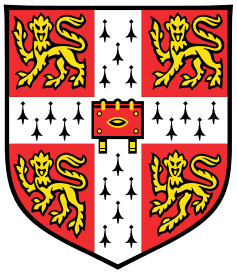
Methods for Determining the Genetic Causes of Rare Diseases



Daniel Greene

MRC Biostatistics Unit
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare College

January 2018

Methods for Determining the Genetic Causes of Rare Diseases

Daniel Greene

Thanks to the affordability of DNA sequencing, hundreds of thousands of individuals with rare disorders are undergoing whole-genome sequencing in an effort to reveal novel disease aetiologies, increase our understanding of biological processes and improve patient care. However, the power to discover the genetic causes of many unexplained rare diseases is hindered by a paucity of cases with a shared molecular aetiology. This thesis presents research into statistical and computational methods for determining the genetic causes of rare diseases. Methods described herein treat important aspects of the nature of rare diseases, including genetic and phenotypic heterogeneity, phenotypes involving multiple organ systems, Mendelian modes of inheritance and the incorporation of complex prior information such as model organism phenotypes and evolutionary conservation.

The complex nature of rare disease phenotypes and the need to aggregate patient data across many centres has led to the adoption of the Human Phenotype Ontology (HPO) as a means of coding patient phenotypes. The HPO provides a standardised vocabulary and captures relationships between disease features. The use of such ontologically encoded data is widespread in bioinformatics, with ontologies defining relationships between concepts in hundreds of subfields. However, there has been a dearth of tools for manipulating and analysing ontological data. I developed a suite of software packages dubbed 'ontologyX' in order to meet these needs, simplify visualisation of such data, and enable them to be incorporated into complex analysis methods. An important aspect of the analysis of ontological data is quantifying the semantic similarity between ontologically annotated entities, which is implemented in the ontologyX software. We employed this functionality in a phenotypic similarity regression framework, 'SimReg', which models the relationship between ontologically encoded patient phenotypes of individuals and rare variation in a given genomic locus. It does so by evaluating support for a model under which the probability that a person carries rare alleles in a locus depends on the similarity between the person's ontologically encoded phenotype and a latent characteristic phenotype which can be inferred from data. A probability of association is computed by comparison of the two models, allowing prioritisation of candidate loci for involvement in disease with respect to a heterogeneous collection of disease phenotypes. SimReg includes a sophisticated treatment of HPO-coded phenotypic data but dichotomises the genetic data at a locus. Therefore, we developed an additional method, 'BeviMed', standing for *Bayesian Evaluation of Variant Involvement in Mendelian Disease*, which evaluates the evidence of association between allele configurations across

rare variants within a genomic locus and a case/control label. It is capable of inferring the probability of association, and conditional on association, the probability of each mode of inheritance and probability of involvement of each variant. Inference is performed through a Bayesian comparison of multiple models: under a baseline model disease risk is independent of allele configuration at the given rare variant sites and under an alternate model disease risk depends on the configuration of alleles, a latent partition of variants into *pathogenic* and *non-pathogenic* groups and a mode of inheritance. The method can be used to analyse a dataset comprising thousands of individuals genotyped at hundreds of rare variant sites in a fraction of a second, making it much faster than competing methods and facilitating genome-wide application. The thesis concludes by describing an analysis pipeline and web application called 'Gene-docs' that utilises ontologyX, SimReg and BeviMed to perform inferences and makes the results and the underlying data available to collaborating biologists and clinicians.

I would like to dedicate this thesis to my wife Danielle, whose love and support has been my motivation thoughout.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations.

Daniel Greene
January 2018

Acknowledgements

I would like to thank my supervisor, Dr. Ernest Turro, who has been consistently supportive, generous with his time and ideas, and an inspiration. Being his student has been a privilege and a joy, and without him none of this would have been possible. I would also like to thank my co-supervisors, Professor Sylvia Richardson and Professor Willem Ouwehand, for all of their support, direction and many fruitful discussions over the years. I am also grateful to my colleague William Astle, who has always made time for technical discussions and had many good suggestions.

I would like to thank my wife, Danielle, and our children, Sean and Lyra, for their love and patience, for keeping my spirits up, and being there for me every day. Lastly, I would like to thank my parents, Brendan and Michele, for all their encouragement and support during the course of my PhD, and for opening all the doors which led me here.

Table of contents

List of figures

List of tables

Chapter 1

Background

It is important that the underlying genetic causes of rare diseases are identified in order to further our understanding of biology, identify affected relatives and improve patient management. Analysis of cosegregation between genetic markers and disease within large disease-affected families has led to the identification of the genes involved in thousands of rare diseases over the last few decades. However, the aetiologies of many diseases have not yet been discovered. Recently, high-throughput sequencing has been used to call genetic variants — specific sequences of nucleotides at specific locations relative to a reference genome — in all genes, Whole-Exome Sequencing (WES), or across the entire genome Whole-Genome Sequencing (WGS), facilitating association-based study designs. Consequently, large genomic sequencing studies attempting to solve unexplained diseases are underway. For example, the 100,000 Genomes project [67], which was established to create a genomic medicine service for the United Kingdom National Health Service, will sequence the genomes of $50,000$ individuals with rare diseases and their close relatives. Genome-wide association studies (GWAS) have identified thousands of common variants marginally associated with common diseases or quantitative traits. However, methods that test for marginal association are underpowered for identifying the determinants of rare diseases. This is because rare diseases are typically caused by combinations of only one or two rare alleles, each of which may be observed in only a very small number of individuals in a study. Borrowing of information across variants, at least those within particular genomic regions — also known as loci — is therefore an important requirement of any good statistical association method for rare diseases. In order to boost power in the context of such genetic heterogeneity, methods that combine genotype data across many variants in a locus have been developed. 'Burden tests' borrow information across variants in a region by aggregating alleles within loci, and test for association between the phenotype and the aggregated allele count. However, the majority of rare variants are neutral with respect to severe disease

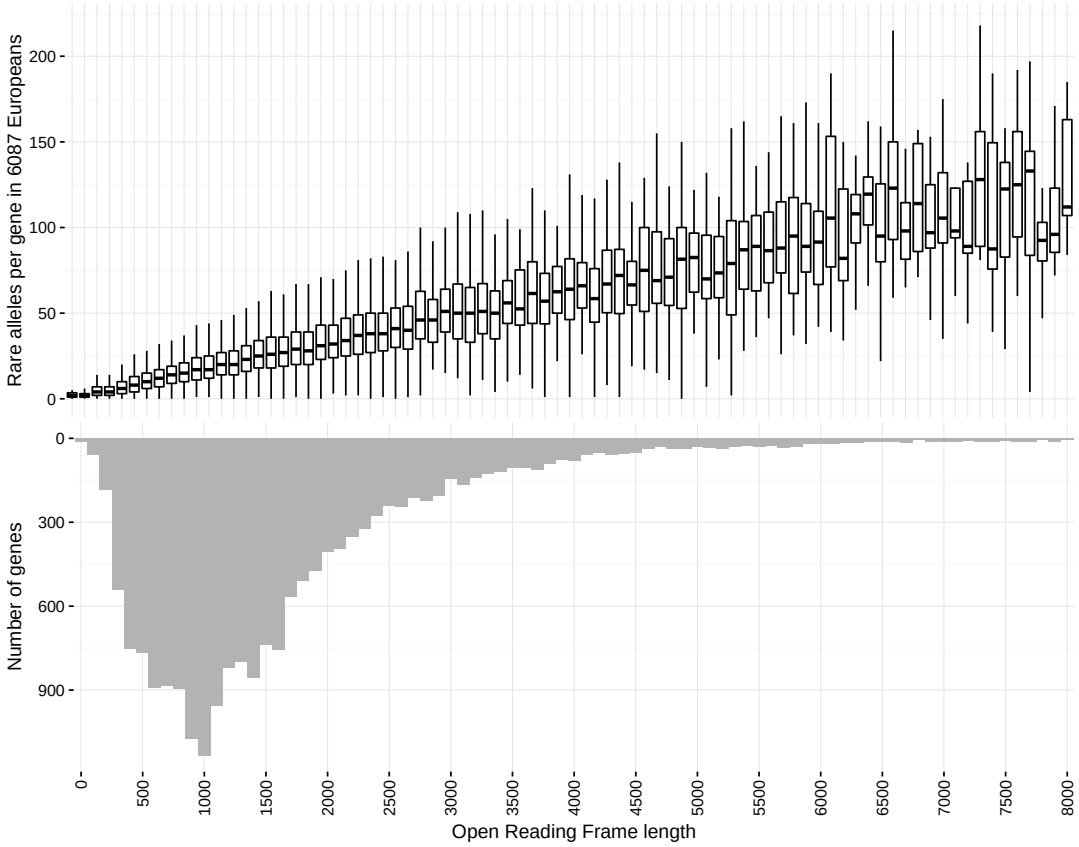


Fig. 1.1 Top: the distribution of the number of rare alleles observed per gene by open reading frame length. There is a clear linear relationship between gene length and number of rare alleles observed, suggesting that the majority of rare variants are benign. Bottom: histogram of open reading frame length for human genes.

(see Figure 1.1), which reduces the power of such methods due to the inclusion of benign variants in the aggregation. There exist statistical methods which are designed for identifying associations between complex traits and rare variants, but these do not explicitly model the modes of inheritance typical of rare diseases. Furthermore, rare disease phenotypes are often composite, involving features affecting multiple organ systems and which manifest stochastically and with variable severity. Thus, many disease phenotypes are not effectively represented by existing models.

1.1 A portrait of rare Mendelian diseases

The majority of rare genetic diseases are 'Mendelian' in the sense that they are caused by the presence of pathogenic alleles on one or two chromosomes, and can therefore be inherited respectively 'dominantly' or 'recessively'. They are so called because the first evidence for traits being transmitted in such a way was presented using the colour and shape of pea pods by Gregor Mendel, an Austrian monk who lived in the 19$^{\text{th}}$ century. There are over 7,000 known Mendelian diseases [77] and a genetic aetiology has been found for at least half of them. Between them, they affect approximately 1 in every 20 people, and all organ systems in the body. Despite their overall prevalence, individual Mendelian diseases are rare due to the effects of negative selection pressure and low mutation rate during meiosis (the *de novo* mutation rate depends on factors including genomic location and paternal age, but is approximately 1.3×10^{-8} nucleotides per generation [26]). Consequently, deleterious alleles responsible for rare diseases have typically arisen recently in the population and presently have a very low population allele frequency. An allele that causes severe disease in childhood in everyone who carries it would result only from *de novo* mutation and would occur with frequency no greater than the prevalence of the disease. On the other hand, a mild disease or a disease with late onset may be caused by alleles with higher frequencies. In recessive diseases, individuals who carry a pathogenic allele on only one chromosome are unaffected, and there is therefore weaker negative selection against the allele. For example, Bernard-Soulier syndrome is an autosomal recessive disease which causes reduced platelet count, increased mean platelet volume and abnormal bleeding tendency. It has a prevalence estimated to be less than one in a million live births, whilst variants which cause the disease can have a population allele frequency as high as 0.06% [17, 55]. Population-specific factors may also play a role. Sickle-cell disease, which causes red blood cells to be malformed and results in anaemia, is also recessively inherited. However, in regions where malaria is endemic, heterozygotes have a reproductive advantage due to the fact that alleles responsible for sickle-cell disease confer malaria resistance. Hence, the most common pathogenic variant has persisted at frequencies in excess of 15% in some West African populations, where prevalence of the disease can be as high as 4% [62].

Rare diseases are often caused by variants in or affecting protein-coding genes — segments of the genome which are transcribed to RNA and then translated to sequences of amino acids from which proteins are formed — because of the potential for major disruption of cell biology which depends on proteins functioning correctly. Variants in genes may have various impacts on protein formation, depending on the how they affect the sequence of 'codons', nucleotide triplets in the DNA which directly translate to amino acids. Due to redundancy of codon translations to amino acids, some variants, termed 'synonymous' do not alter amino

acid sequence and therefore generally do not affect protein formation. 'Non-synonymous' variants generate codons which do lead to an amino acid change, and therefore are considered more likely to be pathogenic. 'Frame-shift' variants cause a shift in the Open Reading Frame (ORF) — the sequence of codons that gets translated — for the gene by inserting or deleting a non-multiple of three nucleotides, therefore leading all subsequent codons to be mistranslated. A 'Stop-gain' variant is a Single Nucleotide Variant (SNV) which results in a codon being replaced with a stop-codon, which discontinues translation, and therefore results in a truncated protein product. Splicing is part of the process whereby RNA molecules transcribed from DNA are converted to mRNA molecules, which can then be translated to generate proteins. It consists of removing the 'introns', untranslated sequences of nucleotides between 'exons', translated sequences of nucleotides, from the RNA. 'Splice-variants' are variants which interfere with splicing, for example causing exons to be omitted from or introns retained abnormally in the resultant mRNA. The regions which lie immediately upstream of the start-codon and downstream of the stop-codon for a gene do not get translated and are known as the $5'$ untranslated region (UTR) and $3'$ UTR respectively. Variants which lie in these regions have the potential to disrupt translation by affecting the binding or unbinding of the translational machinery. There is often substantial genetic heterogeneity amongst individuals with the same rare disease, typically due to different variants in the same gene or functional unit leading to the same consequences. For example, there are over $2,000$ distinct variants in the *CFTR* gene that can cause cystic fibrosis [14], an autosomal recessive disease which progressively damages the lungs.

To discover the cause of disease, affected individuals would ideally be grouped *a priori* into clusters with a shared (though unknown) genetic aetiology, whilst all other individuals could be treated as controls. Unfortunately, the nature of rare diseases and collection of rare disease phenotype data means it is often difficult to separate rare disease cases into such groups. This is because the majority of rare diseases are phenotypically heterogeneous, indeed, most affect more than one organ system as is shown in Figure 1.2. Additionally, many rare diseases exhibit substantial phenotypic variability with respect to the presence and severity of disease features. In the past, this has led to diseases subsequently found to have a common genetic cause being given a multitude of names. For example, *MYH9*-related disease has been called May-Hegglin anomaly, Fechtner syndrome, Sebastian syndrome and Epstein syndrome [97]. Even the same pathogenic variant — a serine to leucine substitution at position 96 of the *MYH9* gene — appears to cause hearing loss, renal dysfunction and cataracts stochastically and with low correlation [74]. Moreover, phenotypic heterogeneity can be sufficiently large to induce significant overlap of clinical phenotypes between different diseases. For example, a low platelet count and abnormal bleeding are both typical

characteristics of *MYH9*-related disease, but they are also present in patients with the aetiologically distinct Wiskott-Aldrich syndrome. Similarly, abnormal bleeding and abnormal platelet granules typically manifest in Wiskott-Aldrich patients, but these phenotypes are also present in cases with the aetiologically distinct Hermansky-Pudlak syndrome. See Feng et al. [23], Anikster et al. [4], Suzuki et al. [111], Zhang et al. [125], Morgan et al. [71], Li et al. [57], Cullinane et al. [18] and Beaulieu et al. [7] for further examples. Furthermore, pathogenic variants may not have complete 'penetrance', that is they do not always cause the symptoms of the disease to be manifested. For example, variants in *BRCA1* responsible for dominantly inherited breast cancer confer only a 54% risk of developing breast cancer by age 60 [20].

The power of case/control association tests is likely to be compromised by misspecification of the disease status label, and thus such tests are not well suited to a heterogeneous collection of rare disease cases. Tests which summarise the clinical manifestations of a disease with a single variable (e.g. see methods described in Lee et al. [51]), lose power when multiple phenotypic traits contain complementary information about the same causal genotype, which is often the case for rare diseases. Although methods for modelling pleiotropy of variants — the property of influencing multiple traits — have proven successful in the context of genome-wide association studies [9, 78, 107], they are ill-suited for rare disease studies in which the phenotype data are often of mixed type, contain correlated components and are collected with variable detail and completeness.

The HPO addresses the need for a standardised vocabulary for rare disease phenotypes and is being used to encode the disease phenotypes of participants for several large international projects [119, 24, 81, 33]. It is a directed acyclic graph (DAG) representing over 10,000 phenotypic abnormalities as nodes connected to each other through 'is-a' relations, represented as edges. The HPO was created with the support of experts in many areas of medicine to accommodate coding of phenotypic data derived from diverse sources, such as laboratory assays, images, graphs and clinical interpretations. Methods exist that compare HPO phenotypes with HPO-coded profiles corresponding to known diseases for the purpose of differential diagnosis [47, 6]. The HPO-coded profiles can be supplemented with functional gene-specific information to prioritise genes [100, 121]. If genotype data are available, these and other methods [93, 124] can be used to prioritise variants and potentially suggest new causes of disease [44, 100, 121]. However, there have been no statistical association methods which share information between the HPO-encoded phenotypes of rare disease cases and relate the phenotype to the genetic variant data.

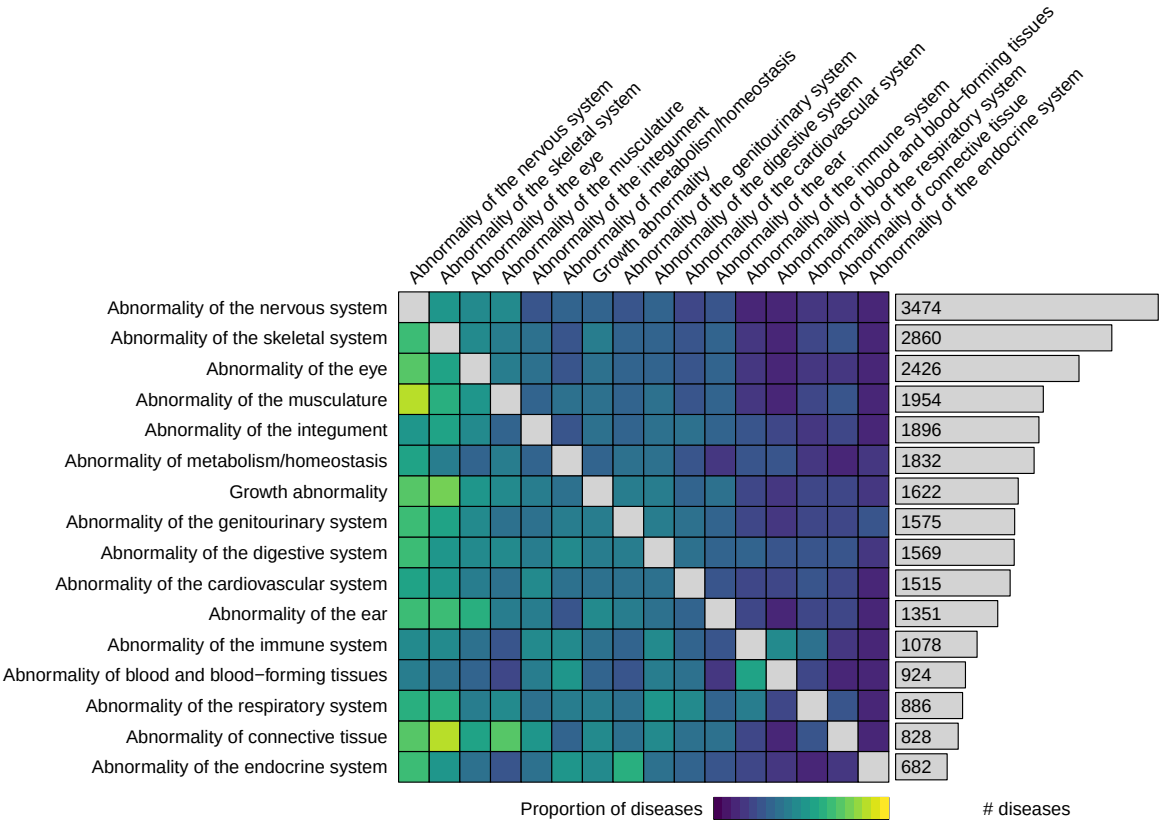


Fig. 1.2 Distribution of organ systems co-affected by rare diseases for each organ system (rows). The bar chart on the right shows the total number of diseases affecting each organ system. A total of 6,911 rare diseases are represented, with phenotype data downloaded from the HPO website [117].

1.2 The impact of high-throughput sequencing on rare disease research

The revolution of high-throughput sequencing technology has seen the cost of sequencing a human genome fall from $1 billion to $1,000 since it began in 2005 with the launch of the Roche 454 sequencing machine. The state of the art in 2017 is the Illumina dye sequencing technique, which was selected for use in the 100,000 Genomes Project. The technology is based on *sequencing by synthesis*, recording the sequence of bases in short single-stranded DNA molecules by detecting flourescently tagged nucleotides as they are incorporated onto the complementary strand. This procedure can be performed in parallel on tens of millions of fragments with the latest Illumina sequencing machines, which are able to sequence 45 human genomes to a depth averaging 30 reads per base in a day at a cost of $1,000 per genome [66]. The output of such high throughput short read sequencing is a set of nucleotide sequences of some given length, corresponding to sections of the genome of the sequenced individual. The reads must then be aligned to the human reference genome to allow comparison between them and for genetic variants to be called.

The use of high-throughput sequencing to study rare diseases began with WES because rare diseases are often caused by exonic variants and exomes are much smaller than genomes (an exome is approximately 60 megabases, which is 2% of the size of a whole genome [75]), therefore making them cheaper to sequence. In WES, sample exonic DNA fragments are selected from amongst genomic DNA fragments using an oligonucleotide hybridisation technique [5]. The selected fragments can then be sequenced using a high-throughput sequencing technique. WES opened up new opportunities for determining the genetic causes of rare diseases as rare diseases are often caused by exonic variants, and unlike with the standard Single Nucleotide Polymorphism (SNP) genotyping arrays, any number of rare exonic variants can be observed. There is therefore the potential to unlock clinical benefits: many individuals with rare diseases suffer years of futile clinical and genetic tests, misdiagnosis, and lack of properly informed treatment in a so-called ‘diagnostic odyssey’. However, approximately 15,000–20,000 variants are observed per sequenced exome [108], motivating prefiltering in order to increase power to identify which variants are pathogenic. Typically, analysis proceeds by filtering variants for allele frequency against databases of variation for the healthy population [65], for example ExAC and gnomAD [55], enabling many variants to be ruled out because their allele frequency is too high to be consistent with their involvement in a rare heritable disorder. Furthermore, interpretation of specific exonic sequence variants is aided by prediction of consequences for protein sequence and formation [69]. For example, synonymous variants are typically removed from consideration.

Depending on the exact parameters of the filtering, this often leaves hundreds of candidate variants. Further analysis consists of either looking at cosegregation of variants with disease amongst exome sequenced family members or by applying tests of association between rare variation and disease status for unrelated people. Rare variants are much more numerous than common variants across the genome, which renders the hypothesis space in rare disease research much larger than in GWAS of complex traits. Therefore, variants in the same locus need to be modelled jointly to achieve adequate power.

In exome sequencing studies of rare diseases, the proportion of affected participants for whom a pathogenic or likely pathogenic variant is identified is usually quite low. For instance, the Deciphering Developmental Disorders (DDD) project is an exome sequencing project which was set up to determine the genetic cause of developmental disorders. It sequenced 13,600 affected individuals and a further 20,000 of their parents, and has revealed 30 novel associations between genes and developmental disorders. As of April 2015, a pathogenic or likely pathogenic variant was identified for only 27% of participants, comprising 1,133 previously investigated undiagnosed children with developmental disorders [120].

The falling cost of WGS has led to it substituting WES as the standard tool for investigating rare diseases as it provides more uniform coverage across exons and has the potential to reveal the involvement of non-exonic variants. However, variation in the non-exonic regions of the genome is comparatively harder to interpret in terms of its biological impact. The non-exonic regions of the genome are also much larger, and millions of variants are called per individual. Cohorts for deriving population allele frequencies are also currently smaller: ExAC contains $123,136$ exome sequences compared with $15,496$ whole-genome sequences for gnomAD as of May 2017.

1.3 The National Institute for Health Research BioResource–Rare Diseases project dataset

The National Institute for Health Research BioResource–Rare Diseases (NBR–RD) project was established as a pilot study for the $100,000$ Genomes project to test the feasibility of applying uniform sequencing to a large number of individuals, obtain molecular diagnoses for participants, and ultimately uncover novel aetiologies of genetic diseases. The data generated by this project were used in the development of the methods presented in this thesis. It has sequenced over $7,000$ probands — original disease-affected members of the families participating in the study — and about $2,000$ additional affected and unaffected relatives as of 2017, with genetic diseases falling into one of 15 broad disease categories (see Table 1.1).

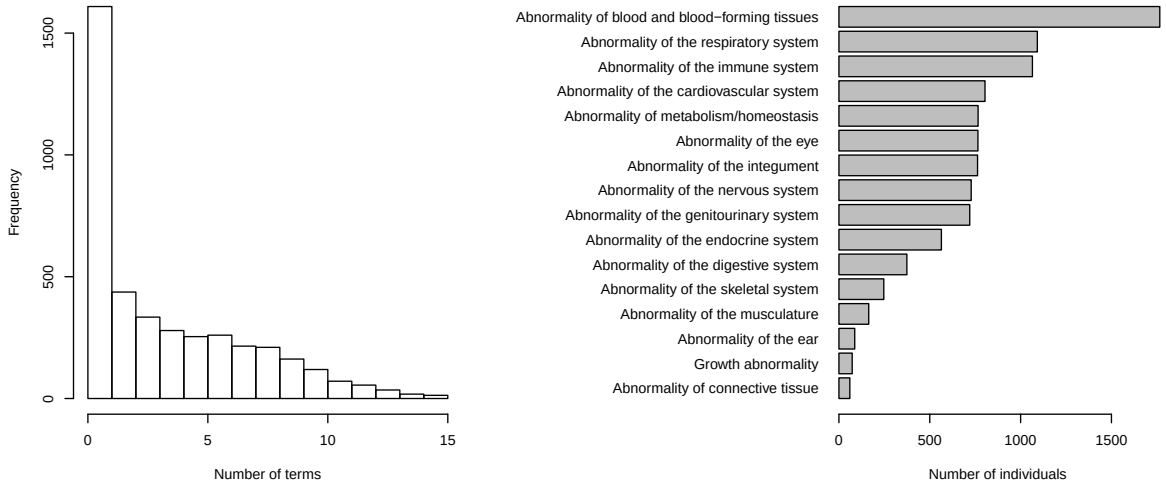


Fig. 1.3 Left: the distribution of the number of HPO terms used to encode the phenotypes of HPO-coded individuals in the NBR–RD project. Right: the number of individuals coded with each high-level phenotypic abnormality term.

The data for each disease category are analysed by a specialised team comprising experts in the particular disease phenotype, geneticists and bioinformaticians. Given a particular disease category, individuals assigned to different disease categories can generally be treated as controls. This is sometimes complicated by phenotypic overlap between projects: for example, thrombocytopenia — a pathologically low platelet count — is a phenotype which manifests in some individuals enrolled to the Primary Immune Disorder (PID) project and some to the Bleeding and Platelet Disorders (BPD) project. Participant phenotypes have been encoded using HPO terms where applicable. Amongst the patients whose phenotypes have been encoded, the number of terms used had a mode of one and mean of four per person. The top-level phenotype abnormalities — HPO terms which descend from the term 'Phenotypic abnormality' — mostly correspond to abstract abnormalities of different organ systems, but also to developmental processes (e.g. Growth abnormality) and tissue types (e.g. Abnormality of connective tissue). Figure 1.3 shows the distribution of the number of terms used per individual and the prevalence of top-level phenotypic abnormalities amongst the cases in the collection. The most frequent top-level phenotypic abnormality term was 'Abnormality of blood and blood-forming tissues', annotated to 1,766 individuals.

The participants were sequenced using Illumina's short read sequencing platform and resultant reads were aligned to the human reference genome hg19 using the Isaac aligner [90]. Variants were called with the 'Northstar' pipeline, involving SNV caller Starling and Structural Variant (SV) callers Manta and Canvas. The variants were then filtered to create a subset for use in downstream analysis by retaining those which had an allele frequency of

at most 0.1% in any of the reference cohorts (including 1000 Genomes [1], UK10K [115], ExAC and gnomAD [55]), were classified as non-synonymous, frame-shift, splice-site or stop-gain/stop-loss, or located in the 5′ UTR or 3′ UTR regions of a gene. After filtering, 3,863,577 variants were retained, observed across 9,632 people.

Testing for association between so many genetic variants in so many individuals affected by one of so many different rare diseases is challenging. Powerful new methods which are robust to phenotypic heterogeneity, genetic heterogeneity within and across loci, and which are computationally efficient will be required if large projects focusing on sequencing and analysing probands such as the NBR–RD project are to yield the best possible results.

Sub project	Sequenced individuals
Bleeding and Platelet Disorders (BPD)	1167
Cerebral small vessel disease (CSVD)	244
Ehlers-Danlos and Ehlers-Danlos like syndromes (EDS)	15
Genomics England pilot (GEL)	1963
Hypertrophic Cardiomyopathy (HCM)	241
Intrahepatic Cholestasis of Pregnancy (ICP)	261
Leber Hereditary Optic Neuropathy (LHON0	71
Multiple Primary Malignant Tumours (MPMT)	521
Neuropathic Pain Disorders (NPD)	168
Pulmonary Arterial Hypertension (PAH)	1131
Primary Immune Disorders (PID)	1308
Primary Membranoproliferative Glomerulonephritis (PMG)	151
Stem Cell and Myeloid Disorders (SMD)	221
Specialist Pathology: Evaluating Exomes in Diagnostics (SPEED)	1384
Steroid Resistant Nephrotic Syndrome (SRNS)	249

Table 1.1 NBR–RD breakdown of sequenced individuals by project

1.4 The ThromboGenomics platform

The ThromboGenomics platform [99] was designed in order to improve the rate of diagnosis of BPDs. It is a high-throughput sequencing assay targeting a panel of 63 genes known to be involved in BPDs selected by expert clinicians and researchers. The panel of genes tested is updated annually based on the results of ongoing research into the genetic causes of BPDs. Targeting such a small genomic area allows for cost-effective high-depth sequencing of the gene panel, leading to highly confident identification of variants. SNV/indels and structural variants are called using GATK [68] and ExomeDepth [83] respectively, and are filtered for low frequency against control populations, at most 10% minor allele frequency amongst other

ThromboGenomics samples to guard against systematic artefacts and predicted consequence according to SnpEff (retaining only those with a predicted translational impact of `MODERATE` or `HIGH`, or lying within a splice region). The filtered variants are then assessed for clinical significance by a multi-disciplinary team (MDT) comprising a mixture of clinical, genetics and bioinformatics experts. The platform was validated against 300 samples from individuals comprising a mixture of BPD cases with a previously identified causal variant, BPD cases with a suspected disorder but without knowledge of causal variants, BPD cases whose phenotypes could not be matched to a known BPD, and unaffected relatives. The average number of potentially significant variants retained affecting the 63 genes per individual across the 300 samples was 5.34. When the targeted sequencing assay was applied to the 159 samples from cases with a previously identified causal variant, all such variants were recapitulated by the platform, demonstrating its efficacy. However, when applied to 76 samples from the cases whose phenotypes could not be matched to a known BPD, only eight molecular diagnoses were reached, emphasising the need for further investigation into the causes of BPDs.

1.5 Overview of thesis

Chapter 2 begins with a description of the software which I developed in order to lay the foundations for analysing ontologically encoded phenotypes. The software, dubbed 'ontologyX' [39], constitutes a suite of R packages which enables sets of ontological terms to be represented, visualised and compared to other sets using semantic similarity measures, operations which were previously difficult to perform. It then continues to describe applications of quantifying the phenotypic similarity of HPO-coded individuals, including testing hypotheses about the involvement of variants harboured by particular genes in disease and using phenotypic similarity to prioritise candidate variants for a given case. It concludes by describing an approach for testing association between phenotypically similar individuals and presence of rare alleles in the same genes by partitioning HPO-coded individuals into clusters based on phenotypic similarity. Unsupervised clustering of phenotypes is used to generate clusters, and the members and non-members of each cluster in turn are respectively treated as the cases and controls upon which association tests can be performed.

Chapter 3 describes work which builds upon the phenotype similarity methods introduced in Chapter 2 by incorporating it into a Bayesian model for the relationship between rare alleles and disease. The method, named 'SimReg', enables prior information to be taken into account whilst testing a specific alternative model, whereby similarity between an individual's HPO phenotype and a latent characteristic disease phenotype predicts disease

risk. The method allows the disease phenotype to be estimated from the data, is powerful to detect pleiotropy, robust to phenotypic variability and is the first disease association method which is applicable to phenotype data encoded using HPO terms. We then show that the method is effective by applying it to over $2,000$ individuals with various rare diseases for whom high-throughput sequencing data was available, recovering numerous true associations between diseases and genes, including a novel disease-gene association [109]. Allele counts at rare variant sites are aggregated in SimReg, and therefore power is reduced by inclusion of non-pathogenic variants in the aggregation. In order to increase power for testing association between a set of rare variants containing both pathogenic and non-pathogenic varieties, and a disease for which cases are clearly distinguishable from controls, we developed another method called 'BeviMed', presented in Chapter 4.

In BeviMed, a Bayesian comparison of a baseline model where disease risk is independent of allele configuration at given rare variant sites and an alternate model where it depends on the configuration of alleles at a latent subset of pathogenic variants is used to infer a probability of association, and conditional on association, the probability of involvement of each variant. We compare BeviMed with other methods for testing association between allele counts at rare variant sites and disease status, and demonstrate that BeviMed is more powerful, particularly for recessive diseases, using a simulation study. We then apply BeviMed to the NBR–RD project data and recover numerous true associations between genes and thrombocytopenia. Chapter 5 describes a web-application which was developed alongside the methods described in earlier chapters, and makes the results of systematic application of these methods and the raw data used to generate them available to clinicians.

The current chapter concludes by introducing the literature on quantifying similarity of sets of ontological terms and describing the various methods for analysing HPO data and statistical association methods for rare variants, which are referred to in later chapters.

Publications

During my PhD I have contributed to several publications, upon which the majority of this thesis is based. For completeness I list them here, including my contribution and where applicable the chapter or section in which they are discussed.

- **D. Greene**, S. Richardson and E. Turro (2017). "ontologyX: A suite of R packages for working with ontological data". *Bioinformatics*, 33(7):1104-1106.
  Methodological development, discussed in Section 2.1.

- S. K. Westbury[†], E. Turro[†], D. Greene[†], C. Lentaigne[†], A. M. Kelly[†], T. K. Bariana[†], et al. (2015). "Human phenotype ontology annotation and cluster analysis to unravel

genetic defects in 707 cases with unexplained bleeding and platelet disorders". *Genome Med*, 7(1):36.
Methodological development and data analysis, discussed in Section 2.2.

- I. Simeoni, J. C. Stephens, F. Hu, S. V. Deevi, K. Megy, T. K. Bariana, ..., **D. Greene** et al. (2016). "A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders". *Blood*, 127(23):2791–2803.
  Data analysis, discussed in Section 2.3.

- E. Turro, **D. Greene**, A. Wijgaerts, C. Thys, C. Lentaigne, T. K. Bariana, et al. (2016). "A dominant gain-of-function mutation in universal tyrosine kinase *SRC* causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies". *Science Translational Medicine*, 8(328):328ra30.
  Data analysis, discussed in Section 2.4.

- **D. Greene**, NIHR BioResource–Rare Diseases Consortium, S. Richardson and E. Turro (2016). "Phenotype similarity regression for identifying the genetic determinants of rare diseases". *The American Journal of Human Genetics*, 98(3):490–499.
  Methodological development and data analysis, discussed in Chapter 3.

- S. Stritt, P. Nurden, E. Turro, **D. Greene**, S. B. Jansen, S. K. Westbury, et al. (2016). "A gain-of-function variant in *DIAPH1* causes dominant macrothrombocytopenia and hearing loss". *Blood*, 127(23):2903–2914.
  Data analysis, discussed in Chapter 3.

- **D. Greene**, NIHR BioResource–Rare Diseases Consortium, S. Richardson and E. Turro (2017). "A fast association test for identifying pathogenic variants involved in rare diseases". *The American Journal of Human Genetics*, 101(1):104–114.
  Methodological development and data analysis, discussed in Chapter 4.

- S. Sivapalaratnam, S. K. Westbury, J. C. Stephens, **D. Greene**, K. Downes, A. M. Kelly, et al. (2017). "Rare variants in *GP1BB* are responsible for autosomal dominant macrothrombocytopenia". *Blood*, 129(4):520–524.
  Data analysis.

- N. Pontikos, J. Yu, I. Moghul, L. Withington, F. Blanco-Kelly, T. Vulliamy, ..., **D. Greene** et al. (2017). "Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data". *Bioinformatics*.
  Data analysis.

- S. K. Westbury, M. Canault, **D. Greene**, E. Bermejo, K. Hanlon, M. P. Lambert, et al. (2017). "Expanded repertoire of *RASGRP2* variants responsible for platelet dysfunction and severe bleeding". *Blood*.
  Data analysis.

- P. Tuijnenburg†, H. L. Allen†, S. O. Burns, **D. Greene**, M. H. Jansen, E. Staples, et al. "Whole-Genome Sequencing identifies *NFKB1* haploinsufficiency as the commonest monogenic cause of Common Variable Immunodeficiency". *Journal of Allergy and Clinical Immunology* (under review).
  Data analysis.

† indicates equal contribution.

1.6 Current methods and software for analysing rare disease data

Ontologically encoded data

Phenotype data are increasingly being encoded using ontological terms, for example from the HPO and Mammalian Phenotype Ontology (MPO). If such data are to be analysed effectively, methods for manipulating ontological terms which account for the structure of the 'ontology' in which they belong are required. In information science, an ontology is a formal description of a set of entities, called 'terms', their properties, and the relationships between them. They therefore support systematic reasoning about the entities they describe. Typically, ontologies specify 'is a'/'subclass of' relations between terms forming a DAG, although arbitrary relations can be represented. Ontologies may be encoded in Open Biomedical Ontologies (OBO) format or Web Ontology Language (OWL) format (with OWL format the more expressive, supporting more complex relationships and reasoning). There exist ontologies describing over 100 biomedical subjects which can be downloaded in either format from the OBO Foundry's website [103].

Ontological annotation is used to describe many biological phenomena, including gene function [34], variant effects [116] and human phenotype abnormalities [48], with many annotation datasets publicly available. Sets of ontological terms annotating the same entity will be assumed to be non-redundant, hereafter referred to as *minimal*. Formally, a set of terms is said to be non-redundant/minimal if and only if it lacks elements implied by other terms in the set through directed edges in the ontology. The terms highlighted blue in Figure 1.4 comprise such a set, as there is no directed path between any pair of blue nodes.

