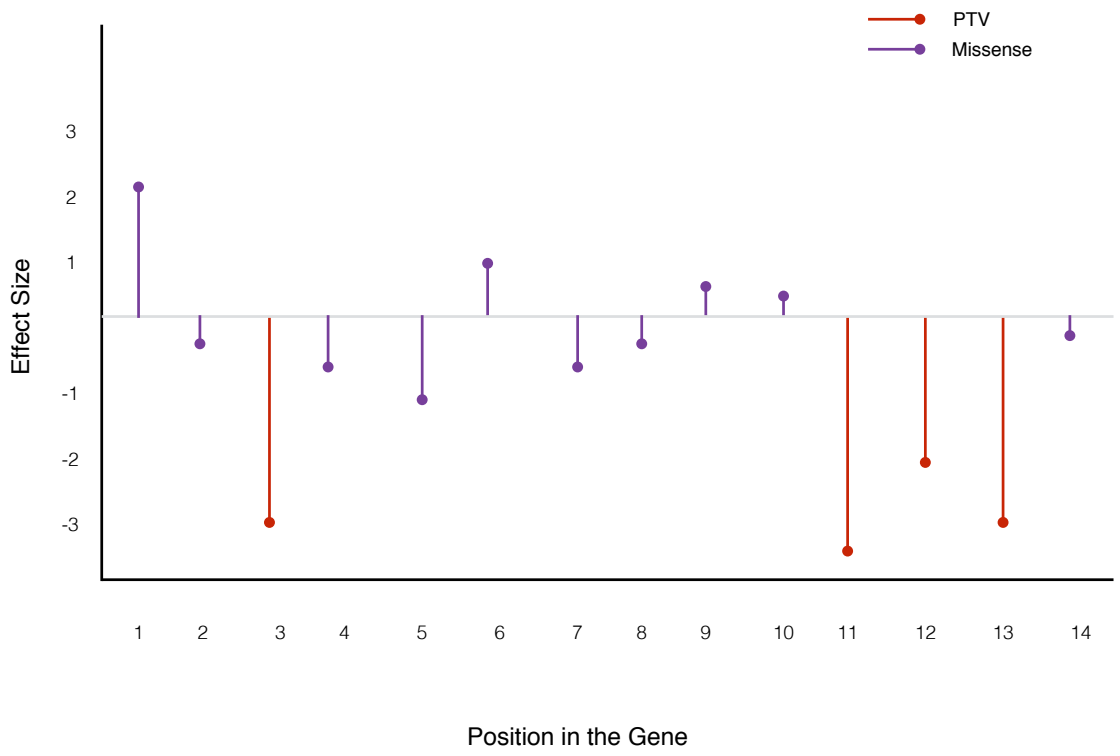


Figure 8.1: A diagram of genome-editing for experimental quantification of the functional consequence of DNA sequence variants. Genome-editing, for example by multiplex homology-directed repair, in conjuncion with functional assay data, for example a splice disruption assay, can be used to obtain an experimental readout of the functional consequences of DNA sequence variants - a critical component in understanding the contribution of rare variants to disease predisposition. Figure adapted from (Findlay et al., 2014).

Genome saturation and integrating functional assay data. More recently, new technologies like TALENs, Zinc finger nucleases, CRISPR/CAS9 (Cong et al., 2013), and experimental techniques have raised the possibility of obtaining experimental data to assess the functional consequences of DNA sequence variation with a functional assay (Patwardhan et al., 2009; Findlay et al., 2014). This experimental approach was recently applied in the setting of type 2 diabetes where the authors used a human adipocyte differentiation assay to assess the functional consequence of rare variants in *PPARG* identified from a large-scale sequencing study of over 19,000 individuals (Majithia et al., 2014). Then, the authors used this data to inform the aggregate testing of rare variants in *PPARG* identifying an association (OR = 7.22, $P = 0.005$) that was missed in the initial analysis without the experimental data (OR = 1.35, $P = 0.17$). In future work I would like to integrate data generated from the application of genome saturation techniques in rare variant association studies.

Data on large number of diseases and relevant quantitative trait measurements are being generated. In addition, as we move to whole genome sequencing data sets and large population biobanks like the UK Biobank I hope that the work presented in this thesis will help advance the understanding of common human diseases and could possibly be used to identify new drug targets for therapeutic intervention.

Appendices

List of Abbreviations

1KG 1000 Genomes
ADPSBQ adipose subcutaneous
AITD autoimmune thyroid disease
ARTTBL artery tibial
ASE allele specific expression
BF Bayes Factor
biLMM bivariate linear mixed model
C-alpha MRP test C-alpha multiple rare variants and phenotypes test
CD Crohn's disease
CeD coeliac disease
CNV copy number variation
CVAS common variant association studies
EBV Epstein-Barr virus
EM Expectation Maximization
eQTL expression Quantitative Trait Loci
ExAC Exome Aggregation Consortium
GEM grouped effects model
GERP Genomic Evolutionary Rate Profiling
GO gene ontology
GoT2D Genetics of Type 2 Diabetes
GPC Genetic Power Calculator
GQ genotype quality
GTEx Genotype Tissue Expression
GTM grouped tissue model
GTM* hierarchical grouped tissue model
GWAS genome wide association studies
GWG Genotyping Working Group
HapMap Haplotype Map

HDL-C high-density lipoprotein cholesterol
HGMD Human Gene Mutation Database
HRTLV heart left ventricle
ITM independent tissue model
kb kilobases
LCL lymphoblastoid cell line
LD linkage disequilibrium
LDL-C low-density lipoprotein cholesterol
locdb locus database
LoF loss of function
LUNG Lung
MAF minor allele frequency
MAMBA Modeling And Mining Big DatA
Mb megabases
MCMC Markov Chain Monte Carlo
MS multiple sclerosis
MSCLSK skeletal muscle tissue
MWW Mann-Whitney-Wilcoxon
NCBI National Center for Biotechnology Information
NERVET nerve tibial
ng nanogram
NGS next-generation sequencing
NHGRI National Human Genome Research Institute
NMD nonsense-mediated decay
NMR nuclear magnetic resonance
OR odds ratio
PD psoriasis
PTVs protein truncating variants
QC quality control
QT quantitative trait
RNA-seq RNA sequencing
RPKM reads per kilobase of transcript per million reads mapped
RVAS rare variant association studies
s.d. standard deviation
SDM splice disruption model
SEM similar effects model

