

A network-based analysis of the cellular
and genetic etiology of disease

Alexander John Cornish

Submitted in part-fulfilment of the requirements for the degree
of Doctor of Philosophy of Imperial College London
and the Diploma of Imperial College London

Department of Life Sciences

Imperial College London

2016

Declaration

The contents of this thesis are my own work unless otherwise specified.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Alexander John Cornish

Abstract

Thousands of disease-associated loci have been identified in genome-wide association studies (GWAS). These loci can span multiple genes and identifying which, if any, of these genes are causal can be challenging. Multiple methods have been developed to identify causal genes, some of which use networks of physical interactions between proteins. The performance of many of these methods may however be limited by their failure to use data specific to the tissues and cell types that manifest each disease. Furthermore, many network-based approaches may be biased towards better-studied genes.

In order to use data specific to a disease-manifesting cell type to identify disease-associated genes, it is first necessary to identify the disease-manifesting cell types. In this thesis, I report the development of the GSC (Gene Set Compactness) and GSO (Gene Set Overexpression) methods, which I use to identify associations between 352 diseases and 73 cell types. The GSC method identifies these associations using cell-type-specific protein-protein interaction (PPI) networks, which I generate by integrating PPI and gene expression data. Using text mining, it is demonstrated that these methods identify a large number of well-characterised disease-cell-type associations and associations that warrant further investigation.

I also describe the development of ALPACA (Analysing Loci using Phenotypic And Cellular Associations), which identifies disease-associated genes using cell-type-specific PPI networks and phenotype data from humans and mice. I demonstrate that by taking a permutation-based approach, ALPACA avoids being biased towards better-studied genes. Furthermore, I demonstrate that using cell-type-specific networks, instead of generic networks, improves method performance.

As the number of available tissue and cell-type-specific data continues to

increase, methods that integrate these data will become increasingly important in understanding disease etiology.

Acknowledgements

I would first like to thank my friends and colleagues, past and present, in the Structural Bioinformatics Group, Imperial College London. I would especially like to thank Michael Sternberg for his supervision, guidance and encouragement over the last three years. My thanks also go to Alessia David, Ioannis Filippis, Suhail Islam, Christopher Yates and Joe Greener for our conversations on all matters bioinformatics.

Acknowledgment must also go to the British Heart Foundation, for providing the funding that allowed me to complete this research.

I would also like to thank my family and friends for the invaluable support they have given me throughout this research. Finally, I thank Hannah, for providing me with plentiful distractions when the work got tough and for reminding me that there are sometimes (occasionally) more important things in life than science.

Contents

1	Introduction	21
1.1	Outline of thesis	21
1.2	Disease variants	22
1.2.1	Identifying disease variants	22
1.2.2	Human variant databases	28
1.2.3	Non-human variant databases	32
1.3	Biological networks	33
1.3.1	Types of biological network	34
1.3.2	PPI databases	39
1.3.3	Context-specific networks	40
1.4	Methods for identifying causal genes	44
1.4.1	Text mining	45
1.4.2	Physical interactions	48
1.4.3	Functional relationships	52
1.4.4	Phenotypic similarities	54
1.4.5	Data from model organisms	57
1.4.6	Context-specific data	59
1.5	Methods for mapping diseases to contexts	62
1.5.1	Text mining	63
1.5.2	Gene expression data	65
1.5.3	Gene expression and PPI data	67
1.5.4	Epigenetic data	68
1.6	Scope of thesis	71
2	Materials and methods	72

2.1	Performance evaluation	72
2.1.1	Performance statistics	72
2.1.2	ROC curves	73
2.2	Statistical tests	74
2.2.1	Fisher’s exact test	74
2.2.2	The Mann-Whitney U test	74
2.3	Adjusting for multiple testing	75
2.3.1	Bonferroni correction	75
2.3.2	Benjamini-Hochberg procedure	75
2.4	Hierarchical clustering	76
2.5	Normalising read counts	77
2.6	Random-walk-based network algorithms	77
2.6.1	Measuring distances	78
2.6.2	Propagating scores	79
3	Identifying associations between diseases and cell types	80
3.1	Introduction	81
3.1.1	Motivation	81
3.1.2	Generating context-specific networks	82
3.1.3	Mapping diseases to contexts	82
3.2	Materials and methods	83
3.2.1	Gene expression data	83
3.2.2	Protein-protein interaction data	94
3.2.3	Generating context-specific PPI networks	94
3.2.4	Disease-gene association data	95
3.2.5	Mapping diseases to contexts	96
3.3	Results	103
3.3.1	Network topology features	103
3.3.2	Disease-associated cell-type-specific sub-networks	105
3.3.3	Parameter selection	106
3.3.4	Effect of gene set size on method performance	108
3.3.5	Comparison of the associations identified by the methods	110