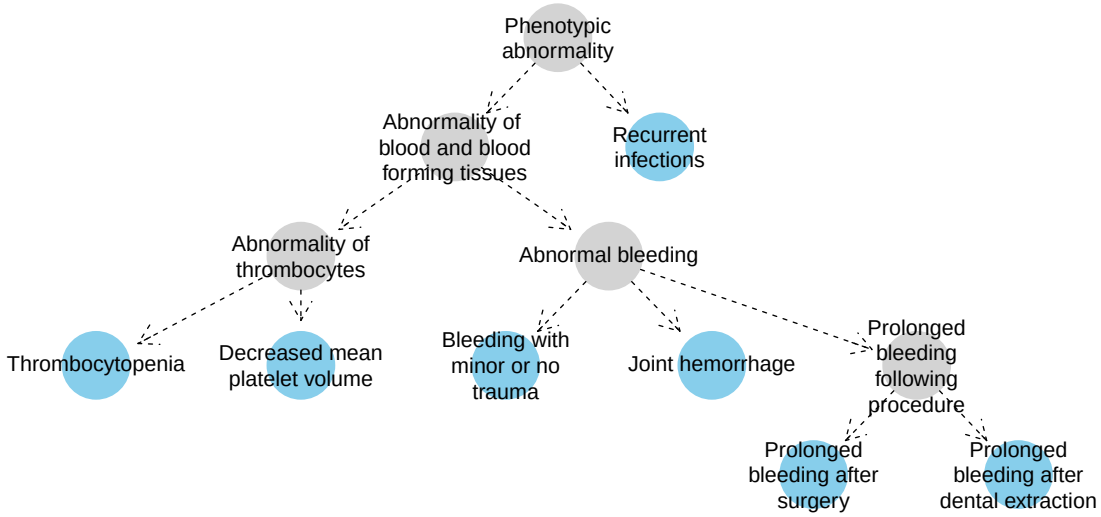


Fig. 1.4 Example HPO coding of the disease phenotype of an individual with Wiskott-Aldrich syndrome. The nodes in blue imply the presence of the more general ancestral phenotypes depicted as grey-filled nodes. No blue node has a directed path to any other, which means that the blue nodes comprise a *minimal* set of HPO terms.

Semantic similarity measures

A semantic similarity measure is a metric on terms or sets of terms where the distance between entities reflects how similar their meanings are. Semantic similarity measures for terms belonging to ontologies are often defined with respect to the structure of their ontologies, so we discuss such measures here. There are various methods for evaluating the similarity of two terms, which typically fall into categories: edge-based similarity measures, based on the edges connecting the terms in the ontology; vector-based similarity measures, for example the Jaccard metric (the size of the intersection divided by the size of the union) and node-based similarity measures. Here we refer to the 'ancestors' of a term in an ontology to mean the set of terms for which it satisfies the transitive 'is-a' relation encoded in the DAG structure of the ontology.

Pesquita et al. [80] argues that edge-based methods, which define similarity between two terms based on the number of edges between them or between them and their common ancestors, are unsuited to biomedical ontologies. Nodes and edges at the same depth (i.e. number of edges away from the root terms — terms with no parents) in biomedical ontologies do not generally have the same specificity and do not represent the same semantic distance respectively. Vector-based methods, although simple to implement, do not leverage the structure of the ontology. We therefore discuss only node-based similarity measures here.

Node-based measures typically define the similarity between two terms in terms of the 'information content'. The information content is defined with respect to the frequency p of the term within some set of ontologically annotated objects:

$$\mathrm{IC}(t) = -\log p$$

Resnik's 'similarity' [91] of two terms is defined as the information content of the most informative common ancestor of two terms:

$$s_{\mathrm{Resnik}}(t_1, t_2) = \max_{t \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \mathrm{IC}(t). \tag{1.1}$$

where $\mathrm{anc}(t)$ denotes the union of the term t and set of ancestor terms of t. Lin et al. [59] proposed an alternative expression for the similarity of two terms. A maximum value of 1 is obtained when two terms are the same (unlike Resnik's measure, where the similarity is just that of their most informative common ancestor).

$$s_{\mathrm{Lin}}(t_1, t_2) = \frac{\max_{t \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \mathrm{IC}(t)}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)} \tag{1.2}$$

There are several ways of defining the frequency of terms, from which the information content can be computed. For example, the frequency could be defined as the frequency of annotation within some collection of ontologically annotated entities, such as the HPO annotated diseases in Online Mendelian Inheritance in Man (OMIM). There exist other methods for evaluating the similarity of two ontological terms, but they are not reviewed here (see Lee et al. [54] for more details).

Similarity between sets of terms

The expressions for the similarity of terms can be extended to give the similarity between sets of terms in different ways. A popular approach is the 'best match' method. This method is based on taking an expression for between-term similarity, s, and for each term in term set A, using it to select the best-matching term in term set B. The similarities of these best matches are then pooled and aggregated, for example by taking the mean, producing a similarity score. The same procedure is then applied to the swapped term sets, and the two aggregated scores are combined by taking the average or product. Thus, the similarity S of term sets A and B can be defined in terms of asymmetric best match function S':

$$
\begin{aligned}
S'(A \to B) &= \frac{1}{|A|} \sum_{t_A \in A} \max_{t_x \in B} s(t_A, t_x) \mathbb{1}_{t_A \in \mathrm{anc}(t_x)}, \\
S_{\mathrm{mean}}(A,B) &= \frac{S'(A \to B) + S'(B \to A)}{2}, \\
S_{\mathrm{product}}(A,B) &= S'(A \to B) \times S'(B \to A)
\end{aligned} \tag{1.3}
$$

The best choice of measure depends on the particular application [80]. For instance, if sensitivity to sharing rare ancestral terms when evaluating the similarity of two sets of terms is desired, a best match average measure based on Resnik's expression for the similarity of two terms is appropriate as it is equal to the information content of the most informative ancestral term, and therefore sharing a rare term contributes substantially to the similarity. On the other hand, a best match average measure based on Lin's expression is more robust, as two terms have a maximum similarity of 1, which occurs if they are the same. Therefore, averaging over the similarities of multiple terms in a set strongly reduces the contribution of rare terms present because of noise.

Software for working with ontological data

The ontoCAT R package enables simple querying and traversal of ontologies, but many of its key functions are slow and dependency on Java hinders its portability. There are software packages enabling manipulation and plotting of graphs, for example the R packages graph [35] and Rgraphviz [40] respectively, which can be used to view sections of ontologies. However, their functions are low level, which makes procedures such as plotting of ontological term sets and fine-grained control of graphical parameters quite involved. There are R packages which provide procedures for computing semantic similarities between terms and sets of terms for specific ontologies [31, 122, 123] but they do not support semantic similarity computation for arbitrary ontologies. The the majority of these packages, including the primary ones GOSim and GOSemSim, depend on the GO.db package [13], which stores information about the Gene Ontology (GO) and GO gene annotations in a precompiled SQLite databases and is accessed through the DBI [89] package. This makes it inflexible — for example, by preventing the user from using a custom information content upon which to base semantic similarity calculations. They also do not allow the computation of semantic similarity between arbitrary term sets, in spite of this being a critical task for many applications, instead, only allowing semantic similarity between annotated genes specified (by default) through Entrez IDs. Furthermore, the methods provided by these packages are quite slow applied large datasets comprising hundreds or thousands of sets of terms.

In summary, there is a paucity of flexible routines for calculating semantic similarities in R, and there has been no software which enables ontological data relating to arbitrary ontologies to be manipulated, easily visualised and used in semantic similarity calculations in mainstream statistical languages.

Methods for analysing ontologically encoded phenotypes

Phenomizer Phenomizer [47] is a differential diagnosis tool which uses semantic similarity to rank HPO encoded disease phenotypes derived from OMIM by their relevance to a query set of HPO terms. The similarities are computed using the best-match average based on Resnik's expression. There is a tendency for semantic similarity measures to exhibit annotation bias whereby high similarity is observed for sets of terms containing many terms. Therefore, instead of using the raw semantic similarity, a p-value is estimated by permutation, i.e. the similarity to the query set of HPO terms is compared to the distribution of similarities between the query set and randomly selected sets of the same number terms. Diseases are then ranked by the p-values obtained.

BOQA BOQA [6], which stands for 'Bayesian Ontology Query Algorithm', is an algorithm for differential diagnosis designed to be noise and imprecision tolerant by modelling a query phenotype as an outcome which depends on a latent disease state. It uses a three-layer Bayesian network: the bottom layer is the configuration of rare diseases for a given individual (i.e. a Boolean vector corresponding to disease status for each disease), the middle layer is the configuration of ontological phenotype terms which depends on the configuration of diseases, and the top layer is the query layer which depends on the second layer. Ontological structure is modelled by propagating parent-child relations through the ontology in the phenotype configuration and query layers. Noise (nuisance terms) and imprecision (missing/less specific terms) are each modelled using a constant probability of occurrence. Marginal probabilities of each disease are then inferred and used to rank differential diagnoses.

Exomiser Exomiser [102] is an application which ranks candidate variants for a disease-affected individual or family by combining various sources of information relevant to the likely pathogenicity. Information used includes a phenotypic relevance score based on semantic similarity between the individual's ontologically encoded phenotype and any phenotype terms associated with the genes in which the candidate variants lie.

Statistical methods for relating rare variants to rare diseases

Probability of loss-of-function intolerance Haploinsufficiency for a particular gene occurs when disease is caused by having only a single functional copy of a gene, although not all genes are prone to haploinsufficiency. The probability of loss-of-function intolerance score (pLI) [55] for a gene is an estimated probability of haploinsufficiency given an observed number of protein truncating alleles across samples. The estimate is based on modelling the protein truncating allele count for each gene using a Poisson distribution whose rate depends on a latent class of the gene (haploinsufficient, recessive or Protein Truncating Variant (PTV) tolerant). The rates are estimated *a priori* based on the observed number of protein truncating alleles for genes whose class is known. Notably, virtually all genes which are known to cause disease through haploinsufficiency score highly (pLI > 0.9), and 72% of genes with pLI > 0.9 have no known disease phenotype, despite the strong evidence of constraint against PTVs.

Burden testing A typical genome differs from the human reference genome at approximately 4 million sites, with $40,000$-$200,000$ of these variants having frequencies below 0.5% [1]. Thus, even after filtering variants (for example based on population allele frequency and predicted consequence as described earlier), there will be many benign candidate variants and allele counts will be too low to have good power by testing marginally for association. Burden tests aggregate rare variants within a genomic locus in order to boost power when testing for association with a disease phenotype. An example of a Burden test is the 'cohort allelic sums test' [72] which uses a genetic score that is equal to 1 if an individual carries at least 1 (or 2) alleles at the filtered variant sites under a dominant (or recessive) inheritance model, and 0 otherwise. A Fisher exact test is then used to compute a p-value of association between a dichotomous phenotype and the genetic score.

SKAT SKAT specifies a random effect for each variant and performs a score test under the null hypothesis that the variance of the random effects is zero. The variance-covariance structure of the random effects under the alternative hypothesis is determined by a kernel function, which would typically be a weighted genetic correlation across the variants in the locus. SKAT can incorporate nuisance covariates, accounts for linkage disequilibrium between variants under consideration and is well-powered for traits whereby many different variants in a locus with varying effect sizes and allele frequencies contribute to the phenotype. The following model for disease risk conditional on genotypes is employed by the test:

$$\text{logit}\,\mathbb{P}(y=1) = \alpha_0 + X\alpha + G\beta,$$

where y is the disease status of n individuals ($y_i = 1$ if individual i is affected, 0 otherwise), X is a matrix of covariates, G is the genotypes of the individuals at the variant sites in some genetic locus (i.e. $G_{ij} \in \{0,1,2\}$ equalling the number of copies of the rare allele at variant site j carried by individual i) for homozygous, α_0 is the background rate of disease, and α and β are random effects. The SKAT test is performed by assuming that the β values follow a distribution with a mean of zero and variance of τ. Testing the null hypothesis that there is no association with disease (i.e. $\beta = 0$) is therefore equivalent to testing $\tau = 0$. This can be done using a variance-component score test, and thus only requires the null model to be fitted. The score statistic is:

$$Q = (y - \hat{\mu})' K (y - \hat{\mu}),$$

where $\hat{\mu}$ is the estimated mean of y under the null, and K is an $n \times n$ matrix whose entries depend on the kernel function K, $K_{ii'} = K(G_{i\cdot}, G_{i'\cdot})$, giving the genetic similarity of individuals i and i'. K should be chosen depending on the kinds of effects which are modelled. By default, the weighted linear kernel is used, i.e. $K(G_{i\cdot}, G_{i'\cdot}) = \sum_{j=1}^{p} w_j G_{ij} G_{i'j}$, where w_j is the weighting for variant j. Epistatic effects can be modelled by using the weighted Identity By State (IBS) kernel, i.e. $K(G_{i\cdot}, G_{i'\cdot}) = \sum_{j=1}^{p} w_j (2 - |G_{ij} - G_{i'j}|)$, and all main effects and pairwise interaction effects can be modelled using the *quadratic* kernel, i.e. $K(G_{i\cdot}, G_{i'\cdot}) = \sum_{j=1}^{p} w_j (1 + G_{ij} G_{i'j})^2$.

The score statistic follows a mixture Chi-square distribution under the null, for which p-values can be computed efficiently using the Davies method [19]. When the weighted linear kernel is used, Q can be computed given only the individual variant test statistics for marginal involvement, S_j:

$$Q = \sum_{j=1}^{p} w_j S_j^2,$$
$$S_j = G_{\cdot j} (y - \hat{\mu}).$$

Thus K need not be evaluated, and the value of the statistic can be computed rapidly. However, if a non-linear kernel is used, this is no longer possible and SKAT quickly becomes computationally expensive, with complexity at least $\mathcal{O}(n^2)$. In the available implementations [53, 95] K is evaluated explicitly, requiring memory for an $n \times n$ dense matrix to be allocated, therefore leading to a prohibitively large memory footprint when the sample size is large.

SKAT is more powerful than the Burden test when the proportion of variants in the analysis which are pathogenic is lower, or when the effect directions of the variants is mixed, and vice-versa when the proportion of pathogenic variants is close to 1 and the effect sizes are in the same direction. SKAT-O [52] maximises power by forming a test statistic which is a linear combination of the SKAT and Burden test statistics, estimated adaptively from the data.

Methods which identify likely pathogenic variants For scientific follow-up, it is important to infer which variants are likely to be pathogenic, conditional on an association being present in a given locus. The *backwards elimination* [43] procedure removes individual variants iteratively from the predictors as long as this increases a test statistic of association (either Burden or SKAT). The *adaptive combination of p-values* (ADA) [60] algorithm ranks variants by p-value obtained using Fisher's exact test and selects a threshold on p-value that maximises an aggregate test statistic. As these algorithms prune variants in a step-wise fashion, they do not explore the full space of possible combinations of pathogenic variants. It is also important that inference can be performed sufficiently quickly to enable application across tens of thousands of regions, with tens to hundreds of variants in each one. ADA and backwards elimination SKAT/Burden rely on permutations to obtain empirical p-values, rendering them too computationally expensive to apply in these scenarios.

In principle, Bayesian inference lends itself well to rare variant association analysis because it provides a coherent framework for sharing information across variants and provides a natural way of incorporating prior information on variant pathogenicity. The *variational Bayes discrete mixture* (vbdm) method [63], the *Bayesian risk index* (BRI) [87], the *integrative Bayesian model uncertainty* method (iBMU) and the *Bayesian rare variant detector* (BRVD) [58] all model a mixture of pathogenic and non-pathogenic variants in a locus. The iBMU method includes a hierarchical model for the probabilities that individual variants contribute to disease risk given a set of predictor-level covariates. However these models employ additive models of disease risk or severity suited to complex diseases rather than rare diseases caused by dominant or recessive inheritance of one or two pathogenic alleles. Typically, Bayesian methods are slow, due to the necessity of averaging over prior distributions, and often depend on Markov chain Monte Carlo (MCMC) inference procedures. This is particularly so in variable selection scenarios, such as selecting a set of involved variants. The vbdm method uses a variational Bayes inference technique, derived by factorising the likelihood of model parameters over the indicators of pathogenicity for each variant, which makes inference rapid. However, the derivation of the algorithm depends on an additive

model for a quantitative trait, hence it cannot be extended to a model for Mendelian modes of inheritance.

Chapter 2

Methods for ontological data and applications

Ontologies are widely used in bioinformatics, but the absence of simple, general tools for manipulating ontological data has made analysis of such data unnecessarily difficult. In particular, in order to develop complex methods which utilise rare disease phenotype data encoded using HPO terms, a solid software layer upon which to build them is required. This chapter begins by presenting 'ontologyX' [39], a suite of R packages which simplify and harmonise the manipulation, analysis and visualisation of ontological data for arbitrary ontologies. The ontologyIndex package enables arbitrary ontologies to be read into R, supports representation of ontological objects by native R types, and provides a parsimonius set of performant functions for querying ontologies. ontologySimilarity and ontologyPlot extend ontologyIndex with functionality for straightforward visualization and semantic similarity calculations respectively. Methods for analysing collections of ontologically annotated entities are then discussed, including assessment of the statistical significance of ontological similarity within-group ontological similarity of groups of term sets, prioritisation of variants based on similarity between HPO-encoded phenotype and disease HPO profiles, comparison of ontologically encoded rare disease phenotypes with those of model organisms and analysis of clusters generated using unsupervised clustering algorithms.

The description of the ontologyX software given in this Chapter is based on that given in: D. Greene, S. Richardson and E. Turro (2017). "ontologyX: A suite of R packages for working with ontological data", *Bioinformatics*, 33(7):1104-1106.

2.1 ontologyX

ontologyIndex

ontologyIndex is an R package which I developed in order to provide a low-level set of functions for exploiting the structure of ontologies. Ontologies can be read into R from files in OBO format, with most commonly used ontologies available in this format on the OBO Foundry's website [103]. Ontologies which are only available in a OWL format may be used by first converting them into OBO format, for example using the ROBOT command line tool [79]. OWL format allows expression of relationships which can't be expressed in OBO format, so therefore information may be lost in conversion to OBO format. For example, OWL enables quantified relationships between terms and relationships with arbitrary cardinalities to be represented, for example, 'insect `has 6` legs'. However, the functionality exposed by ontologyIndex only depends on the is-a relationship being defined, so does not depend on any of these more complex relationships, and hence it is always sufficient to convert to OBO format beforehand. A custom internal representation of ontologies — the `ontology_index` class — stores properties of terms. An `ontology_index` object is a `list` containing named slots corresponding to term properties. Each slot is then a `vector` or `list` containing one value per term, depending on the cardinality of the property: `vectors` are used where the cardinality of the property is exactly one, for example `$id`, as each term has exactly one ID, and `lists` otherwise. This enables properties of individual terms or sets of terms to be looked up in the `ontology_index` using base R functions, in a way which is identical to using `data.frames`, which are widely used amongst R users. The `get_ontology` function is used to read ontologies into R. By default, only the minimal properties used by the main functions in the package are read in from a given OBO file. These are the `id`, `name`, `parents`, `children` and `ancestors` properties (the latter two derived from the `parents` field). However, all properties which are present in the file can be read by passing the argument `extract_tags="everything"` to `get_ontology`. Although the `ancestors` slot depends on the `parents` slot, storing the ancestors of each term enables considerable speed-up of ontological operations that depend on term ancestry, such as collecting the full set of ancestors or descendants for a set of terms (see Table 2.1). Ontologies are usually 'wide' in the sense that the average number of child terms is relatively high compared to the average number of parent terms and therefore the additional memory required to store the ancestor terms for each term is relatively low. For example, the average number of child terms per term with children in the HPO is 3.72 whilst the average number of parent terms for terms with parents is 1.32.

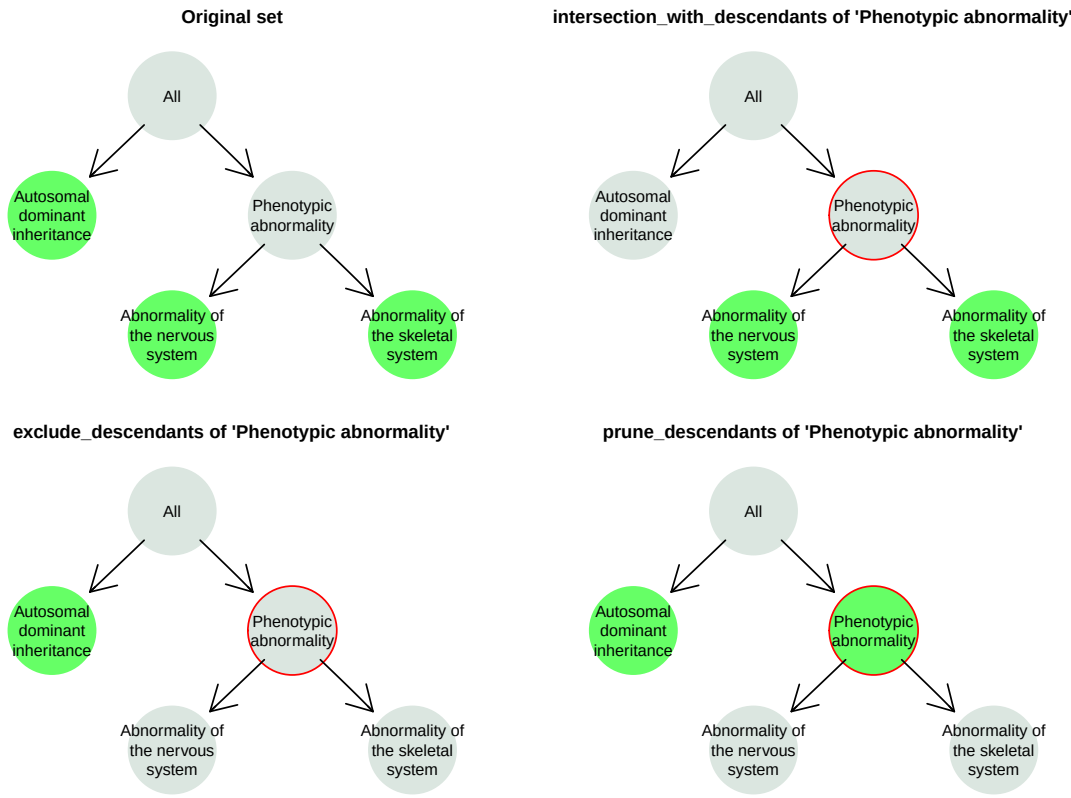


Fig. 2.1 Visualisation of ontological set operations in ontologyIndex. In this diagram, the set of terms 'Autosomal dominant inheritance', 'Abnormality of the nervous system' and 'Abnormality of the skeletal system' — coloured green in the top left panel — are subsetted in three different ways with respect to a given root term (here 'Phenotypic abnormality', highlighted with a red circle). The remaining terms after applying the operations are shown in the other three panels labelled by the corresponding operation name.

ontologyIndex uses native R types to represent ontological terms and sets of terms (i.e. `character` vectors of term IDs), enabling simple integration with R's features, high-level functions, and other packages. It includes functions for performing set operations respecting the structure of the ontology, for example: `exclude_descendants`, which given term sets A and B, excludes terms in B and their descendants from set A; `prune_descendants`, which preserves terms in B that are ancestors of terms in A after applying `exclude_descendants`), and `minimal_set`, which maps a set of ontological terms onto a non-redundant set. Figure 2.1 depicts example outcomes of applying ontologyIndex functions.

ontologyIndex is lightweight, fast, and readily extended by other packages. For example, my R package gsEasy [36] facilitates gene-set enrichment analysis [110] using the `get_ancestors` function to propagate parent-child relations through the GO.

Table 2.1 Mean execution time for retrieving descendants and ancestors for individual terms in the Human Phenotype Ontology.

	Descendants (ms)	Ancestors (ms)
ontoCAT	11.99	12.75
ontologyIndex	0.38	0.14

gsEasy

Gene Set Enrichment Analysis (GSEA) is a procedure for assessing whether a given set of genes is over-represented within the highest ranks of a larger containing set of genes with respect to a given gene scoring, e.g. the log fold change in expression. p-values are calculated based on an 'enrichment score' (ES), representing over-representation of the given set of genes, by comparing it to the null distribution of the score derived by permutation. ES is obtained by computing the maximum value of $P_{hit} - P_{miss}$, defined as:

$$
\begin{aligned}
P_{hit}(S,i) &= \sum_{j \in S, j \leq i} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p \\
P_{miss}(S,i) &= \sum_{g_j \notin S, j \leq i} \frac{1}{N - |S|}, \\
ES(S) &= \max_i P_{hit}(S,i) - P_{miss}(S,i),
\end{aligned}
$$

where r_i is a score for the gene with the i^{th} highest score amongst all included N genes, S is the set of ranks for the genes in the test set, and p is a fixed constant which powers the scores.

In spite of its conceptual simplicity, its considerable popularity (the introductory paper, Subramanian et al. [110], has been cited over 11,000 times), and the ubiquity of R amongst bioinformaticians, it has been difficult to perform in R because of the absence of simple functions which execute rapidly and difficulty in defining gene sets corresponding to GO annotation. gsEasy [36] is a software package which I developed in order to make GSEA straightforward, flexible and fast in R. It extends functionality in the ontologyIndex package in order to derive sets of genes which are annotated with particular ontological terms. Use of native R types as parameters, independence from external databases and implementation in C++ facilitates simple, effective application across large numbers of gene sets using R's high level functions.

It has a function, `gset` for calculating p-values based on ES, employing the same terminology and variable naming for its function arguments as are used in the original paper

[110]. In order to test a set of five genes which appeared at the top five ranks out of 1000, the command to calculate an enrichment p-value using `gset` can be given as:

```
> gset(S=1:5, N=1000)
[1] 9.9999e-06
```

Note that if the total number of genes `N` is specified in place of the scores for the genes `r`, then the value of r_i is set to $\frac{N-i+1}{N}$.

In contrast to other GSEA packages, it does not require laborious manipulations of input data. Sets of genes annotated with the same GO term are derived from the GO term annotations of genes using ontologyIndex [39] by selecting the set of genes which are annotated explicitly or implicitly using each term. To achieve this, gsEasy has a function `get_ontological_gene_sets` for creating lists of gene sets defined by ontological annotations of genes, which also works for ontologies other than the GO.

ontologyPlot

ontologyPlot extends ontologyIndex with functions which simplify the task of plotting sets of terms and the 'is-a' relations between terms contained by them through allowing the user to pass only an `ontology_index` and a vector of term IDs as the `terms` argument in the plotting function, `onto_plot`. ontologyPlot utilises the Rgraphviz package's interface to the graphviz [32] graphical layout engine. ontologyPlot also allows graphs to be exported in standard DOT format using the function `write_dot` without constraining the graphical parameters included. Thus, users can take full advantage of options in any rendering software (e.g. `graphviz`). Possible graphical options which the user can set using the standard Rgraphviz engine include colour of node borders (`color`), colour of nodes (`fillcolor`), node labels (`label`), label font size (`fontsize`), node shape (`shape`), and node size (`width`). Arguments controlling the graphical parameters can be given as single values (setting all nodes to have the same value for the parameter), vectors (where the components correspond to the values of the parameter for each term), or functions, which determine how to set the parameters after the plotting function is called in the context of the other arguments given. Such functions supplied with the package include `label_by_term_set` which sets the labels of the nodes to the name of the term and a comma-separated list of the names of sets of terms the term is included in, either explicitly (because the term belongs to the set) or implicitly (because a descendant term belongs to the set), in the `term_sets` argument (if one is given). This makes it simple to generate informative plots for a set of ontologically annotated entities. Another function, `width_by_significance` can automatically set the

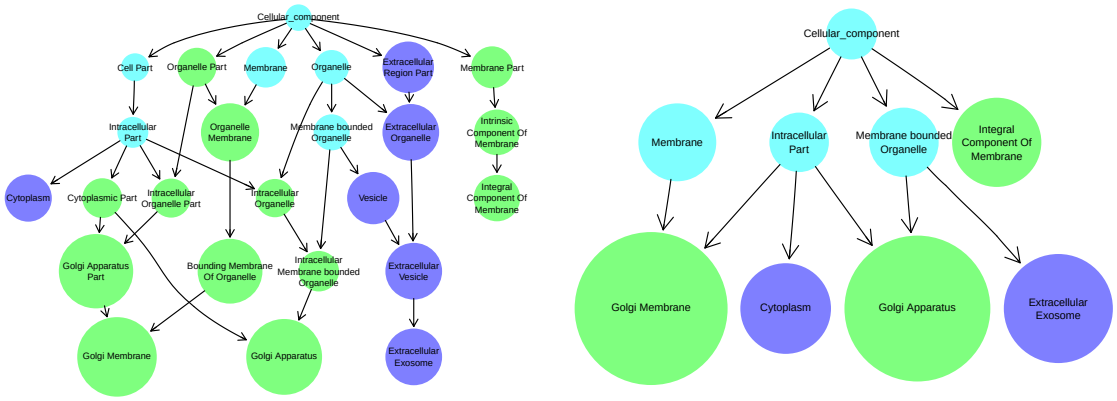


Fig. 2.2 Plot of terms descending from the `cellular_component` term in the GO, extracted using the `exclude_descendants` function from ontologyIndex, for genes *QPCTL* and *CRNN* using ontologyPlot. The left panel shows the full set of ancestral terms used in the annotation of the genes, while the right panel shows only those remaining after `remove_uninformative_terms` has been called. Terms annotated to both genes, either implicitly or explicitly, are shown in light blue, while those annotated only *QPCTL* and *CRNN* are shown in green and dark blue respectively. The size of the nodes has been set to be proportional to the information content (i.e. negative log frequency) of the terms with respect to gene annotation downloaded from the GO website.

size of terms. It requires a list of term sets, `term_sets`, and a vector of term population frequencies, `frequencies` to also be supplied to `onto_plot`. The widths of the nodes are then set to the minus log p-values of their corresponding terms' significance of occurrence amongst `term_sets` under the null hypothesis that the frequency within the term sets is the same as the population frequency. This enables visualisations which highlight the most significant terms to be generated.

It includes several functions for transforming sets of terms to distill the important features for particular visualisations. For example, given a set of ontologically annotated objects, the function `remove_uninformative_terms` removes terms whose children are annotated to the same objects, leading to simpler diagrams. Another such function is `remove_links`, which given a set of terms, removes those terms which only link two other terms in the set together, i.e. terms which have exactly one parent and one child within the given set. Figure 2.2 demonstrates how ontologyPlot can be used to visualise GO annotation for *QPCTL* and *CRNN*, and the effect of using `remove_uninformative_terms` to simplify the figure.

ontologySimilarity

Semantic similarity quantifies similarity between ontological terms and sets of terms. ontologySimilarity extends ontologyIndex to enable similarities between ontological objects to be computed given an `ontology_index` and sets of term IDs. It facilitates the calculation of similarity at three levels: between ontological terms (ID strings), between ontologically annotated objects (ID string vectors), and within groups of ontologically annotated objects (lists of ID string vectors). It implements Resnik's [91] and Lin's [59] expressions for the similarity of terms. Unlike other packages for calculating semantic similarities, ontologySimilarity does not depend on static, pre-built SQLite databases or Bioconductor annotation packages and works with arbitrary term annotations. Furthermore, it offers inferential procedures such as `get_sim_p`, which assesses the strength of similarity between groups of objects (see Section 2.2). Flexible functions facilitate use in complex methods, for example as in my R package SimReg [37] described in Chapter 3, which implements a semantic similarity based regression algorithm. All similarity routines are written in C++ and called from R [21], and the user can balance performance and memory usage for downstream analysis by selecting whether to store similarities between terms or term sets, or store an index for fast similarity lookups. I compared the performance of ontologySimilarity against other packages offering functions for calculating pairwise term and gene similarities, the results of which are shown in Table 2.2. The results indicate that ontologySimilarity executes substantially faster, in some scenarios thousands of times so, making it much more convenient for use with large datasets.

Table 2.2 Execution times for computing pairwise similarity matrices for 1000 randomly selected GO terms and 100 randomly selected sets of GO terms annotating genes.

	Term sim (s)	Gene sim (s)
GOSim	1075.43	298.34
GOSemSim	1.71	116.72
ontologySimilarity	0.31	0.06
ontologySimilarity (indexed)		0.04

The following versions of software packages were used to generate the results presented in this section: ontologyIndex 2.2, ontologyPlot 1.4, ontologySimilarity 2.1, GOSim 1.11, GOSemSim 1.99.4 and ontoCAT 1.26.0.

2.2 Statistical significance of within-group ontological similarity

Work described in this Section is based on content from: S. K. Westbury†, E. Turro†, D. Greene†, C. Lentaigne†, A. M. Kelly†, T. K. Bariana†, et al. (2015). "Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders". *Genome Med*, 7(1):36 († indicates equal contribution).

In this section I describe methods for evaluating the statistical significance of the similarity between members in a group of ontologically annotated entities. The methods for computing the within-group ontological similarity of a group of term sets presented here will be based on expressions for the similarity of pairs of terms (Equations 1.1 and 1.2) and for the similarity of pairs of term sets (Equation 1.3). The within-group ontological similarity of the group will then be the mean or minimum similarity across the pairs of objects. The similarity S of group of term sets g with respect to between-term set similarity s could then be evaluated respectively using the expressions:

$$S_{\text{mean}} = \binom{|g|}{2}^{-1} \sum_{a,b \in g} s(a,b),$$

$$S_{\text{min}} = \min_{a,b \in g} s(a,b).$$

Using the minimum, S_{min} as opposed to the mean, S_{mean}, leads to a high similarity only if all pairs of term sets in the group have high similarity. The group similarity does not give an indication of the significance of the similarity of a group of term sets. I therefore developed a statistical test for assessing the within-group ontological similarity of a group, g, contained in a larger collection of term sets, G, through the computation of a p-value by permutation test. The method works by comparing the group similarity of g to the group similarity of the subsets of G with the same size as g. Typically $|G|$ would be large enough so as to make computing the group similarity of all $\binom{|G|}{|g|}$ prohibitively computationally expensive. Hence, in practice the p-value would be approximated using Monte Carlo simulation: N subsets of G with size $|g|$, A, are randomly sampled, and the proportion with at least as high group similarity as g is an unbiased estimate of the p-value:

$$\frac{1}{N} \sum_{g' \in A} \mathbb{1}_{S(g') \geq S(g)}. \tag{2.1}$$

This procedure is implemented in the ontologySimilarity package by the function `get_sim_p`. For efficiency in the situation where only statistically significant results are of interest, `get_sim_p` allows the user to configure a threshold triggering the sampling to stop early, should the strength of evidence against it being significant become strong enough. To be precise, this is done by specifying the minimum and maximum number of samples to draw, the maximum p-value considered 'significant' and a threshold probability of significance, below which sampling stops. The probability that the function would return a significant p-value is computed using a normal approximation to the binomial.

Application to bleeding and platelet disorders

One of the subprojects of the NBR–RD is the BPD study. Seven hundred and seven participants were recruited from over 30 different centres worldwide under the inclusion criteria of unexplained bleeding and/or platelet abnormality affecting count, volume, morphology or aggregation. Eighty novel HPO terms were created in order to capture phenotypic abnormalities of those recruited, with the majority being subclasses of 'Abnormality of blood and blood-forming tissue' [119]. The phenotypic data gathered for each participant includes the HPO-encoded phenotype, full blood counts including 15 quantitative measurements (including platelet, leukocyte and red-blood cell counts), and family history. Diagnosis of BPD currently requires detailed clinical evaluation, but a defective coagulation factor or specific platelet pathway is implicated in only 40-60% of cases [88] and the unknown genetic basis of many BPDs means the rate of genetic diagnosis is low. BPD disease phenotypes are quite heterogeneous, frequently involving combinations of abnormalities of platelet aggregation, blood cell counts, and blood vessel walls as well as other organ systems, for example, the skeleton in thrombocytopenia absent radius syndrome. Figure 2.3 shows the distribution of comorbidities across individuals with each qualifying phenotype. Hence, establishing high-throughput sequencing as a diagnostic tool for BPDs presents an interesting and important challenge for methodological development.

In order to demonstrate that ontological similarity measures provide useful quantification of phenotypic similarity, we calculated the ontological similarity between the phenotypes of all pairs of BPD cases. The Resnik expression for the similarity of terms (Equation 1.1) and 'best match average' expression for the similarity of term sets were used in order to be more sensitive to the co-occurrence of rare terms within groups. Figure 2.4 shows a symmetric heatmap of case-case phenotype similarity, where the rows and columns represent cases grouped by family and clinically diagnosed syndrome where available. Inspection of the main diagonal indicates members of the same group, either based on pedigree or provisional clinical diagnosis, tend to have increased phenotypic similarity. Significance of

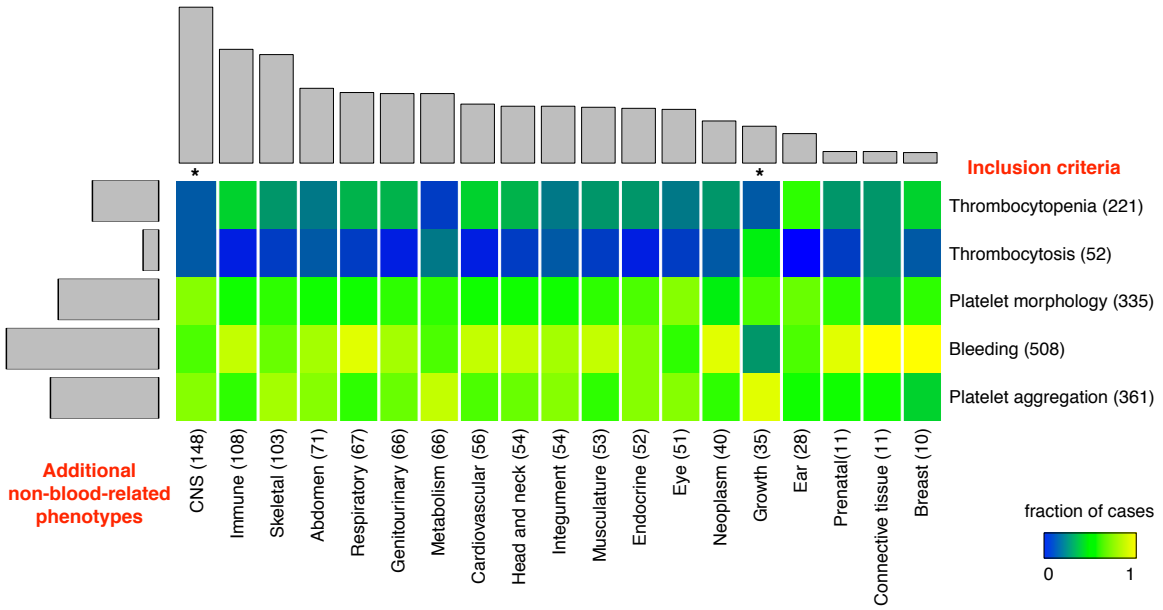


Fig. 2.3 Heatmap of the relative frequencies of HPO terms related to bleeding and platelet abnormalities annotating patients with abnormalities affecting other organ systems for 707 BPD index cases [119]. The numbers in the brackets and the bar plots indicate the number of index cases with at least one HPO term pertinent to abnormality in the organ or disease area after removal of overlapping terms. *indicates that the distribution of terms pertinent to enrolment for a particular column is significantly different compared to the sum (along rows) of all other columns (p-value < 0.05 after Bonferroni correction by chi-squared test). The columns are ordered by the number of cases having a term in each abnormality class.