SEMCC similar effects model for case-control

SFS site frequency spectra

SKAT Sequence Kernel Association Test

SKINS skin sun exposed

SMP statistics/matrix/permutation

SNPs single nucleotide polymorphisms

SNVs single nucleotide variants

T1D type 1 diabetes

T2D type 2 diabetes

T2D-GENES Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples

TG triglycerides

THYROID thyroid

Ti/Tv transition to tansversion ratio

TNF tumor necrosis factor

VCF Variant Call Format

VEP Variant Effect Predictor

VQSR Variant Quality Score Recalibration

WHLBLD whole blood

Supplementary Tables

Gene Set	Genes
Adipocytokines	*ADIPOQ,AGT,APLN,AZGP1,C3,FGF21,IL10,IL13,IL1B,IL6, IL8,LEP,CCL2,MIF,NAMPT,RETN,TGFB1,TNF,VEGFA*
Adiponectin	*ACACA,ACACB,ADIPOQ,ADIPOR1,ADIPOR2,APPL1, SLC2A4,PRKAA1,PRKAA2,PRKAB1,PRKAB2,PRKAG1, PRKAG2,PRKAG3,RAB5A,STK11*
Calcium-influx	*ABCC8,KCNJ11,SSTR2,SSTR3,KCNJ3,KCNJ6,NALCN, CACNA1H,CACNA1D,CACNA1A,CACNA1C,CACNA1E, SCN8A,SCN9A,SCN1B,SCN3B,KCNMA1,KCNB1,KCNB2, KCNQ1,KCNH2,KCNJ12,KCNJ4,KCNJ15,KCNN3,KCNN4, KCNN1,RYR2,ITPR1,ITPR2,ITPR3,CLCN3*
Cell cycle	*CCND1,CCND2,CCND3,CDK4,CDK6,RB1,RBL1,RBL2, ABL1,HDAC1,HDAC2,E2F1,E2F2,E2F3,E2F4, E2F5,TFDP1,TFDP2,GSK3B,TGFB1,TGFB2,TGFB3, SMAD2,SMAD3,SMAD4,MYC,ZBTB17,CDKN2A, CDKN2B,CDKN2C,CDKN2D,CDKN1B,CDKN1C, CDKN1A,CCNE1,CCNE2,CDK2,SKP1,CUL1,RBX1, SKP2,CCNA2,CCNA1,CDC6,CDC45,CDC7,DBF4,CDK1, CCNB1,CCNB2,CCNB3,CDC25B,CDC25C, YWHAZ,YWHAB,YWHAQ,YWHAE,YWHAH,YWHAG, PLK1,WEE1,WEE2,PKMYT1,CCNH,CDK7, ANAPC1,ANAPC2,CDC27,ANAPC4,ANAPC5,CDC16, ANAPC7,CDC23,ANAPC10,ANAPC11,CDC26, ANAPC13,CDC20,PTTG1,PTTG2,ESPL1,SMC1A, SMC1B,SMC3,STAG2,STAG1,RAD21,TTK,BUB1,BUB3, BUB1B,MAD1L1,MAD2L1,MAD2L2,FZR1,CDC14B, CDC14A,ATR,ATM,TP53,CHEK1,CHEK2,CREBBP, EP300,PRKDC,MDM2,GADD45A,GADD45B,GADD45G, PCNA,SFN,CDC25A,ORC1,ORC2,ORC3,ORC4, ORC5,ORC6,MCM2,MCM3,MCM4,MCM5,MCM6,MCM7*

CREBBP coTF	*SREBF1,IRF9,KLF5,SERTAD1,AC091153.1,NCOA2, TDG,FHL2,ZBTB2,RBBP4,ONECUT1,PLAGL1, VDR,HDAC3,EBF1,ACVR1,CSNK2A1P,HOXB4,NCOA6, NCOA3,SPIB,HOXB7,RELA,HOXA11,IKBKG, MAML2,IFNAR2,CCNC,CREM,GPBP1,IRF7,ZBTB17, CRX,RUVBL1,SOX9,FGFR1,EWSR1,SMARCB1, RPS6KA5,FOXO1,MAML3,KAT2A,AP1B1,TP53, NFIC,SS18L1,HOXB1,HLF,H3F3A,DDX5,CUX1,RUNX1, GABPA,SRF,CDC25B,MSH6,SREBF2,TP73,MED6, CREBBP,STAT4,NFATC4,ABCC9,NMI,PROX1,HMGA1, ATF1,MED1,MAF,GAK,HOXA10,POU2F3,CHUK, SERTAD2,CDK8,IKBKB,RARA,SMAD1,EIF2B1,SMARCA4, CITED4,NPAS2,KLF13,CNOT3,HDAC1,CITED2, MYBL2,MED21,GMEB1,TCF7L2,RXRG,CTBP1,KLF4,UBTF, NLK,PIAS1,SERTAD3,NKX21,XRCC6,CEBPB,KAT5, NAP1L1,HTT,BRCA1,MAML1,EGR1,TAF7L,RBPJ,KHDRBS1, MDC1,ALX1,TGS1,RPA2,MED24,RPS6KA1,FOS, PHOX2A,PIAS3,MDM2,HNF1A,ZNF639,ELK1,FOXM1,ETS2, MTDH,CRTC2,NUP98,HOXA9,HIF1A,ING1,PCMT1, AIRE,SUV39H1,TCF3,PPARGC1A,RPS6KA3,GMEB2, CDX2,ATF4,ETS1,KAT2B,TRERF1,SND1,SH3GL1, PAX5,NCOA1,CREB1,TCF12,EP300,HSF1,DAXX, TRIP10,HIPK2,AR,PPARG,NOTCH1,HOXD4,E2F3, POLR2A,NOTCH3,MAST1,JUN,CENPJ,NFE2L2,RXRA, N4BP2,DACH1,PPARA,SMAD4,SRCAP,CTNNB1, SMAD3,POU1F1,MYOD1,NEUROG1,NFATC2,PRKCD,STAT2, H3F3B,SNW1,MGMT,MYBL1,MLL,KLF1,STAT3, TRIM21,GATA1,HOXD10,MYC,MECOM,RBCK1,THRA,CSNK2A2, DHX9,CDH2,NR3C1,E2F1,ATF3,PRRX2,E2F5, WRB,NFE2,ESR1,CDKN1A,KPNA2,MED15,YWHAH,NOTCH2,GLI3, RPS6KA2,MYB,HNF1B,HNF4A,CSK,NFIA,SPI1, AP2A2,MSH2,CARM1,TACC2,GTF2B,HOXB2,RBBP7,MBD2,WT1, HOXB3,IRF3,HOXB6,FOXO4,ATF2,NOTCH4, HOXB9,STAT1,HMX3,STAT5A,MSX1,GCM1,STAT5B,GATA2,SMAD2, ABCA1,STAT6,CITED1,BCL3,MCM7,CDK5RAP3, CAMK4,PML,SNIP1,TRIP4,TRAM2,ACTA2,MAFG*
ER Stress	*EIF2AK3,ERN1,ATF6,CCDC88B,EIF2S1,XBP1, ATF4,ATF5,PPP1R15A,DDIT3,BCL2,MAPK8,MBTPS1, MBTPS2,ATF6B,NR0B2,DERL1,UCHL1,EIF2AK4, EIF2AK1,EIF2AK2,SYVN1,WFS1,HSPA5,ATP2A2,PDX1, MAFA,SEC61A1,SEC61A2,SEC61B,SEC61G,SIAH1,SIAH2*
Inflammatory Cytokines	*AGT,GPR77,CCR2,CCR5,CD74,CHUK,CRADD,CXCR1, CXCR2,CXCR3,CXCR4,ECSIT,FADD,FGF21,FGFR1, FGFR2,FGFR3,IKBKB,IKBKE,IKBKG,IL10,IL10RA, IL10RB,IL13,IL13RA1,IL13RA2,IL1B,IL1R1,IL6,IL6R,IL6ST, IL8,IRAK1,IRAK2,JAK1,JAK2,JAK3,KDR,KLB,LAP3, LTBP1,LTBP2,LTBP3,LTBP4,MADD,MAP2K1,MAP2K2, MAP3K1,MAP3K14,MAP4K1,MAPK14,MAPK8,CCL2, NAMPT,NFKB1,NFKBIA,NMNAT1,NMNAT2,NMNAT3, PIK3C2A,PIK3C2B,PIK3CA,RHOA,RIPK1,RIPK2,RIPK3, RIPK4,SHC1,SHC2,SIRT1,SIRT2,SIRT3,SIRT4,SIRT5, SIRT6,SIRT7,SMAD2,SMAD3,SMAD4,SMAD6,SMAD7, STAT3,TAB1,MAP3K7,TGFB1,TGFB2,TGFB3,LEFTY2, TGFBR1,TGFBR2,TNF,TNFRSF1A,TNFRSF1B,TOLLIP, TRADD,TRAF2,TRAF6,TYK2,VEGFA,XIAP*
Insulin	*STX1A,STX1B,SNAP25,VAMP2,SYT7,SYT5,SENP1, ATF6,XBP1,ERN1,EIF2AK3,ATF4,EIF2A*
Leptin	*MAPK1,MAPK3,GRB2,IRS1,JAK1,JAK2,LEP,LEPR, PTPN11,SOCS3,STAT3,STAT5A*
Mtor	*AKT1,EIF3A,EIF4A1,EIF4A2,EIF4B,EIF4E,EIF4EBP1, EIF4G1,EIF4G2,EIF4G3,FKBP1A,MKNK1,MTOR,PDK2, PDPK1,PIK3CA,PIK3R1,PPP2CA,PTEN,RPS6,RPS6KB1, TSC1,TSC2*