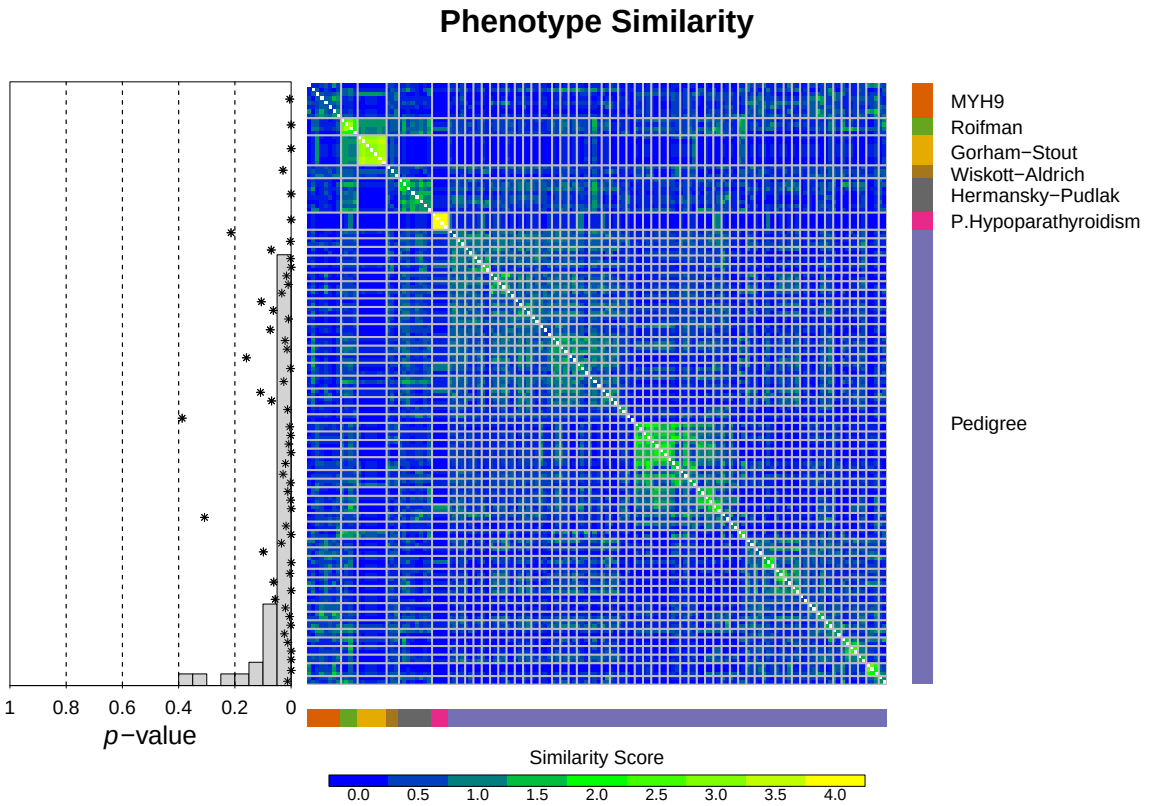


Fig. 2.4 Symmetric heatmap of similarity between pairs of BPD cases grouped by pedigree and clinical diagnosis (labelled on y-axis). The scatter plot on the left shows the p-values (x-axis) for similarity of these groups (y-axis) computed using the expression in Equation **??**. The histogram superimposed shows the distribution of these p-values.

group similarity for each group was assessed by computation of p-values using the expression in Equation 2.1. These p-values are shown in the histogram on the left in Figure 2.4. The majority (43/56) are less than 0.05. Indeed, a meta-analysis using Fisher's method to assess whether these groups clustered closely together as a whole yielded a p-value less than the numerical resolution of the analysis software.

2.3 Rare variant prioritisation for genetic diagnostics

Work described in this Section is based on content from: I. Simeoni, J. C. Stephens, F. Hu, S. V. Deevi, K. Megy, T. K. Bariana, ..., D. Greene et al. (2016). "A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders". *Blood*, 127(23):2791–2803.

The MDT meetings which are held to interpret results generated by the ThromboGenomics platform, described in Section 1.4, are time-consuming and expensive. We therefore used a sample of the output data to assess whether HPO-based phenotypic similarity could help prioritise variants identified by targeted sequencing assays. We used phenotypic similarity measures to compare phenotypes linked to the diseases associated with the 63 BPD genes curated by a team of clinical experts and transcribed into HPO terms, and the phenotypes of 109 cases under investigation with the platform were HPO-encoded. The phenotypic similarities between the curated gene HPO profiles and the observed case HPO phenotypes were computed using Lin's expression for the similarity between terms (Equation 1.2) and the 'best match average' expression for the similarity between term sets. Here, Lin's expression was employed instead of Resnik's in order to achieve an even weighting for the importance of terms in the curated gene HPO profile, rather than have it strongly depend on the information content. For each case, the gene in which they carried a variant passing the ThromboGenomics filters having the highest phenotypic similarity to the case phenotype was selected. As an example, a case with Bernard-Soulier syndrome was coded with six HPO terms and subsequently found to carry candidate variants in four genes. The homozygous variant in *GP1BB*, identified independently by the MDT as likely pathogenic, was chosen by the prioritisation as the top candidate because the HPO profile linked to *GP1BB* was more similar to the HPO phenotype of the case than the curated gene HPO profiles of any of the other three genes in which the case had a candidate variant (left-hand panel of Figure 2.5). The right-hand panel of Figure 2.5 shows the overall results, which indicate that in 85% (93/109) of cases, the correct gene, as previously identified by the MDT, scored the highest similarity to the case phenotype out of all the candidate variants ($p < 10^{-6}$). Furthermore, where the top-ranked gene did not correspond to the MDT-designated gene, the difference tended to be smaller than when there was concordance, as shown in Figure 2.6. Thus, prioritisation based on phenotypic similarity may be a useful tool for supporting the MDT review process.

2.4 Similarity to model organism phenotypes

Work described in this Section is based on content from: E. Turro, D. Greene, A. Wijgaerts, C. Thys, C. Lentaigne, T. K. Bariana, et al. (2016). "A dominant gain-of-function mutation in universal tyrosine kinase *SRC* causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies". *Science Translational Medicine*, 8(328):328ra30.

Variants in orthologous genes often manifest in similar phenotypes, therefore model organisms are used to study phenotype-gene associations. These associations are typically studied using mouse models, as the mouse is an established experimental tool and almost

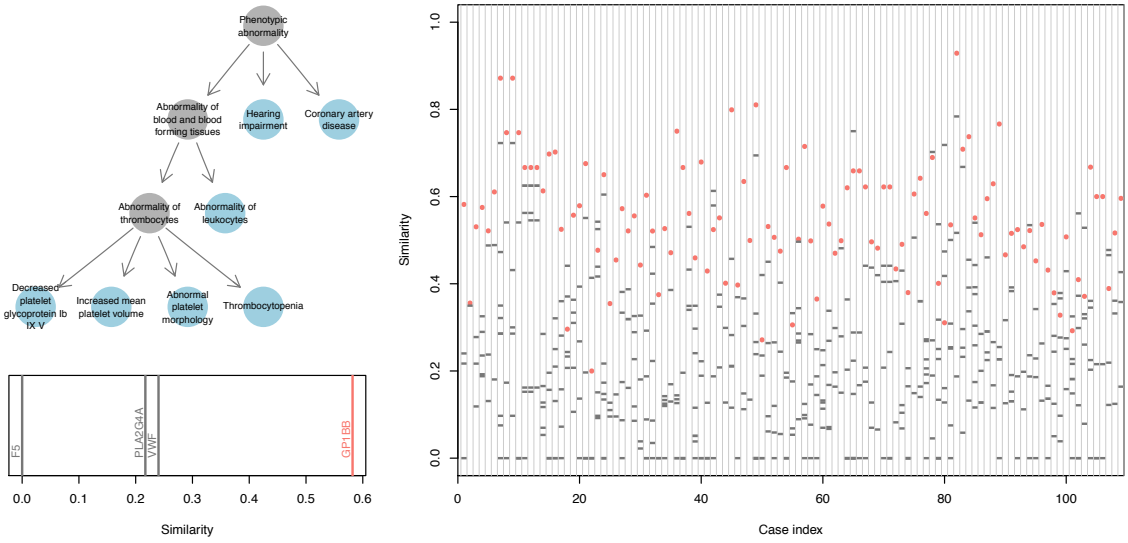


Fig. 2.5 Left: the HPO profile of a specific individual with *GP1BB*-related Bernard-Soulier syndrome, and beneath, the phenotypic similarity to the disease profiles of each gene in which the individual carried a filtered variant. Right: the phenotypic similarities between 109 thrombogenomics cases and the disease profiles corresponding to genes in which they carried filtered variants. The similarities to disease profiles of the genes containing the variants determined to cause individuals' diseases are shown in red, whilst the similarities to the disease profiles of other genes in which individuals carried filtered variants are shown in grey.

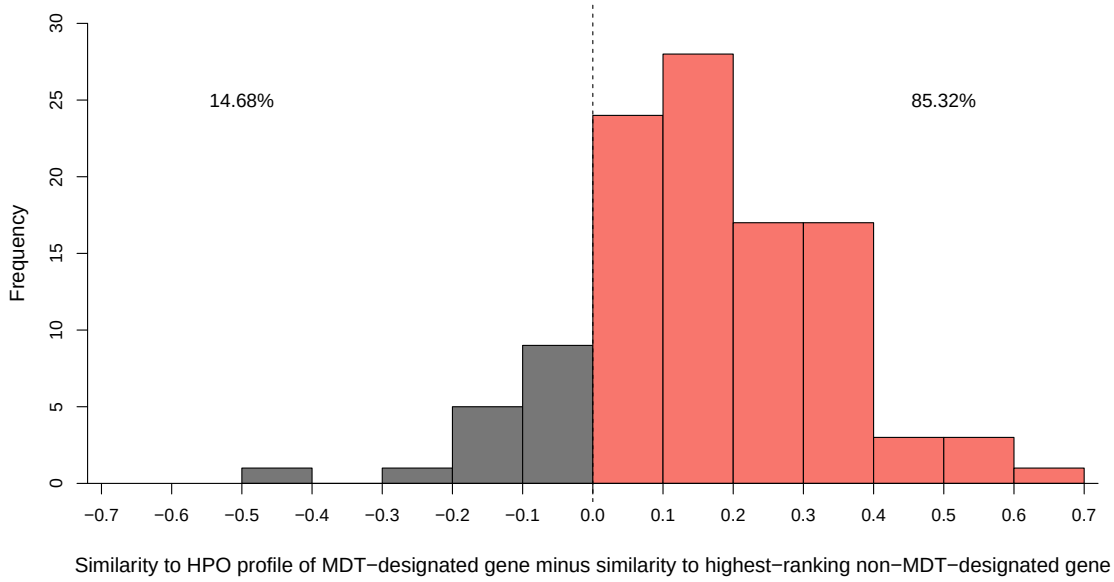


Fig. 2.6 Histogram summarising the data shown in Figure 2.5. The x-axis represents the difference of the similarities between the case HPO phenotype and the HPO profiles of the 'MDT designated' and 'most similar non-MDT designated' genes.

all mouse genes have a human ortholog, with an average of 85% sequence identity [42]. Genes are inactivated or 'knocked-out' by artificially introducing mutations which lead to dysfunctional gene products. Attempts to generate comprehensive datasets on mouse models are currently underway by various organisations. The Knockout Mouse Project (KOMP) is an American initiative set up to create cell lines containing null mutations in all genes in the mouse genome. The International Mouse Phenotyping Consortium (IMPC) is an international organisation which, amongst other objectives, aims to phenotype all of these mouse lines, systematically recording phenotype abnormalities using MPO terms. The Monarch Initiative [73] is a web-based platform which aggregates model organism data from many sources (including the IMPC), relating genotype and phenotype across multiple species and enabling researchers to explore genotype-phenotype associations across species. Data exposed by the platform include the phenotypic annotation of model organisms for orthologs of 14,779 of the 19,008 protein coding human genes (78%), where only half of these have phenotypic annotation for humans.

Using cross-species phenotype ontologies [46], one can map these MPO encoded mouse knockout phenotypes to HPO profiles, facilitating similarity comparisons with human phenotypes. Thus, given candidate variants for explanation of a disease affecting a given family, the variants can be prioritised by phenotypic similarity between the HPO profile of the knockout

mouse model for the genes harbouring the variants and the HPO encoded phenotypes of the affected family members.

This approach was used to analyse candidate variants for a three-generation family presenting with dominantly inherited thrombocytopenia, myelofibrosis, bleeding, platelet dysfunction with abnormal α-granules, and bone pathologies [114]. DNA from two of the affected cases was sequenced and a set of candidate variants was formed by retaining shared rare variants (with respect to frequency in reference collections) predicted to affect amino acid sequence. The resulting set consisted of 67 plausible causal variants in 67 candidate genes.

To obtain the ranking by phenotypic relevance of the candidate variants, we constructed HPO profiles for each of the corresponding genes and computed the mean phenotypic similarity to three HPO-coded affected individuals in the family. The gene-based HPO profiles were constructed by combining HPO terms associated with each gene (genes were assigned terms belonging to the HPO profiles of diseases [92] with which there were associated through OMIM) and HPO terms derived from mapping MPO terms associated with mouse models [8] for the knocked-out orthologous gene using a cross-species ontology [46]. The phenotypic similarity measure used was the Resnik best match average approach, where the information content assigned to each term was derived from its frequency among 856 unrelated HPO-coded individuals from the NBR–RD BPD subproject, in order to be sensitive to the case phenotypes sharing rare phenotype terms with the gene HPO profiles. ontologySimilarity was used to compute the phenotypic similarities between the HPO-coded phenotypes and the gene HPO profiles.

This approach ranked *SRC* at the top of the 31 candidate genes for which phenotype term data were available (shown in Figure 2.7B). *SRC* encodes the proto-oncogene tyrosine protein kinase SRC. The affected family members carried a variant encoding amino acid substitution E527K, and modelling the effect of the substitution on protein structure suggested it to be a gain-of-function variant due to the predicted loss of SRC's self-inhibitory capacity, later confirmed by *in vitro* studies.

Knockout of its highly homologous ortholog in mice, *Src*, showed no bleeding and apparently normal platelets [105]. However, the mouse model presented with increased bone density (osteopetrosis) (see the HPO-translated phenotypic profile in Figure 2.7C, generated using ontologyPlot), which is the opposite of the osteoporosis of the pedigree cases [50]. The semantic similarity of these two related but opposing terms contributed towards *SRC* scoring the highest phenotypic similarity (Figure 2.7 B and C). The *SRC* variant had a Combined Annotation Dependent Depletion (CADD) Phred score [45] of 34 (Figure 2.7D) and was among only 24 of the 67 candidates to be unobserved in the ExAC (Exome Aggregation

Consortium) database and in 2,974 further subjects from our in-house collection. Results of sequencing of RNA from blood stem and progenitor cells, including megakaryocytes (MKs) [15], showed that the *SRC* transcript ranked among the highest according to the probability of being overexpressed in MKs compared to the other seven cell types (posterior probability = 0.47) (Figure 2.7E). These independent sources of evidence contributed to establishing the variant coding glutamic acid (E) 527 lysine (K) in *SRC*'s kinase domain as the primary causative candidate for this novel syndrome.

Table 2.3 Abbreviations used for HPO terms in Figure 2.7

Abbreviation	HPO term name
Abdom. Orgs.	Abnormality of the abdominal organs
AG	Abnormal α-granules
AGD	Abnormal α-granule distribution
BBFT	Abnormality of blood and blood-forming tissues
Bleeding	Abnormal bleeding
BMT	Bleeding with minor or no trauma
Bone Min. Density	Abnormality of bone mineral density
Endo. Sys.	Abnormality of the endocrine system
Erythrocytes	Abnormality of erythrocytes
Face	Abnormality of the face
Gen. Sys.	Abnormality of the genital system
Giant Plts.	Giant platelets
Hemoglobin	Abnormal hemoglobin
IMPV	Increased mean platelet volume
Integument	Abnormality of the integument
MCLBM	Abnormality of multiple cell lineages in the bone marrow
NAG	Abnormal number of α-granules
OCS	Abnormal surface-connected open canalicular system
PA	Phenotypic abnormality
Plt Morph.	Abnormal platelet morphology
Plt Shape	Abnormal platelet shape
Recur. Frac.	Recurrent fractures
Skel. Sys.	Abnormality of the skeletal system
Spleen	Abnormality of the spleen
Subcut. Hem.	Subcutaneous hemorrhage
TCP	Thrombocytopenia
Teeth	Abnormality of the teeth

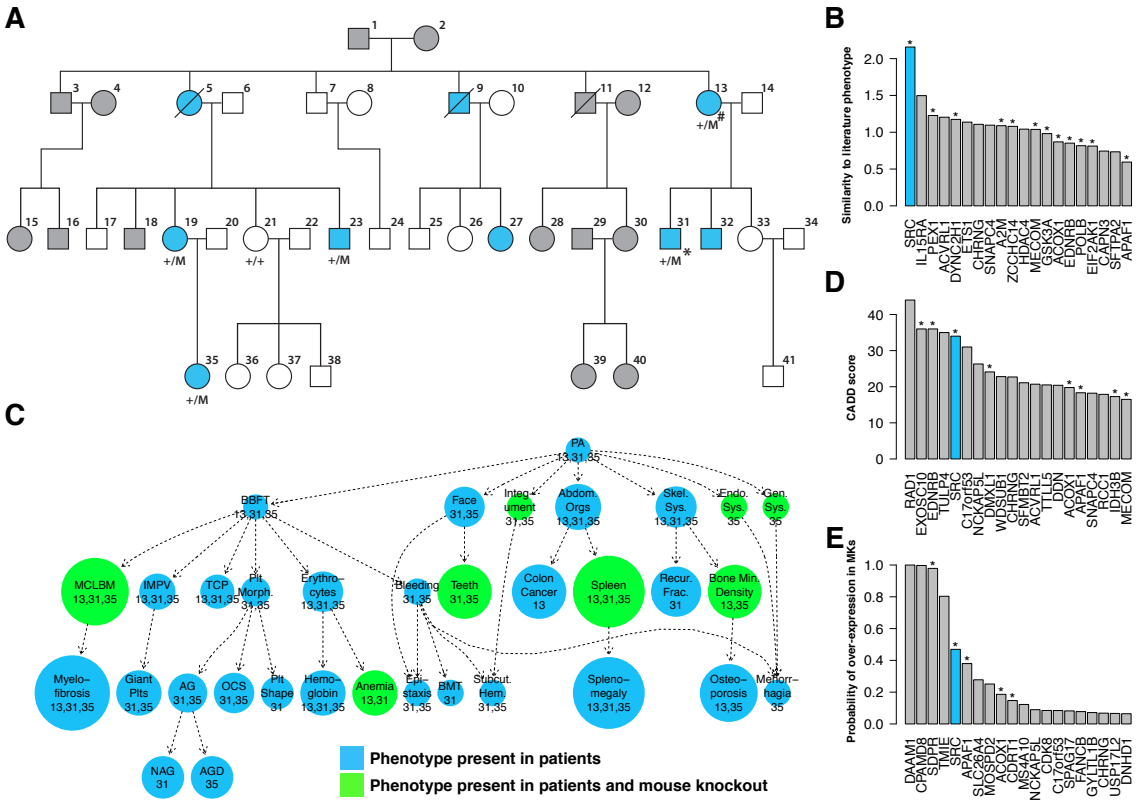


Fig. 2.7 Selection of the c.1579G>A mutation in *SRC* as a candidate pathogenic variant. (A) Pedigree showing male (square) and female (circle) members who have macrothrombocytopenia (blue), are unaffected (empty), or are without clinical information (grey), some of whom are deceased (slash). Cases carry the c.1579G>A mutation in *SRC* (M) in variant calling done by whole-genome sequencing (WGS) (#), whole-exome sequencing (WES) (*), and/or Sanger sequencing (all genotyped subjects). (B) Bar plot showing the mean phenotypic similarity of cases 13, 31, and 35 to OMIM/Mouse Phenotype Ontology (MPO) phenotypes associated with each gene, truncated at the top 20 genes, with novel variants absent from control data indicated by * and *SRC* highlighted in blue (as in D and E). (C) The numbers inside each node indicate which cases were coded with the corresponding HPO term. Abbreviations given in Table 2.3. Terms in green are also present in the OMIM/MPO entries for *SRC*/*Src*. The size of each node is determined by its contribution to the mean phenotype similarity score between the three cases and the OMIM/MPO terms for *SRC*/*Src*. Abbreviations given in main text. The cases of this pedigree were the only ones enrolled in the BRIDGE Bleeding and Platelet Disorders (BRIDGE-BPD) study coded with "Thrombocytopenia", "Myelofibrosis" and "Abnormality of the skeletal system." (D) Bar plot showing the CADD Phred score of the rare variant for each candidate gene. (E) Probability that each candidate gene is specifically overexpressed in MKs compared to blood stem cells and six other hematopoietic progenitors.

2.5 Unsupervised clustering of ontological phenotypes

In order to compare cases and controls statistically to identify candidate genes and variants in a collection of phenotypically heterogeneous HPO-coded cases one must first group the cases based on a potential shared genetic aetiology. Manually partitioning many hundreds or thousands of cases into such groups is difficult and time consuming. Unsupervised clustering methods based on pairwise phenotypic similarity could therefore be useful in generating such groups. Many such methods are available, some of which are sketched here, where 'distance matrix' refers to the negative of the similarity matrix.

Partitioning Around Medoids (PAM) Given a number of clusters and a similarity matrix, and initialising with an arbitrary partition of entities into clusters and an arbitrarily chosen 'exemplar' for each cluster, the examplars are iteratively updated by a pair of steps: (1) assign each point to a cluster corresponding to its most similar exemplar, and (2) update exemplars to minimise the total intra-cluster distance. These steps are repeated until convergence is achieved.

Affinity Propagation (AP) Clustering [27] Similar to PAM in that it selects exemplars to represent clusters. It does not require the number of clusters to be specified *a priori*.

Hierarchical Clustering Iteratively create a hierarchy of objects by agglomerating the most similar pairs of objects and or groups of objects, based on a pairwise similarity matrix. Different methods of measuring the similarity of a single object to a group are possible. One such method is the 'averaging' approach, where the average similarity of the object to objects in the group is used. Another method is the 'complete linkage' approach, where the minimum similarity of the object to the objects in the group is used. The method is implemented as the `hclust` function in R's stats package.

Given a partition of the collection into phenotype clusters, we are faced with the challenge of searching for associations between membership of clusters and presence of rare variants in genes. One could test for associations between each gene and each cluster. For example, a Burden test could be used to assess whether presence of a rare variant in a gene is associated with membership of a cluster. However, with approximately twenty thousand genes to test, and potentially tens or hundreds of clusters, the number of competing hypotheses is huge. Here, I outline an approach which incorporates model organism data that is more powerful for detecting associations between genes and clusters when models for orthologous genes have phenotypes overlapping phenotypes enriched amongst members of a cluster. The procedure uses a clustering algorithm to generate a partition of HPO-encoded phenotypes of unrelated cases, selects a limited number of 'key' terms enriched in each cluster, then

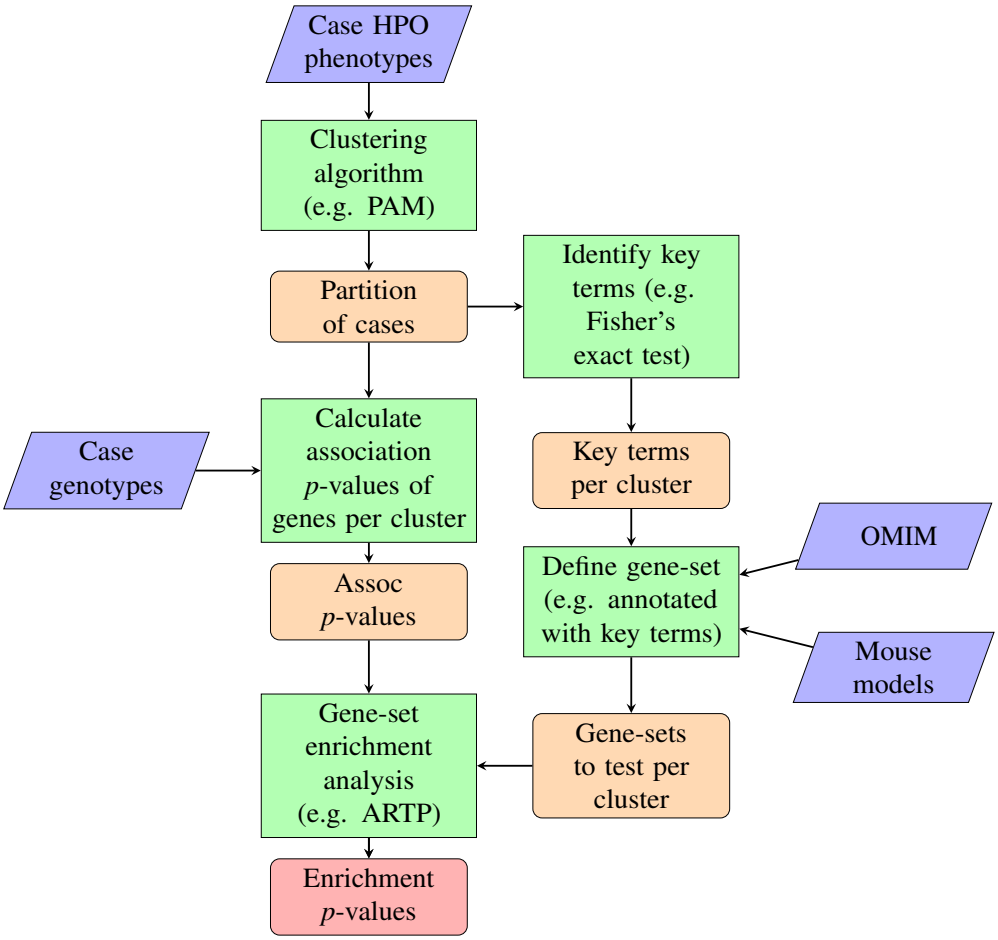


Fig. 2.8 Schematic representation of enrichment test strategy for phenotype clustering

uses an enrichment set analysis (e.g. GSEA), to test whether the set of genes known to be associated with the term through the literature (i.e. based on aggregating ontological annotation from OMIM and model organism databases) is enriched amongst those genes most strongly associated with the cluster (i.e. using a test statistic, from the Burden test for example). A flow-chart depicting the approach is shown in Figure 2.8.

Each of the three clustering methods described at the start of this section were used to generate clusters. Initially, a pairwise phenotype similarity matrix based on the HPO-encoded phenotypes of 940 sequenced probands belonging to the BPD project was generated. Lin's expression for the similarity of terms and the best match product expression was employed to evaluated the similarity between term sets. Information content was set with respect to the 940 probands. The number of clusters for the PAM and hierarchical clustering methods was set to 30, as approximately this many clusters led to the most clinically consistent clusters being produced [119]. The number of clusters generated by the AP algorithm was 88. The

Fisher's exact test was used to rank phenotypic terms by the strength of their associations with each cluster, and the top three terms were selected for each. Fisher's exact test was then used again to assign p-values for each gene/cluster pair, where the test was for association between the cluster membership and presence of at least one rare allele in the gene. The Adaptive Rank-Truncated Product (ARTP) was used to evaluate the enrichment of gene sets amongst the scored genes in each cluster, in preference to the standard GSEA [110]. This was because the ARTP method has been shown to be more powerful when the number of involved genes is small [22], and in this scenario only a handful of genes are likely to be involved in the diseases represented within each cluster.

To calculate the ARTP p-value for a given set of genes G with respect to a given cluster, the Rank-Truncated Product (RTP) statistics for the one, two, three, four and five most strongly associated genes were computed by applying Fisher's combined p-value statistic to their p-values of association. This relies upon independence of the p-values, which is a reasonable assumption in this scenario because of the low linkage between rare variants. The minimum of these five statistics is then used as the enrichment score S for G. The p-value p_G for enrichment is then calculated by permutation of gene labels for N permutations, $G'_1, G'_2, ..., G'_N$ ($N = 1000$ was used).

$$S(G) = \min_{i \in \{1,..,5\}} \left[-2 \sum_{j=1}^{i} \log g^{(j)} \right],$$

$$p_G = \sum_{i=1}^{N} \mathbb{1}_{S(G'_i) \leq S(G)},$$

where $g^{(j)}$ is the Fisher exact test's p-value of association between presence of rare alleles in the j^{th} ranked gene and membership of the given cluster. The ARTP p-values were calculated for each of the key terms in each cluster. The sets of clusters generated by the hierarchical and PAM clustering algorithms led to 2 and 1 significant cluster/phenotype term associations being found at a significance threshold of 0.05. The AP clustering results contain numerous significant associations, also containing several known ones, including associations between 'Reduced factor IX activity' and *F9* ($p = 0.002$), 'Epistaxis' and *F7* ($p = 0.049$), 'Increased mean platelet volume' and *GP9* ($p = 0.012$), and 'Abnormal platelet shape' and *TUBB1* ($p = 0.012$).

The method of using data from model organisms gives a substantial increase in power to detect true associations in comparison with testing association with each gene marginally. A drawback of the method is that it can only identify associations between genes and clusters when the gene is already associated with some of the enriched terms in the cluster through

model organisms. Furthermore, due to the multiple intuition-based steps in this procedure, it is not clear what alternative hypotheses the method is powerful under. We therefore developed a model-based approach which would allow us to harness phenotypic similarity of HPO phenotypes, restrict against a carefully defined alternative model and enable more control over how prior information is incorporated. This development led to the 'Phenotype similarity regression' method, which is described in the following Chapter.

Chapter 3

Phenotype similarity regression

Rare genetic diseases are often caused by high-penetrance rare variants and characterised by abnormalities spanning multiple organ systems ascertained with variable clinical precision. Existing methods for identifying genes with variants responsible for rare diseases summarise phenotypes with unstructured binary or quantitative variables. The Human Phenotype Ontology allows composite phenotypes to be represented systematically but association methods accounting for the ontological relationship between HPO terms do not exist. This chapter describes a Bayesian method, called SimReg, which models the association between HPO-encoded disease phenotypes and genotype. The method enables inference of the probability of an association together with a latent characteristic HPO phenotype for the disease. It thus formalises a clinical approach to phenotyping that is lacking in standard regression techniques for rare disease research. We demonstrate the power of the method by uncovering a number of true associations in a large collection of genome-sequenced and HPO-coded cases with rare diseases.

The description of the method given in this Chapter is based on that given in: D. Greene, NIHR BioResource–Rare Diseases Consortium, S. Richardson and E. Turro (2016). "Phenotype similarity regression for identifying the genetic determinants of rare diseases". *The American Journal of Human Genetics*, 98(3):490–499.

3.1 Introduction

Modelling sparse and ontologically structured phenotype data as a response directly is difficult due the logical dependency between terms induced by the structure of the ontology. Additionally, terms can be correlated due to biological factors. SimReg overcomes these difficulties by treating the HPO-encoded phenotypes of the subjects as explanatory variables, and their corresponding genotypes as the response. This is an example of 'inverse regression',

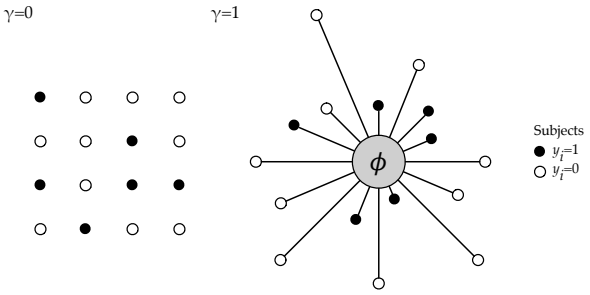


Fig. 3.1 Cartoons depicting the two models being compared. Under the baseline model ($\gamma = 0$), rare genotype carriers (filled nodes) are distributed at random with respect to individual phenotypes. Here the grid is used as an abstract representation, and its geometry has no interpretation. Under the alternative model ($\gamma = 1$), individuals are more likely to carry a rare genotype if they are phenotypically similar (as indicated by short edges) to the characteristic phenotype ϕ than if they are dissimilar to it (as indicated by long edges).

an idea also employed in the context of genome-wide association studies O'Reilly et al. [78], and is adequate in our setting because we are not interested in interpreting the regression coefficients per se but only on evaluating the probability of association. We define a subject's 'genotype' y as a binary label that can take on the values 'rare' (1) or 'common' (0) according to a pre-specified function of the genetic data. For example, we could define the label 'rare genotype' to mean that there is at least one rare variant in a particular gene (dominant inheritance), or at least two rare variants in a particular gene (recessive inheritance).

Our method then seeks to compare two models for the data, indexed by γ:

- Under the baseline model ($\gamma = 0$), the probability of observing the rare genotype is the same for each case.
- Under the alternate model ($\gamma = 1$), the probability of observing the rare genotype depends on the phenotypic similarity S (to be defined later) of the case to a latent *characteristic* HPO phenotype ϕ.

Figure 3.1 gives a graphical representation of the models. We adopt a Bayesian inference framework, where the model selection indicator γ and characteristic phenotype ϕ are estimated through their posterior distributions. Of particular interest is the posterior mean of γ, which represents the probability that $\gamma = 1$, thus indicating the strength of evidence for an association. A crucial element of our approach is the construction of an appropriate function for quantifying the semantic similarity of the characteristic phenotype ϕ to the phenotypes of the subjects. The choice of function is motivated by the need to optimally discriminate between subjects having clinical features which are pertinent to a disease from those having overlapping or unrelated phenotypes due to a different disease. To achieve this, we have

chosen a function which accounts for the ontological structure of the HPO and induces a parsimonious characteristic phenotype: it selects the required terms to distinguish patient groups while avoiding overfitting and is robust to coding of patients with spurious or sporadic terms. Importantly, the function is flexible with respect to the phenotypic variability of disease and robust to the HPO coding practices of clinicians.

Our Bayesian approach provides a natural means of incorporating information from the scientific literature into our prior belief about the characteristic phenotype. In this work, we focus on gene-specific inference and up-weight the prior probability of characteristic phenotypes that are similar to clinical [48] and murine phenotypes [8] relevant to the gene.

We demonstrate the effectiveness of our method in identifying associations between genotype and phenotype through a simulation study, whereby phenotypes are simulated given genotypes in such a way as to emulate the effect of a hypothetical set of pathogenic variants. We go on to apply our inference procedure to data representing over 2,000 unrelated individuals enrolled to subprojects of the NBR–RD project. We show that our method, implemented in the SimReg software package, can identify genes with rare variants responsible for a diverse set of pathologies in a single application and can estimate recognised disease phenotypes.

3.2 Model specification

We use a logistic regression framework to specify the two models under comparison:

$$\begin{aligned} & y_i \sim \text{Bernoulli}(p_i), \\ & \gamma = 0: \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \hat{h}_i, \\ & \gamma = 1: \quad \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \hat{h}_i + \beta S(\phi, x_i). \end{aligned} \tag{3.1}$$

Here, $y_1, \ldots, y_N$ are the genotypes of the N subjects in the collection, where $y_i = 1$ if subject i possesses the rare genotype and $y_i = 0$ if subject i possesses the common genotype. $x_1, \ldots, x_N$ are the corresponding phenotypes of the subjects, where x_i comprises the minimal set of HPO terms required to describe the phenotypic abnormalities of subject i. The parameter $\hat{h}_i$ is a log odds offset for each individual which can take into account batch effects and factors affecting the background rate of rare genotypes.

The term $S(\phi, x_i)$ denotes a chosen measure of phenotypic similarity between the characteristic phenotype and subject i's phenotype. Note that our response and predictor are inverted compared to classical regression methods to avoid having to treat sparse and structured HPO

data as the response. Under the baseline model, the intercept α is the global rate of rare genotypes. Under the alternate model, there is an additional parameter β, which is strictly positive and captures the effect of a unit increase in phenotypic similarity to the characteristic phenotype ϕ on the log odds of having the rare genotype. Thus, the probability that $\gamma = 1$ is greater in expectation when $S(\phi, x_i)$ is larger if $y_i = 1$ than if $y_i = 0$.

Similarity measure

Our chosen similarity measure S is built with consideration for:

- quantification of the similarity of terms,
- quantification of the similarity of a patient phenotype x_i to the characteristic phenotype ϕ,
- flexible transformation of the similarity between phenotypes.

We use Lin's [59] similarity function to compare two different terms (see Equation 1.2), hereafter represented by function s. As terms cannot have a higher information content than their descendants, the similarity s between two terms can range between zero and one. Next, we consider asymmetric measures of similarity between a case phenotype and ϕ inspired by the best match average function described in Section 1.6. The expressions used for semantic similarity between the case phenotype and ϕ are:

$$S_\phi(\phi \to x_i) = \frac{1}{|\phi|} \sum_{t_\phi \in \phi} \max_{t_x \in x_i} s(t_\phi, t_x) \mathbb{1}_{t_\phi \in \text{anc}(t_x)},$$

$$S_x(x_i \to \phi) = \frac{1}{|x_i|} \sum_{t_x \in x_i} \max_{t_\phi \in \phi} s(t_x, t_\phi) \mathbb{1}_{t_\phi \in \text{anc}(t_x)}.$$

The standard best match average function does not include the indicator variable above, which evaluates to 1 only if the node in ϕ is amongst the ancestors of the node in x_i. We prefer to include this restriction, which penalises similarity to ϕ when it includes over-specific terms, in order to concentrate the posterior weight of ϕ preferentially on nodes which are present amongst the subjects. The presence of a term in ϕ that is absent from x_i has the effect of lowering S_ϕ, while the presence of a term in x_i that is absent from ϕ has the effect of lowering S_x. Summation of two asymmetric similarities, as used in the best match average, would allow reasonably high overall similarities to be obtained even when one of the two asymmetric similarities is close to zero. We prefer to multiply rather than add up the two similarity measures to obtain an expression for the overall similarity function used in Equation 3.1,

as it ensures that the overall similarity can only be high when there is a high asymmetric similarity in both directions. However, as the values of S_x and S_ϕ are influenced by factors such as how frequent terms are in the reference database (which affects nodal IC) and the structure of the HPO graph, there is no guarantee that a linear function of their product optimally distinguishes subjects with objectively distinct clinical features. For example if two distinct phenotypic features always manifest together with a given disease caused by a particular rare genotype, presence of only one of these features would contribute positively to S_ϕ, but would not be informative of presence of the rare genotype. To ensure the model is robust to the choice of S, we allow modulation of the shapes of the similarity parameters, S_ϕ and S_x, through transformations f and g, respectively. A reasonable choice for f and g is the beta cumulative distribution function (CDF), as it maps $[0,1]$ to $[0,1]$ monotonically (i.e. enforcing that S increases as S_x and S_ϕ increase), and allows a wide variety of shapes:

$$\begin{aligned} f(z,a_f,b_f) &= I_z(a_f,b_f), \\ g(z,a_g,b_g) &= I_z(a_g,b_g), \\ I_z(a,b) &= \frac{\int_0^z t^{a-1}(1-t)^{b-1}\mathrm{d}t}{\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t} \end{aligned}$$

where I_z is the regularised incomplete beta function, and a_f, a_g, b_f, b_g are unknown parameters to be estimated. Finally, the overall similarity function is given by:

$$S(\phi,x_i) = f(S_\phi(\phi \to x_i), a_f, b_f) \cdot g(S_x(x_i \to \phi), a_g, b_g). \tag{3.2}$$

Priors

We propose the following prior distributions for the model indicator and the regression parameters:

$$\begin{aligned} \gamma &\sim \text{Bernoulli}(\pi), \\ \alpha &\sim \text{Normal}(\text{mean} = 0, \text{sd} = 5), \\ \log\beta &\sim \text{Normal}(\text{mean} = 2, \text{sd} = 1). \end{aligned}$$

The value of π indicates how likely we believe *a priori* that there is a true association. All of the analyses in this paper assume $\pi = 0.05$. We place a vague prior on α around 0. The prior distribution on β is positive because the probability of $y_i = 1$ increases with $S(\phi,x_i)$, given $\gamma = 1$. Different diseases have a wide variety of genetic heterogeneity across loci. If

only variants in one locus can cause a particular disease, then β could be very high, and the converse would be true if many different loci contain variants causal for the same disease. Thus, the prior variance of β allows for a wide range of effect sizes given the range of S.

Prior on similarity transformations f and g We choose the prior for f in order to favour parsimonious characteristic phenotypes and the prior for g to allow for an indeterminate number of nodes appearing sporadically among patients. The presence of a term in the characteristic phenotype ϕ that is absent from the patient phenotype x_i has the effect of lowering S_ϕ, while the presence of a term in x_i that is absent from ϕ has the effect of lowering S_x (see Equation 3.2). For example, if ϕ has one HPO term and it is also present in x_i, then $S_\phi = 1$. However, the presence of one or two additional spurious terms can reduce S_ϕ to as low as 0.5 or 0.33 respectively. In order to discourage non-parsimonious characteristic phenotypes, we place a high prior weight on f transformations whose corresponding probability density functions have means above 0.5 (i.e. $\frac{a_f}{a_f+b_f} > 0.5$) as this ensures that a value of $f(S_\phi)$ cannot be obtained if the value of S_ϕ is low. Specifically, we specify the priors on the parameters of f. The resultant distribution of transformations f and g are represented in Figure 3.2 (left). However, in order to allow for patients coded with sporadic terms that are not part of the core disease phenotype, we specify a more flexible prior distribution on g than we do on f, as illustrated in Figure 3.2 (right).

The prior distribution for the transformation f is specified by independent priors for the logarithms of the ratio and sum of its parameters, a_f and b_f: $\log \frac{a_f}{b_f}$ and $\log(a_f + b_f)$, which respectively correspond to the logit of the mean (i.e. $\text{logit}\frac{a_f}{a_f+b_f}$) and a measure of the steepness of the resultant beta CDF. The prior distribution for the transformation g is specified equivalently. The following priors for the log ratio and sum of a_f, b_f and a_g, b_g for similarity transformations f and g respectively are used:

$$\begin{aligned}
\log \frac{a_f}{b_f} &\sim \text{Normal}(\mu_f, \sigma_f^2), \\
\log(a_f + b_f) &\sim \text{Normal}(\mu_{f'}, \sigma_{f'}^2), \\
\log \frac{a_g}{b_g} &\sim \text{Normal}(\mu_g, \sigma_g^2), \\
\log(a_g + b_g) &\sim \text{Normal}(\mu_{g'}, \sigma_{g'}^2)
\end{aligned} \tag{3.3}$$

Our choice of hyperparameter values was informed by a sensitivity analysis assessing the model's performance on data for *ACTN1*. We found that not using a transformation at all (i.e. not modulating the similarity with f and g, which is equivalent to using the identity function obtained by setting $a_f = b_f = a_g = b_g = 1$), or using an overly flexible prior on

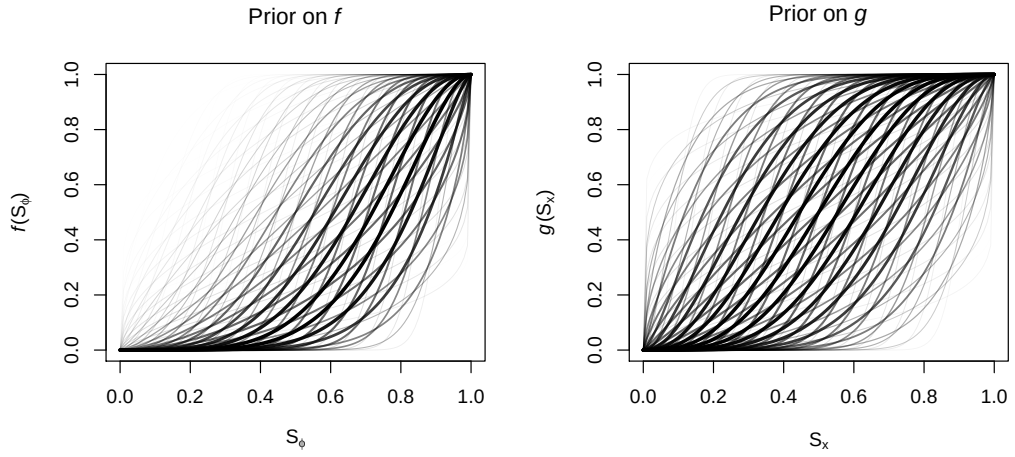


Fig. 3.2 The distribution of shapes for the incomplete beta function transformation of phenotype similarities S_ϕ and S_x for select values of the parameters, given the hyperparameter values in Equation 3.3. The thickness and opacity of each line is proportional to the prior probability of the corresponding parameterisation of the transformation.

f, discourages inclusion of the essential 'Thrombocytopenia' term relative to inclusion of spurious alternative terms, conditional on inclusion of the other essential term, 'Increased mean platelet volume'. This occurs because if the value of $\frac{a_f}{a_f+b_f}$ has high posterior weight near 0.5, then spurious terms can be accommodated by mapping values near 0.5 to near 1. As more prior weight is shifted to f transformations with a value of $\frac{a_f}{a_f+b_f}$ greater than 0.5, the probability of joint inclusion of the two key nodes of this disease is increased (Figure 3.3).

Prior probability for ϕ By default, our prior distribution on the characteristic phenotype ϕ places a uniform prior probability on all minimal sets of terms of size less than or equal to k. In consultation with clinical colleagues, we chose $k = 3$ on the grounds that three nodes should adequately distinguish between the primary features of most rare diseases. If y is set based on variants in a particular feature, such as a gene, then our prior can up-weight HPO phenotypes comprising terms annotated to that feature on the basis of reports in the scientific literature. Thus, the prior on ϕ is given by

$$\mathbb{P}(\phi) = \begin{cases} \frac{1}{|\Phi^{(k)}|} & \text{No literature phenotype} \\ \frac{S'(M \to \phi)}{\sum_{\psi \in \Phi^{(k)}} S'(M \to \psi)} & \text{Literature phenotype } M \end{cases} \tag{3.4}$$

Fig. 3.3 Graphical representation of the posterior distribution of ϕ when no f/g transformations are used and for different values of μ_f (with $\sigma_f^2 = 1, \mu_g = 0, \sigma_g^2 = 1.5$) using the data for *ACTN1*. Each node shows the marginal probability of inclusion in ϕ. Without the f/g transformations, the essential 'Thrombocytopenia' term carries low posterior weight. If the f/g transformations are included, as the value of μ_f is increased, from 0 through 0.5 to 1, the probability of inclusion of the term 'Thrombocytopenia' increases.

where $\Phi^{(k)}$ denotes the set of all minimal sets of up to k HPO terms and S' is an unstandardised similarity function:

$$S'(M \to \tau) = \frac{1}{|\tau|} \sum_{t \in \tau} \max_{m \in M} \exp\left(s_{\text{Resnik}}(m,t)\right),$$

where s_{Resnik} is Resnik's definition for the similarity between terms, defined in Equation 1.1. Resnik's definition was chosens in order to attach greater prior weight to rare phenotype terms shared with the literature phenotype. In practice, the literature phenotype could be

obtained from OMIM or from the Mouse Genome Informatics (MGI) database [8] after mapping murine phenotypes coded using the 'Mammalian Phenotype Ontology' [104] to HPO terms through a cross-species phenotype ontology [46].

Our inference procedure, described in Section 3.3, requires sampling model parameters from their posterior distributions. Conditional on $\gamma = 1$, it is not straightforward to sample from the space of minimal sets $\Phi^{(k)}$, because not all possible HPO term combinations comprise such a minimal set. To overcome this difficulty, we propose an unrestricted vector of k HPO terms $\tilde{\phi}$, and then derive the associated underlying phenotype ϕ by applying a mapping function υ. We therefore need to impose a prior distribution on the unrestricted space which is compatible with the desired prior for ϕ (Equation 3.4) on the restricted space. To be precise, the prior on $\tilde{\phi}$ is given by:

$$\mathbb{P}(\tilde{\phi}) = \frac{\mathbb{P}\left(\upsilon(\tilde{\phi})\right)}{\left|\left\{\tilde{\phi}' \in \mathrm{H}^k : \upsilon(\tilde{\phi}') = \upsilon(\tilde{\phi})\right\}\right|},$$

where H^k is the space of all vectors of k HPO terms and υ maps an arbitrary such vector of terms to its corresponding minimal set. The denominator accounts for the number of unrestricted vectors that map to the same minimal set.

In order to calculate $p(\phi)$ when using a uniform distribution over $\Phi^{(k)}$, we need to calculate the number of distinct minimal sets $\left|\Phi^{(k)}\right|$. This is trivial when $k = 1$, as $\left|\Phi^{(k)}\right| = |\mathrm{H}|$. However it becomes more computationally intensive as k increases, so in our implementation we use the approximation $\binom{|\mathrm{H}|}{k}$. This approximation works well in practice when k is small. It has no effect on the update of the $\tilde{\phi}$ parameter, as the $\left|\Phi^{(k)}\right| = |\mathrm{H}|$ expression cancels out in the acceptance probability for $\tilde{\phi}'$, but it does affect the update of γ as it penalises the model $\gamma = 1$ slightly by overestimating the size of $\left|\Phi^{(k)}\right|$.

When using an informative prior distribution, weighted by similarity to the literature phenotype, we need to calculate $\sum_{\psi \in \Phi^{(k)}} S'(M \to \psi)$. In order to avoid having to sum over the entire space $\Phi^{(k)}$, we employ the approximation $\left|\Phi^{(k)}\right| \times k \times \mathrm{mean}_{\psi \in \mathrm{H}} S'(M \to \psi)$.

Finally, to compute $p(\tilde{\phi})$, we also need to calculate the number of alternative unrestricted vectors that map to the same minimal set, i.e. $\left|\left\{\tilde{\phi}' \in \mathrm{H}^k : \upsilon(\tilde{\phi}') = \upsilon(\tilde{\phi})\right\}\right|$, where ν maps an unrestricted vector to a minimal set. We use the following expression for the number of representations:

$$\left|\bigcup_{t \in \upsilon(\tilde{\phi})} \mathrm{anc}(t)\right|^k + \sum_{i=1}^{|\upsilon(\tilde{\phi})|} (-1)^i \binom{|\upsilon(\tilde{\phi})|}{i} \left(\left|\bigcup_{t \in \upsilon(\tilde{\phi})} \mathrm{anc}(t)\right| - i\right)^k$$

3.3 Inference

We perform model comparison using the Markov chain Monte Carlo (MCMC) based method of Carlin and Chib [12].

The method of Carlin and Chib is a means of inferring the parameters in two models and computing a Bayes factor comparing them. Instead of targeting the posterior distribution of each model individually, the following function is targeted:

$$(\gamma, \theta^{(0)}, \theta^{(1)}) \mapsto (1-\gamma)L_y^{(0)}(\theta^{(0)})p_0(\theta^{(0)})f_1(\theta^{(1)}) + \gamma L_y^{(1)}(\theta^{(1)})p_1(\theta^{(1)})f_0(\theta^{(0)}).$$

Here, $\theta^{(0)}$ and $\theta^{(1)}$ are vectors of the parameters of models 0 and 1 respectively and $p_0(\theta^{(0)})$ and $p_1(\theta^{(1)})$ are their respective priors. The likelihood functions under model 0 and 1 are given by $L_y^{(0)}(\theta^{(0)})$ and $L_y^{(1)}(\theta^{(1)})$ respectively. The functions $f_0(\theta^{(0)})$ and $f_1(\theta^{(1)})$ are arbitrary probability density functions called 'pseudopriors', representing the conditional probability distributions of the parameters of one model given the alternate model is true, i.e. $\theta^{(0)}|\gamma = 1$ and $\theta^{(1)}|\gamma = 0$ respectively. The conditional posterior distributions of the parameters $\theta^{(0)}|\gamma = 0$ and $\theta^{(1)}|\gamma = 1$ can be estimated from MCMC samples made at iterations when $\gamma = 0$ and $\gamma = 1$, respectively, and the posterior probability that model 1 is true can be estimated from the proportion of iterations in which $\gamma = 1$.

Let α^* be the intercept parameter under $\gamma = 0$ and α be the intercept parameter under $\gamma = 1$ so that they may be distinguished. For convenience, we perform inference of $\tilde{\phi}$, which is on the unrestricted space of vectors of HPO terms, rather than ϕ, because it is difficult to propose uniformly from the space of minimal sets $\Phi^{(k)}$. However, ϕ can be recovered from $\tilde{\phi}$ easily by mapping to the corresponding minimal set. The MCMC algorithm proceeds to target the following distribution:

$$\begin{aligned}
&\mathbb{P}(\gamma, \alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}|y) \propto \\
&(1-\gamma)L_y^{(0)}(\alpha^*)p_0(\alpha^*)f_1(\alpha)f_1(\beta)f_1(a_f)f_1(b_f)f_1(a_g)f_1(b_g)f_1(\tilde{\phi}) \\
&\quad +\gamma L_y^{(1)}(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p_1(\alpha)p_1(\beta)p_1(a_f)p_1(b_f)p_1(a_g)p_1(b_g)p_1(\tilde{\phi})f_0(\alpha^*).
\end{aligned}$$

For optimal mixing of the Markov chain, the pseudopriors should approximate the respective conditional posterior distribution given the model, that is, $f_0(\theta^{(0)}) \propto L_y^{(0)}(\theta^{(0)})p_0(\theta^{(0)})$ and $f_1(\theta^{(1)}) \propto L_y^{(1)}(\theta^{(1)})p_1(\theta^{(1)})$. Without tuning the pseudopriors, accepting proposed updates to γ will happen very infrequently due to a poor fit of the data to the alternate model. To achieve this, we tune the pseudopriors using empirical summary statistics obtained by running initial Markov chains under each model separately. For parameters α^* and α, Normal pseudopriors are used, while for the strictly positive parameters β, a_f, b_f, a_g

and b_g, Log-Normal pseudpriors are used. The hyperparameters of these pseudopriors are obtained using maximum likelihood estimation based on the MCMC samples. We compose a pseudoprior for $\tilde{\phi}$ by counting the number of appearances of HPO terms in any of the k slots of $\tilde{\phi}$ throughout the tuning iterations:

$$\mathbb{P}(t) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{k}\mathbb{1}(\tilde{\phi}_{ij} = t) + \varepsilon}{Ik + \varepsilon|H|}, \tag{3.5}$$

where I is the number of MCMC tuning iterations, t is a term in the set of HPO terms H and $\tilde{\phi}_{ij}$ is the j^{th} element of $\tilde{\phi}$ in the i^{th} iteration. We allow a non-zero probability of inclusion of terms which have not been sampled at all during the tuning batch by setting $\varepsilon = 1$. Using the above expression, we define the pseudoprior on $\tilde{\phi}$ as

$$f_1\left(\tilde{\phi}\right) = \prod_{j=1}^{k}\mathbb{P}\left(\tilde{\phi}_j\right).$$

The updates for each iteration of the MCMC algorithm are given below. Note that here we drop the model subscript on the true prior distributions p and pseudoprior distributions f because each parameter belongs to only one model.

1. An update of α^*:

   $\gamma = 0$ Propose an update of α^* by drawing from

   $$\alpha^{*'} \sim \text{Normal}(\alpha^*, s_\alpha^2)$$

   and accepting with probability

   $$\min\left(1, \frac{L_y^{(0)}(\alpha^{*'})p(\alpha^{*'})}{L_y^{(0)}(\alpha^*)p(\alpha^*)}\right).$$

   $\gamma = 1$ Sample $\alpha^{*'}$ from the pseudoprior distribution for α^*:

   $$\alpha^{*'} \sim \text{Normal}(\hat{\mu}_{\alpha^*}, \text{sd} = \hat{\sigma}_{\alpha^*}^2).$$

2. An update of α:

   $\gamma = 0$ Sample α' from the pseudoprior distribution for α:

   $$\alpha' \sim \text{Normal}(\hat{\mu}_\alpha, \hat{\sigma}^2_\alpha).$$

   $\gamma = 1$ Propose an update of α by drawing from

   $$\alpha' \sim \text{Normal}(\alpha, s^2_\alpha)$$

   and accepting with probability

   $$\min\left(1, \frac{L_y^{(1)}(\alpha', \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\alpha')}{L_y^{(1)}(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\alpha)}\right).$$

3. An update of β:

   $\gamma = 0$ Sample $\log\beta'$ from the pseudoprior distribution for $\log\beta$

   $$\log\beta' \sim \text{Normal}(\hat{\mu}_\beta, \hat{\sigma}^2_\beta).$$

   $\gamma = 1$ Propose an update of $\log\beta$ by drawing from

   $$\log\beta' \sim \text{Normal}(\beta, s^2_\beta)$$

   and accepting with probability

   $$\min\left(1, \frac{L_y^{(1)}(\alpha, \beta', a_f, b_f, a_g, b_g, \tilde{\phi})p(\beta')}{L_y^{(1)}(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\beta)}\right).$$

4. An update of the shape parameters a_f, b_f, a_g and b_g analagously as is done for β.

5. An update of $\tilde{\phi}$:

   $\gamma = 0$ Sample $\tilde{\phi}$ from the pseudoprior distribution for $\tilde{\phi}$ by sampling all k terms independently from the distribution described in Equation 3.5.

   $\gamma = 1$ Propose updating $\tilde{\phi}$ to $\tilde{\phi}'$ by setting component $t = \left[\tilde{\phi}\right]_j$ (where j is chosen at random from $1,\ldots,k$) to a random term t', selected with probability π_t. Hence $\tilde{\phi}'$ can be specified as

$$\left[\tilde{\phi}'\right]_h = \begin{cases} t' & h = j \\ \left[\tilde{\phi}\right]_h & \text{otherwise.} \end{cases}$$

   The proposal is accepted with probability

$$\min\left(1, \frac{L_y^{(1)}(y|\alpha,\beta,a_f,b_f,a_g,b_g,\tilde{\phi}')p(\tilde{\phi}')\pi_t}{L_y^{(1)}(\alpha,\beta,a_f,b_f,a_g,b_g,\tilde{\phi})p(\tilde{\phi})\pi_{t'}}\right).$$

   We set the proposal distribution of the new term $\{\pi_t : t \in \mathrm{H}\}$ to equal that of the individual components of $\tilde{\phi}$ under its pseudoprior (see Equation 3.5). An alternative approach that does not rely on a tuning chain is to propose a new term proportionally to the number of subjects having $y_i = 1$ whose phenotypes include the term or one of its descendants in the HPO:

$$\pi_t \propto \sum_{i=1}^{N} \mathbb{1}_{y_i=1}\mathbb{1}_{t\in\bigcup_{t'\in x_i}\mathrm{anc}(t')}.$$

6. An update of γ by Gibbs sampling:

$$\gamma' \sim \text{Bernoulli}\left(\frac{\omega^{(1)}}{\omega^{(0)}+\omega^{(1)}}\right),$$

   where

$$\omega^{(0)} = (1-\pi)L_y^{(0)}(\alpha^*)f(\alpha)f(\beta)f(a_f)f(b_f)f(a_g)f(b_g)f(\tilde{\phi})p(\alpha^*),$$
$$\omega^{(1)} = \pi L_y^{(1)}(\alpha,\beta,a_f,b_f,a_g,b_g,\tilde{\phi})p(\alpha)p(\beta)p(a_f)p(b_f)p(a_g)p(b_g)p(\tilde{\phi})f(\alpha^*),$$

   where π is the prior probability that $\gamma = 1$.

3.4 Simulation study

We assessed the performance of SimReg by analysing datasets generated under two scenarios, labelled by $\tilde{\gamma}$. Under $\tilde{\gamma} = 1$, the HPO phenotypes $x_{1,\ldots,N}$ were simulated conditional on the genotypes $y_{1,\ldots,N}$ of N individuals while under $\tilde{\gamma} = 0$ they were simulated independently of the genotypes. The phenotypes are simulated conditional on the genotypes, rather than by generating genotype data conditional on phenotypes. Therefore, we only assess our estimates of the model indicator γ in relation to $\tilde{\gamma}$. Although the simulation setup does not match the model, it has the advantage of more closely resembling the natural process by which phenotypes depend on genotypes according to easily interpretable parameters. When $\tilde{\gamma} = 1$, phenotypes for all subjects having $y_i = 1$ were formed by selecting terms from an arbitrarily chosen disease template ('Decreased mean platelet volume', 'Thrombocytopenia' and 'Autism'). Each term was selected with a pre-specified probability r, termed 'expressivity', and m further noise terms drawn at random from a set of approximately 1,000 HPO terms were appended, where $m \sim \text{Poisson}(\lambda = 5)$. The set of terms from which the noise terms were drawn was created by selecting 200 HPO terms at random, taking the union with the disease template terms, and then aggregating all the ancestral terms. Phenotypes for subjects having $y_i = 0$ were drawn at random using terms from the above set with $\lambda = 8$, and then mapped to minimal sets. When $\tilde{\gamma} = 0$, all phenotypes were sampled from the noise term set with $\lambda = 8$. This ensures that on average individuals have approximately 8 terms, irrespective of y_i and $\tilde{\gamma}$. The simulation was performed with the set of disease template terms and set of noise terms fixed but with different numbers of individuals carrying the rare genotype ($\sum_i y_i \in \{2,4,6,8,10,20\}$ out of $N = 1,000$) and varied levels of expressivity $r \in \{\frac{1}{3}, \frac{2}{3}, 1\}$. The low expressivity set-ups capture situations in which a fraction of the individuals having a rare genotype can be considered to carry a neutral variant with respect to the disease in question because they have none of the template terms. For the same reason, they capture scenarios of incomplete penetrance of a subset of the underlying rare variants. Furthermore, a degree of genetic heterogeneity is built into our simulation setup, as there is a non-zero probability of a template phenotype term being randomly allocated to an individual with the common genotype.

The results of repeating the simulation 64 times for each value of $\tilde{\gamma}$ and combination of r and $\sum_i y_i$, depicted in Figure 3.4, show that power to detect a true association, as assessed by the posterior mean of γ, increases with the expressivity of the disease terms r and also with the frequency of the rare genotype in the study sample $\sum_i y_i$ (red points). Under $\tilde{\gamma} = 0$, the posterior mean of γ remains very close to zero in all circumstances (grey dots). Specifically, we find that 2, 6, and 20 cases out of 1,000 subjects are sufficient to obtain perfect or near perfect discrimination between the two models when the expressivity is 1, $\frac{2}{3}$

and $\frac{1}{3}$ respectively. When the number of subjects with the rare genotype is equal to 6 and the expressivity is $\frac{2}{3}$, which implies that any two individuals with the rare genotype only have a 0.17 chance of having exactly the same template terms, our method can achieve a positive predictive value of 1, even when the negative predictive value is as high as 0.95, by thresholding at $\mathbb{P}(\gamma = 1|y) \geq 0.25$. Under this set-up, we expect 1.78 of the 6 individuals with the rare genotype to have none of the template terms at all, which indicates that the method has some resilience to the presence of $y_i = 1$ induced by neutral rather than pathogenic variants.

Genetic heterogeneity We performed a different version of the simulation study described above to assess the performance of our method when genetic heterogeneity is controlled explicitly. Here, we vary a parameter representing the extent of genetic heterogeneity, v, so that for each individual having $y_i = 1$, there were v additional individuals with phenotypes simulated from the same distribution but having $y_i = 0$.

We applied the inference to datasets generated with $v \in \{0, 1, 3, 9\}$. Thus, the simulations where $v = 0$ correspond to the scenario of the simulations described above, and those where $v = 9$ represent situations where only one tenth of the cases having a phenotype arising from the disease template have $y_i = 1$.

The results of the simulation, given as box plots of the estimated posterior means of γ under the various scenarios (Figure 3.5), demonstrate that although power goes down as genetic heterogeneity increases, the sensitivity of the method, thresholding on $\gamma > 0.25$, approaches 100% when expressivity r is 1 and $\sum_i y_i$ is at least 6, and also when expressivity r is $\frac{2}{3}$ and $\sum_i y_i$ is at least 10, irrespective of v. When $v = 3$ and $r = \frac{1}{3}$, which means the HPO terms have very low expressivity and only a quarter of individuals drawn from the template phenotype carry the rare genotype, γ exceeded 0.25 in 87.5% of our simulations as long as 20 out of $1,000$ individuals carried the genotype. Thus we conclude that our method is powerful even in challenging scenarios in which there is substantial genetic heterogeneity and low phenotypic expressivity.

Specificity The simulation study shows that if the phenotypes are homogeneously selected from a wide range of HPO nodes, then our method is unlikely to produce high posterior estimates of γ. However, the simulation study presented above is based on only 64 repetitions for each simulation set-up (shown as 64 grey dots in each panel). In order to more accurately assess the specificity of our method we simulated 20,000 independent sets of phenotypes, simulated with a total of 6 cases having the rare genotype. The distribution of the posterior mean values of γ inferred for the datasets are shown in the left panel of Figure 3.6. There

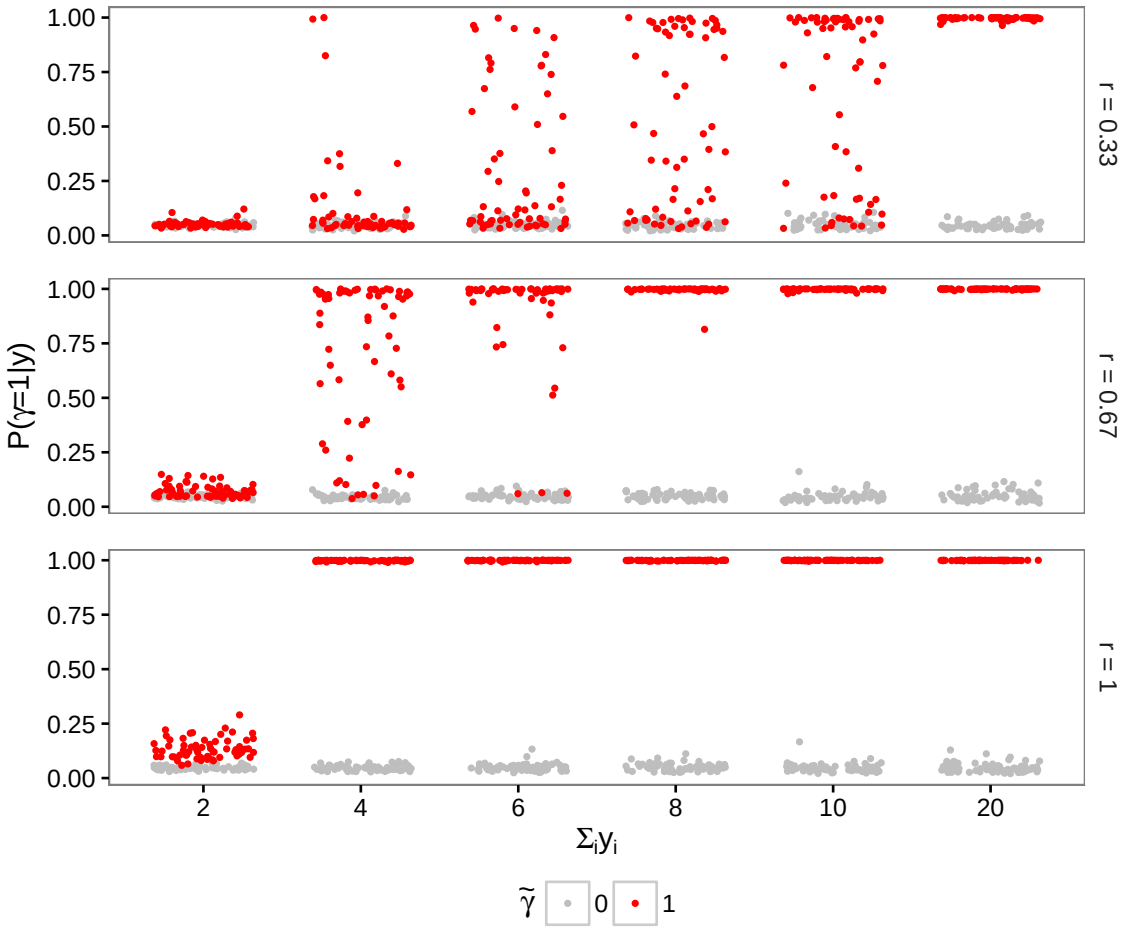


Fig. 3.4 Results of inference using SimReg on simulated data. Phenotype data were simulated using three levels of expressivity r of the disease terms. The plots within each panel correspond to different frequencies $\sum_i y_i$ of the rare genotype. In each plot, the red dots mark the estimated posterior mean of γ for 64 datasets simulated under $\tilde{\gamma} = 1$ and the grey dots show an equivalent set of estimates for datasets simulated under $\tilde{\gamma} = 0$ (i.e. whereby phenotypes for subjects having $y_i = 1$ are sampled from the same distribution as for subjects having $y_i = 0$). The position of points on the x-axis within a plot is arbitrary.

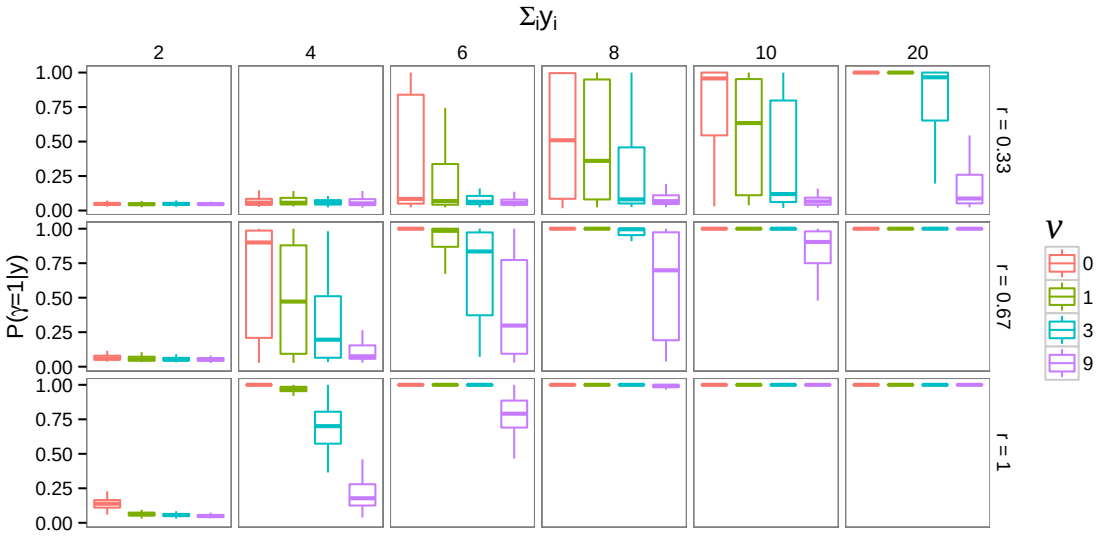


Fig. 3.5 Box plots showing the distribution of the posterior estimates of γ at various levels of phenotypic expressivity, r, for various sample frequencies of the rare genotype, $\sum_i y_i$ and with different levels of genetic heterogeneity as captured by ν. The boxes contain the inter-quartile range, with whiskers extending to the extreme values up to 1.5 times the inter-quartile range from the box.

were a total of 7 simulated datasets for which the value was greater than 0.25, with the highest estimate being equal to 0.86, which equates to a specificity of 99.97%. The dataset for which the highest value was obtained contained four (out of six) individuals with the rare genotype, labelled, 3–6, who had a high mean posterior similarity (> 0.3) to the characteristic phenotype (middle panel of Figure 3.6). By chance, these four individuals had been assigned highly specific terms relating to bone ossification, the toe and long bone morphology (right panel of Figure 3.6). This coincidental sharing of HPO terms by these individuals who also carried the rare genotype led to the abnormally high posterior estimate of γ. However, this is a desirable property of our method because in practice it is not possible to know whether such a correlation is causal or spurious.

Overall, the results of our simulation study show that our method produces accurate results even in the presence of significant phenotypic or genetic heterogeneity and low expressivity of the rare genotype's characteristic terms. As these are typical hallmarks of many rare disease studies, our evaluation substantiates the utility of our approach.

Computational performance We applied the inference procedure to simulated datasets to assess the performance of SimReg. We varied the number of phenotyped individuals

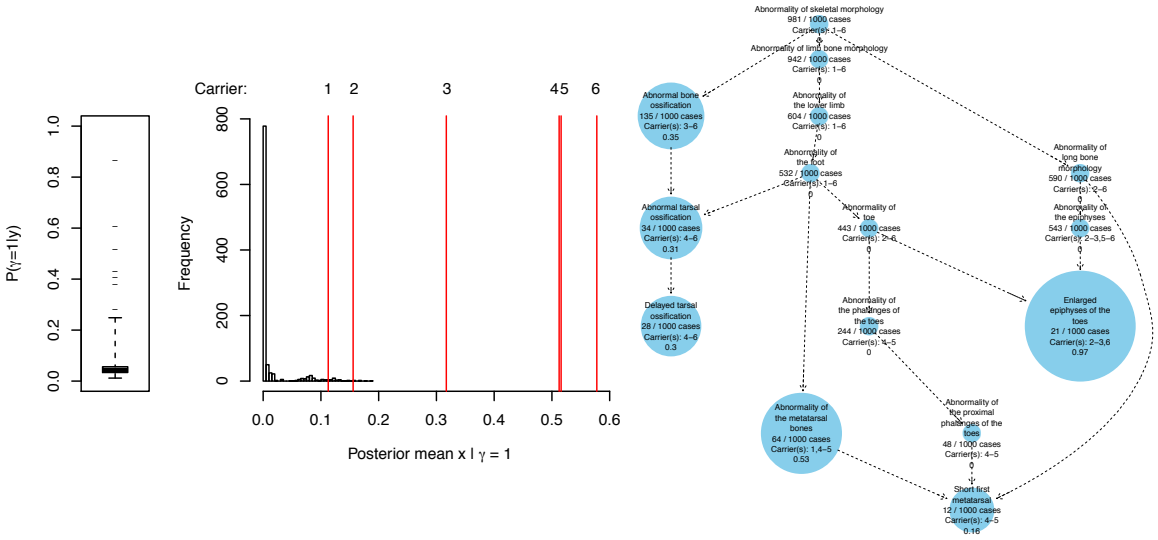


Fig. 3.6 The distribution of posterior γ for applications of the inference to the 20,000 repeats of the simulation is shown as a box plot. The box contains the inter-quartile range, with whiskers extending between the lowest posterior γ obtained and 0.25. For the simulated dataset for which the highest posterior mean value of γ was inferred, the posterior mean similarities to ϕ, x_i, for the 1,000 simulated patient phenotypes are shown as a histogram, with those of the individuals with the rare genotype, $i|y_i = 1$, marked by red lines. The inferred characteristic phenotype ϕ for this dataset is shown as a graph. Each node is labelled with a) the HPO term b) the number of simulated individuals out of 1,000 who had the term c) which individuals with the rare genotype had the term (as labelled in the middle panel) and d) the posterior probability of inclusion in ϕ conditional on $\gamma = 1$ (also represented by node size).

and the number of terms (sampled from a preset collection of approximately 1,000 terms) allocated to each individual, and programmed the algorithm to generate 20,000 MCMC samples (of which 10,000 are tuning iterations). The results of the performance test are shown in Table 3.1.

3.5 Application to real data

Our dataset comprises HPO phenotypes and corresponding variant call data for 2,045 unrelated individuals enrolled to the NBR–RD study. At the time the analysis was performed, detailed HPO data were available only for cases enrolled to the BPD project [119]. A set of genes within which variants are known to be implicated in each class of diseases was provided by BRIDGE collaborators to assess the performance of the model.

N	2 terms	4 terms	8 terms
100	7.31	9.03	11.55
1,000	47.38	65.70	91.74
10,000	451.54	627.75	879.78

Table 3.1 Completion times in seconds for applications of the SimReg procedure. The rows indicate the total number of individuals included in the inference, and the columns indicate the number of HPO phenotype terms allocated to each individual. These results were obtained by running SimReg on a single CPU of a computer with 2.40GHz processors.

We used variant call data from 686 sequenced exomes and 1,359 sequenced whole genomes. To account for biases that may alter the baseline rate of rare genotypes (e.g. sequencing platform), we use a plug-in offset in the regression equations (3.1), estimated *a priori* (see Appendix). Variants were retained only if they were predicted to alter protein sequence and were either absent from ExAC [55] or had an allele frequency therein below 1/1,000 or 1/10,000 respectively when a recessive or dominant mode of inheritance was assumed in the analysis. Rare variants were aggregated within genes to account for genetic heterogeneity and increase power. We defined the binary genotypes y based on three different aggregation approaches corresponding to the following hypothetical modes of inheritance:

- **Dominant:** presence of at least one rare allele,
- **Recessive:** presence of at least two rare alleles,
- **High-impact dominant:** presence of at least one rare allele predicted [16] to introduce a splice site aberration, frameshift, start loss or stop gain.

Estimation of the offset $\hat{h}_i$ In order to accommodate prior beliefs about the background rate of observing the rare genotype for a particular gene, we obtained point estimates of the effects of gene length and sequencing platform on the log odds of observing the rare genotype. We fitted a logistic regression model linking these variables to the genotype data across all genes for all 2,045 sequenced unrelated individuals. The model used was:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \lambda l_j + \omega^T z_{i\cdot}, \end{aligned}$$

where $y_{ij} = 1$ indicates presence of the rare genotype in gene j for individual i, which occurs with probability p_{ij}, l_j is the length in base pairs of the coding region of gene j and $z_{ik} = 1$ if individual i was sequenced on sequencing platform k and 0 otherwise. Thus, λ is

interpretable as the effect size of gene length and $\omega_1, \ldots, \omega_K$ as the effect sizes of sequencing platforms $1, \ldots, K$. We found that certain sequencing platforms led to gene-specific biases in variant calls. To ensure robustness to these biases, we only used data for genes having a Fisher exact p-value of association between the rare genotype and the sequencing platform greater than 0.05. Under a model of no association, the offset for gene j is given by:

$$y_i \sim \text{Bernoulli}(p_i),$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \hat{h}_i,$$

where $\hat{h}_i = \hat{\lambda} l_j + \hat{\omega}^T z_{i\cdot}$. The $\hat{h}_i$ was obtained for all genes in all the hypothetical modes of inheritance described above.