RAS	*DIRAS1,DIRAS2,DIRAS3,ERAS,GEM,HRAS,KRAS,MRAS, NKIRAS1,NKIRAS2,NRAS,RALA,RALB,RAP1A,RAP1B, RAP2A,RAP2B,RAP2C,RASD1,RASD2,RASL10A,RASL10B, RASL11A,RASL11B,RASL12,REM1,REM2,RERG,RERGL, RRAD,RRAS,RRAS2,RASSF1,RASSF2,RASSF3,RASSF4, RASSF5,RASSF6,RASSF7,RASSF8,RASSF9,RASSF10*
Stimulation of insulin secretion	*SLC2A1,SLC2A3,GCK,PKLR,PKM2,DLAT,DLD,PDHA1, PDHB,PDHX,PDP1,CS,ACO1,ACO2,IDH2,OGDH, DLST,SUCLA2,SUCLG1,SUCLG2,SDHA,SDHB,SDHC, SDHD,FH,MDH1,MDH2*
Wnt	*APC,AXIN1,BTRC,CCND1,CREBBP,CSNK1A1,CSNK1D, CSNK2A1,CTBP1,CTNNB1,DVL1,FRAT1, FZD1,GSK3B,HDAC1,LEF1,MAP3K7,MYC,NLK,PPARD, PPP2CA,SMAD4,TAB1,TLE1,WIF1,WNT1*
Beta cell GPCRs	*GPR119,ADRB2,ADRA2A,MTNR1A,MTNR1B,HTR2B, HTR1D,CHRM3,SSTR2,GLP1R,FFAR1, GCGR,ADRB1,GIPR,FFAR3,FFAR2,O3FAR1, ADCYAP1R1,VIPR2,NPY1R,GHSR,KISS1R, CCKAR,CNR1,P2RY12,ADRA2B*
Cell cycle G1 S	*ABL1,ATM,ATR,CCNA1,CCND1,CCNE1,CDC25A, CDK1,CDK2,CDK4,CDK6,CDKN1A,CDKN1B, CDKN2A,CDKN2B,DHFR,E2F1,GSK3B,HDAC1, RB1,SKP2,SMAD3,SMAD4,TFDP1,TGFB1,TGFB2, TGFB3,TP53*
Monogenic All	*ABCC8,AGPAT2,AIRE,AKT2,ALMS1,ARL6,BBS1, BBS10,BBS12,BBS2,BBS4,BBS5,BBS7,BBS9,BLK,BSCL2, DCAF17,CAV1,CEL,CEP290,CFTR,CIDEC,CISD2, DMPK,EIF2AK3,FOXP3,GATA4,GATA6,GCK,GLIS3,GLUD1, HADH,HNF1A,HNF1B,HNF4A,IER3IP1,INS,INSR, ISL1,KCNJ11,KLF11,LEP,LEPR,LMNA,MC4R,MKKS,MKS1, MNX1,NEUROD1,NEUROG3,PAX4,PAX6,PCNT, PCSK1,PDX1,PIK3CA,PIK3R1,PLAGL1,PLIN1,POLD1,POMC, PPARG,PPP1R3A,PRSS1,PTEN,PTF1A,PTRF, RFX6,SH2B1,SIRT1,SLC16A1,SLC19A2,SLC2A2,SPINK1,TRIM32, TTC8,UCP2,WFS1,WRN,ZFP57,ZMPSTE24*
Monogenic Primary	*ABCC8,ALMS1,CEL,CISD2,EIF2AK3,GATA4, GATA6,GCK,GLIS3,GLUD1,HADH,HNF1A,HNF1B, HNF4A,INS, INSR,KCNJ11,LMNA,MNX1,NEUROD1,NEUROG3, PDX1,PPARG,PTF1A,RFX6,SLC2A2,UCP2,WFS1*
Monogenic OMIM	*NEUROD1,KCNJ11,GCK,HNF4A,BLK,HNF1B,KLF11,CEL,HNF1A,PDX1,ABCC8,PAX4*