ACTN1 as exemplar gene

We now describe the properties of SimReg's output by focussing on the results for *ACTN1*, a gene that has recently been reported to harbour rare variants responsible for reduced platelet number and increased platelet size (macrothrombocytopenia) [49]. We note that data for *ACTN1* were used to inform and motivate our choice for the similarity measure given in Equation 3.2. Once learnt on the *ACTN1* data, this choice has then been used universally for all genes. We observe strong evidence that the rare genotype for *ACTN1* is associated with similarity to a characteristic phenotype ($\mathbb{P}(\gamma = 1|y) = 1$), as expected. The estimated characteristic phenotype focuses primarily on phenotypes that include 'Thrombocytopenia' and 'Increased mean platelet volume' (Figure 3.7), which together correspond to macrothrombocytopenia. The slightly more general terms 'Abnormal platelet count' and 'Abnormal platelet volume' also have substantial marginal posterior weight while the rest of the nodes in the HPO have a marginal posterior probability of inclusion less than 0.02. As can be seen in a two-dimensional matrix of the marginal posterior on pairs of terms, there is a high degree of co-occurrence of the two primary terms representing the *ACTN1*-related phenotype, which implies that they are not alternatives but rather complements that together produce a good model fit.

DIAPH1 and *RASGRP2*

Under the high impact dominant mode of inheritance described above, one of the genes with the highest estimated value of γ that also has a BPD-like inferred phenotype is *DIAPH1* ($\gamma = 0.87$). We recently showed, through an application of our similarity regression approach, that

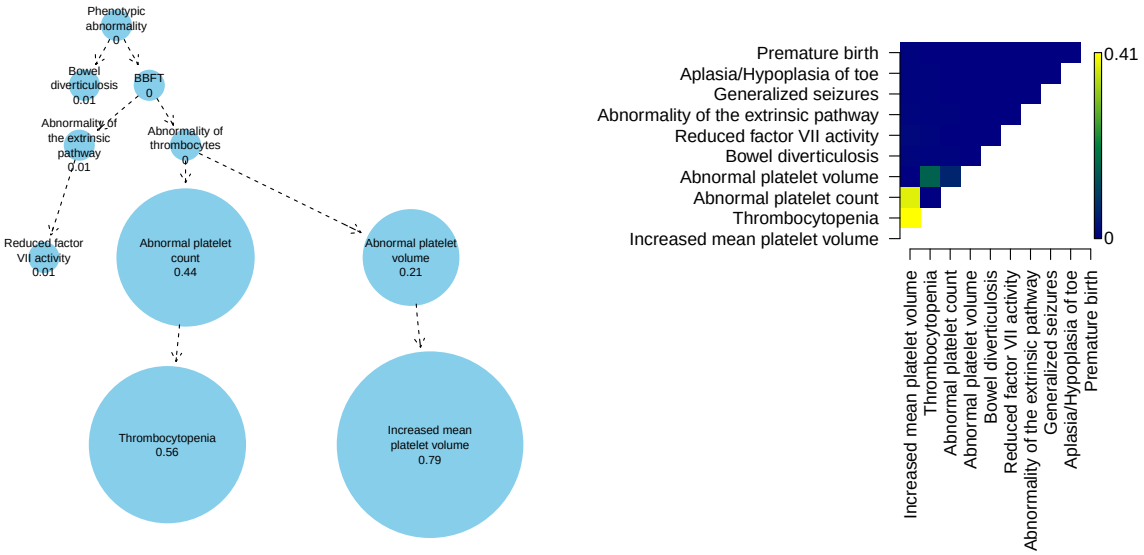


Fig. 3.7 Results of applying SimReg to phenotype data for all individuals and corresponding genotype data for *ACTN1*. There were 43 individuals in our dataset coded with the rare genotype for this gene, of which 22 were coded with 'Thrombocytopenia' and 'Increased mean platelet volume'. The graph shows the estimated probabilities of inclusion of individual terms in ϕ (only the 7 terms with the highest probabilities of inclusion and their ancestors are shown). The acronym 'BBFT' refers to Abnormality of blood and blood-forming tissues'. The heat map shows the estimated probabilities of pairs of terms co-occurring in ϕ, for pairs composed from the 10 most frequently included individual terms.

the introduction of a premature stop codon present in two unrelated individuals in the BPD project truncates DIAPH1's $3'$ auto-inhibitory domain and causes macrothrombocytopenia, hearing loss and mild bleeding [109]. As shown in Figure 3.8 (left), the salient terms in ϕ relate to hearing impairment and abnormality of blood and blood-forming tissues, with the latter driven mainly by thrombocytopenia and bleeding. The high posterior estimate of γ was obtained in part because a sensorineural hearing loss phenotype had previously been reported in the literature [64], which up-weighted hearing abnormality terms in the prior for ϕ (Table 3.2). However, even without using an informative prior on ϕ, a high posterior probability of an association ($\gamma = 0.59$) could be found for *DIAPH1*.

RASGRP2 was recently implicated in a new form of Glanzmann's-like thrombasthenia based on data from a single pedigree [11]. Glanzmann's is characterised by impaired platelet aggregation, leading to excessive bleeding. Under a recessive mode of inheritance, our similarity regression successfully detects an association ($\gamma = 0.75$) for *RASGRP2* and estimates a characteristic phenotype concentrated around 'Abnormal platelet aggregation' (Figure 3.8). It is characteristic of Glanzmann's that platelet aggregation is impaired in response to multiple agonists because their common downstream effect—the binding of platelets to fibrinogen—is impeded by the presence of reduced numbers of fibrinogen receptors. Here we also observe this phenomenon but only collagen-induced platelet aggregation carries significant weight in the characteristic phenotype because it is the only specific aggregation term that is shared by all the cases of this recently discovered disorder. There is also a very low probability of inclusion of two rare terms which are not related to the disease—'Atypical scarring of skin' and 'Intracranial meningioma'—because of a chance comorbidity in one of the affected cases.

Overall results

Finally, we turn our attention to the overall results of applying the inference procedure to data for all genes under the three modes of inheritance considered, subject to $\sum_i y_i \geq 2$. In total, we applied the inference to 19,573, 3,134 and 9,733 genes respectively for the dominant, recessive and high impact dominant modes of inheritance. The estimates of $\mathbb{P}(\gamma = 1|y)$ are shown as vertical density plots in Figure 3.9. For the majority of genes (65%), $\mathbb{P}(\gamma = 1|y) < \mathbb{P}(\gamma = 1) = 0.05$, which implies that no characteristic phenotype can be found that helps distinguish carriers of the rare genotype from other subjects. This result is consistent with the expectation that variants in only a small proportion of genes are implicated in these rare diseases and indicates that specificity is largely controlled.

Strikingly, under all three assumed modes of inheritance, most of the highly confident results (i.e. the genes for which the estimates of $\mathbb{P}(\gamma = 1|y)$ are close to one) are for genes

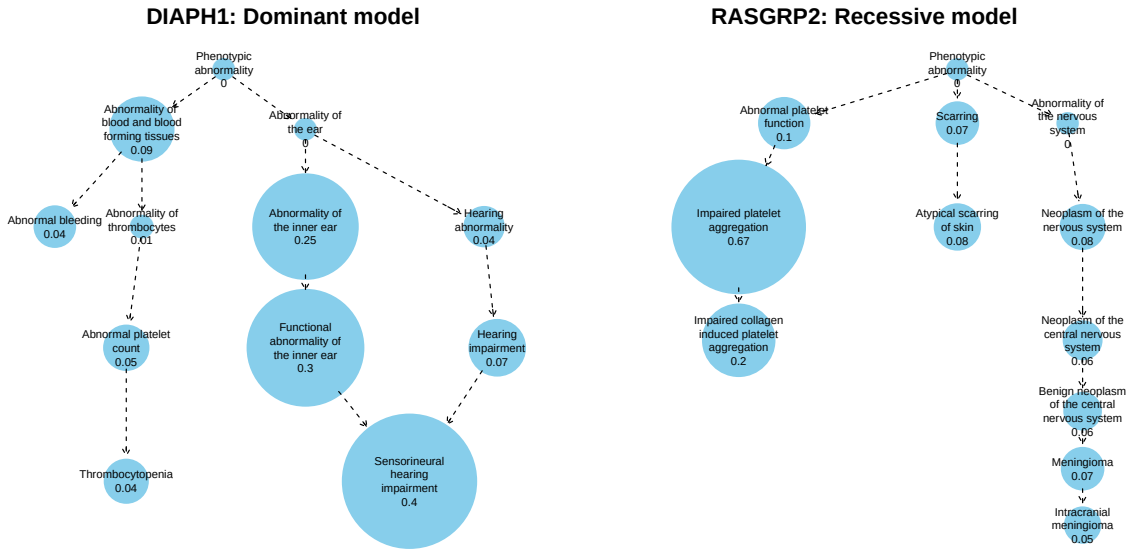


Fig. 3.8 Estimated posterior probabilities of individual terms being included in the characteristic phenotype ϕ using phenotype data for all subjects and variant data for *DIAPH1* ($\sum_i y_i = 2$) encoded under a high impact dominant model and *RASGRP2* ($\sum_i y_i = 7$) encoded under a recessive model. The 10 terms with the highest marginal posterior probability are shown. The estimated posterior probability that $\gamma = 1$ is equal to 0.872 and 0.750 for *DIAPH1* and *RASGRP2* respectively.

known to be relevant to the pathologies of the patients (indicated by red labels). In all but one case (*KIF1A*), where a gene had $\mathbb{P}(\gamma = 1|y) > 0.25$ and was in one of the projects' set of known genes, a characteristic phenotype similar to the known phenotype was inferred (Table 3.2). Above a threshold of $\mathbb{P}(\gamma = 1|y) = 0.25$, there was a significant enrichment for known genes (Fisher exact test $p = 2.39 \times 10^{-4}, 1.98 \times 10^{-4}$ and 2.23×10^{-7} for the dominant, recessive and high impact dominant modes of inheritance, respectively). Some of the inferred known genes are highlighted more than once across the three modes of inheritance in Figure 3.9 because there is power to detect the association even when the mode of inheritance is misspecified. For example, *RASGRP2*-related Glanzmann's is recessive, yet $\mathbb{P}(\gamma = 1|y) > 0.25$ even if a high impact dominant mode of inheritance is assumed.

The black dashes in Figure 3.9 correspond to unknown genes for which the inferred $\mathbb{P}(\gamma = 1)$ is greater than 0.25, of which there were 8, 1 and 5 found for the dominant, recessive and high impact dominant model of inheritance, respectively. These candidates are genes with potentially novel roles in disease and are being actively explored.

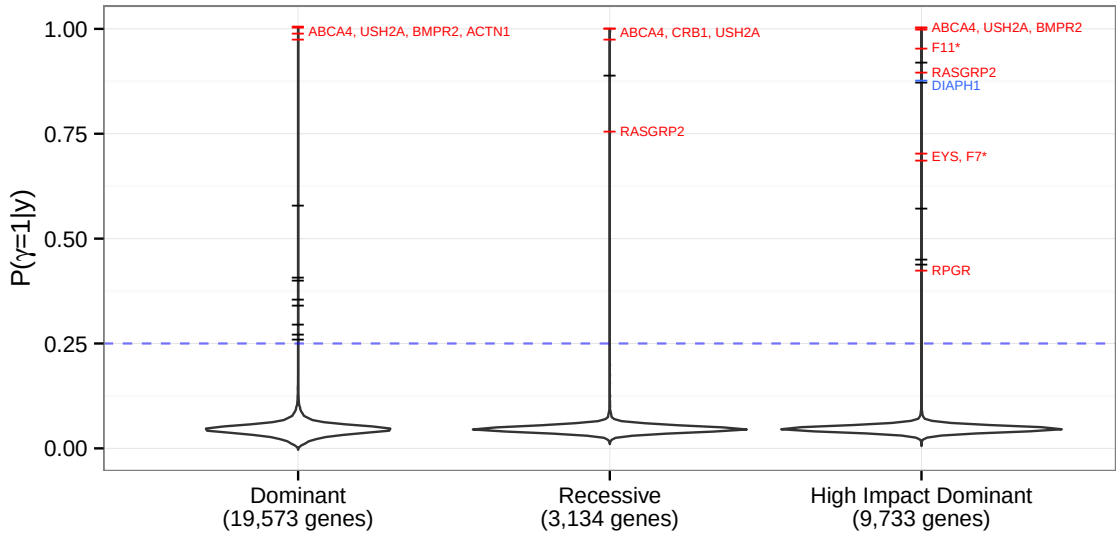


Fig. 3.9 Distributions of the estimated posterior means of γ obtained by applying the SimReg method to each gene under three different modes of inheritance. The estimated values greater than 0.25 are marked by dashes. The genes known to be associated with each of the NBR–RD subproject disorders having $\mathbb{P}(\gamma = 1|y) > 0.25$ and a compatible inferred phenotype are labelled and coloured in red. An asterisk indicates that a posterior mean of γ greater than 0.25 was estimated only with the use of a prior on ϕ that was informed by the literature of human and murine heritable disorders.

Gene	Mode of Inheritance	Known	$\mathbb{P}(\gamma = 1\|y)$	Highest Marginal Posterior Probability Terms in ϕ
ACTN1	Dominant	BPD	1.00	Increased Mean Platelet Volume (0.79), Thrombocytopenia (0.56), Platelet Count (0.44)
BMPR2	Dominant	PAH	1.00	Pulmonary Hypertension (0.34), Elevated Pulmonary Artery Pressure (0.31), Pulmonary Artery (0.11)
ABCA4	Recessive	RD	0.99	Retinal Dystrophy (0.22), Retina (0.22), Fundus (0.16)
USH2A	Recessive	RD	0.99	Retina (0.23), Retinal Dystrophy (0.2), Fundus (0.17)
CRB1	Recessive	RD	0.97	Retinal Dystrophy (0.21), Retina (0.18), Fundus (0.18)
F11	High Impact Dominant	BPD	0.95	Reduced Factor XI Activity (0.89), Intrinsic Pathway (0.11), Platelet Aggregation (0.07)
RASGRP2	Recessive	BPD	0.75	Platelet Aggregation (0.67), Collagen-induced Platelet Aggregation (0.2), Platelet Function (0.1)
EYS	High Impact Dominant	RD	0.70	Retinal Dystrophy (0.2), Retina (0.17), Fundus (0.14)
F7	High Impact Dominant	BPD	0.68	Extrinsic Pathway (0.5), Reduced Factor VII Activity (0.46), White Hair (0.1)
RPGR	High Impact Dominant	RD	0.42	Retina (0.2), Retinal Dystrophy (0.17), Posterior Segment Of The Eye (0.16)

Table 3.2 Genes for which $\mathbb{P}(\gamma = 1|y) > 0.25$ and the inferred phenotype was compatible with the known disease. We display the mode of inheritance under which the association was found, the known disease, the probability of association and the top three HPO terms in the inferred phenotypes. The marginal posterior probability of inclusion in the characteristic phenotype is shown in brackets next to each term. When an association was found under multiple modes of inheritance, only the true mode is shown. Note that the inferred phenotypes are influenced by prior phenotypic information in the form of OMIM and MGI annotations. The following abbreviations were used for the known disease classes: BPD - Bleeding and Platelet Disorder, RD - Retinal Dystrophy, PAH - Pulmonary Arterial Hypertension.

3.6 Discussion

We have described a method for identifying the genetic determinants of rare diseases that does not require the disease phenotype to be specified *a priori*. The method uncovers associations between rare genotypes and the similarities between patient phenotypes and a latent characteristic phenotype. Throughout this paper, rare variants have been aggregated within genes according to a hypothesised mode of inheritance in order to define presence or absence of a rare genotype. However, the unit of analysis could be a set of interacting domains or any other arbitrary genomic grouping. As the inference would typically be applied to each unit independently, it is important to set the priors on the model indicator γ with care in order avoid generating unrealistically high posterior probabilities of association. During final review of this work, a prioritisation procedure was proposed that combines a standard measure of strength of phenotypic clustering amongst individuals having two loss-of-function variants in a gene and the probability of the variants appearing in opposite haplotypes in an outbred population [2]. In contrast, our inference procedure is based on statistical principles and the formulation of a model which is flexible with regards to phenotypic expressivity and genetic architecture, and robust to noisy clinical coding and moderate genetic heterogeneity. Our Bayesian model naturally accounts for prior evidence of disease phenotypes associated with variants in particular genes by differentially weighting the prior probability of inclusion of HPO terms in the characteristic phenotype. Our finding that variants in *DIAPH1* can cause macrothrombocytopenia is an example of how this up-weighting can improve the inference.

The approach we have described is a natural and powerful way of modelling many rare disease phenotypes because it accounts for phenotypic abnormalities across all organ systems encoded with variable precision. Studies of syndromic diseases in particular may benefit from this way of uncovering associations. Our model can also be used for predicting the log odds of the rare genotype using solely phenotype data by means of a function implemented in our SimReg software. This could be used to aid diagnosis by indicating which of a patient's genes should be prioritised for sequencing based on his of her HPO terms. Finally, our regression approach may prove useful for performing inference using notions of similarity between terms in other ontologies where a binary response can be encoded.

In summary, our work represents an advancement in the statistical modelling of ontological heterogeneity which may prove useful at a time in which large collections of deeply phenotyped and sequenced cases are being assembled to uncover hitherto elusive causes of rare heterogeneous diseases. Although our method improves significantly on modelling of phenotypic heterogeneity, our treatment of genetic heterogeneity can still be refined, as we currently rely on aggregation of genetic information into single binary variables. In the following Chapter we explore improved modelling of genetic heterogeneity, in which the possibility of a mixture of pathogenic and neutral variants is accounted for explicitly.

Chapter 4

Bayesian evaluation of variant involvement in Mendelian disease

In this chapter, we present a fast and powerful inference procedure for identifying loci associated with rare hereditary diseases using Bayesian model comparison. Under a baseline model, disease risk is fixed across all individuals in a study. Under an association model, disease risk depends on a latent bipartition of rare variants in a locus into pathogenic and non-pathogenic, the number of pathogenic alleles that each individual carries and the mode of inheritance. A parameter indicating presence of an association and the parameters representing the pathogenicity of each variant and the mode of inheritance can be inferred in a Bayesian framework. Variant-specific prior information derived from allele frequency databases, consequence prediction algorithms or genomic datasets, can be integrated into the inference. Association models can be fitted to different subsets of variants in a locus and compared using a model selection procedure. This procedure can improve inference if only a particular class of variants confers disease risk and suggest particular disease etiologies related to that class. We show that our method, called BeviMed, is more powerful and informative than existing rare variant association methods in the context of dominant and recessive diseases. The high computational efficiency of BeviMed makes it feasible to test for associations in many loci, including regions in the large non-exonic fraction of the genome. We have applied BeviMed to whole-genome sequencing data from 6,586 individuals enrolled to the NBR–RD study. We show that it can identify multiple loci involved in rare diseases, while correctly inferring the modes of inheritance, the likely pathogenic variants and the variant classes responsible.

The description of the method given in this Chapter is based on that given in: D. Greene, NIHR BioResource–Rare Diseases Consortium, S. Richardson and E. Turro (2017). “A fast

association test for identifying pathogenic variants involved in rare diseases". *The American Journal of Human Genetics*, 101(1):104–114.

4.1 Introduction

The statistical association methods required to identify loci relevant to rare diseases need to fulfil several criteria. Firstly, they need to allow some sharing of information across variants because rare diseases are often genetically heterogeneous. Secondly, they need to account for the presence of pathogenic rare variants that act upon disease risk in a dominant or a recessive manner alongside benign rare variants that do not affect disease risk. Thirdly, they must be capable of integrating prior information into the inference regarding the plausibility of a locus being implicated in a disease and variant-level co-data on pathogenicity. Such co-data can be derived from population allele frequency databases, consequence predictions, conservation-based predictions or genomic datasets, for example. Lastly, methods need to have efficient implementations that enable fast application across a large number of regions in the genome.

Section 1.6 discusses currently available methods for rare variant association testing in the context of rare disease: Burden tests lose power due to the inclusion of neutral variants in aggregated rare variant counts; methods modelling complex traits lose power due to not modelling Mendelian modes of inheritance and methods which infer the likely pathogencity of variants are not applicable to dichotomous phenotypes or too slow to be applied to large genomic regions for large case/control cohorts.

Here we present a Bayesian model in which disease risk depends on the genotypes at rare variants in a locus, a latent mode of inheritance and a latent partition of variants into pathogenic and non-pathogenic subsets. Different modes of inheritance are modelled by conditioning the probability of case status on the number of pathogenic alleles and the ploidy for each individual at the variants. Thus, disease risk due to compound heterozygosity or X-linked inheritance is explicitly accommodated. Prior belief on variant pathogenicity can be incorporated in the form of shifts in the log odds of pathogenicity relative to a global mean. By placing a vague prior distribution on the scale of these shifts, the usefulness or otherwise of these co-data is accounted for flexibly to maximise power.

For a given set of variants, inference is performed by comparing the model described above with a baseline model in which disease risk is independent of the genotypes. The mode of inheritance and the pathogenicity of each variant, conditional on an association, can be inferred through the posterior distributions of parameters in the model. Particular classes of variants in a locus may be the only ones which confer disease risk. For example, only

variants in the 5′ UTR region or only high-impact coding variants may be involved. Our method can compare models fitted to different classes of variants in order to infer which ones are responsible for disease. Typically the inference process would be repeated over many sets of variants selected from different loci throughout the genome. The procedures are implemented in an efficient R package called BeviMed, which stands for *Bayesian evaluation of variant involvement in Mendelian disease*.

4.2 Model specification

Let y be a binary vector of length N indicating whether individual i is a case ($y_i = 1$) or a control ($y_i = 0$) with respect to a particular disease. Suppose k rare variants are under consideration (typically in a particular genomic region) and the genotype for individual i at variant j is coded in the ith row and jth column of the genotype matrix G. A genotype of 0 or 2 denotes homozygosity for the common or minor allele respectively and a genotype of 1 denotes heterozygosity. Under a baseline model, labelled $\gamma = 0$, y is independent of G and all individuals have a probability of being a case τ_0. Under the association model, labelled $\gamma = 1$, individuals either have or do not have a *pathogenic configuration of alleles*, and have probabilities of being a case π and τ respectively. Note that this leads to a slightly different interpretation of model parameters τ and τ_0: τ is the probability of observing case status given that there is no association with the locus, and τ_0 is the probability of observing the case status given that there is an association with the locus, but conditional upon the individual having a non-pathogenic configuration of alleles. Whether or not an individual has a pathogenic configuration of alleles depends on a function f of the genotypes $G_{i\cdot}$ of that individual, a latent binary vector z indicating which of the k variants are pathogenic, a value s_i equal to the ploidy of the individual at the variant sites, and a variable m representing the mode of inheritance governing the disease etiology though the k variants:

$$\begin{aligned} &\gamma = 0 : \mathbb{P}(y_i = 1) = \tau_0, \\ &\gamma = 1 : \mathbb{P}(y_i = 1) = \begin{cases} \tau & \text{if } f(G_{i\cdot}, z, s_i, m) = 0, \\ \pi & \text{if } f(G_{i\cdot}, z, s_i, m) = 1. \end{cases} \end{aligned} \tag{4.1}$$

The function f can represent a dominant inheritance model or a recessive inheritance model that accounts for sex-dependent differences in ploidy on the X chromosome (i.e.

X-linked recessive inheritance), depending on variable $m \in \{m_{\text{dom}}, m_{\text{rec}}\}$:

$$f(G_{i\cdot}, z, s_i, m_{\text{dom}}) = \mathbb{1}_{\sum_j G_{ij} z_j \geq 1},$$
$$f(G_{i\cdot}, z, s_i, m_{\text{rec}}) = \mathbb{1}_{\sum_j G_{ij} z_j \geq s_i}.$$

Thus, the interpretation of z depends on the mode of inheritance. In order to have a pathogenic allele configuration, individual i requires at least 1 allele at a variant for which $z_j = 1$ under a dominant model, but s_i alleles under a recessive model, where s_i would typically be inferred beforehand from the sequencing data. If genotypes are phased, then a requirement that the s_i pathogenic alleles are on different haplotypes can be imposed. Recent relatedness is a potential confounder because it is correlated with both case/control status and genotype and, therefore, only unrelated individuals should be included in the model.

We place beta priors on all three parameters representing risk of disease:

$$\tau_0 \sim \text{Beta}(\alpha_0, \beta_0),$$
$$\tau \sim \text{Beta}(\alpha_\tau, \beta_\tau),$$
$$\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi).$$

The mean risk of disease for individuals without a pathogenic combination of alleles in the variants under consideration is uncertain under both models, and thus we place uniform priors on τ_0 and τ by default. However, as pathogenic combinations of alleles typically confer a high disease risk, we suggest setting the hyperparameters for π to $\alpha_\pi = 6$ and $\beta_\pi = 1$. This prior has a mean of 6/7 and almost 90% of the support lies in the interval between 0.5 and 1, allowing for penetrance levels as low as 50% and as high as 100%. However, the prior mean could be adapted, for example, to reflect the consistency with which the disease manifests within families. Our model comparison framework is preferable to logistic regression with a spike-and-slab prior on the regression coefficient because it allows for specification of prior distributions and inference of posterior distributions conditional on γ.

We adopt a logistic regression framework (similar to the probit framework described in Quintana and Conti [86]) for the prior probability that the variants are pathogenic. The logit of the prior probabilities are shrunk towards a common mean, ω. If prior information which discriminates between the likely pathogenicity of variants is available, it can be incorporated in the form of a covariate c with regression coefficient ϕ in the regression equation:

$$z_j \sim \text{Bernoulli}(p_j),$$
$$\text{logit}\, p_j = \omega + \phi c_j.$$

One would typically place a Gaussian prior on the intercept ω but, for computational purposes, we prefer to use a logit-beta prior with hyperparameters α_ω and β_ω (see Section 4.3). The prior mean of ω should reflect the expected proportion of variants that are pathogenic, conditional on an association, and may depend on the filtering procedures used to select the variants to include in the model. By default, $p(\omega)$ reflects a prior expectation that 20% of variants are pathogenic and a prior probability of only 0.01 that the proportion of pathogenic variants exceeds 0.54. This prior is well suited to missense variants but a distribution with a higher mean should be specified if most variants are expected to be pathogenic. This would be the case if the variants under consideration are all protein-truncating and thought to be functionally equivalent to each other. To ensure that ω can be interpreted as the global mean log odds of pathogenicity, the c are required to sum to zero. Thus, any user-supplied weights, $\tilde{c}_j$, are centred such that $c_j = \tilde{c}_j - \frac{1}{k}\sum_l \tilde{c}_l$. By default, if not all weights are identical, they are also standardised by dividing by their standard deviation. By standardising the weights, a prior on the regression coefficient phi can be chosen so that the offsets to the log odds, ϕc_j, have realistic magnitudes given the capabilities of standard methods to predict the relative impacts of different rare variants on disease risk. We place a log-normal prior on the regression coefficient ϕ to force the effects of the c_j to be the same as their signs. The prior mean of ϕ is set to 1 so that the c_j are interpretable as prior shifts in the log odds of pathogenicity relative to the mean. A prior variance on ϕ of 0.35 ensures that there is enough probability mass close to zero so that the effect of the co-data can be diminished if the co-data are not informative and increased if they improve the model fit.

Finally, the prior probability on the mode of inheritance parameter m and the model indicator parameter γ need to be specified. By default, we set the prior probabilities for each mode of inheritance given an association to be the same, i.e. $\mathbb{P}(m = m_{\text{dom}}|\gamma = 1) = 0.5$, and we assume that there is only a 1% chance *a priori* of an association, i.e. $\mathbb{P}(\gamma = 1) = 0.01$. However, for a particular set of variants, the choice of values for these parameters could be based on the scientific literature or reference variant databases, for example.

4.3 Inference

Inference on presence of an association is based on the posterior probability of the model indicator γ, which can be derived from the evidence under each model and the prior on γ:

$$\mathbb{P}(\gamma = 1|y) = \frac{\mathbb{P}(y|\gamma = 1)\mathbb{P}(\gamma = 1)}{\sum_{u\in\{0,1\}} \mathbb{P}(y|\gamma = u)\mathbb{P}(\gamma = u)}.$$

Closed-form expressions exists for the evidence under models $\gamma = 0$ and $\gamma = 1$. The evidence for model $\gamma = 0$ can be computed efficiently, irrespective of y, using the beta function:

$$\frac{B(\alpha_0 + \sum_i y_i, \beta_0 + N - \sum_i y_i)}{B(\alpha_0, \beta_0)}.$$

The evidence for model $\gamma = 1$, $\mathbb{P}(y|\gamma = 1)$, can be expressed by conditioning on the different modes of inheritance and summing:

$$\mathbb{P}(y|\gamma = 1) = \mathbb{P}(\mathbb{P}(y|\gamma = 1, m = m_{\text{dom}}) + \mathbb{P}(y|\gamma = 1, m = m_{\text{rec}}),$$

The evidence under $\gamma = 1$ conditioning on mode of inheritance m, $\mathbb{P}(y|\gamma = 1, m)$, contains a sum over every possible value of z:

$$\sum_{z \in \{0,1\}^k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y|z,m) p(z|\omega,\phi) p(\omega) p(\phi) \, \mathrm{d}\omega \, \mathrm{d}\phi.$$

For brevity, we have omitted the hyperparameters α_τ, β_τ, α_π and β_π from the conditioning in $p(y|z,m)$ above.

The likelihood $p(y|z,m)$ factorises into two components corresponding to individuals with and without a pathogenic combination of alleles, where the rate parameters τ and π can be integrated out analytically. Thus the likelihood can be expressed in closed form:

$$\frac{B\left(\alpha_\tau + \sum_i y_i \left(1 - x_i^{(z)}\right), \beta_\tau + \sum_i (1 - y_i) \left(1 - x_i^{(z)}\right)\right)}{B(\alpha_\tau, \beta_\tau)} \times \frac{B\left(\alpha_\pi + \sum_i y_i x_i^{(z)}, \beta_\pi + \sum_i (1 - y_i) x_i^{(z)}\right)}{B(\alpha_\pi, \beta_\pi)}, \tag{4.2}$$

where $x_i^{(z)} = f(G_{i\cdot}, z, s_i, m)$.

As noted in the model specification, we use a logit-beta prior on ω, that is:

$$\text{expit}(\omega) \sim \text{Beta}(\alpha_\omega, \beta_\omega).$$

Thus, when $c_j = 0 \ \forall j \in \{1, \ldots, k\}$, z is independent of ϕ, and both ω and ϕ can be integrated out:

$$p(z) = \frac{B(\alpha_\omega + \sum_j z_j, \beta_\omega + k - \sum_j z_j)}{B(\alpha_\omega, \beta_\omega)}.$$

By default, $\alpha_\omega = 2$ and $\beta_\omega = 8$.

The space of z grows exponentially with the number of variants k, which can run into the dozens or hundreds. Therefore, despite the formulation of the model enabling many of the parameters to be integrated out analytically, the expression for the integrated likelihood cannot be evaluated in practice.

To tackle this problem, we reviewed various methods for estimating the evidence of a model [30] and chose the method of power posteriors [29], which enables the evidence to be estimated by Markov chain Monte Carlo (MCMC) sampling. In this method, the MCMC is tempered, which is helpful in a variable selection setting such as ours because it makes exploration of the space of sets of pathogenic variants more efficient. Samples are drawn from a series of related distributions of the parameters called power posteriors, $\mathbb{P}(z|y,m)^t\mathbb{P}(z)$, at temperatures $t \in \{t_0 = 0, \ldots, t_L = 1\}$ using MCMC. Let $\tilde{z}_l^{(b)}$ be the bth sample drawn at temperature t_l. These samples can be combined to obtain an estimate of the log integrated likelihood, $\log \mathbb{P}(y|\gamma = 1, m)$:

$$\sum_{l=0}^{L-1} \log \left(\frac{1}{|\tilde{z}_l|} \sum_{b=1}^{|\tilde{z}_l|} e^{(t_{l+1}-t_l)\log \mathbb{P}\left(y|\tilde{z}_l^{(b)},m\right)} \right). \tag{4.3}$$

Running Markov chains at different temperatures concurrently allows exchanges of state between chains at adjacent temperatures, which encourages good mixing. If the Kullback-Leibler divergence between adjacent power posterior distributions is large, the resulting estimates of $\mathbb{P}(y|\gamma = 1, m)$ may be susceptible to substantial numerical error. The method that minimises this error involves tuning the temperatures using a procedure such as interval bisection [28] and subsequently re-generating the chains to allow mixing between them. By default we use a pre-selected set of temperatures $t = \left(\frac{l}{6}\right)^2$ for $l \in \{0, 1, \ldots, 6\}$, and draw $1,000$ samples from each chain. This works well in practice and avoids the need to discard an initial set of MCMC samples for tuning the temperatures.

The use of MCMC to tackle this overall inference problem is in contrast to other methods designed for similar purposes [61, 51], probably because of the stringent requirements for computational speed. However, our algorithm contains features which makes MCMC sampling efficient.

In each chain, Gibbs sampling is used to update each individual component of z in turn. An update to z_j consists of sampling from its full conditional distribution:

$$\text{Bernoulli}\left(\left(1 + \frac{p(y|z^{(0)},m)p(z^{(0)}|\omega,\phi)}{p(y|z^{(1)},m)p(z^{(1)}|\omega,\phi)} \right)^{-1} \right),$$

where $z^{(0)}_{j'} = z^{(1)}_{j'} = z_{j'}$ for $j' \neq j$ and $z^{(0)}_j = 0, z^{(1)}_j = 1$. During the course of the algorithm, we keep track of $x_i = f(G_{i\cdot}, z, s_i, m)$ for each individual. Given an update of a single component of z, only individuals for whom $G_{ij} > 0$ need to have their corresponding value of x_i updated. G is often sparse as it typically represents rare allele counts, allowing this operation to be performed quickly. If values for c are specified then ω and ϕ are updated using a Metropolis Hastings within Gibbs.

Averaging over the space of all variant/variable sets using MCMC is a daunting challenge, in particular in circumstances such as these where non-additive models for the interaction effects of the variables are used. However, in practice, there is little collinearity between rare variant allele counts in unrelated individuals, rare allele count matrices are sparse because each individual is unlikely to have more than a handful of rare alleles in a given locus, and the interaction effects in dominant and recessive inheritance are quite simple, leading to low correlation between the elements of z. This means that the sampling procedure can explore the space of z efficiently.

However, when k is large and the mode of inheritance is recessive, with some cases being compound heterozygous, mixing of the MCMC sampler could potentially be poor if only one element of z is updated at a time. In particular, it could be very rare for the Markov chain to transition from a state satisfying $z_{j_1} = z_{j_2} = 0$ for some truly pathogenic variants j_1, j_2, to a state where $z_{j_1} = z_{j_2} = 1$, as there may be no intermediate state that would lead to an increase in likelihood. This is particularly problematic if the prior on ω is concentrated near 0. Thus, under $m = m_{\text{rec}}$, we propose updates to elements of z corresponding to variants occurring in the same individuals in tandem, which overcomes the potential rarity of sampling a state with a high likelihood.

The likelihood shown in Equation 4.2 can be expressed in terms of ratios of gamma functions with arguments which differ by integer amounts less than or equal to the number of individuals. Hence, the differences between all possible return values of $\log\Gamma$ that are required by the procedure can be computed before commencing the sampling and stored to avoid evaluating the $\log\Gamma$ function repeatedly. Evaluating $\log\Gamma$ is computationally expensive, so replacing it with look-ups in the pre-computed values tables results in significant speed-ups.

To improve computational efficiency whilst maintaining adequate precision, the implementation provides an option to stop sampling once the estimated evidence lies within a given confidence interval, or once there is sufficient confidence that the log evidence is below a given threshold. This behaviour is implemented based on the method of *consistent batch means* [25]. The log evidence is the sum of the logarithms of expectations taken with respect

to the power posteriors (Equation 4.3), so the central limit theorem does not apply and we estimate the confidence interval by simulation.

By default, if variant weights are either all zero or not specified, ϕ and ω are integrated out. Otherwise, they are estimated using Metropolis-within-Gibbs sampling, based on a random walk proposals with fixed variance. The BeviMed package includes options allowing the user to control whether and how the proposal variance is adapted in order to achieve a desired acceptance ratio. Fixed-length sequences of values for ϕ and or ω are drawn from the Markov chain using fixed proposal variances. Based on the proportion of proposals which were accepted in the sequence, the proposal variance is then increased or decreased in order to direct the acceptance rate towards a given interval (see Appendix C for more details).

The posterior probability of γ provides a natural means of ranking sets of variants from different loci across the genome. The marginal posterior probability of z given γ and m can be obtained directly from the samples drawn from the MCMC routine at the temperature $t = 1$ and used for ranking variants by their likely pathogenicity. The expected posterior number of cases whose disease risk was due to their pathogenic configuration of alleles and the expected number of pathogenic variants contributing to the explanation of those cases, given $\gamma = 1$, can also be computed using the samples drawn at temperature $t = 1$. The expected number of cases is computed using the expression $\mathbb{E}_{z|y} \sum_i y_i x_i^{(z)}$, and the expected number of variants involved is computed using the expression $\mathbb{E}_{z|y} \sum_{j \in \left\{ j' : \sum_i G_{ij'} y_i x_i^{(z)} > 0 \right\}} z_j$.

The model $\gamma = 1$ assumes that the prior probabilities of variant pathogenicity are conditionally independent. However, particular classes of variants in a locus may confer disease risk, while others may be benign. We can impose a prior correlation structure on the z reflecting these competing hypotheses by fitting a different association model for each class of variant. If one of the hypotheses matches the true etiology of disease, then this modelling approach can improve model fit and thus increase power. Let $\gamma \in \{1, 2, \ldots, g\}$ index the association models and let I_{uv} indicate whether variant v is included in association model u. Then, we can compute the probability of association across the competing models as:

$$\mathbb{P}(\gamma > 0 | y, G, c, I) = \frac{\sum_{u=1}^{g} \mathbb{P}(y | \gamma = u, G^{(u)}, c^{(u)}) \mathbb{P}(\gamma = u)}{\sum_{u=0}^{g} \mathbb{P}(y | \gamma = u, G^{(u)}, c^{(u)}) \mathbb{P}(\gamma = u)},$$

where $G^{(u)} = G_{\cdot \{v : I_{uv} = 1\}}$ and $c^{(u)} = c_{\{v : I_{uv} = 1\}}$. The prior on the model indicator, $\mathbb{P}(\gamma)$, can be informed by external data. For example, if a gene has a high probability of loss of function intolerance [55], then the prior corresponding to a model of high-impact variants in that gene could be up-weighted relative to competing models. We can also compute the posterior probability of variant pathogenicity averaged over all association models using the following

expression:

$$\mathbb{P}(z|\gamma>0,y,G,c,I)=\frac{\sum_{u=1}^{g}\mathbb{P}(z|\gamma=u,y,G^{(u)},c^{(u)})\,\mathbb{P}(\gamma=u|y,G^{(u)},c^{(u)})}{\sum_{u=1}^{g}\mathbb{P}(\gamma=u|y,G^{(u)},c^{(u)})}.$$

Other quantities of interest, such as the expected posterior number of cases explained by pathogenic variants, can be averaged over models in the same way.

4.4 Simulation study

We conducted a simulation study under different scenarios and using different methods in order to evaluate power to detect associations, assess accuracy in variant pathogenicity classification and investigate the effect of integration of variant-level co-data on inference. We generated random allele count matrices for $1,000$ individuals at k rare variant sites with allele frequencies of 0.0017 and 0.03 respectively for the dominant and recessive modes of inheritance. We used $k=25$ for the main simulation study. We labelled the first five variants pathogenic and the remaining variants non-pathogenic. The case/control labels were simulated using the expression in Equation 4.1, assuming $s_i=2$ (i.e. diploidy), a particular mode of inheritance (either dominant or recessive) and a particular combination of values for τ and $\pi\in\{0,\frac{1}{10},\frac{2}{10},...,1\}$ such that $\pi>\tau$. Our selection of τ and π is comprehensive but for rare diseases we would expect $\tau<0.5$ and $\pi\gg 0.5$. For each combination of mode of inheritance, value of τ and value of π, $5,000$ allele count matrices were generated and $5,000$ corresponding case/control vectors were generated. The $5,000$ datasets were copied and the case/control labels corresponding to the copied set were permuted to break the association between the genotypes and the phenotype. Thus, under each scenario, we had a pool of $10,000$ datasets, of which half were generated under a model of association and half were generated under a model of no association.

In order to assess the performance of different methods in a realistic setting, we evaluated their ability to rank non-permuted datasets amongst a large set of permuted datasets. Under each simulation scenario, we generated mixtures of ten non-permuted and 990 permuted datasets selected at random from the corresponding pool. We then applied each method and computed the mean positive predictive value (PPV), over $10,000$ repetitions, at 80% power. The PPV, which is equal to one minus the false discovery rate (FDR), is inversely related to power. Thus, a higher PPV for a given power implies greater power for a given FDR. We preferred to evaluate PPV rather than power because it changes monotonically as the rank threshold for declaring a positive result is lowered, while the empirical FDR does not.

We selected the methods ADA, CAST and SKAT for comparison as they represent diverse approaches: ADA enables individual variant-level inference, CAST is based on the popular Burden test but can account for either dominant or recessive inheritance modes, and SKAT is a popular and flexible method designed for rare variants affecting complex traits. The other methods mentioned above were either inapplicable (e.g. vbdm requires a continuous response), too computationally expensive to incorporate into the simulation study (BRI), unavailable (BRVD) or shown to be inferior to ADA in a previous publication [61]. Note that ADA p-values were computed using $10,000$ permutations instead of the default $1,000$ in order to reduce the number of ties.

The results were ranked based on the posterior probability that $\gamma = 1$ for BeviMed and the negative log p-value of association for the other methods. Variants were ranked according to BeviMed's marginal posterior probability for the components of z and according to inclusion in ADA's variant selection. The other methods do not provide variant-level inference. Although the backwards elimination procedure is implemented for SKAT, it is so slow as to make its use impractical in even a moderately sized study such as this.

Under a dominant model, BeviMed had a slightly higher PPV than the other methods while, under a recessive model, it greatly outperformed competing methods: when $\pi = 0.8$ and $\tau = 0.2$, BeviMed had a PPV of 100%, whilst SKAT, CAST and ADA had a PPV of 42%, 8% and 2% respectively (Figure 4.1A, B). This favourable performance was achieved despite using the same priors for model parameters τ and π, irrespective of the values of τ and π used to simulate the data. We note that BeviMed's performance for $\tau = 0.2$ was approximately the same for the following three pairs of values for the hyperparameters α_ω and β_ω): (2,8), which is the default, (1,1), which places a uniform prior on expit(ω) and (2,1), which places higher prior weight on values of expit(ω) near 1.

For $\tau = 0.2$ and high π, BeviMed was able to provide accurate rankings of variants by estimated pathogenicity, particularly under a dominant mode of inheritance (area under the curve = 0.97 at $\pi = 0.9$, Figure 4.1C). ADA's average classification of variant pathogenicity at $\pi = 0.9$ gave a true positive rate of 0.78 and a false positive rate of 0.063, while BeviMed's true positive rate at that same false positive rate was 0.88.

We evaluated the performance of BeviMed in relation to the most competitive alternative method, SKAT, using the same parameters described above, but increasing the total number variants k to 50, 100, 150 and 200. Power decreased for both methods as the total number of variants increased, but the discrepancy in power between BeviMed and SKAT increased (Figure 4.1E). For example, under the dominant model, BeviMed's PPV at $k = 200$ and $\pi = 1.0$ was 83% while SKAT's PPV was only 34%.

To demonstrate the effect of including prior information regarding variant pathogenicity on BeviMed's inference, we conducted a further study whereby we simulated datasets with $m = m_{\text{dom}}$, $\tau = 0.2$ and $\pi = 0.85$ and modified the values of the variant-specific co-data $\tilde{c}_j$ as follows. The values of $\tilde{c}_j$ for all variants was set to either 1 or 0. The number of truly pathogenic variants which were assigned the value 1 was set to $0, 1, 2, 3, 4$ or 5, and the number of truly non-pathogenic variants which were assigned the value 1 was set to $0, 4, 8, 12, 16$ or 20. Thus, the proportions of correctly and incorrectly up-weighted variants were varied between 0 and 1 in increments of 0.2. In the extreme, the co-data could support the true classification perfectly or support an entirely incorrect classification perfectly. As SKAT can incorporate variant-specific relative weights, we applied it to the same simulated data, setting SKAT's weights for up-weighted variants to twice that of the others. This choice of up-weighting factor was as low as possible whilst ensuring that, when the weights were perfectly concordant with the true pathogenicity of the variants, the PPV was approximately the same for SKAT as for BeviMed. Expected PPV at 80% power was estimated as described above, based on $5,000$ truly associated and $5,000$ permuted datasets, for BeviMed and SKAT under each combination of proportions of correctly up-weighted and incorrectly up-weighted variants.

The results show that BeviMed is substantially more robust to deleterious weightings (Figure 4.1D, left). When the co-data matched the truth perfectly, the power for BeviMed and SKAT was approximately the same (by design), but when the co-data was entirely counter-productive, BeviMed's PPV was 0.46 and SKAT's PPV was 0.06. BeviMed's advantage was achieved naturally in our Bayesian setting through modulation of ϕ, which had a posterior expectation of 1.93 when the co-data was most useful but only 0.46 when it was least useful (Figure 4.1D, right).

Computational performance

We compared the execution times of the different association tests, including SKAT with backwards elimination, on simulated datasets generated as described above using $N \in \{1000, 5000, 100000\}$, $k \in \{25, 100, 1000\}$ and allele frequency of $1/1,000$. The results, displayed in Table 4.2, show that CAST is the fastest method, as it uses a straightforward Fisher's exact test. However, CAST is substantially less powerful than BeviMed under both dominant and recessive models (Figure 1). BeviMed has comparable execution time to SKAT for small datasets and surpasses it for large datasets, as BeviMed's complexity scales with $\sum_{i,j} G_{ij} > 0$, which typically increases only linearly with N. SKAT was also run using the other kernels available in the R package [53]: *identity by state*, *quadratic* and *two-way interactions*. All three modes were substantially slower than BeviMed and linear SKAT

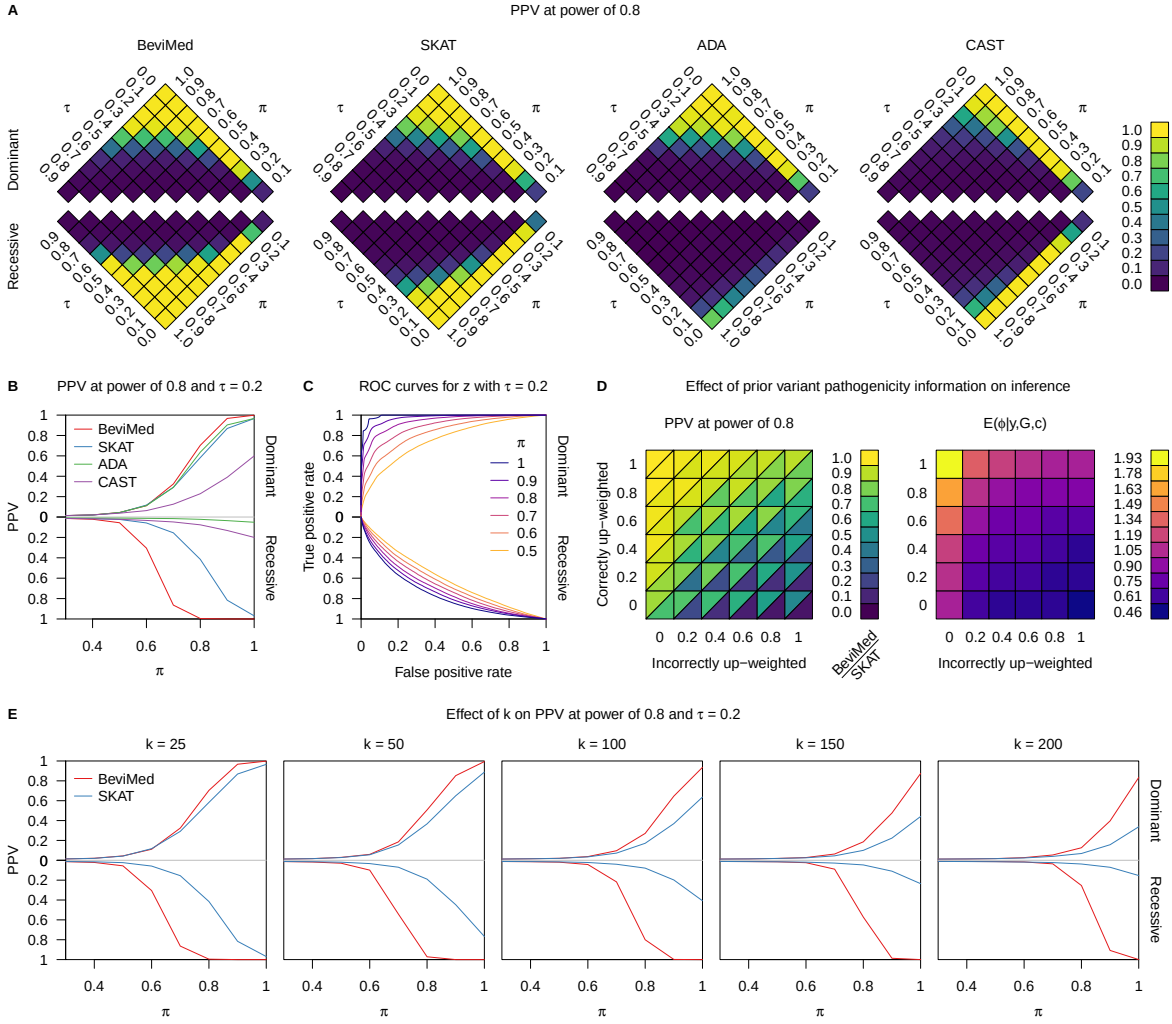


Fig. 4.1 Results of the simulation study. **A**, **B** Mean PPV at power of 80% over repeat simulation of the BeviMed, SKAT, ADA, and CAST rare variant association tests for data simulated using the expression in Equation 4.1 for various combinations of values of τ and π. **C** Receiver operating characteristic (ROC) curves for the classification of variants as pathogenic by BeviMed for different values of π. **D** Left: mean PPV at power of 80% for BeviMed and SKAT at $\tau = 0.2$ and $\pi = 0.85$, for varying proportions of pathogenic and non-pathogenic variants being up-weighted in the co-data variables. Right: posterior mean of ϕ corresponding to the applications of BeviMed on the left-hand grid. **E** Mean PPV at power of 80% over repeat simulation of the BeviMed and SKAT association tests for different values of k.

(Table 4.2), less powerful than BeviMed in the simulation study and less powerful than linear SKAT under at least one mode of inheritance. BeviMed has vastly superior performance to the other methods which can infer the pathogenicity of variants, whilst also reporting posterior uncertainty in pathogenicity status. The complete set of applications to the data from the NBR–RD, which comprises two phenotypes, 35,205 loci, two modes of inheritance and four variant classes, took 7 hours to complete using 16 CPU cores.

Software used in comparisons The source code for ADA was downloaded from http://homepage.ntu.edu.tw/~linwy/ADAprioritized.html. Parts of the ADA code were re-implemented in C++ in order to complete our simulation study in a reasonable amount of time.
The implementation of the SKAT method from the R package SKAT [53] (version 1.2.1) was used in the simulation study.
The source code for the BE-SKAT method was downloaded from http://www.columbia.edu/~ii2135/ (package version 1.0).

4.5 Application to real data

The NBR–RD has generated whole genome sequencing data for 6,586 unrelated individuals with diverse rare diseases in an effort to identify novel genetic etiologies. We applied BeviMed to the data, setting the case/control status based on two dichotomous phenotypes represented in the study: pathologically low numbers of platelets in the blood stream (thrombocytopenia) with absence of syndromic features ($\sum_i y_i = 184$), and Roifman syndrome ($\sum_i y_i = 3$).

Hereditary thrombocytopenia can be caused by variants in a large number of genes with diverse functions, including genes encoding transcription factors, cytoskeletal proteins and membrane proteins [56]. Severe thrombocytopenia is typically monogenic and non-syndromic forms are usually dominant. Roifman syndrome is a rare autosomal recessive disease with symptoms including growth retardation, spondyloepiphyseal dysplasia and cognitive delay, initially described by Roifman et al [94]. Last year, variants in the non-coding gene *RNU4ATAC* were identified as the cause of this disease on the basis of pedigree studies involving six cases [70]. Within the bleeding and platelet disorders branch of the NBR–RD dataset, three unrelated cases with Roifman were enrolled because they presented with immune thrombocytopenia. It has recently been described that defects in *RNU4ATAC* lead to abnormal minor intron retention in hundreds of genes [41].

For each gene, we considered single nucleotide variants (SNVs) and short insertions/deletions (indels) with an allele frequency in ExAC [55] and the whole-genome sequencing component of UK10K [115] less than $1/1,000$ and large deletions overlapping exons with an internal frequency less than $1/1,000$. SNVs and indels had to have a `HIGH` or `MODERATE` Variant Effect Predictor (VEP)[69] impact or they had to have a VEP Sequence Ontology-coded consequence that included `non_coding_transcript_exon_variant`, `5_prime_UTR_variant` or `3_prime_UTR_variant`. If a variant had consequences in relation to multiple transcripts of the same gene, only the highest-impact consequence was retained. In total, we considered $1,338,830$ variants in $35,205$ gene loci, each containing between 1 and $2,615$ variants.

We set $\mathbb{P}(m = m_{\text{dom}}|\gamma = 1)$ to 0.8 for thrombocytopenia and 0.1 for Roifman syndrome, to reflect the belief that these diseases tend to be dominantly and recessively inherited, respectively. For each locus, we considered four association models corresponding to four classes of variants:

- "High": large deletions and variants with a `HIGH` impact annotation,
- "Moderate": variants with a `MODERATE` or `HIGH` impact annotation or a consequence including `non_coding_transcript_exon_variant` but none of the UTR-related consequences,
- "5′ UTR": variants without a `MODERATE` or `HIGH` impact annotation and a consequence including `5_prime_UTR_variant`,
- "3′ UTR": variants without a `MODERATE` or `HIGH` impact annotation and a consequence including `3_prime_UTR_variant`.

The hyperparameters were assigned default values except for α_ω and β_ω, which we set to $(2,1)$ instead of $(2,8)$ under the "High" model. This reflects a belief that a greater proportion of variants are likely to be pathogenic under the "High" model than under the other three models. When we fitted the "Moderate" model, we up-weighted the variants that were also included in the "High" class relative to the others by setting their uncentred weights $\tilde{c}_j$ to one rather than zero. For coding loci, we assigned prior probabilities of 0.004, 0.003, 0.002 and 0.001 to the four models above, respectively, in order to reflect the relative biological plausibility of the different classes of variants being involved in disease. For non-coding genes, we assigned a prior probability of the "Moderate" model equal to 0.01. Thus, $\mathbb{P}(\gamma > 0) = 0.01$ for all genes. For completeness, we also applied SKAT to the four classes of variants described above separately, using default settings, and retained the result with the lowest p-value for each locus.

Identifying associations with thrombocytopenia

The median value of the posterior probability of association with thrombocytopenia across all gene loci was 0.0064. The independent gene loci for which the posterior probability of association exceeded 0.9 are shown in Table 4.1. We show the posterior probability of association, the posterior probability of the mode of inheritance parameter, the estimated number of cases explained by the pathogenic variants, the estimated number of pathogenic variants which are present in the cases and the total number of variants considered. These results corroborate established knowledge of platelet disorders. *ACTN1* related macrothrombocytopenia is a dominant bleeding and platelet disorder recognised since 2013 [49]. *GP1BB* has traditionally been linked to a recessive bleeding and platelet disorder called Bernard-Soulier syndrome [96], but earlier this year we reported a dominant mode of inheritance resulting in a milder platelet phenotype [101]. The posterior on the mode of inheritance parameter strongly favoured dominance in this case, which is consistent with an absence of Bernard-Soulier cases in our dataset. *RUNX1* encodes a transcription factor linked with a dominant platelet disorder with associated myeloid malignancy. *MYH9* harbours variants responsible for *MYH9*-related disorder, which is characterised by macrothrombocytopenia and occasional Döhle-like inclusion bodies in neutrophils and pathologies of the ear, eye, kidney or liver. Finally, variants in the 5′ UTR of *ANKRD26* were reported to result in non-syndromic macrothrombocytopenia in 2011 [82] after an initial erroneous report that variants in the neighbouring gene *ACBD5* were responsible [85]. The association we have identified is driven by variants in the 5′ UTR, despite this class of variants being down-weighted relative to the classes comprising variants with missense or high-impact predicted consequences on the translated gene product. The variant level results of the inference shown in Figure 4.2 indicate the high posterior probability of association for the first eight variants in the 5′ UTR. It is noteworthy that one of the variants, which encodes c.−113A>C and was reported in a follow-up [76] to the original 2011 paper, does not appear to be pathogenic, as five out of the six individuals with the variant, including one homozygous for the alternate allele, do not have a bleeding or platelet disorder.

There were four additional loci having $\mathbb{P}(\gamma = 1|y) > 0.9$. They all tagged a true association in Table 4.1 but were labelled with the names of neighbouring genes and had lower posterior probabilities of association. A missense variant encoding Ile308Thr in *WAC* was in linkage disequilibrium with one of the 5′ UTR variants in *ANKRD26*, inducing $\mathbb{P}(\gamma = 1|y) = 0.993$ for the *WAC* locus. The other three results were induced by the presence of deletions in *RUNX1* spanning three neighbouring RNA genes or pseudogenes: *AF015262.2* ($\mathbb{P}(\gamma = 1|y) = 0.978$), *RPL34P3* ($\mathbb{P}(\gamma = 1|y) = 0.976$) and *EZH2P1* ($\mathbb{P}(\gamma = 1|y) = 0.974$).

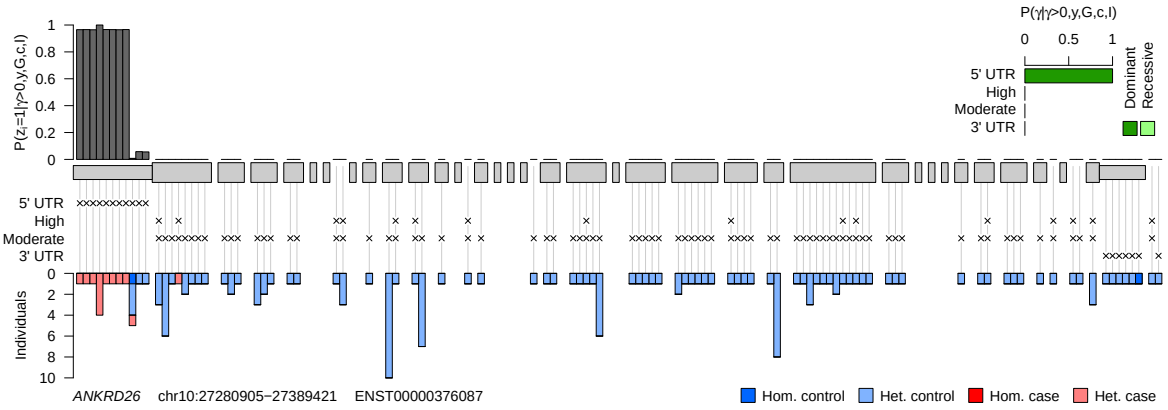


Fig. 4.2 Results obtained by applying our inference procedure to rare allele counts in *ANKRD26* against the thrombocytopenia case/control label. Exons are represented by grey blocks starting from the 5′ UTR on the left and ending with the 3′ UTR on the right. The bar chart above shows the marginal posterior probabilities of pathogenicity conditional on dominant inheritance for individual rare variants observed amongst the cases and controls. A small amount of jitter due to Monte Carlo sampling error can be observed. The bar chart beneath shows the breakdown of heterozygous and homozygous carriers of the variants between cases and controls.

The alternate method which was most powerful based on the results of the simulation, SKAT, did not rank the loci listed above as highly, even when only the variant class with the lowest p-value was retained for each gene. *RUNX1*, *ANKRD26*, *MYH9* and *ACTN1* and *GP1BB* had ranks of 1,3, 8, 16 and 74, respectively, with none of the other loci in the top 20 ranks being known to be implicated in thrombocytopenia.

Identifying variants responsible for Roifman syndrome

The locus with the highest posterior probability of association with the Roifman syndrome case label was *RNU4ATAC* ($\mathbb{P}(\gamma = 1) = 1.000$), driven by four different single nucleotide variants in this non-coding gene. Two of the cases were compound heterozygous, including for a variant observed in 6 controls, and one was homozygous. As all but one of the variants were seen only in heterozygosity, the posterior probability of variant pathogenicity conditional on recessive inheritance was relatively high across the gene but markedly lower than the causal variants observed in the cases, which had a posterior probability of pathogenicity very close to 1 (Figure 4.3). All other genes had a posterior probability of association less than 0.9 and the expected number of cases explained by the variants in other loci was less than 2. SKAT assigned *RNU4ATAC* a p-value of zero, but this was also the case for 34 other genes, which were tied in the top rank.

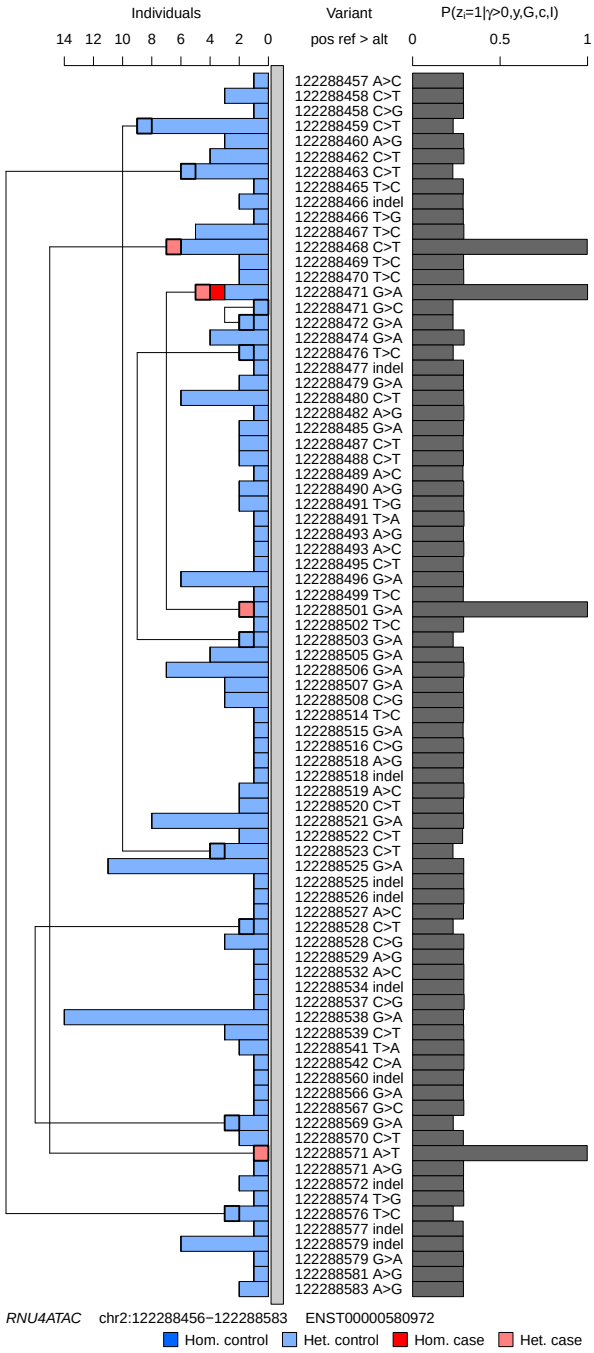


Fig. 4.3 Results of applying the inference procedure to rare allele counts in *RNU4ATAC* against the Roifman syndrome case label. The bar chart on the left shows the break down of heterozygous and homozygous carriers of the variants in cases and controls. Compound heterozygous individuals with two rare alleles in *RNU4ATAC* were observed, and for each such individual a line is drawn linking the two variants. The bar chart on the right shows the marginal posterior probabilities of pathogenicity conditional on recessive inheritance for each rare variant. A small amount of jitter due to Monte Carlo sampling error can be observed.

Locus	Posterior probability of association	Posterior probability of dominance	Modal model	Estimated number of explained cases	Estimated number of explaining variants	Number of variants
ANKRD26	1.000	1.000	5′ UTR	10.792	7.792	87
RUNX1	1.000	1.000	Moderate	8.153	8.191	214
MYH9	1.000	1.000	Moderate	10.964	9.116	141
GP1BB	0.999	1.000	Moderate	8.223	7.228	69
ACTN1	0.999	1.000	Moderate	9.867	7.867	121

Table 4.1 Independent loci with a posterior probability of association with thrombocytopenia greater than 0.9.

4.6 Discussion

We have presented a Bayesian genetic association method for rare diseases that is more powerful than existing methods, particularly for the recessive mode of inheritance, and provides summary statistics on variant-level pathogenicity and mode of inheritance very efficiently. It enables mode of inheritance to be integrated out or inferred from the data. Indeed, we were able to determine a dominant mode of inheritance for variants in a gene, *GP1BB*, that has been associated only with a recessive disease for more than thirty years. Given an association under a particular mode of inheritance, our method also estimates the number of cases explained by pathogenic variants and the number of variants that are pathogenic.

Prior information specific to a particular set of variants under consideration can modulate the evidence of association, which can be critical when the number of cases with a shared genetic etiology is small. For example, the prior on the model indicator γ can be adjusted to reflect locus-specific genomic and epigenomic knowledge in order to encourage regions with higher prior plausibility of involvement in the disease phenotype to rank more highly than if the same prior had been used across all regions. As inference would typically be applied to many regions, the priors on γ across different applications should be chosen with care to avoid a high overall FDR for a given threshold on the posterior probability of γ. The prior probability of pathogenicity for a particular variant, given the association model, can be modulated by knowledge about the variant, such as predicted consequence, allele frequency or conservation. These variant weights are interpretable as prior shifts in the log odds of

Method	N=1000, k=25	N=1000, k=100	N=5000, k=25	N=5000, k=100	N=100000, k=1000
Association tests with variant identification					
BeviMed	0.03	0.09	0.07	0.23	38.90
ADA	3.69	11.30	18.76	59.23	–
BE-SKAT	53.46	175.18	137.39	799.80	–
Association tests without variant identification					
CAST	0.01	0.01	0.02	0.05	1.77
SKAT	0.02	0.05	0.09	0.30	140.82
SKAT (IBS)	3.73	4.38	548.47	675.62	–
SKAT (quadratic)	4.08	4.12	571.69	598.96	–
SKAT (2wayIX)	4.09	4.37	580.90	575.67	–

Table 4.2 Execution times in seconds of different association tests for datasets with different N and k. BE-SKAT refers to SKAT with backwards elimination of variants. SKAT (IBS) refers to application of SKAT using the weighted identity by state kernel function. The p-values for ADA and BE-SKAT were computed using their default number of permutations, respectively $1,000$ and 300. Dashes indicate that the method took longer than 1 hour to compute.

pathogenicity, which provides an intuitive basis for assigning particular values to them. In sharp contrast to frequentist approaches, we use a flexible prior on the effect sizes of the weightings that reflects the uncertainty in their utility.

The results of inference on different subsets of rare variants in a locus (selected, for example, on the basis of their predicted consequences) can be interpreted and combined easily in a Bayesian framework using a model selection procedure. The posterior probability of variant pathogenicity and other quantities of interest can be averaged over models. In addition to increasing statistical power if particular classes of variants in a locus are the only ones that confer disease risk, this feature also allows inference of the kind of variants responsible for disease, which may suggest particular genetic etiologies. In our applications, we were able to identify a set of variants in the $5'$ UTR of a gene that causes a platelet disorder. The high posterior probability of pathogenicity of variants in the $5'$ UTR to the exclusion of coding variants, even those observed only in cases, was made possible by our model selection procedure.

Variants highlighted by a method such as ours would usually undergo assessment by a multidisciplinary diagnostic team and it would resolve increasing numbers of cases over time. In our application to real data, we have kept the case/control labels the same for each application. However, in the context of genetically heterogeneous diseases, we would recommend relabelling any case whose phenotype has been fully accounted for by pathogenic or likely pathogenic variants in a different locus as a control. This boosts specificity as it makes it less likely for a non-pathogenic rare variant carried by a case to induce a high probability of association.

The model assumes that relatedness between individuals is sufficiently low as not to be associated with either case/control status or the genotypes. In practice, we recommend removal of any first, second and third degree relatives. Our method is designed to be applied to up to thousands of rare variants at a time and efforts should be made to ensure all potentially implicated variants in a locus are included in the model, or the set of models, under comparison. Rare variants would typically be unlinked within a locus but may occasionally be linked across loci. For example, large deletions may span multiple genes and certain pairs of rare variants could be in linkage disequilibrium. In these situations, a non-pathogenic rare variant in one locus linked to a pathogenic variant in another locus could induce a non-causal association. Such associations can either be filtered post-hoc through comparison of inference results in nearby loci or avoided altogether by joint modelling of variants across multiple nearby loci.

Although Bayesian inference is typically thought of as slow, our implementation can handle data from over a million variants spread across tens of thousands of regions called in thousands of samples in a few hours. BeviMed is thus capable of handling with ease the demands of modern genomic datasets in the coding and the regulatory regions of the genome.

Chapter 5

Gene-docs: a web application for browsing phenotypic and genetic data

I have contributed to the ongoing analysis of the phenotypic and genetic data generated by the BPD arm of the NBR–RD project as part of analysis team with a variety of backgrounds, including statistical analysts and clinicians. The raw data relating to individuals recruited to the NBR–RD project are stored at a secure data centre. Remote access to the data requires experience with unix software, particularly for connecting to remote servers securely. The data are typically stored in large, compressed files and require experience with specialised software for browsing and querying. It is therefore difficult for collaborators without bioinformatics training to make use of the data in their research. Furthermore, as there are many different diseases represented amongst the probands, cumulatively carrying millions of different rare variants, many separate analyses are required in order to infer which hypotheses are most probable and should be followed up in further work. Sharing up-to-date results of such inference in an interpretable way is therefore an important part of the research.

In order to improve access to the data and share the results of analysis, I have developed Gene-docs, a web application which provides a user interface for querying the genetic and phenotypic information so that it can be reviewed effectively, and providing prioritised results of analysis. It has been deployed successfully for the data generated by the NBR–RD project, proving its efficacy for a large rare disease sequencing project. The underlying functionality, responsible for querying the variant call and phenotype data and performing analysis, is written in R. The presentation layer is written in a combination of languages: php, javascript and html. Part of this work comprised development of a computational pipeline to run systematic analyses of the data, including BeviMed and SimReg.

In order to apply BeviMed to the collection, probands must first be clustered into groups with similar clinical phenotypes suggestive of a possible shared genetic aetiology so that

probands in the each group can be treated as cases whilst the rest of the probands can be treated as controls. The task is challenging because of the extensive heterogeneity of phenotypes and diversity of clinical coding practices, and therefore cannot be undertaken by unsupervised computational methods alone. Guided by expert biological and clinical opinion, a set of rules were developed to classify cases into groups of with potentially shared disease aetiology based on their phenotypes. These 'case group rules' are used to automatically link new cases with group labels based solely on their phenotype, without specific input required from clinicians or analysts. In a scenario where new data are made available steadily, this reduces the time spent assigning people to case groups, reduces the chance that mistakes are made, and means that new results can be forwarded for further investigation sooner after data are released.

SimReg is applied in a similar fashion to variants in each gene, except that instead of using case/control labels, the HPO phenotype of each proband is used. The results of systematically applying SimReg and BeviMed are presented as tables, where gene/gene-phenotype associations respectively with the highest posterior probabilities of association are display and arranged in descending order of posterior probability of association. For SimReg, this table contains columns containing the gene involved, the posterior probability of association, and a summary of the inferred characteristic phenotype giving the three most marginally associated HPO terms. Each result is then linked to a page which gives the full details of the inference, including the raw data and inferred characteristic phenotype as an ontological plot. For BeviMed, the table of results includes columns containing the gene, probability of association, probability of dominant inheritance given association, modal model (i.e. corresponding to a particular class of variants as described in Section 4), expected number of cases explained and expected number of variants involved. Results are linked to pages showing the detailed results of the inference and raw data, including the table of variants in the region to which the analysis was applied annotated with bar charts giving the marginal posterior probability of involvement of each variant (see Figure 5.2). Tabulating the most significant results of analysis and linking them to the raw data significantly speeds up the process of deciding on whether new associations are biologically plausible and worth following up.

Gene-docs has multiple forms designed to handle different kinds of queries (see Figure 5.1). A form enabling data lookup by gene generates tables of variants within the query gene, listing individuals harbouring each variant. Users can select filters dictating which variant to include based on functional class, predicted consequence, allele frequency thresholds based on ExAC [55] and CADD Phred score thresholds [45]. Users have the option of specifying a dichotomous disease phenotype, which triggers BeviMed to run in real time

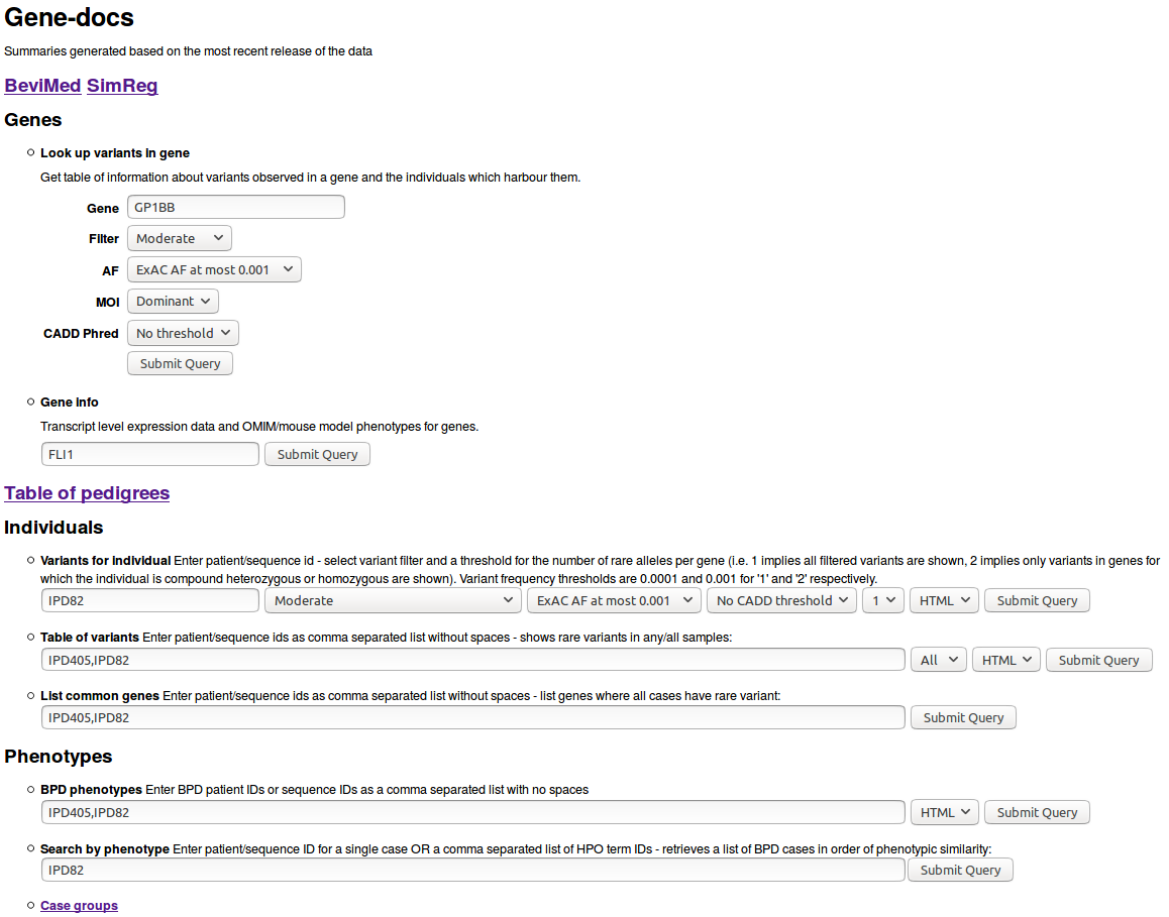


Fig. 5.1 Screenshot of the Gene-docs main page showing links to pages containing the results for the systematic application of BeviMed and SimReg, and forms for submitting various queries against the phenotypic and genetic data. Forms are organised into sections: the 'Genes' section contains forms which allow data pertinent to given genes to be looked up; the 'Table of pedigrees' links to a table of statistics about all pedigrees in the BPD study with links to detailed individual pedigree pages; the 'Individuals' section contains forms which allow genetic data relevant to given samples or sets of samples to be looked up and the 'Phenotypes' section contains a form which allows phenotypic data to be looked up by individual, a form which allows individuals to be sorted in order of phenotypic similarity to a given individual, and a link to detailed phenotypic information on the members of each case group.