Table 1: Type 2 diabetes premium gene sets

Metabolite Label	Description
Alb	Albumin
XXL-VLDL-PL	Phospholipids-in-chylomicrons-and-extremely-large-VLDL
XXL-VLDL-L	Total-lipids-in-chylomicrons-and-extremely-large-VLDL
XXL-VLDL-P	Concentration-of-chylomicrons-and-extremely-large-VLDL-particles
XL-VLDL-PL	Phospholipids-in-very-large-VLDL
XL-VLDL-TG	Triglycerides-in-very-large-VLDL
XL-VLDL-L	Total-lipids-in-very-large-VLDL
XL-VLDL-P	Concentration-of-very-large-VLDL-particles
L-VLDL-C	Total-cholesterol-in-large-VLDL
L-VLDL-FC	Free-cholesterol-in-large-VLDL
L-VLDL-PL	Phospholipids-in-large-VLDL
L-VLDL-TG	Triglycerides-in-large-VLDL

L-VLDL-CE	Cholesterol-esters-in-large-VLDL
L-VLDL-L	Total-lipids-in-large-VLDL
L-VLDL-P	Concentration-of-large-VLDL-particles
M-VLDL-C	Total-cholesterol-in-medium-VLDL
M-VLDL-FC	Free-cholesterol-in-medium-VLDL
M-VLDL-PL	Phospholipids-in-medium-VLDL
M-VLDL-TG	Triglycerides-in-medium-VLDL
M-VLDL-CE	Cholesterol-esters-in-medium-VLDL
M-VLDL-L	Total-lipids-in-medium-VLDL
M-VLDL-P	Concentration-of-medium-VLDL-particles
S-VLDL-C	Total-cholesterol-in-small-VLDL
S-VLDL-FC	Free-cholesterol-in-small-VLDL
S-VLDL-PL	Phospholipids-in-small-VLDL
S-VLDL-TG	Triglycerides-in-small-VLDL
S-VLDL-L	Total-lipids-in-small-VLDL
S-VLDL-P	Concentration-of-small-VLDL-particles
XS-VLDL-PL	Phospholipids-in-very-small-VLDL
XS-VLDL-TG	Triglycerides-in-very-small-VLDL
XS-VLDL-L	Total-lipids-in-very-small-VLDL
XS-VLDL-P	Concentration-of-very-small-VLDL-particles
IDL-FC	Free-cholesterol-in-IDL
IDL-PL	Phospholipids-in-IDL
IDL-L	Total-lipids-in-IDL
IDL-P	Concentration-of-IDL-particles
L-LDL-C	Total-cholesterol-in-large-LDL
L-LDL-FC	Free-cholesterol-in-large-LDL
L-LDL-PL	Phospholipids-in-large-LDL
L-LDL-CE	Cholesterol-esters-in-large-LDL
L-LDL-L	Total-lipids-in-large-LDL
L-LDL-P	Concentration-of-large-LDL-particles
M-LDL-C	Total-cholesterol-in-medium-LDL
M-LDL-PL	Phospholipids-in-medium-LDL
M-LDL-CE	Cholesterol-esters-in-medium-LDL
M-LDL-L	Total-lipids-in-medium-LDL
M-LDL-P	Concentration-of-medium-LDL-particles
S-LDL-C	Total-cholesterol-in-small-LDL
S-LDL-L	Total-lipids-in-small-LDL
S-LDL-P	Concentration-of-small-LDL-particles
XL-HDL-C	Total-cholesterol-in-very-large-HDL
XL-HDL-FC	Free-cholesterol-in-very-large-HDL
XL-HDL-PL	Phospholipids-in-very-large-HDL
XL-HDL-TG	Triglycerides-in-very-large-HDL
XL-HDL-CE	Cholesterol-esters-in-very-large-HDL
XL-HDL-L	Total-lipids-in-very-large-HDL
XL-HDL-P	Concentration-of-very-large-HDL-particles
L-HDL-C	Total-cholesterol-in-large-HDL
L-HDL-FC	Free-cholesterol-in-large-HDL
L-HDL-PL	Phospholipids-in-large-HDL
L-HDL-CE	Cholesterol-esters-in-large-HDL
L-HDL-L	Total-lipids-in-large-HDL
L-HDL-P	Concentration-of-large-HDL-particles
M-HDL-C	Total-cholesterol-in-medium-HDL
M-HDL-FC	Free-cholesterol-in-medium-HDL
M-HDL-PL	Phospholipids-in-medium-HDL

M-HDL-CE	Cholesterol-esters-in-medium-HDL
M-HDL-L	Total-lipids-in-medium-HDL
M-HDL-P	Concentration-of-medium-HDL-particles
S-HDL-TG	Triglycerides-in-small-HDL
S-HDL-L	Total-lipids-in-small-HDL
S-HDL-P	Concentration-of-small-HDL-particles
XXL-VLDL-TG	Triglycerides-in-chylomicrons-and-extremely-large-VLDL
VLDL-TG	Triglycerides-in-VLDL
IDL-TG	Triglycerides-in-IDL
IDL-C	Total-cholesterol-in-IDL
LDL-C	Total-cholesterol-in-LDL
HDL-C	Total-cholesterol-in-HDL
Serum-TG	Serum-total-triglycerides
Serum-C	Serum-total-cholesterol
VLDL-D	Mean-diameter-for-VLDL-particles
LDL-D	Mean-diameter-for-LDL-particles
HDL-D	Mean-diameter-for-HDL-particles
VLDL-TG-eFR	Triglycerides-in-VLDL
IDL-C-eFR	Total-cholesterol-in-IDL
LDL-C-eFR	Total-cholesterol-in-LDL
HDL2-C	Total-cholesterol-in-HDL2
ApoA1	Apolipoprotein-A-I
ApoB	Apolipoprotein-B
ApoBtoApoA1	Apolipoprotein-B-by-apolipoprotein-A-I
HDL3-C	Total-cholesterol-in-HDL3
bOHBut	3-hydroxybutyrate
Ace	Acetate
AcAce	Acetoacetate
Ala	Alanine
MobCH2	CH2-groups-of-mobile-lipids
MobCH3	CH3-groups-of-mobile-lipids
Cit	Citrate
Crea	Creatinine
MobCH	Double-bond-protons-of-mobile-lipids
Glc	Glucose
Gln	Glutamine
Glycerol	Glycerol
Gly	Glycine
Gp	Glycoprotein-acetyls,-mainly-a1-acid-glycoprotein
His	Histidine
Ile	Isoleucine
Lac	Lactate
Leu	Leucine
Phe	Phenylalanine
Pyr	Pyruvate
Tyr	Tyrosine
Urea	Urea
Val	Valine
EstC	Esterified-cholesterol
FreeC	Free-cholesterol
FAw3	Omega-3-fatty-acids
FAw6	Omega-6-fatty-acids
FAw-9S	Omega-7,-omega-9-and-saturated-fatty-acids
TotFA	Total-fatty-acids

LA	18:2,-linoleic-acid
otPUFA	Other-polyunsaturated-fatty-acids-than-18:2
DHA	22:6,-docosahexaenoic-acid
MUFA	Monounsaturated-fatty-acids;-16:1,-18:1
TotPG	Total-phosphoglycerides
PC	Phosphatidylcholine-and-other-cholines
SM	Sphingomyelins
FAw3toFA	Ratio-of-omega-3-fatty-acids-to-total-fatty-acids
FAw6toFA	Ratio-of-omega-6-fatty-acids-to-total-fatty-acids
FAw79StoFA	Ratio-of-omega-7,-omega-9-and-saturated-fatty-acids-to-total-fatty-acids
CH2inFA	Average-number-of-methylene-groups-in-a-fatty-acid-chain
TGtoPG	Ratio-of-triglycerides-to-phosphoglycerides
CH2toDB	Average-number-of-methylene-groups-per-a-double-bond
DBinFA	Average-number-of-double-bonds-in-a-fatty-acid-chain
BIStoDB	Ratio-of-bisallylic-groups-to-double-bonds
BIStoFA	Ratio-of-bisallylic-groups-to-total-fatty-acids
FALen	Description-of-average-fatty-acid-chain-length,-not-actual-carbon-number

Table 2: Description of metabolites used in the analysis

Bibliography

AC't Hoen, P., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., Laros, J. F., Buermans, H. P., Karlberg, O., Brännvall, M., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology*.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9.

Akbari, M. R., Lepage, P., Rosen, B., McLaughlin, J., Risch, H., Minden, M., and Narod, S. A. (2013). *PPM1D* mutations in circulating white blood cells and the risk for ovarian cancer. *Journal of the National Cancer Institute*, page 323.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683.

Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44:293–308.

Aslanidis, C., Ries, S., Fehringer, P., Büchler, C., Klima, H., and Schmitz, G. (1996). Genetic and biochemical evidence that CESD and Wolman disease are distinguished by residual lysosomal acid lipase activity. *Genomics*, 33(1):85–93.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, pages 11–10.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619.

Bateman, J. F., Freddi, S., Nattrass, G., and Savarirayan, R. (2003). Tissue-specific RNA surveillance? Nonsense-mediated mRNA decay causes collagen X haploinsufficiency in Schmid metaphyseal chondrodysplasia cartilage. *Human Molecular Genetics*, 12(3):217–225.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24.

Baudat, F., Manova, K., Yuen, J. P., Jasin, M., and Keeney, S. (2000). Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Molecular Cell*, 6(5):989–998.

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11):2692–2703.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D., and Smith, K. (2011). Cython: the best of both worlds. *Computing in Science Engineering*, 13(2):31–39.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Borg, I. and Groenen, P. (1997). Modern Multidimensional Scaling. *NY Springer*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. *Statistics Department University of California Berkeley, CA, USA*.

Burge, C. B., Tuschl, T., and Sharp, P. A. (1999). 20 Splicing of Precursors to mRNAs by the Spliceosomes. *Cold Spring Harbor Monograph Archive*, 37:525–560.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning*, pages 161–168. ACM.

Casadio, A., Longman, D., Hug, N., Delavaine, L., Vallejos Baier, R., Alonso, C. R., and Cáceres, J. F. (2015). Identification and characterization of novel factors that act in the nonsense-mediated mRNA decay pathway in nematodes, flies and mammals. *EMBO Reports*, 16(1):71–78.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chehade, J. M., Gladysz, M., and Mooradian, A. D. (2013). Dyslipidemia in type 2 diabetes: prevalence, pathophysiology, and management. *Drugs*, 73(4):327–339.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology*, 40(6):1652–1666.

Clarke, G. M., Rivas, M. A., and Morris, A. P. (2013). A flexible approach for the analysis of rare variants allowing for a mixture of effects on binary or quantitative traits. *PLoS Genetics*, 9(8):e1003694.

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., and Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nature Genetics*, 37(2):161–5.

Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H., and Hobbs, H. H. (2006). Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272.

Cohen, J. C., Stender, S., and Hobbs, H. H. (2014). *APOC3*, coronary disease, and complexities of mendelian randomization. *Cell Metabolism*, 20(3):387–389.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., et al. (1998). New goals for the US human genome project: 1998-2003. *Science*, 282(5389):682–689.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823.

Consortium, I. H. et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Consortium, I. H. G. S. et al. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913.

Corsini, A., Fantappie, S., Granata, A., Bernini, F., Catapano, A., Fumagalli, R., Romano, L., and Romano, C. (1989). Binding-defective low-density lipoprotein in family with hypercholesterolaemia. *The Lancet*, 333(8638):623.

Corvelo, A., Hallegger, M., Smith, C. W., and Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. *PLoS Computational Biology*, 6(11):e1001016.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1:131.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*.

Crunkhorn, S. (2012). Trial watch: PCSK9 antibody reduces LDL cholesterol. *Nature Reviews Drug Discovery*, 11(1):11.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics*, pages 323–333.

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–1223.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8.

Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., Merker, J. D., Goldfeder, R. L., Enns, G. M., David, S. P., et al. (2014). Clinical interpretation and implications of whole-genome sequencing. *JAMA*, 311(10):1035–1045.

Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, 45(11):1345–1352.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Faustino, N. A. and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes & Development*, 17(4):419–437.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, 513(7516):120–123.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Flanagan, S. E., Patch, A.-M. M., and Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, 14(4):533–537.

Flannick, J., Beer, N. L., Bick, A. G., Agarwala, V., Molnes, J., Gupta, N., Burtt, N. P., Florez, J. C., Meigs, J. B., Taylor, H., et al. (2013). Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nature Genetics*, 45(11):1380–1385.