Gene	CHROM	POS	ID	REF	ALT	BeviMed	Cases	Type	Amino.Acid
GP1BB	22	19711095	.	G	C		***A001105*** *(GP1BB)*	start_lost	ENSP00000383382.2:p.Met1?
GP1BB	22	19711380	.	C	T		C003039	missense_variant	ENSP00000383382.2:p.Pro5Leu
GP1BB	22	19711427	.	CGCCCGGCCGCAGGTT	C		***A012351*** *(GP1BB)*	inframe_deletion	ENSP00000383382.2:p.Ala24_Ala28del
GP1BB	22	19711493	.	G	C		E010668	missense_variant	ENSP00000383382.2:p.Gly43Arg
GP1BB	22	19711506	.	C	T		C002057, F006354, R014584	missense_variant	ENSP00000383382.2:p.Ala47Val
GP1BB	22	19711514	.	C	A		G001399	missense_variant	ENSP00000383382.2:p.Pro50Thr
GP1BB	22	19711569	.	C	T		C002439, ***A007322*** *(GP1BB)*	missense_variant	ENSP00000383382.2:p.Thr68Met
GP1BB	22	19711581	.	C	T		C003994*	missense_variant	ENSP00000383382.2:p.Pro72Leu
GP1BB	22	19711629	.	C	T		U012296	missense_variant	ENSP00000383382.2:p.Ala88Val
GP1BB	22	19711673	.	T	C		***A014177*** *(GP1BB)*	missense_variant	ENSP00000383382.2:p.Trp103Arg
GP1BB	22	19711682	.	G	A		F011414	missense_variant	ENSP00000383382.2:p.Gly106Ser
GP1BB	22	19711686	.	G	A		F009752*, F009934*	missense_variant	ENSP00000383382.2:p.Arg107His
GP1BB	22	19711730	.	C	T		**N013595**	missense_variant	ENSP00000383382.2:p.Pro122Ser
GP1BB	22	19711755	.	C	A		R011982	missense_variant	ENSP00000383382.2:p.Pro130His
GP1BB	22	19711761	.	T	A		***A007385*** *(GP1BB)*	missense_variant	ENSP00000383382.2:p.Leu132Gln
GP1BB	22	19711769	.	G	A		L009406, L009898*, **A010811**	missense_variant	ENSP00000383382.2:p.Asp135Asn
GP1BB	22	19711823	.	G	A		C003968, F010353, Y014680	missense_variant	ENSP00000383382.2:p.Ala153Thr
GP1BB	22	19711890	.	T	C		N013582*, N013617*, N013628*	missense_variant	ENSP00000383382.2:p.Leu175Pro
GP1BB	22	19711910	.	G	T		E001385, F010048	missense_variant	ENSP00000383382.2:p.Ala182Ser
GP1BB	22	19711932	.	G	A		**A001088**	missense_variant	ENSP00000383382.2:p.Arg189Gln
GP1BB	22	19711964	.	C	T		E009211	stop_gained	ENSP00000383382.2:p.Arg200Ter
GP1BB, SEPT5	22	19711634	.	C	T		S012384	missense_variant	ENSP00000383382.2:p.Pro90Ser
MICAL3, XXbac-B476C20.13, RHEBP3 (332 more)	22	18475385	.				**A007271**	large_deletion	
SEPT5, GP1BB	22	19711413	.	T	C		*A007261 (GP1BB)**, ***A007262*** *(GP1BB)*, ***H011038*** *(GP1BB)*	missense_variant	ENSP00000383382.2:p.Leu16Pro
SEPT5, GP1BB	22	19711493	.	G	T		***H011483*** *(GP1BB)*, *B200402 (GP1BB)**	missense_variant	ENSP00000383382.2:p.Gly43Trp
SEPT5, GP1BB	22	19711503	.	G	A		***A001106*** *(GP1BB)*, ***A010146*** *(GP1BB)*	stop_gained	ENSP00000383382.2:p.Trp46Ter
SEPT5, GP1BB	22	19711776	.	T	C		G013449	missense_variant	ENSP00000383382.2:p.Leu137Pro
SEPT5, GP1BB	22	19711808	.	TG	T		***A007380*** *(GP1BB)*	frameshift_variant	ENSP00000383382.2:p.Ala150ArgfsTer43

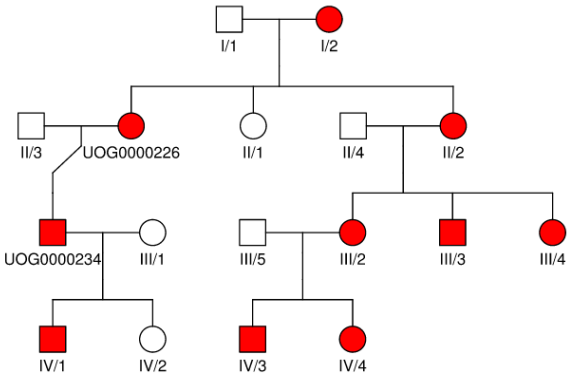
Fig. 5.2 Screenshot of a table of variants from gene specific page for *GP1BB* with BeviMed annotation of variant posterior probabilities of pathogenicity. Each row corresponds to one variant in the locus, with the sample IDs of individuals carrying alleles for the variant in the 'Cases' column. Bold text indicates a case, whilst regular text indicates a control. Asterisks indicate that people were excluded from the analysis because they were either unaffected or relatives of a proband, and gene names in brackets next to a sample ID indicate that a variant in the named gene has been identified as the causative variant. In addition to the information which identifies each variant, i.e. the chromosome, position, reference sequence and alternate sequence, the table also contains the dbSNP [98] ID (ID), the VEP Sequence Ontology-coded consequence (Type), the amino acid change (Amino Acid) and unshown columns the affected transcript identifier [3], the CADD Phred score and allele frequency in various reference cohorts.

allowing annotation of the variants with probabilities of pathogenicity given association with the gene (Figure 5.2). javascript enabling the variants to be filtered based on probability of pathogenicity is included in the page. Further to the table of variants, a table of additional information for each individual with alleles in the given gene is generated. Information included in the table includes sequence/patient IDs, enrolling centre names, gender, ethnicity, clinical notes and phenotypic information.

Another form allows lookup by sample ID(s), bringing up detailed phenotypic information about the query samples, including graphs showing the HPO encoded phenotype rendered using ontologyPlot. There is also a form for looking up families of individuals by pedigree ID. On the pedigree page, the `R` package `kinship2` [112] is used to draw an annotated pedigree diagram (Figure 5.3). In addition to the standard phenotypic and metadata for pedigree members, where pedigrees have affected sequenced members, variant call data are analysed to bring up a table of rare variants which cosegregate with affection status within the pedigree, ordered by the phenotypic similarity to the ontological phenotype profiles of model organisms for genes orthologous to the ones harbouring the variants.

Family UOG0000226

Family



Case statistics

Patient ID	Seq ID	Centre ID	Pedigree	Gender	Status	OrphaName	M.O.I. Terms	Labels
UOG0000226	A009770	Griefswald	UOG0000226	F				StraightforwardThrombocytopenia,RedCellMPV
UOG0000234	A009769	Griefswald	UOG0000226	M				StraightforwardThrombocytopenia

Fig. 5.3 Screenshot of top part of Gene-docs pedigree specific page showing pedigree diagram and table of statistics about participating members of the pedigree, including identifiers, gender, diagnostic status (Status), names of applicable syndromes (OrphaName), HPO terms relevant to mode of inheritance (M.O.I. Terms) and the names of case groups to which the individual has been assigned (Labels).

For each case group rule, Gene-docs generates a page tabulating detailed meta and phenotypic information for each affected ‘case’ (with respect to the given rule). The table of phenotypic information is sorted by ascending order of mean phenotypic similarity to the other cases in the group in order to ease the process of identifying individuals who do not belong in the group. The phenotype based rules are useful in a project such as the NBR–RD project where affected individuals are regularly recruited because the groups can be derived autonomously.

Another form allows searching by phenotype, generating a list of HPO coded cases ordered by phenotypic similarity to the query phenotype. The query phenotype can be specified as an individual if generating a list of cases phenotypically similar to the individual is desired.

Chapter 6

Conclusions

Historically, linkage-based analysis of pedigrees has been used to identify genes which contain variants responsible for rare diseases. The falling cost of high-throughput sequencing has led to the initiation of international whole-genome rare disease sequencing projects, increasing power to detect association between genomic loci and disease across large cohorts. Despite these advances, methodological development for rare diseases has not kept pace: methods for analysing such data have not properly taken account of the composite nature of rare disease phenotypes and the mixture of pathogenic and benign variants in conjunction with Mendelian modes of inheritance. Furthermore, frequentist methods have been preferred to Bayesian methods to test for association between disease status and genomic loci [51], as they often allow p-values to be computed rapidly. However, there are many sources of prior information which are relevant to rare disease, for example tissue-specific expression data and mouse model phenotypes. Including such information in the analysis could be critical to boost signal over a critical threshold when analysing diseases with very low case numbers. It is therefore important that Bayesian methods are developed which appropriately account for key aspects of rare diseases and which are scalable to modern datasets comprising thousands or hundreds of thousands of indviduals.

Rare disease phenotypes are now frequently encoded using terms from the HPO. A lack of software for representing such data in mainstream statistical programming environments has made such ontologically encoded data difficult to analyse. In Chapter 2 I described ontologyX, a suite of R packages which enable ontological data for arbitrary ontologies to be manipulated and analysed effectively: ontologyIndex provides a simple set of functions for reading ontologies, transforming ontological annotation and supporting extension to higher-level functionality; ontologySimilarity enables semantic similarity computation for annotation sets belonging to arbitrary ontologies and is thousands of times faster than existing R packages, and ontologyPlot makes visualising sets of ontological terms and relationships

between them straightforward. I then showed how ontological similarity can be used to analyse rare disease phenotypes, including testing for significant phenotypic similarity of subgroups of rare disease cases in a larger cohort, and how using ontological similarity between mouse models for gene knock-outs and the ontologically encoded disease of a family can be used to support prioritisation of candidate causal variants for a family with a rare disease. Finally, I showed how the software can be used to construct clusters of rare disease cases based on unsupervised clustering of ontologically encoded phenotypes, and that by using ontologically encoded model organism phenotypes to inform a subsequent gene-set enrichment analysis of clusters, good power for detecting associations can be achieved.

In Chapter 3 I described SimReg [37], a method based on a statistical model which links disease risk to ontologically encoded phenotypes, in which a latent disease phenotype is inferred from the data. We demonstrated effectiveness of this method by applying it to over 2,000 rare disease cases belonging to the NBR–RD project and inferred many genuine associations between diseases and genes. Prior knowledge about the kinds of disease phenotype involved is incorporated into the prior distribution for the latent phenotype by up-weighting the prior probability mass on disease phenotypes by a factor which depends on the similarity to the ontological profile for the gene derived from the literature. This approach boosted the evidence for the association between *DIAPH1* and hearing impairment, leading to the discovery of a novel association between the gene and bleeding abnormality and thrombocytopenia [109]. Since publication, SimReg has also been incorporated into an open-source web application for phenotypic and genetic data called 'Phenopolis' [84], where it is used to augment summaries of data relevant to genes with inference summaries including the characteristic phenotype.

In Chapter 4 I described 'BeviMed' [38], a method based on a statistical model which links disease risk to an individual's configuration of alleles at rare variant sites in a locus. It explicitly models a mixture of pathogenic and non-pathogenic alleles, and dominant and recessive inheritance by presence of at least one and two pathogenic alleles respectively. It is faster and more powerful than comparable existing methods and is feasible for application to modern datasets. It has been applied to the data generated by the NBR–RD project and shortlisted numerous true associations and implicated many novel variants in disease [113, 118]. The output of the method applied across multiple genomic loci can be tabulated, prioritised by the probability of association, displaying information about inferred aetiology (i.e. the posterior distribution on the class of variants involved), inferred mode of inheritance, and estimated number of cases explained and variants involved.

In Chapter 5, I described a computational analysis pipeline and web application for presenting genetic data, phenotypic data and analytical results together in a way which

enables collaborators without bioinformatics training to fine-grained access to raw data and browse significant results of analysis. During the course of my PhD studies, development of the application has been guided by feedback from clinical researchers and ongoing development of new methods, with many features added over time. Consequently it is widely used amongst researchers involved in the NBR–RD project and has helped streamline the discovery and publication process on several occasions.

The methods presented in this thesis formalise essential properties of the genetic basis of rare diseases in statistical models, and have been applied to large datasets, resulting in the discovery of novel aetiologies of disease. However, further methodological development will be necessary in order to fully exploit the data generated by large sequencing studies for rare diseases. Association methods which model the relationship between HPO encoded phenotypes and rare variation while accounting for genetic heterogeneity within a locus could increase power to detect genomic regions in many unsolved rare diseases. In future work I intend to develop such a method, which extends the model for phenotypic heterogeneity presented in Chapter 3 in order to account for local genetic heterogeneity and Mendelian inheritance as is done by the model presented in Chapter 4. The challenge will be modelling the response as a vector of genotypes at rare variant sites rather than a single aggregate genotype, while conditioning on a latent parameter representing the mode of inheritance, a latent parameter representing the pathogenicity of each variant and the latent parameter representing the characteristic HPO phenotype. This will substantially increase the size of the model space and make inference computationally challenging, requiring the development of new algorithms.

Variation in DNA sequence at non-coding loci can influence gene expression regulation, for example by affecting the propensity of proteins called transcription factors for binding to DNA and effecting their normal role in transcription. Genetic variation can also effect the folding of DNA into three-dimensional structures, affecting accessibility of the DNA to transcription machinery (including transcription factors) and the ability of distal regulatory DNA elements to interact with that machinery. Recent results reveal the role of non-exonic DNA sequences in a variety of diseases through their effects on transcription binding sites and the three-dimensional structure of DNA [106]. However, the size of the non-coding genome relative to the number of WGS samples sequenced so far means that naively testing segments of the genome for association with disease is underpowered. Therefore methods for identifying which regions of the genome are more likely to be involved in gene expression regulation can be informative. For instance, in ATAC-seq [10], which stands for ‘Assay for Transposase-Accessible Chromatin Sequencing’, determining which loci of the genome are potentially involved in regulation for a particular tissue type. Additionally, mechanisms

depending on non-coding loci for maintaining normal gene expression may not be as sensitive to the specific sequence of nucleotides as normal protein structure is. Therefore models which up-weight the importance other aspects of DNA sequences in non-coding loci — for example the size of copy number variations, which could play a role in the three-dimensional structure of DNA — may be helpful for investigating involvement in disease.

The work presented in this thesis will prove useful for analysing variation affecting the regulatory regions of the genome, and serve as a good starting point for further methodological development in the field.

References

[1] 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[2] Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A. F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T. W., Foulds, N., Francis, R., Gabriel, G., Gerety, S. S., Goodship, J., Hobson, E., Jones, W. D., Joss, S., King, D., Klena, N., Kumar, A., Lees, M., Lelliott, C., Lord, J., McMullan, D., O'Regan, M., Osio, D., Piombo, V., Prigmore, E., Rajan, D., Rosser, E., Sifrim, A., Smith, A., Swaminathan, G. J., Turnpenny, P., Whitworth, J., Wright, C. F., Firth, H. V., Barrett, J. C., Lo, C. W., FitzPatrick, D. R., and Hurles, M. E. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet.*, 47(11):1363–1369.

[3] Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J.-H., White, S., Zadissa, A., Flicek, P., and Searle, S. M. J. (2016). The ensembl gene annotation system. *Database*, 2016:baw093.

[4] Anikster, Y., Huizing, M., White, J., Shevchenko, Y. O., Fitzpatrick, D. L., Touchman, J. W., Compton, J. G., Bale, S. J., Swank, R. T., Gahl, W. A., and Toro, J. R. (2001). Mutation of a new gene causes a unique form of Hermansky-Pudlak syndrome in a genetic isolate of central Puerto Rico. *Nat. Genet.*, 28(4):376–380.

[5] Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755.

[6] Bauer, S., Kohler, S., Schulz, M. H., and Robinson, P. N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*, 28(19):2502–2508.

[7] Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *The American Journal of Human Genetics*, 94(6):809–817.

[8] Blake, J. A., Bult, C. J., Eppig, J. T., Kadin, J. A., and Richardson, J. E. (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, 42(Database issue):D810–7.

[9] Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., Yengo, L., Wild, P. S., Schillert, A., Ziegler, A., et al. (2013). GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS genetics*, 9(8):e1003657.

[10] Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology*, pages 21–29.

[11] Canault, M., Ghalloussi, D., Grosdidier, C., Guinier, M., Perret, C., Chelghoum, N., Germain, M., Raslova, H., Peiretti, F., Morange, P. E., Saut, N., Pillois, X., Nurden, A. T., Cambien, F., Pierres, A., van den Berg, T. K., Kuijpers, T. W., Alessi, M. C., and Tregouet, D. A. (2014). Human CalDAG-GEFI gene (*RASGRP2*) mutation affects platelet function and causes severe bleeding. *J. Exp. Med.*, 211(7):1349–1362.

[12] Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

[13] Carlson, M. (2016). *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 3.3.0.

[14] Castellani, C., Cuppens, H., Macek, M., Cassiman, J., Kerem, E., Durie, P., Tullis, E., Assael, B., Bombieri, C., Brown, A., et al. (2008). Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *Journal of cystic fibrosis*, 7(3):179–196.

[15] Chen, L., Kostadima, M., Martens, J. H., Canu, G., Garcia, S. P., Turro, E., Downes, K., Macaulay, I. C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. *Science*, 345(6204):1251033.

[16] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92.

[17] Cooper, D. N., Ball, E. V., and Krawczak, M. (1998). The human gene mutation database. *Nucleic acids research*, 26(1):285–287.

[18] Cullinane, A. R., Curry, J. A., Carmona-Rivera, C., Summers, C. G., Ciccone, C., Cardillo, N. D., Dorward, H., Hess, R. A., White, J. G., Adams, D., Huizing, M., and Gahl, W. A. (2011). A BLOC-1 mutation screen reveals that *PLDN* is mutated in Hermansky-Pudlak Syndrome type 9. *Am. J. Hum. Genet.*, 88(6):778–787.

[19] Davies, R. B. (1980). The distribution of a linear combination of x2 random variables. *Applied Statistics*, 29(3):323–333.

[20] Easton, D. F., Ford, D., and Bishop, D. T. (1995). Breast and ovarian cancer incidence in *BRCA1*-mutation carriers. Breast Cancer Linkage Consortium. *American journal of human genetics*, 56(1):265.

[21] Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., and Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

[22] Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., and Dudbridge, F. (2012). Comparison of methods for competitive tests of pathway analysis. *PloS one*, 7(7):e41018.

[23] Feng, L., Seymour, A. B., Jiang, S., To, A., Peden, A. A., Novak, E. K., Zhen, L., Rusiniak, M. E., Eicher, E. M., Robinson, M. S., Gorin, M. B., and Swank, R. T. (1999). The β3A subunit gene (*Ap3b1*) of the Ap-3 adaptor complex is altered in the mouse hypopigmentation mutant pearl, a model for Hermansky-Pudlak syndrome and night blindness. *Hum. Mol. Genet.*, 8(2):323–330.

[24] Fitzgerald, T. W., Gerety, S. S., Jones, W. D., van Kogelenberg, M., King, D. A., McRae, J., Morley, K. I., Parthiban, V., Al-Turki, S., Ambridge, K., Barrett, D. M., Bayzetinova, T., Clayton, S., Coomber, E. L., Gribble, S., Jones, P., Krishnappa, N., Mason, L. E., Middleton, A., Miller, R., Prigmore, E., Rajan, D., Sifrim, A., Tivey, A. R., Ahmed, M., Akawi, N., Andrews, R., Anjum, U., Archer, H., Armstrong, R., Balasubramanian, M., Banerjee, R., Baralle, D., Batstone, P., Baty, D., Bennett, C., Berg, J., Bernhard, B., Bevan, A. P., Blair, E., Blyth, M., Bohanna, D., Bourdon, L., Bourn, D., Brady, A., Bragin, E., Brewer, C., Brueton, L., Brunstrom, K., Bumpstead, S. J., Bunyan, D. J., Burn, J., Burton, J., Canham, N., Castle, B., Chandler, K., Clasper, S., Clayton-Smith, J., Cole, T., Collins, A., Collinson, M. N., Connell, F., Cooper, N., Cox, H., Cresswell, L., Cross, G., Crow, Y., D'Alessandro, M., Dabir, T., Davidson, R., Davies, S., Dean, J., Deshpande, C., Devlin, G., Dixit, A., Dominiczak, A., Donnelly, C., Donnelly, D., Douglas, A., Duncan, A., Eason, J., Edkins, S., Ellard, S., Ellis, P., Elmslie, F., Evans, K., Everest, S., Fendick, T., Fisher, R., Flinter, F., Foulds, N., Fryer, A., Fu, B., Gardiner, C., Gaunt, L., Ghali, N., Gibbons, R., Gomes Pereira, S. L., Goodship, J., Goudie, D., Gray, E., Greene, P., Greenhalgh, L., Harrison, L., Hawkins, R., Hellens, S., Henderson, A., Hobson, E., Holden, S., Holder, S., Hollingsworth, G., Homfray, T., Humphreys, M., Hurst, J., Ingram, S., Irving, M., Jarvis, J., Jenkins, L., Johnson, D., Jones, D., Jones, E., Josifova, D., Joss, S., Kaemba, B., Kazembe, S., Kerr, B., Kini, U., Kinning, E., Kirby, G., Kirk, C., Kivuva, E., Kraus, A., Kumar, D., Lachlan, K., Lam, W., Lampe, A., Langman, C., Lees, M., Lim, D., Lowther, G., Lynch, S. A., Magee, A., Maher, E., Mansour, S., Marks, K., Martin, K., Maye, U., McCann, E., McConnell, V., McEntagart, M., McGowan, R., McKay, K., McKee, S., McMullan, D. J., McNerlan, S., Mehta, S., Metcalfe, K., Miles, E., Mohammed, S., Montgomery, T., Moore, D., Morgan, S., Morris, A., Morton, J., Mugalaasi, H., Murday, V., Nevitt, L., Newbury-Ecob, R., Norman, A., O'Shea, R., Ogilvie, C., Park, S., Parker, M. J., Patel, C., Paterson, J., Payne, S., Phipps, J., Pilz, D. T., Porteous, D., Pratt, N., Prescott, K., Price, S., Pridham, A., Procter, A., Purnell, H., Ragge, N., Rankin, J., Raymond, L., Rice, D., Robert, L., Roberts, E., Roberts, G., Roberts, J., Roberts, P., Ross, A., Rosser, E., Saggar, A., Samant, S., Sandford, R., Sarkar, A., Schweiger, S., Scott, C., Scott, R., Selby, A., Seller, A., Sequeira, C., Shannon, N., Sharif, S., Shaw-Smith, C., Shearing, E., Shears, D., Simonic, I., Simpkin, D., Singzon, R., Skitt, Z., Smith, A., Smith, B., Smith, K., Smithson, S., Sneddon, L., Splitt, M., Squires, M., Stewart, F., Stewart, H., Suri, M., Sutton, V., Swaminathan, G. J., Sweeney, E., Tatton-Brown, K., Taylor, C., Taylor, R., Tein, M., Temple, I. K., Thomson, J., Tolmie, J., Torokwa, A., Treacy, B., Turner, C., Turnpenny, P., Tysoe, C., Vandersteen, A., Vasudevan, P., Vogt, J., Wakeling, E., Walker, D., Waters, J., Weber, A., Wellesley, D., Whiteford, M., Widaa, S., Wilcox, S., Williams, D., Williams, N., Woods, G., Wragg, C., Wright, M., Yang, F., Yau, M., Carter, N. P., Parker, M., Firth, H. V., FitzPatrick, D. R., Wright, C. F., Barrett, J. C., and Hurles, M. E. (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228.

[25] Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260.

[26] Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Van Duijn, C. M., Swertz, M., Wijmenga, C., Van Ommen, G., et al. (2015). Genome-wide patterns and properties of *de novo* mutations in humans. *Nature genetics*, 47(7):822–826.

[27] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315:972–977.

[28] Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723.

[29] Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

[30] Friel, N. and Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308.

[31] Fröhlich, H., Speer, N., Poustka, A., and Beißbarth, T. (2007). GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC bioinformatics*, 8(1):1.

[32] Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, 30(11):1203–1233.

[33] GEL (2015). The Genomics England Limited 100,000 Genomes Project. http://www.genomicsengland.co.uk (2015).

[34] Gene Ontology Consortium (2015). Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056.

[35] Gentleman, R., Whalen, E., Huber, W., and Falcon, S. (2016). *graph: A package to handle graph data structures*. R package version 1.51.0.

[36] Greene, D. (2016). *gsEasy: Gene Set Enrichment Analysis in R*. R package version 1.2.

[37] Greene, D., NIHR BioResource–Rare Diseases Consortium, Richardson, S., and Turro, E. (2016). Phenotype similarity regression for identifying the genetic determinants of rare diseases. *The American Journal of Human Genetics*, 98(3):490–499.

[38] Greene, D., NIHR BioResource–Rare Diseases Consortium, Richardson, S., and Turro, E. (2017a). A fast association test for identifying pathogenic variants involved in rare diseases. *The American Journal of Human Genetics*, 101(1):104–114.

[39] Greene, D., Richardson, S., and Turro, E. (2017b). ontologyx: a suite of r packages for working with ontological data. *Bioinformatics*, 33(7):1104–1106.

[40] Hansen, K. D., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., and Sarkar, D. (2016). *Rgraphviz: Provides plotting capabilities for R graph objects*. R package version 2.17.0.

[41] Heremans, J., Turro, E., Thys, C., Wouters, C., Van Geet, C., consortium, B.-B., Meyts, I., and Freson, K. (2017). RNA sequencing of Roifman syndrome megakaryocytes reveals a role for a small nuclear RNA in platelet and granule biology. *Under submission*.

[42] Institute, N. H. G. R. (2017). NIH webpage describing importance of mouse models. https://www.genome.gov/10001345/.

[43] Ionita-Laza, I., Capanu, M., De Rubeis, S., McCallum, K., and Buxbaum, J. D. (2014). Identification of rare causal variants in sequence-based studies: methods and applications to *VPS13B*, a gene involved in cohen syndrome and autism. *PLoS Genet*, 10(12):e1004729.

[44] Javed, A., Agrawal, S., and Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods*, 11(9):935–937.

[45] 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.

[46] Köhler, S., Doelken, S. C., Ruef, B. J., Bauer, S., Washington, N., Westerfield, M., Gkoutos, G., Schofield, P., Smedley, D., Lewis, S. E., Robinson, P. N., and Mungall, C. J. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, 2:30.

[47] Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464.

[48] Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., et al. (2016). The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, page gkw1039.

[49] Kunishima, S., Okuno, Y., Yoshida, K., Shiraishi, Y., Sanada, M., Muramatsu, H., Chiba, K., Tanaka, H., Miyazaki, K., Sakai, M., et al. (2013). *ACTN1* mutations cause congenital macrothrombocytopenia. *The American Journal of Human Genetics*, 92(3):431–438.

[50] Lazner, F., Gowen, M., Pavasovic, D., and Kola, I. (1999). Osteopetrosis and osteoporosis: two sides of the same coin. *Human molecular genetics*, 8(10):1839–1846.

[51] 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.

[52] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Team, E. L. P., Christiani, D. C., Wurfel, M. M., Lin, X., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237.

[53] Lee, S., with contributions from Larisa Miropolsky, and Wu, M. (2016). *SKAT: SNP-Set (Sequence) Kernel Association Test*. R package version 1.2.1.

[54] Lee, W.-N., Shah, N., Sundlass, K., and Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. In *AMIA annual symposium proceedings*, volume 2008, page 384. American Medical Informatics Association.

[55] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

[56] Lentaigne, C., Freson, K., Laffan, M. A., Turro, E., and Ouwehand, W. H. (2016). Inherited platelet disorders: toward DNA-based diagnosis. *Blood*, 127(23):2814–2823.

[57] Li, W., Zhang, Q., Oiso, N., Novak, E. K., Gautam, R., O'Brien, E. P., Tinsley, C. L., Blake, D. J., Spritz, R. A., Copeland, N. G., Jenkins, N. A., Amato, D., Roe, B. A., Starcevic, M., Dell'Angelica, E. C., Elliott, R. W., Mishra, V., Kingsmore, S. F., Paylor, R. E., and Swank, R. T. (2003). Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1). *Nat. Genet.*, 35(1):84–89.

[58] Liang, F. and Xiong, M. (2013). Bayesian Detection of Causal Rare Variants under Posterior Consistency. *PloS one*, 8(7):e69633.

[59] Lin, D. et al. (1998). An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304. Citeseer.

[60] Lin, W.-Y. (2014). Adaptive combination of p-values for family-based association testing with sequence data. *PloS one*, 9(12):e115971.

[61] Lin, W.-Y. (2016). Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study. *Scientific reports*, 6.

[62] Livingstone, F. B. (1985). *Frequencies of hemoglobin variants: thalassemia, the glucose-6-phosphate dehydrogenase deficiency,* G6PD *variants, and ovalocytosis in human populations*. Oxford University Press, USA.

[63] Logsdon, B. A., Dai, J. Y., Auer, P. L., Johnsen, J. M., Ganesh, S. K., Smith, N. L., Wilson, J. G., Tracy, R. P., Lange, L. A., Jiao, S., et al. (2014). A variational Bayes discrete mixture test for rare variant association. *Genetic epidemiology*, 38(1):21–30.

[64] Lynch, E. D., Lee, M. K., Morrow, J. E., Welcsh, P. L., Leon, P. E., and King, M. C. (1997). Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the *Drosophila* gene *diaphanous*. *Science*, 278(5341):1315–1318.

[65] MacArthur, D., Manolio, T., Dimmock, D., Rehm, H., Shendure, J., Abecasis, G., Adams, D., Altman, R., Antonarakis, S., Ashley, E., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.

[66] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

[67] Marx, V. (2015). The DNA of a nation. *Nature*, 524(7566):503–505.

[68] 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.

[69] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, 17(1):122.

[70] Merico, D., Roifman, M., Braunschweig, U., Yuen, R. K., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., et al. (2015). Compound heterozygous mutations in the noncoding *RNU4ATAC* cause Roifman Syndrome by disrupting minor intron splicing. *Nature communications*, 6.

[71] Morgan, N. V., Pasha, S., Johnson, C. A., Ainsworth, J. R., Eady, R. A., Dawood, B., McKeown, C., Trembath, R. C., Wilde, J., Watson, S. P., and Maher, E. R. (2006). A germline mutation in *BLOC1S3*/Reduced Pigmentation causes a novel variant of Hermansky-Pudlak syndrome (HPS8). *Am. J. Hum. Genet.*, 78(1):160–166.

[72] Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56.

[73] Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722.

[74] Murayama, S., Akiyama, M., Namba, H., Wada, Y., Ida, H., and Kunishima, S. (2013). Familial cases with *MYH9* disorders caused by *MYH9* S96L mutation. *Pediatr. Int.*, 55(1):102–104.

[75] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., et al. (2009). Targeted capture and massively parallel sequencing of twelve human exomes. *Nature*, 461(7261):272.

[76] Noris, P., Perrotta, S., Seri, M., Pecci, A., Gnan, C., Loffredo, G., Pujol-Moix, N., Zecca, M., Scognamiglio, F., De Rocco, D., et al. (2011). Mutations in *ANKRD26* are responsible for a frequent form of inherited thrombocytopenia: analysis of 78 patients from 21 families. *Blood*, 117(24):6673–6680.

[77] OMIM (1985). Online Mendelian Inheritance in Man. https://omim.org/.

[78] O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, R., Coin, L. J. M., O'Reilly, P. F., and Jarvelin, M.-R. (2012). MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS ONE*, 7(5):e34861.

[79] Overton, J. A., Dietze, H., Essaid, S., Osumi-Sutherland, D., and Mungall, C. J. (2015). ROBOT: A command-line tool for ontology development. In *Lisbon, Portuga: 5th International Conference on Biomedical Ontology*.

[80] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS comput biol*, 5(7):e1000443.

[81] Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O., den Dunnen, J. T., Firth, H. V., Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., Holm, I. A., Huang, L., Hurles, M. E., Hutton, B., Krier, J. B., Misyura, A., Mungall, C. J., Paschall, J., Paten, B., Robinson, P. N., Schiettecatte, F., Sobreira, N. L., Swaminathan, G. J., Taschner, P. E., Terry, S. F., Washington, N. L., Zuchner, S., Boycott, K. M., and Rehm, H. L. (2015). The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum. Mutat.*, 36(10):915–921.

[82] Pippucci, T., Savoia, A., Perrotta, S., Pujol-Moix, N., Noris, P., Castegnaro, G., Pecci, A., Gnan, C., Punzo, F., Marconi, C., et al. (2011). Mutations in the 5' UTR of *ANKRD26*, the ankirin repeat domain 26 gene, cause an autosomal-dominant form of inherited thrombocytopenia, THC2. *The American Journal of Human Genetics*, 88(1):115–120.

[83] Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754.

[84] Pontikos, N., Yu, J., Moghul, I., Withington, L., Blanco-Kelly, F., Vulliamy, T., Wong, T. L., Murphy, C., Cipriani, V., Fiorentino, A., Arno, G., Greene, D., Jacobsen, J. O., Clark, T., Gregory, D. S., Nemeth, A., Halford, S., Inglehearn, C. F., Downes, S., Black, G. C., Webster, A. R., Hardcastle, A. J., UKIRDC, and Plagnol, V. (2017). Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics (Oxford, England)*.

[85] Punzo, F., Mientjes, E., Rohe, C., Scianguetta, S., Amendola, G., Oostra, B., Bertoli-Avella, A., and Perrotta, S. (2010). A mutation in the acyl-coenzyme A binding domain-containing protein 5 gene (*ACBD5*) identified in autosomal dominant thrombocytopenia. *Journal of Thrombosis and Haemostasis*, 8(9):2085–2087.

[86] Quintana, M. and Conti, D. (2013). Integrative variable selection via bayesian model uncertainty. *Statistics in medicine*, 32(28):4938–4953.

[87] Quintana, M. A., Berstein, J. L., Thomas, D. C., and Conti, D. V. (2011). Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genetic epidemiology*, 35(7):638–649.

[88] Quiroga, T., Goycoolea, M., Panes, O., Aranda, E., Martínez, C., Belmont, S., Muñoz, B., Zúñiga, P., Pereira, J., and Mezzano, D. (2007). High prevalence of bleeders of unknown cause among patients with inherited mucocutaneous bleeding. A prospective study of 280 patients and 299 controls. *Haematologica*, 92(3):357–365.

[89] R Special Interest Group on Databases (R-SIG-DB), Wickham, H., and Müller, K. (2016). *DBI: R Database Interface*. R package version 0.5.

[90] Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., et al. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, page btt314.

[91] Resnik, P. et al. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130.

[92] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

[93] Robinson, P. N., Köhler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., and Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, 24(2):340–8.

[94] Roifman, C. (1997). Immunological aspects of a novel immunodeficiency syndrome that includes antibody deficiency with normal immunoglobulins, spondyloepiphyseal dysplasia, growth and developmental delay, and retinal dystrophy. *Can. J. Allergy Clin. Immunol*, 2:94–98.

[95] Sanchez, G. (2013). *AssotesteR: Statistical Tests for Genetic Association Studies*. R package version 0.1-10.

[96] Savoia, A., Pastore, A., De Rocco, D., Civaschi, E., Di Stazio, M., Bottega, R., Melazzini, F., Bozzi, V., Pecci, A., Magrin, S., et al. (2011). Clinical and genetic aspects of Bernard-Soulier syndrome: searching for genotype/phenotype correlations. *Haematologica*, 96(3):417–423.

[97] Seri, M., Cusano, R., Gangarossa, S., Caridi, G., Bordo, D., Lo Nigro, C., Ghiggeri, G. M., Ravazzolo, R., Savino, M., Del Vecchio, M., D'Apolito, M., Iolascon, A., Zelante, L. L., Savoia, A., Balduini, C. L., Noris, P., Magrini, U., Belletti, S., Heath, K. E., Babcock, M., Glucksman, M. J., Aliprandis, E., Bizzaro, N., Desnick, R. J., and Martignetti, J. A. (2000). Mutations in *MYH9* result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. The May-Hegglin/Fechtner Syndrome Consortium. *Nat. Genet.*, 26(1):103–105.

[98] 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.

[99] Simeoni, I., Stephens, J. C., Hu, F., Deevi, S. V., Megy, K., Bariana, T. K., Lentaigne, C., Schulman, S., Sivapalaratnam, S., Vries, M. J., et al. (2016). A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders. *Blood*, 127(23):2791–2803.

[100] Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., Durtschi, J., Eilbeck, K., Reese, M. G., Jorde, L. B., Huff, C. D., and Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification

of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.*, 94(4):599–610.

[101] Sivapalaratnam, S., Westbury, S. K., Stephens, J. C., Greene, D., Downes, K., Kelly, A. M., Lentaigne, C., Astle, W. J., Huizinga, E. G., Nurden, P., et al. (2017). Rare variants in *GP1BB* are responsible for autosomal dominant macrothrombocytopenia. *Blood*, 129(4):520–524.

[102] Smedley, D., Jacobsen, J. O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*, 10(12):2004–2015.

[103] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.

[104] Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*, 1(3):390–399.

[105] Soriano, P., Montgomery, C., Geske, R., and Bradley, A. (1991). Targeted disruption of the c-src proto-oncogene leads to osteopetrosis in mice. *Cell*, 64(4):693–702.

[106] Spielmann, M. and Mundlos, S. (2016). Looking beyond the genes: the role of non-coding variants in human disease. *Human molecular genetics*, 25(R2):R157–R165.

[107] Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, 8(7):e65245.

[108] Stitziel, N. O., Kiezun, A., and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome biology*, 12(9):227.

[109] Stritt, S., Nurden, P., Turro, E., Greene, D., Jansen, S. B., Westbury, S. K., Petersen, R., Astle, W. J., Marlin, S., Bariana, T. K., Kostadima, M., Lentaigne, C., Maiwald, S., Papadia, S., Kelly, A. M., Stephens, J. C., Penkett, C. J., Ashford, S., Tuna, S., Austin, S., Bakchoul, T., Collins, P., Favier, R., Lambert, M. P., Mathias, M., Millar, C. M., Mapeta, R., Perry, D. J., Schulman, S., Simeoni, I., Thys, C., Gomez, K., Erber, W. N., Stirrups, K., Rendon, A., Bradley, J. R., van Geet, C., Raymond, F. L., Laffan, M. A., Nurden, A. T., Nieswandt, B., Richardson, S., Freson, K., Ouwehand, W. H., and Mumford, A. D. (2016). A gain-of-function variant in *DIAPH1* causes dominant macrothrombocytopenia and hearing loss. *Blood*, 127(23):2903–2914.

[110] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[111] Suzuki, T., Li, W., Zhang, Q., Karim, A., Novak, E. K., Sviderskaya, E. V., Hill, S. P., Bennett, D. C., Levin, A. V., Nieuwenhuis, H. K., Fong, C. T., Castellan, C., Miterski, B., Swank, R. T., and Spritz, R. A. (2002). Hermansky-Pudlak syndrome is caused by mutations in *HPS4*, the human homolog of the mouse light-ear gene. *Nat. Genet.*, 30(3):321–324.

[112] Therneau, T. M. and Sinnwell, J. (2015). *kinship2: Pedigree Functions*. R package version 1.6.4.

[113] Tuijnenburg, P., Allen, H. L., Burns, S. O., Greene, D., Jansen, M. H., Staples, E., Stephens, J., Carss, K. J., Biasci, D., Baxendale, H., Thomas, M., Chandra, A., Kiani-Alikhan, S., Longhurst, H. J., Oksenhendler, E., Simeoni, I., Bree, G. J. d., Tool, A. T., Leeuwen, E. M. v., Ebberink, E. H., Meijer, A. B., Tuna, S., Whitehorn, D., Brown, M., NIHR BioResource–Rare Diseases Consortium, Turro, E., Thrasher, A. J., Smith, K. G. C., Thaventhiran, J. E., and Kuijpers, T. W. ((under review)). Whole-Genome Sequencing identifies *NFKB1* haploinsufficiency as the commonest monogenic cause of Common Variable Immunodeficiency. *Journal of Allergy and Clinical Immunology*.

[114] Turro, E., Greene, D., Wijgaerts, A., Thys, C., Lentaigne, C., Bariana, T. K., Westbury, S. K., Kelly, A. M., Selleslag, D., Stephens, J. C., Papadia, S., Simeoni, I., Penkett, C. J., Ashford, S., Attwood, A., Austin, S., Bakchoul, T., Collins, P., Deevi, S. V. V., Favier, R., Kostadima, M., Lambert, M. P., Mathias, M., Millar, C. M., Peerlinck, K., Perry, D. J., Schulman, S., Whitehorn, D., Wittevrongel, C., De Maeyer, M., Rendon, A., Gomez, K., Erber, W. N., Mumford, A. D., Nurden, P., Stirrups, K., Bradley, J. R., Lucy Raymond, F., Laffan, M. A., Van Geet, C., Richardson, S., Freson, K., and Ouwehand, W. H. (2016). A dominant gain-of-function mutation in universal tyrosine kinase *SRC* causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. *Science Translational Medicine*, 8(328):328ra30.