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B., Grarup, N., Burtt, N. P., Mahajan, A., Fuchsberger, C., Atzmon, G., Benediktsson, R., et al. (2014). Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nature Genetics*, 46(4):357–363.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Friedewald, W. T., Levy, R. I., and Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18(6):499–502.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., ODonovan, M. C., Owen, M. J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, 91(4):597–607.

Fu, W., OConnor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–83.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145.

Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*.

Glueck, C., Fallat, R., Millett, F., Gartside, P., Elston, R., and Go, R. (1975). Familial hyper-alpha-lipoproteinemia: studies in eighteen kindreds. *Metabolism*, 24(11):1243–1265.

Goh, L. and Yap, V. B. (2009). Effects of normalization on quantitative traits in association test. *BMC Bioinformatics*, 10(1):415.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M.,

Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., and Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774.

Hartl, D. L., Clark, A. G., et al. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland.

Hegele, R. A. (2001). Monogenic dyslipidemias: window on determinants of plasma lipoprotein metabolism. *American Journal of Human Genetics*, 69(6):1161.

Heij, C., De Boer, P., Franses, P. H., Kloek, T., Van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Human Molecular Genetics*, 18(R1):R1–R8.

Hentze, M. W. and Kulozik, A. E. (1999). A perfect message: RNA surveillance and nonsense-mediated decay. *Cell*, 96(3):307–310.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Hofker, M. H. (2010). APOC3 null mutation affects lipoprotein profile APOC3 deficiency: from mice to man. *European Journal of Human Genetics*, 18(1):1–2.

Holbrook, J. A., Neu-Yilik, G., Hentze, M. W., and Kulozik, A. E. (2004). Nonsense-mediated decay approaches the clinic. *Nature Genetics*, 36(8):801–808.

Hopper, J. L., Southey, M. C., Dite, G. S., Jolley, D. J., Giles, G. G., McCredie, M. R., Easton, D. F., and Venter, D. J. (1999). Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in *BRCA1* and *BRCA2*. *Cancer Epidemiology Biomarkers & Prevention*, 8(9):741–747.

Huang, N., Lee, I., Marcotte, E. M., and Hurles, M. E. (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics*, 6(10):e1001154.

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M., Calvo, F., Eerola, I., Gerhard, D. S., et al. (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

Hunt, K. A., Mistry, V., Bockett, N. A., Ahmad, T., Ban, M., Barker, J. N., Barrett, J. C., Blackburn, H., Brand, O., Burren, O., et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*.

Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L.-P., Oksala, N., Laurila, P.-P., Kangas, A. J., Soininen, P., Savolainen, M. J., Viikari, J., et al. (2012). Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genetics*, 8(8):e1002907.

Isidor, B., Lindenbaum, P., Pichon, O., Bezieau, S., Dina, C., Jacquemont, S., Martin-Coignard, D., Thauvin-Robinet, C., Le Merrer, M., Mandel, J. L., David, A., Faivre, L., Cormier-Daire, V., Redon, R., and Le Caignec, C. (2011). Truncating mutations in the last exon of *NOTCH2* cause a rare skeletal disorder with osteoporosis. *Nature Genetics*, 43(4):306–8.

Isis Pharmaceuticals (2014). Isis Pharmaceuticals Initiates Phase 3 Study of ISIS-APOCIII Rx in Patients with FCS. *Isis Pharmaceuticals*.

Jocelyn Kaiser (2007). A plan to capture human diversity in 1000 genomes. *Science*, 21:1842.

Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R., Martins, R. A., Kennedy, B. A., Hassell, R. G., Visser, M. E., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, 42(8):684–687.

Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G., and Tybjærg-Hansen, A. (2014). Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. *New England Journal of Medicine*, 371(1):32–41.

Kaiser, J. (2014). The hunt for missing genes. *Science*, 344(6185):687–689.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., et al. (2008). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics*, 41(1):56–65.

Keeney, S., Giroux, C. N., and Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, 88(3):375–384.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743.

Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44(3):269–276.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Kukurba, K. R., Zhang, R., Li, X., Smith, K. S., Knowles, D. A., Tan, M. H., Piskol, R., Lek, M., Snyder, M., MacArthur, D. G., et al. (2014). Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genetics*, 10(5):e1004304.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–81.

Laakso, M. (1999). Hyperglycemia and cardiovascular disease in type 2 diabetes. *Diabetes*, 48(5):937–942.

Lademann, U., Kallunki, T., and Jäättelä, M. (2001). A20 zinc finger protein inhibits TNF-induced apoptosis and stress response early in the signaling cascades and independently of binding to TRAF2 or 14-3-3 proteins. *Cell Death and Differentiation*, 8(3):265–272.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, 274(5287):536–539.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z.-Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *The American Journal of Human Genetics*, 94(2):233–245.

Lappalainen, T., Sammeth, M., Friedlander, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Consortium, T. G., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Hasler, R., Syvanen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–75.

Lejeune, F. and Maquat, L. E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Current Opinion in Cell Biology*, 17(3):309–315.

Lewis, G. F., Carpentier, A., Adeli, K., and Giacca, A. (2002). Disordered fat storage and mobilization in the pathogenesis of insulin resistance and type 2 diabetes. *Endocrine Reviews*, 23(2):201–229.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Lim, E. T., Liu, Y. P., Chan, Y., Tiinamaija, T., Käräjämäki, A., Madsen, E., Altshuler, D. M., Raychaudhuri, S., Groop, L., Flannick, J., et al. (2014a). A novel test for recessive contributions to complex diseases implicates bardet-biedl syndrome gene *BBS10* in idiopathic type 2 diabetes and obesity. *The American Journal of Human Genetics*.

Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al. (2014b). Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genetics*, 10(7):e1004494.

Linde, L., Boelz, S., Neu-Yilik, G., Kulozik, A. E., and Kerem, B. (2007a). The efficiency of nonsense-mediated mRNA decay is an inherent character and varies among different cells. *European Journal of Human Genetics*, 15(11):1156–1162.

Linde, L., Boelz, S., Nissim-Rafinia, M., Oren, Y. S., Wilschanski, M., Yaacov, Y., Virgilis, D., Neu-Yilik, G., Kulozik, A. E., Kerem, E., et al. (2007b). Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin. *Journal of Clinical Investigation*, 117(3):683–692.

Liptak, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*, 46(2):200–204.

Liuwantara, D., Elliot, M., Smith, M. W., Yam, A. O., Walters, S. N., Marino, E., McShea, A., and Grey, S. T. (2006). Nuclear factor-κb regulates β-cell death a critical role for A20 in β-cell protection. *Diabetes*, 55(9):2491–2501.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Lusis, A. J. and Pajukanta, P. (2008). A treasure trove for lipoprotein biology. *Nature Genetics*, 40(2):129–130.

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., Handsaker, R. E., Rosenfeld, J. A., Fromer, M., Jin, M., Mu, X. J., Khurana, E., Ye, K., Kay, M., Saunders, G. I., Suner, M. M., Hunt, T., Barnes, I. H., Amid, C., Carvalho-Silva, D. R., Bignell, A. H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D. N., Xue, Y., Romero, I. G., Wang, J., Li, Y., Gibbs, R. A., McCarroll, S. A., Dermitzakis, E. T., Pritchard, J. K., Barrett, J. C., Harrow, J., Hurles, M. E., Gerstein, M. B., and Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–8.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384.

Mahajan, A., Sim, X., Ng, H. J., Manning, A., Rivas, M. A., Highland, H. M., Locke, A. E., Grarup, N., Im, H. K., Cingolani, P., et al. (2015). Identification and functional characterization of *G6PC2* coding variants influencing glycemic traits define an effector transcript at the *G6PC2-ABCB11* locus. *PLoS Genetics*, 11(1):e1004876–e1004876.

Majithia, A. R., Flannick, J., Shahinian, P., Guo, M., Bray, M.-A., Fontanillas, P., Gabriel, S. B., Rosen, E. D., Altshuler, D., et al. (2014). Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with

increased risk of type 2 diabetes. *Proceedings of the National Academy of Sciences*, 111(36):13127–13132.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294–1301.

Manninen, V., Tenkanen, L., Koskinen, P., Huttunen, J., Mänttäri, M., Heinonen, O., and Frick, M. (1992). Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study. Implications for treatment. *Circulation*, 85(1):37–45.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Maquat, L. E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Reviews Molecular Cell Biology*, 5(2):89–99.

Maquat, L. E., Tarn, W.-Y., and Isken, O. (2010). The pioneer round of translation: features and functions. *Cell*, 142(3):368–374.

Mardia, K. and Kent, J. (1979). Multivariate analysis. *Academic Press*.

McAnally, A. A. and Yampolsky, L. Y. (2010). Widespread transcriptional autosomal dosage compensation in Drosophila correlates with gene expression level. *Genome Biology and Evolution*, 2:44–52.

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., Donnelly, P., et al. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3):26.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070.

Metzker, M. L. (2009). Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Moltke, I., Grarup, N., Jørgensen, M. E., Bjerregaard, P., Treebak, J. T., Fumagalli, M., Korneliussen, T. S., Andersen, M. A., Nielsen, T. S., Krarup, N. T., et al. (2014). A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature*.

Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J. F., de Rijk, P., Dewit, O., Finkel, Y., Gassull, M. A., Goossens, D., Laukens, D., Lemann, M., Libioulle, C., O'Morain, C., Reenaers, C., Rutgeerts, P., Tysk, C., Zelenika, D., Lathrop, M., Del-Favero, J., Hugot, J. P., de Vos, M., Franchimont, D., Vermeire, S., Louis, E., and Georges, M. (2011). Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nature Genetics*, 43(1):43–7.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., Ananda, G., Howie, B., Karczewski, K. J., Smith, K. S., et al. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research*, 23(5):749–761.

Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E. T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genetics*, 7(7):e1002144.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628.

Myocardial Infarction Genetics Consortium (2014). Inactivating mutations in *NPC1L1* and protection from coronary heart disease. *The New England Journal of Medicine*.

Nagy, E. and Maquat, L. E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in Biochemical Sciences*, 23(6):198–199.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. A. (2009). Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104.

Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890.

Neyman, J. and Scott, E. (1966). On the use of C-alpha optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, 41:477–497.

Nigro, V., de Sá Moreira, E., Piluso, G., Vainzof, M., Belsito, A., Politano, L., Puca, A. A., Passos-Bueno, M. R., and Zatz, M. (1996). Autosomal recessive limbgirdle muscular dystrophy, *LGMD2F*, is caused by a mutation in the δ–sarcoglycan gene. *Nature Genetics*, 14(2):195–198.

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., et al. (2001). A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature*, 411(6837):603–606.

Ollier, W., Sprosen, T., and Peakman, T. (2005). UK Biobank: from concept to reality. *Pharmacogenomics*, 6(6):639–646.

Panousis, N., Gutierrez-Arcelus, M., Dermitzakis, E., and Lappalainen, T. (2014). Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biology*, 15(9):467.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27(12):1173–1175.

Pe'er, I., de Bakker, P. I., Maller, J., Yelensky, R., Altshuler, D., and Daly, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, 38(6):663–667.

Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., Brody, J. A., Khetarpal, S. A., Crosby, J. R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics*, 94(2):223–232.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 450:7–15.

Pirinen, M. (2015). Estimating and interpreting genetic correlations. *In Preparation*.

Pirinen, M., Lappalainen, T., Zaitlen, N. A., Dermitzakis, E. T., Donnelly, P., McCarthy, M. I., Rivas, M. A., Consortium, G., et al. (2015). Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics*, page btv074.

Plenge, R. M., Scolnick, E. M., and Altshuler, D. (2013). Validating therapeutic targets through Human Genetics. *Nature Reviews Drug Discovery*, 12(8):581–594.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Pollin, T. I., Damcott, C. M., Shen, H., Ott, S. H., Shelton, J., Horenstein, R. B., Post, W., McLenithan, J. C., Bielak, L. F., Peyser, P. A., Mitchell, B. D., Miller, M., O'Connell, J. R., and Shuldiner, A. R. (2008). A null mutation in human

APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science*, 322(5908):1702–5.

Popp, M. W.-L. and Maquat, L. E. (2013). Organizing principles of mammalian nonsense-mediated mRNA decay. *Annual Review of Genetics*, 47:139.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, 86(6):832–8.

Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant or not? *Human Molecular Genetics*, 11(20):2417–2423.

Pulai, J. I., Neuman, R. J., Groenewegen, A. W., Wu, J., and Schonfeld, G. (1998). Genetic heterogeneity in familial hypobetalipoproteinemia: Linkage and non-linkage to the apoB gene in caucasian families. *American Journal of Medical Genetics*, 76(1):79–86.

Purcell, S., Cherny, S. S., and Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–150.

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., ODushlaine, C., Chambert, K., Bergen, S. E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*.

Qiu, X., Wu, H., and Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, 14(1):124.

R Development Core Team (2005). R: A language and environment for statistical computing. *R foundation for Statistical Computing*.

Rahalkar, A. R. and Hegele, R. A. (2008). Monogenic pediatric dyslipidemias: classification, genetics and clinical spectrum. *Molecular Genetics and Metabolism*, 93(3):282–294.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., et al. (2006). *PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature Genetics*, 39(2):165–167.