[115] UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

[116] Vihinen, M. (2014). Variation Ontology for annotation of variation effects and mechanisms. *Genome research*, 24(2):356–364.

[117] Website, H. (2008). Human Phenotype Ontology website. http://human-phenotype-ontology.github.io/downloads.html.

[118] Westbury, S., Canault, M., Greene, D., Bermejo, E., Hanlon, K., Lambert, M., Millar, C., Nurden, P., Obaji, S., Revel-Vilk, S., Van Geet, C., Downes, K., Papadia, S., Tuna, S., Watt, C., NIHR BioResource–Rare Diseases Consortium, Freson, F., Laffan, M., Ouwehand, W., Alessi, M.-C., Turro, E., and Mumford, A. (2017). Expanded repertoire of *RASGRP2* variants responsible for platelet dysfunction and severe bleeding. *Blood*.

[119] Westbury, S. K., Turro, E., Greene, D., Lentaigne, C., Kelly, A. M., Bariana, T. K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., et al. (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome medicine*, 7(1):36.

[120] Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., King, D. A., Ambridge, K., Barrett, D. M., Bayzetinova, T., et al. (2015).

Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *The Lancet*, 385(9975):1305–1314.

[121] Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods*, 12(9):841–843.

[122] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

[123] Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609.

[124] Zemojtel, T., Kohler, S., Mackenroth, L., Jager, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., Oien, N. C., Schweiger, M. R., Kruger, U., Frommer, G., Fischer, B., Kornak, U., Flottmann, R., Ardeshirdavani, A., Moreau, Y., Lewis, S. E., Haendel, M., Smedley, D., Horn, D., Mundlos, S., and Robinson, P. N. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*, 6(252):252ra123.

[125] Zhang, Q., Zhao, B., Li, W., Oiso, N., Novak, E. K., Rusiniak, M. E., Gautam, R., Chintala, S., O'Brien, E. P., Zhang, Y., Roe, B. A., Elliott, R. W., Eicher, E. M., Liang, P., Kratz, C., Legius, E., Spritz, R. A., O'Sullivan, T. N., Copeland, N. G., Jenkins, N. A., and Swank, R. T. (2003). Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky-Pudlak syndrome types 5 and 6. *Nat. Genet.*, 33(2):145–153.

Appendix A

ontologyX comparisons and examples

This appendix contains the code used to generate the comparisons and examples given for ontologyIndex, ontologyPlot and ontologySimilarity in Chapter 2.
The following versions of software packages were used to generate the results presented in this thesis:
ontologyIndex 2.2, ontologyPlot 1.4, ontologySimilarity 2.2, GOSim 1.11, GOSemSim 1.99.4 and ontoCAT 1.26.0.

Performance comparison with ontoCAT

```
library(ontoCAT)
library(ontologyIndex)

set.seed(1)

if (!exists("hpo_file_path")) stop("Download␣HPO␣from:␣http
   ://purl.obolibrary.org/obo/hp.obo␣and␣set␣hpo_file_path␣
   variable␣to␣its␣path")
#original computation based on HPO data-version: releases
   /2016-04-01

hpo_ontologyIndex <- get_OBO(hpo_file_path)
hpo_ontoCAT <- getOntology(hpo_file_path)

ancestors <- expression(
```

```
        ontoCAT=ontoCAT::getAllTermParentsById(object=
            hpo_ontoCAT, id=term),
        ontologyIndex=ontologyIndex::get_ancestors(
            hpo_ontologyIndex, terms=term)
)

descendants <- expression(
        ontoCAT=ontoCAT::getAllTermChildrenById(object=
            hpo_ontoCAT, id=term),
        ontologyIndex=ontologyIndex::get_descendants(
            hpo_ontologyIndex, roots=term)
)

n_terms <- length(hpo_ontologyIndex$id)

desc_times <- lapply(descendants, function(x) { system.time
    (sapply(hpo_ontologyIndex$id, function(term) eval(x)))
    })
anc_times <- lapply(ancestors, function(x) { system.time(
    sapply(hpo_ontologyIndex$id, function(term) eval(x))) })

#print results table
print(data.frame(check.names=FALSE, stringsAsFactors=FALSE,
     "Descendants␣(ms)"=sapply(desc_times, function(x) 1000*
    as.numeric(x["elapsed"])/n_terms), "Ancestors␣(ms)"=
    sapply(anc_times, function(x) 1000*as.numeric(x["elapsed
    "])/n_terms)))
```

Using ontologyPlot to visualise *QPCTL* and *CRNN* GO annotation

```
library(ontologyIndex)
library(ontologyPlot)
library(ontologySimilarity)
data(go)
data(gene_GO_terms)
data(GO_IC)

cc <- go$id[go$name == "cellular_component"]
cc_anno <- lapply(gene_GO_terms[c("QPCTL","CRNN")],
        function(x) get_ancestors(go,
           intersection_with_descendants(go, roots=cc, x)))

all_terms <- unique(unlist(use.names=FALSE, cc_anno))
QPCTL <- cc_anno$QPCTL
CRNN <- cc_anno$CRNN

terms <- remove_uninformative_terms(go, list(QPCTL, CRNN))

pdf(file="GO-plot.pdf", width=10.5, height=4)
par(mfrow=c(1, 2))
onto_plot(go, terms=all_terms,
        width=calibrate_sizes(GO_IC[all_terms], low=1, high
           =3),
        fontsize=50,
        fillcolor=rgb(0, all_terms %in% QPCTL, all_terms %
           in% CRNN, 0.5))

onto_plot(go, terms=terms,
        width=calibrate_sizes(GO_IC[terms], low=1, high=3),
        fillcolor=rgb(0, terms %in% QPCTL, terms %in% CRNN,
            0.5))
dev.off()
```

Comparing performance of semantic similarity computation

```
library(ontologyIndex)
library(ontologySimilarity)
library(GOSemSim)
library(GOSim)
set.seed(1)

initialise <- expression(
        ontologySimilarity={ data(go); IC_os <-
           descendants_IC(go) },
        GOSim=GOSim:::initialize(),
        GOSemSim=hsGO <- godata("org.Hs.eg.db", keytype = "
           SYMBOL", ont="MF", computeIC=TRUE)
)

# GOSim uses Entrez Gene IDs which we need to map to HGNC
   symbols for comparison with GOSemSim and
   ontologySimilarity. We obtained the mapping using the
   table downloaded from the HUGO Gene Nomenclature
   Committee website, using the custom download page (http
   ://www.genenames.org/cgi-bin/download) and selecting the
    "Approved Symbol" and "Entrez Gene ID" columns, and
   saving it as "symbol.txt" in the working directory.
stopifnot(file.exists("symbol.txt"))
symbol_map <- read.table(file="symbol.txt", header=TRUE,
   sep="\t", comment.char="", stringsAsFactors=FALSE)

env <- environment()

ini_times <- lapply(initialise, function(x) { cat(
   as.character(x), "\n"); system.time(eval(x, envir=env))
   })

IC <- slot(hsGO, "IC")
```

```
gene_anno <- with(data=slot(hsGO, "geneAnno"), split(f=
   SYMBOL, x=GO))

n_genes <- 100
n_terms <- 1000

genes <- names(gene_anno)[1:n_genes]

terms <- names(which(IC < Inf & IC > 0))[1:n_terms]

#create expressions for calculation of term/gene similarity
    matrices for comparison
#note that different values are obtained in the matrices as
    different sources of information content are used.

term_similarity_matrix <- expression(
        GOSim=getTermSim(terms, method="Lin", verbose=FALSE
           ),
        GOSemSim=mgoSim(terms, terms, semData=hsGO, measure
           ="Lin", combine=NULL),
        #ontologySimilarity uses Lin's expression for term
           similarity by default
        ontologySimilarity=get_term_sim_mat(go,
           information_content=IC_os[get_ancestors(go,
           terms)], row_terms=terms, col_terms=terms),
        "ontologySimilarity (indexed)"=NA
)

gene_similarity_matrix <- expression(
        GOSim=getGeneSim(as.character(na.omit(symbol_map$
           Entrez.Gene.ID[match(genes, symbol_map$
           Approved.Symbol)])),similarity="funSimMax",
           similarityTerm="Lin",verbose=FALSE),
        GOSemSim=mgeneSim(genes, semData=hsGO, measure="Lin
           ", combine="BMA", verbose=FALSE),
```

```
        ontologySimilarity=get_sim_grid(ontology=go,
            information_content=IC_os, term_sets=gene_anno[
            genes]),
        "ontologySimilarity (indexed)"=local({ term_sim_mat
            <- get_term_sim_mat(go, information_content=
            IC_os[get_ancestors(go, unique(unlist(use.names=
            FALSE, gene_anno[genes])))]); get_sim_grid(
            term_sim_mat=term_sim_mat, term_sets=gene_anno[
            genes]) })
)

calc_times <- lapply(list(term_sim=term_similarity_matrix,
    gene_sim=gene_similarity_matrix), function(calc_type)
    lapply(calc_type, function(x) if (is.na(x)) NA else
    system.time(eval(x))))

#print results table
print(data.frame(check.names=FALSE, stringsAsFactors=FALSE,
     "Term sim (s)"=sapply(calc_times$term_sim, function(x)
    as.numeric(x["elapsed"])), "Gene sim (s)"=sapply(
    calc_times$gene_sim, function(x) as.numeric(x["elapsed"
    ]))))
```

Appendix B

SimReg manual

Introduction

SimReg has a function `sim_reg` for performing 'Bayesian Similarity Regression'. Its output consists of a posterior probability of association between ontological term sets and a binary response variable, and, conditional on the association, a 'characteristic ontological profile' such that ontological similarity to the profile increases the probability of the binary variable taking the value `TRUE`. The procedure has been used in the context of linking an ontologically encoded phenotype (as HPO terms) to a binary genotype (indicating the presence or absence of a rare variant within a given gene), and this guide will adopt the same type of application.

The function accepts arguments that include the `logical` response variable `y`, the ontologically encoded predictor variable `x`, and additional arguments for tuning the compromise between execution speed and precision for the procedure. It returns an object of class `sim_reg_output`, which contains the results of the inference. The probability of association, i.e. the probability that model selection indicator `gamma` = 1, can be extracted from the output object using the `prob_association` function. The posterior distribution of the characteristic ontological profile `phi` may be of interest, which can be extracted using the function `get_term_marginals`.

In this example, we will apply `sim_reg` to randomly generated data, including ontological term sets comprising terms from the HPO. Random HPO term sets will be generated in two ways: independently of and conditional upon a logical vector, `y` — which is set to 10 `TRUE`s followed by 90 `FALSE`s — and stored respectively in the variables `x_independent` and `x_dependent`. `sim_reg` will then be applied to each of the randomly generated sets of term sets and `y`, and the results will be compared.

Initially, an HPO template for the typical phenotype of a hypothetical disease is created. The template is stored in the variable `template` and set to include HPO terms `HP:0005537`,

HP:0000729 and HP:0001873, corresponding to phenotype abnormalities 'Decreased mean platelet volume', 'Autistic behavior' and 'Thrombocytopenia' respectively. A set of terms which includes those terms in template is then generated at random and stored in the variable terms, to be used as a pool from which to draw random term sets for use in the analysis. To generate the independent HPO term sets, x_independent, 100 sets of five terms are sampled from terms, and each is mapped to its corresponding minimal set. To generate the HPO term sets which depend on y, x_dependent, 100 sets of terms were generated as for x_independent, but the sets of terms corresponding to components of y taking the value TRUE were augmented with two terms randomly selected from template.

```
library(ontologyIndex)
library(SimReg)
data(hpo)
set.seed(1)

template <- c("HP:0005537", "HP:0000729", "HP:0001873")
terms <- get_ancestors(hpo, c(template, sample(hpo$id, size=50)))

y <- c(rep(TRUE, 10), rep(FALSE, 90))
x_independent <- replicate(simplify=FALSE, n=100,
    expr=minimal_set(hpo, sample(terms, size=5)))

no_assoc <- sim_reg(ontology=hpo, x=x_independent, y=y)
prob_association(no_assoc)

## [1] 0.0005632741

x_dependent <- lapply(y, function(y_i) minimal_set(hpo, c(
    sample(terms, size=5), if (y_i) sample(template, size=2))))

assoc <- sim_reg(ontology=hpo, x=x_dependent, y=y)
prob_association(assoc)

## [1] 0.9999682
```

We note that we infer a higher probability of association inferred when applying the inference to the set of terms generated dependently on y. Note that by default, the probability of an association has a prior of 0.05, but this can be set by passing a gamma_prior_prob argument. We can also visualise the estimated characteristic ontological profile, using the function

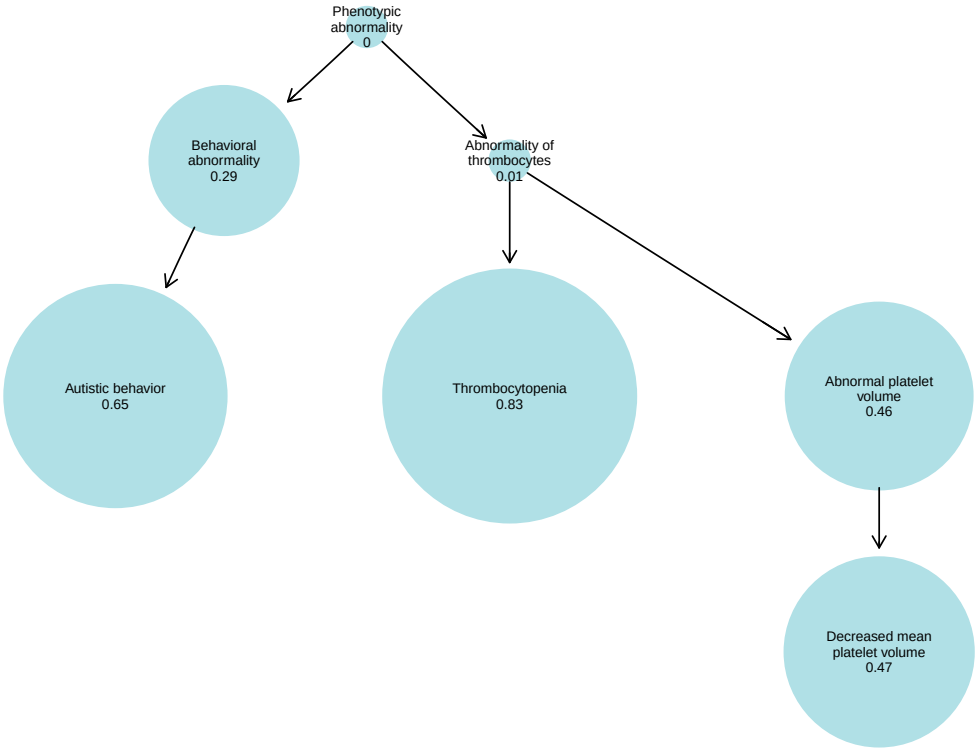


Fig. B.1 The output of using the function `plot_term_marginals` to plot the marginal probabilities of inclusion of terms in ϕ.

`plot_term_marginals`, and note that the inferred characteristic phenotype corresponds well to `template` (see Figure B.1).

```
plot_term_marginals(hpo, get_term_marginals(assoc), max_terms=8)
```

Including prior phenotypic information

This section demonstrates how to set the prior on the variable `phi`. This is done in order to incorporate prior information about likely disease phenotypes, for example, based on the phenotype of a mouse model for an orthologous gene knock-out. The information is supplied to the function `sim_reg` as a parameter called `term_weights` and should be a numeric vector of relative weights for terms included in the sample space of `phi` (by default, the sample space of `phi` includes all terms present amongst the terms in `x` and their ancestors).

```
library(ontologyIndex)
library(ontologySimilarity)
```

```
library(SimReg)
data(hpo)
set.seed(1)
terms <- get_ancestors(hpo, c(hpo$id[match(
    c("Abnormality of thrombocytes","Hearing abnormality"),
    hpo$name)], sample(hpo$id, size=50)))
```

To illustrate the approach, we will construct a scenario where there is moderate evidence for an association between 'Hearing abnormality' and a binary summary of genetic data within a locus, and where there is also a prior expectation that the disease phenotype associated with pathogenic variants in the locus would involve hearing abnormality. We randomly generate data to reflect the moderate evidence for association, and then apply the inference with and without including the prior information and observe the effect on the posterior probability of association.

```
hearing_abnormality <- hpo$id[match("Hearing abnormality", hpo$
    name)]
genotypes <- c(rep(TRUE, 2), rep(FALSE, 98))

#give all subjects 5 random terms and add 'hearing abnormality'
#for those with y_i=TRUE
phenotypes <- lapply(genotypes, function(y_i) minimal_set(hpo, c(
if (y_i) hearing_abnormality else character(0), sample(terms, size
    =5))))
#note that there are three cases with the rare variant
#(i.e. having y[i] == TRUE) and all of them have the 'Hearing
    abnormality' HPO term.

sim_reg(ontology=hpo, x=phenotypes, y=genotypes)

## Probability of association:  0.47
## ------------------------------------------------------------
##                                                Name    p
##                                  Hearing abnormality 0.63
##                                Abnormality of the ear 0.37
##                                  Abnormality of digit 0.03
##                                  Abnormality of limbs 0.03
##                  Abnormality of limb bone morphology 0.03
##                        Abnormality of the upper limb 0.02
##                             Abnormality of limb bone 0.02
```

```
##                Abnormality of the phalanges of the toes 0.02
##                                     Abnormality of toe 0.02
##  Aplasia/hypoplasia involving bones of the extremities 0.02
## ------------------------------------------------------------
```

We construct the `term_weights` parameter to capture our knowledge about the gene from the model organism by setting weights for all terms to one, except those which involve hearing, which are assigned weights of 10. Note that one must set the `names` of the `term_weights` vector.

```
#we set the prior weight of all terms which have the word 'hearing
    '
#in to ten times that of terms which don't.
term_weights <- ifelse(grepl(x=hpo$name, ignore=TRUE,
    pattern="hearing"), 10, 1)
names(term_weights) <- hpo$id
```

Prior information about phenotype (hereafter referred to as the 'literature phenotype') may be available as a set of ontological terms (for example, it may be available as MPO terms from the Mouse Genome Informatics (MGI) website, http://www.informatics.jax.org/). If so, another method of setting the weights is to use a phenotypic similarity function to obtain a numeric vector of similarities for use as prior weights for inclusion of terms in `phi`. This may be more convenient, particularly when dealing with large numbers of genes. In the SimReg paper, the vector is set using the Resnik-based similarities of terms to the terms in the literature phenotype. In order to calculate the similarities based on Resnik's similarity measure, we must first compute the information content for the terms, equal to the negative log frequency. The frequencies can be calculated with respect to different collections of phenotypes. Here, we will calculate it with respect to the frequencies of terms within our collection, `phenotypes`, by calling the `get_term_info_content` function from the ontologyIndex package. Note, it could also be calculated with respect to the frequency of the term amongst the HPO annotation of OMIM diseases (available from the HPO website, http://human-phenotype-ontology.github.io). The function `get_term_set_to_term_sims` in the package ontologySimilarity can then be used to calculate the similarities between the terms in the sample space of `phi` and the literature phenotype. It calculates a matrix of similarities between the individual terms in the literature phenotype and terms in the sample space. In the following example, the hypothetical model organism phenotype will be set to abnormality of the thrombocytes and hearing abnormality.

```
thrombocytes <- hpo$id[match("Abnormality␣of␣thrombocytes", hpo$
    name)]
```

```
literature_phenotype <- c(hearing_abnormality, thrombocytes)
info <- get_term_info_content(hpo, phenotypes)

term_weights_resnik <- apply(get_term_set_to_term_sims(
    ontology=hpo, information_content=info, terms=names(info),
    term_sim_method="resnik",
    term_sets=list(literature_phenotype)), 2, mean)
```

This can then be passed to `sim_reg` through the `term_weights` parameter.

```
sim_reg(
    ontology=hpo,
    x=phenotypes,
    y=genotypes,
    term_weights=term_weights_resnik
)
```

```
## Probability of association:  0.8
## ------------------------------------------------------------
##                                                Name    p
##                                  Hearing abnormality 0.67
##                                Abnormality of the ear 0.33
##                                  Abnormality of digit 0.02
##                                  Abnormality of limbs 0.01
##                    Abnormality of limb bone morphology 0.01
##                Abnormality of the phalanges of the toes 0.01
##                                Abnormality of limb bone 0.01
##                           Abnormality of the upper limb 0.01
##                                       Abnormality of toe 0.01
##  Aplasia/hypoplasia involving bones of the extremities 0.01
## ------------------------------------------------------------
```

Note that including the `term_weights` parameter has increased the mean posterior value of `gamma`.

Appendix C

BeviMed manual

Introduction

BeviMed is a procedure for evaluating the evidence of association between allele configurations across rare variants, typically within a genomic locus, and a case/control label. It is capable of inferring the posterior probability of association, and conditional on association, the probability of each mode of inheritance and probability of involvement of each variant. It works by applying Bayesian model comparison between two models indexed by γ. Under the model labelled $\gamma = 0$, the probability of case status is independent of allele configuration at the given rare variant sites. Under the model labelled $\gamma = 1$, the probability of case status is linked to the configuration of alleles, and a latent partition of variants into *pathogenic* and *non-pathogenic* groups conditional on a mode of inheritance. It can also perform model selection on multiple alternative models, each of which includes a different subset of variants. This has the effect of inducing a prior correlation on the variant pathogenicities, which can boost power if only a particular (though unknown) class of variants is responsible for disease.

The aim of the package is to facilitate prioritisation of large numbers of sets of variants, typically drawn from different loci, by rapid inference of the posterior distributions of γ, mode of inheritance parameter m, and indicator of pathogenicity across variants, z. This guide describes the interface of the package in detail, relating it to the underlying statistical model and including some implementation details. See the 'BeviMed Introduction' vignette for a quick start guide. Unless otherwise stated, N refers to the number of individuals, k refers to the number of rare variants, m refers to the mode of inheritance (either m_{dom} or m_{rec}), and 'evidence' refers to the integrated likelihood of the data under a given model. The acronym 'MOI' will often be used to refer to mode of inheritance.

Functions and classes

BeviMed has functions for evaluating models $\gamma = 0$, $\gamma = 1$ with $m = m_{\text{dom}}$, and $\gamma = 1$ with $m = m_{\text{rec}}$ with respect to the data — a logical length N vector of case/control labels y, and an $N \times k$ integer matrix of allele counts G:

- `gamma0_evidence`, which computes the evidence for model $\gamma = 0$, $\mathbb{P}(y|\gamma = 0)$.
- `bevimed_m`, which samples from the posterior distribution of the model $\gamma = 1$ conditional on a given mode of inheritance. The output yields the evidence $\mathbb{P}(y|m, \gamma = 1, G)$ and probabilities of variant pathogenicity, $\mathbb{P}(z_j|m, \gamma = 1, y, G)$ for $j = 1, \ldots, k$.
- `bevimed`, which evaluates all three models in turn by calling `gamma0_evidence` and `bevimed_m` with respect to each mode of inheritance. By allowing prior probabilities and computing the evidence for each model, it allows the posterior quantities of interest to be computed using Bayes' theorem.
  - model indicator γ, $\mathbb{P}(\gamma = 1|y, G)$
  - mode of inheritance m given association, $\mathbb{P}(m|\gamma = 1, y, G)$.
- `bevimed_polytomous`, which evaluates model $\gamma = 0$ and an arbitrary number of association models using `bevimed_m`, corresponding to different subsets of variants within G and given modes of inheritance.

`bevimed` is simple to apply:

```
obj <- bevimed(y=y, G=G)
```

It returns an object of class BeviMed, which contains the whole output of the inference. A summary of the inference can be printed by evaluating the object in an interactive session:

```
> obj <- bevimed(y=y, G=G)

## -----------------------------------------------------
## Posterior probability of association:
##  0.038 [prior: 0.01]
## Posterior probability of dominance given association:
##  0.799 [prior: 0.5]
## -----------------------------------------------------
##      Model  Post Prior Cases Variants
```

```
##   dominant 0.799   0.5  4.56     2.44
##  recessive 0.201   0.5  1.13     1.61
##
## Post: posterior probability of model given association
## Prior: prior probability of model given association
## Cases: posterior expected number of cases explained
## Variants: posterior expected number of variants involved
## ---------------------------------------------------------
## Probabilities of pathogenicity for individual variants
##
##  Var   Probability pathogenic
##    1 [0.77 =============    ]
##    2 [0.60 ==========       ]
##    3 [0.68 ============     ]
##    4 [0.16 ===              ]
##    5 [0.39 =======          ]
## ---------------------------------------------------------
```

An object of class BeviMed is a list containing slots:

- `"parameters"`, a list of parameter values used to call the function.
- `"models"`, a list of `BeviMed_m` objects returned by the `bevimed_m` function, one for each association model - typically one per mode of inheritance (i.e. dominant and recessive). The `BeviMed_m` class is a list containing samples from the posterior distributions of model parameters conditional on a given mode of inheritance (see help page `?bevimed_m` for more details). As a list, the model specific results can be looked up by model using the `$` operator, e.g. `x$models$dominant`.

The function `bevimed_m` uses an MCMC algorithm to sample from the posterior distribution of the parameters in model $\gamma = 1$. Each individual has an associated 'minimum number of alleles at pathogenic variant sites' required to have a *pathogenic configuration of alleles*. This is determined by the `min_ac` argument (defaulting to 1), and can be set to reflect the desired mode of inheritance. For example, in dominant inheritance, at least one pathogenic allele would render an allele configuration pathogenic, and thus the `min_ac` argument could be set to 1. In X-linked recessive inheritance, at least 1 and 2 pathogenic alleles would be required for a pathogenic configuration respectively for males and females, and thus the `min_ac` argument could be given as a numeric vector of length N with 1s for males and 2s

for females. `bevimed` accepts a `ploidy` argument: an integer vector the same length as `y` which specifies the ploidy of each individual in the locus (defaulting to 2). Internally, it uses this argument to set `min_ac` automatically when it calls `bevimed_m` based on mode of inheritance.

Objects of class BeviMed typically consume a large amount of memory. Summarising objects of class BeviMed with `summary` — which retains important summary statistics as a list — may be useful when performing multiple applications. Specific summary statistics can be obtained by looking them up in these summary lists (see help page `?summary.BeviMed` for names used for each statistic).

`bevimed` passes arguments to `bevimed_m` through the '`...`' argument. However, sometimes it is preferable to pass different arguments to `bevimed_m` depending on mode of inheritance. `bevimed` therefore allows mode of inheritance specific arguments to be passed through `dominant_args` and `recessive_args`, which should be named lists of arguments then only used in the corresponding calls to `bevimed_m`. For example, it might be thought that a smaller proportion of variants would be linked to disease given a dominant mode of inheritance than would given recessive inheritance, in which case `dominant_args` could be used to pass a prior with a lower mean for the parameter `omega` to the dominant application of `bevimed_m`.

Similarly, `bevimed_polytomous` evaluates association models depending on variants corresponding to those in the given `variant_sets` argument: a list of `integer` vectors, one for each model, each indexing variants with respect to their column position in G. Arguments are passed to `bevimed_m` through the '`...`' argument, as with `bevimed`, but model specific arguments are passed using the parameter `model_specific_args`: a list of the same length as `variant_sets`. The mode of inheritance parameter for each association model can be specified using a character vector the same length as `variant_sets` containing elements "dominant" or "recessive" (defaults to `"dominant"`). The prior probability of association for each model can also be specified as a numeric vector of probabilities using the argument `prior_prob_association`.

In this example we use `bevimed_polytomous` to compare two models in the situation where the disease status depends on only the allele configuration at the first variant site from amongst 5 columns: one depending on just the first variant and one depending on all the variants.

```
> bevimed_polytomous(y=G[,1] > 0, G=G,
+     variant_sets=list(first=1, all=1:ncol(G)))
```

```
## ------------------------------------------------------
## Posterior probability of association:
##  0.132 [prior: 0.01]
## Posterior probability of dominance given association:
##  1 [prior: 1]
## ------------------------------------------------------
##  Model  Post Prior Cases Variants
##  first 0.721   0.5  2.91    0.969
##    all 0.279   0.5  2.82    0.941
##
## Post: posterior probability of model given association
## Prior: prior probability of model given association
## Cases: posterior expected number of cases explained
## Variants: posterior expected number of variants involved
## ------------------------------------------------------
## Probabilities of pathogenicity for individual variants
##
##  Var   Probability pathogenic
##    1 [0.96 ================ ]
##    2 [0.01                  ]
##    3 [0.00                  ]
##    4 [0.01                  ]
##    5 [0.01                  ]
## ------------------------------------------------------
```

Priors on model parameters

The user can control the prior distributions of the model parameters when applying the inference functions `bevimed`, `bevimed_polytomous`, `bevimed_m` and `gamma0_evidence` as listed below.

- The probability of association, $\mathbb{P}(\gamma = 1|y)$, with argument `prior_prob_association` in the `bevimed` function (defaults to 0.01).
- The probability of dominant inheritance given association, $\mathbb{P}(m = m_{\text{dom}})$, with the `prior_prob_dominant` in the `bevimed` function (defaults to 0.5).

- The hyper parameters of the beta prior for the probability τ_0 of observing the case label under model $\gamma = 0$. Values for the hyper parameters can be passed to the `bevimed` and `gamma0_evidence` functions as the `tau0_shape` argument (defaults to a vague parameterisation of $\alpha = \beta = 1$).

- The hyper parameters of the beta prior for τ and π, respectively the probabilities of observing the case label for individuals with non-pathogenic and pathogenic allele configurations under model $\gamma = 1$. The default for τ is the same as for τ_0, but the default for π has a mean close to 1, as typically for rare diseases the variants are high penetrance, i.e. have a high probability of causing the disease phenotype. Values for these hyper parameters can be passed as arguments `tau_shape` and `pi_shape` to the `bevimed` and `bevimed_m` functions.

- The prior on the indicators of variant pathogenicity, z. By default, all variants have a shared prior on their probability of pathogenicity, $z_j \sim \text{Bernoulli}(\omega)$ with $\omega \sim \text{beta}(\alpha = 2, \beta = 8)$. The hyper parameters for ω can be specified by the user using the parameter `omega_shape`. However the user can also control the prior on pathogenicity for individual variants. This is done using the `variant_weights` parameter, a numeric vector of length k labelled c in the model specification. The effect of the c values is given by the logistic equation:

$$\begin{aligned} z_j &\sim \text{Bernoulli}(p_j), \\ \text{logit}\, p_j &= \omega + \phi c_j, \\ \log \phi &\sim \text{N}(\mu_\phi, \sigma_\phi^2), \end{aligned}$$

  where ϕ is the scaling factor for c. By default, c is centralised on 0 so that ω is interpretable as the global rate of pathogenicity in the locus, and ϕ has a mean of 1, so c_j is interpretable as a shift in the log odds on the prior probability of variant j being pathogenic. Thus, one could for example of use the untransformed CADD Phred score for each variant as a weight. The raw values of c as given in the `variant_weights` arguments will be used if the parameter `standardise_weights` is set to `FALSE`. The hyper parameters μ_ϕ and σ_ϕ for the prior distribution of $\log \phi$ are respectively represented by arguments `log_phi_mean` and `log_phi_sd`. Hyper parameters for ω and ϕ and the values for c can be passed to functions `bevimed` and `bevimed_m`.

  Estimating the scaling factor ϕ in this way has the advantage of maintaining power even when the weights are counter-productive, as ϕ can take values close to 0 making

the weights redundant. However, it is possible to make the effect of variant weights c fixed by setting the parameter `estimate_phi` to `FALSE`, in which case ϕ is fixed at 1.

Application to real data

It is an assumption of model $\gamma = 1$ that variants are not linked across loci. We therefore recommend removal of any first, second and third degree relatives, and filtering variants are for low allele frequency across all ethnic groups before applying the function. Various software is available for performing these tasks: as an example 'SAMtools' and 'KING' can be used for variant filtering and inferring relatedness respectively. There is also various software for reading VCF files into `R`. The 'BeviMed with VCFs' vignette contains instructions on how to read allele counts across variants in a given locus into `R` from a VCF file directly as a matrix using simple functions depending on the program 'tabix'. However, although this method could be effective for testing a single locus, typically testing association between a disease and multiple loci is required, in which case reading variants belonging to multiple loci at the same time is likely to be more efficient. Often, it will be most effective to read data for as many variants as possible into memory (e.g. breaking up the VCF by chromosome), and looping through loci one at a time, applying `bevimed` the allele count matrix of its variants.

Typically loci would correspond to genes, but it is also applicable to non-coding loci. In order to increase power, variants which are unlikely to be involved in disease can be filtered out, or have their probability of pathogenicity down-weighted using the `variant_weights` parameter. For example, synonymous variants could be removed, and loss-of-function variants could be up-weighted. It is also straightforward to apply the inference to multiple sets of variants corresponding to different classes of variants for a single locus using the `bevimed_polytomous` function. For example, a set containing only loss-of-function variants could be used to evaluate evidence for association between the disease and a 'knocked-out' gene. This increases power if the only variants of a particular class increase disease risk. Prior probabilities of association with each model set can then be combined with the evidence to obtain the posterior probability of association with each model/variant set.

Although typically testing association between a disease and multiple loci is required, BeviMed only provides procedures for dealing with a single locus. This is because most of the time such an analysis is computationally expensive due to the large number of applications required or large quantity of genetic data which must be loaded, and full control is required in order to best exploit the resources available. Here we provide a simple example script which applies the inference to multiple loci and tabulates the results with columns for gene

name, posterior probability of association and probability of dominant inheritance given the association. Let `chr1genes` be a `data.frame` of chromosome 1 genes with columns for name, start position and end position (the 'biomaRt' package could be used to obtain such a table), and `y` be a logical vector indicating disease status, the same length as the number of samples in the VCF.

```
source(paste0(system.file(package="BeviMed", "/scripts/vcf.R")))

all_variants <- vcf2matrix("my-vcf.vcf.gz",
    chr="1", from=1, to=1e9, include_variant_info=TRUE)

row_indices_per_gene <- lapply(1:nrow(chr1genes),
    function(i) {
        which(all_variants$info$POS >= chr1genes$start[i] &
            all_variants$info$POS <= chr1genes$end[i])
    })
names(row_indices_per_gene) <- chr1genes$gene

results <- mclapply(
    mc.cores=16L,
    X=chr1genes$gene,
    FUN=function(gene) {
        G <- all_variants$G[variant_inds[[gene]],,drop=FALSE]
        c(
            list(gene=gene),
            summary(bevimed(y=y, G=G))) })

results_table <- do.call(what=rbind, lapply(results, function(x) data.frame(
    Gene=x[["gene"]],
    prob_assoc=sum(x[["prob_association"]]),
    prob_dominance=x[["prob_association"]]["dominant"]/
        sum(x[["prob_association"]]),
    check.names=FALSE,
    stringsAsFactors=FALSE
)))
```

Performance and tuning

As an MCMC based procedure, statistics produced from `bevimed` have Monte Carlo error. In the implementation in the BeviMed package, z is the only parameter which is sampled and is updated using Gibbs sampling of each component z_j in turn. If variant weights are included, ω and ϕ are also sampled using Metropolis-Hastings within Gibbs steps, causing estimates of the evidence to have higher variance for the same number of samples. By default, `bevimed` draws $1,000$ samples from each of 7 tempered chains in the MCMC algorithm, running at temperatures $t = \left(\frac{l}{6}\right)^2$ for $l \in \{0, 1, \ldots, 6\}$. We have found that this parameterisation leads to quick execution and stable results for sample sizes up to $5,000$ and loci containing over $2,000$ variants, also allowing for the inclusion of variant weights. However, if much larger sample sizes or larger numbers of variants are used — particularly if variant weights are included — it may become necessary to modify the parameters controlling the sampling routine in order to improve the accuracy of the results. Strategies for doing this include:

- increase the number of samples drawn per tempered chain using the `samples_per_chain` argument,
- increase the number tempered chains or change the distribution of temperatures using the `temperatures` argument,
- pass the `tune_temps` argument to `bevimed`, specifying the number of temperatures to select by interval bisection for use in the final application,
- if estimating ϕ and ω, set `tune_omega_and_phi_proposal_sd=TRUE` in the call to `bevimed` in order to adaptively tune the standard deviations of the Metropolis-Hastings proposal distributions so that the acceptance rate falls within a given range, defaulting to $[0.3, 0.7]$. If this option is used, a tuning run of the MCMC algorithm is applied, which estimates a proposal standard deviation for each temperature using successive blocks of `tune_block_size` samples until the desired acceptance rate is obtained.

It is also possible to instruct `bevimed_m` to halt sampling once the estimated evidence lies within a given confidence interval, or once there is sufficient confidence that the evidence is greater than some threshold. The latter might be useful, for instance, if many regions were being tested for association and only those with very strong evidence for association were of interest). By default, `bevimed_m` does not attempt to stop sampling, and always draws `samples_per_chain` samples for each tempered chain. In terms of the argument names, by setting `stop_early=TRUE`, `bevimed_m` draws up to `blocks` batches of `samples_per_chain` samples, stopping as soon as the estimated log evidence lies within

a confidence interval of width `tolerance` (defaults to 1) with confidence of `confidence` (defaults to 0.95) based on `simulations` simulations (defaults to $1,000$), or as soon as there is `confidence` confidence that it is below `log_evidence_threshold`.

By default the function `bevimed_m` stores the complete set of samples drawn during the MCMC process (after burn-in samples are removed) and therefore this function typically uses lots of memory. The vast majority of memory usage is expended storing the trace of samples of z, the indicator of pathogenicity for each variant) and x, the indicator of having a pathogenic configuration for each individual. Storing these traces enables useful summary statistics — for example the expected number of explained cases — to be computed from the output. However, users may only be interested in the probability of association, for example, if prioritising many sets of variants. In this situation, the arguments `return_x_trace` and `return_z_trace` can be set to `FALSE` when calling `bevimed_m` in order to conserve memory, but still allow the evidence and probabilities of association to be computed.

The inference functions `bevimed_m`, `bevimed` and `bevimed_polytomous` always sample from the posterior distribution of pathogenicity of each variant represented in allele count matrix G. However, G does not necessarily contain data relevant to pathogenicity for all variants which are represented in it. This occurs when the allele counts for a variant are zero for all individuals, or when conditioning on recessive inheritance and alleles for the variant are only present for individuals whose total allele count is less than their ploidy. The function `subset_variants` can be used to remove such variants, returning either a transformed matrix or the indices of the variants in the original set for which there is data relevant to pathogenicity in G. Typically, G would only contain variants which were observed in at least one of the individuals, so using this function *a priori* is not likely to result in a speed up when applied conditioning on dominant inheritance, as no variants would be removed. However, it is often the case that only a small number of variants are observed in compound heterozygotes/homozygotes, so it is likely to result in a speed up conditioning on recessive inheritance.