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510.

Reuters (2014). Cholesterol drug halves heart attack and stroke in early test. *New York Times*.

Rice, J. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Rip, J., Nierman, M. C., Ross, C. J., Jukema, J. W., Hayden, M. R., Kastelein, J. J., Stroes, E. S., and Kuivenhoven, J. A. (2006). Lipoprotein lipase S447X a naturally occurring gain-of-function mutation. *Atherosclerosis, Thrombosis, and Vascular Biology*, 26(6):1236–1245.

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, 43(11):1066–1073.

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., Maller, J. B., Kukurba, K. R., DeLuca, D. S., Fromer, M., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348(6235):666–669.

Rivas, M. A., Pirinen, M., Neville, M. J., Gaulton, K. J., Moutsianas, L., Lindgren, C. M., Karpe, F., McCarthy, M. I., and Donnelly, P. (2013). Assessing association between protein truncating variants and quantitative traits. *Bioinformatics*, 29(19):2419–2426.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D., Renwick, A., Seal, S., Ramsay, E., Duarte, S., Rivas, M., and et al (2012). Mosaic *PPM1D* mutations are associated with predisposition to breast and ovarian cancer. *Nature*.

Russell, J. C. and Proctor, S. D. (2006). Small animal models of cardiovascular disease: tools for the study of the roles of metabolic syndrome, dyslipidemia, and atherosclerosis. *Cardiovascular pathology*, 15(6):318–330.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnstrom, K., Mallick, S., Kirby, A., Wall, D., MacArthur, D. G., Gabriel, S. B., Depristo, M., Purcell, S. M., Palotie, A., Boernwinkle, E., Buxbaum, J. D., Cook, E. H., Gibbs, R., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M., and Daly, M. J. (2014). A framework for the interpretation of *de novo* mutations in human disease. *Nature Genetics*.

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317–320.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336.

Shendure, J., Mitra, R. D., Varma, C., and Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, 5(5):335–344.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.-P., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Genetics*.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 21(10):1728–1737.

Sleat, D. E., Donnelly, R. J., Lackland, H., Liu, C.-G., Sohar, I., Pullarkat, R. K., and Lobel, P. (1997). Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis. *Science*, 277(5333):1802–1805.

Sorlie, P. D., Sharrett, A. R., Patsch, W., Schreiner, P. J., Davis, C. E., Heiss, G., and Hutchinson, R. (1999). The relationship between lipids/lipoproteins and atherosclerosis in African Americans and whites: the Atherosclerosis Risk in Communities Study. *Annals of Epidemiology*, 9(3):149–158.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507.

Stehelin, D., Varmus, H. E., Bishop, J. M., and Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*.

Stein, E. A., Mellis, S., Yancopoulos, G. D., Stahl, N., Logan, D., Smith, W. B., Lisbon, E., Gutierrez, M., Webb, C., Wu, R., et al. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *New England Journal of Medicine*, 366(12):1108–1118.

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D., and Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS One*, 8(7):e65245.

Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14(1):43–59.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305.

Sullivan, D., Olsson, A. G., Scott, R., Kim, J. B., Xue, A., Gebski, V., Wasserman, S. M., and Stein, E. A. (2012). Effect of a monoclonal antibody to PCSK9 on low-density lipoprotein cholesterol levels in statin-intolerant patients: the GAUSS randomized trial. *JAMA*, 308(23):2497–2506.

Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G. R., Xifara, D. K., Matchan, A., Hatzikotoulas, K., Rayner, N. W., Chen, Y., et al. (2013). A rare functional cardioprotective *APOC3* variant has risen in frequency in distinct population isolates. *Nature Communications*, 4.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.

The 1000 Genomes Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute (2014). Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *New England Journal of Medicine*, 371(1):22–31.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, 11(1):276–280.

Thompson, D. and Easton, D. (2001). Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. *The American Journal of Human Genetics*, 68(2):410–419.

Thormaehlen, A. S., Schuberth, C., Won, H.-H., Blattmann, P., Joggerst-Thomalla, B., Theiss, S., Asselta, R., Duga, S., Merlini, P. A., Ardissino, D., et al. (2015). Systematic cell-based phenotyping of missense alleles empowers rare variant association studies: a case for LDLR and myocardial infarction. *PLoS Genetics*, 11(2):e1004855–e1004855.

Timpson, N. J., Walter, K., Min, J. L., Tachmazidou, I., Malerba, G., Shin, S.-Y., Chen, L., Futema, M., Southam, L., Iotchkova, V., et al. (2014). A rare variant in *APOC3* is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*, 5.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., and Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine*, 24(19):2911–2935.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Tsygankov, A. Y. (2009). TULA-family proteins: an odd couple. *Cellular and Molecular Life Sciences*, 66(17):2949–2952.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*. Springer Science & Business Media.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Warnick, G. R., Knopp, R. H., Fitzpatrick, V., and Branson, L. (1990). Estimating low-density lipoprotein cholesterol by the Friedewald equation is adequate for classifying patients on the basis of nationally recommended cutpoints. *Clinical Chemistry*, 36(1):15–19.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006.

Whitcomb, D. C., Gorry, M. C., Preston, R. A., Furey, W., Sossenheimer, M. J., Ulrich, C. D., Martin, S. P., Gates, L. K., Amann, S. T., Toskes, P. P., et al. (1996). Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature Genetics*, 14(2):141–145.

White, N. J., Nacheva, E., Asimakopoulos, F., Bloxham, D., Paul, B., and Green, A. (1994). Deletion of chromosome 20q in myelodysplasia can occur in a multipotent precursor of both myeloid cells and B cells. *Blood*, 83(10):2809–2816.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93.

Zelterman, D. and Chen, C.-F. (1988). Homogeneity tests against central-mixture alternatives. *Journal of the American Statistical Association*, 83(401):179–182.

Zetoune, A. B., Fontanière, S., Magnin, D., Anczuków, O., Buisson, M., Zhang, C. X., and Mazoyer, S. (2008). Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genetics*, 9(1):83.

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J.-H., Aach, J., Leproust, E. M., et al. (2009). Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods*, 6(8):613–618.

Zhang, L., Chen, L. H., Wan, H., Yang, R., Wang, Z., Feng, J., Yang, S., Jones, S., Wang, S., Zhou, W., et al. (2014). Exome sequencing identifies somatic gain-of-function *PPM1D* mutations in brainstem gliomas. *Nature Genetics*.

Zhou, J., Lemos, B., Dopman, E. B., and Hartl, D. L. (2011). Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. *Genome Biology and Evolution*, 3:1014–1024.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*.