3.3.6	Disease-cell-type associations identified by the GSC method .	117
3.3.7	Cell-type-based diseasesomes	121
3.4	Method implementation and data availability	123
3.5	Discussion	128
3.6	Conclusions	130
4	Prioritising genes in trait-associated loci	132
4.1	Introduction	132
4.1.1	Motivation	132
4.1.2	Data sources	133
4.1.3	Generating association scores	134
4.2	Materials and methods	135
4.2.1	The ALPACA method	136
4.2.2	Defining trait-associated loci	141
4.2.3	Human disease variant data	145
4.2.4	Human disease phenotype terms	148
4.2.5	Mouse phenotype data	150
4.2.6	Measuring phenotype similarity	151
4.2.7	Protein-protein interaction data	155
4.2.8	Evaluating method performance	156
4.3	Results	159
4.3.1	Study bias in PPI databases	159
4.3.2	Calibrating trait-associated loci	163
4.3.3	Network propagation parameter selection	166
4.3.4	Type-1 error rate analysis	168
4.3.5	Effect of study bias on gene prioritisation	168
4.3.6	Comparison of method performance	171
4.3.7	Performance using data from multiple species	173
4.3.8	Performance using context-specific networks	175
4.3.9	Comparison of edge weighting methods	177
4.3.10	Case study	178
4.4	Discussion	181

4.5	Conclusions	183
5	Discussion and future work	185
5.1	Discussion	185
5.1.1	Moving towards a dynamic picture of the interactome	185
5.1.2	Understanding the context-specific effects of disease genes . .	187
5.1.3	Identifying causal genes using trans-acting regulatory elements	189
5.2	Future work	190
5.2.1	Using data from additional species	190
5.2.2	Applying ALPACA to different network types	191
5.2.3	Prioritising disease variants	192
5.2.4	Making ALPACA available for use	193
6	Conclusions	194
	Appendix	225

List of Figures

1.1	The case-control GWAS setup	25
1.2	Imputing SNPs in unrelated individuals	27
1.3	The Y2H method	35
1.4	The TAP-MS method	36
1.5	Inferring edges using the spoke and matrix models.	38
1.6	Methods for generating context-specific PPI networks	43
1.7	Network distance measures	51
1.8	The MICA semantic similarity measure	56
3.1	FANTOM5 project sample normalisation and combination pipeline . . .	87
3.2	Clustering of cell samples of different potencies in cell type facets . . .	90
3.3	Number of FANTOM5 project samples in each cell type facet	91
3.4	The GSO method	97
3.5	The GSC method	99
3.6	The text mining method	101
3.7	Cell-type-specific network edge weight distribution	104
3.8	Cell-type-specific disease sub-networks	107
3.9	Support between the GSC and GSO methods	114
3.10	Overlap between the GSC and GSO methods	115
3.11	A subset of the associations identified by the GSC method	118
3.12	Cell-type-based diseasome generated using two connections	124
3.13	Cell-type-based diseasome generated using three connections	125
3.14	Cell-type-based diseasome generated using four connections	126
3.15	Cell-type-based diseasome generated using five connections	127
4.1	The ALPACA method	137

4.2	Defining trait-associated loci and the genes they contain	142
4.3	Integrating disease-gene associations	147
4.4	Integrating disease-phenotype-term mappings	150
4.5	Integrating data to generate gene-phenotype associations	152
4.6	The simGIC semantic similarity metric	154
4.7	P-values generated by ALPACA when applied to simulated null GWAS	169
4.8	Performance of ALPACA and PRINCE in cross-validation	171
4.9	Performance of ALPACA using data from different organisms	174
A.1	The disease-cell-type associations identified by the GSC method	228

List of Tables

1.1	Human disease variant databases	29
1.2	Model organism genotype and phenotype databases	33
1.3	PPI databases	39
1.4	Context-specific networks	42
1.5	List of gene prioritisation methods	46
1.6	Studies that have systematically mapped diseases to contexts	64
3.1	Effect of the processing steps on the gene expression data.	84
3.2	The number of samples discarded using different tree cut values	88
3.3	Vertex betweenness scores in the cell-type-specific networks.	105
3.4	GSC performance with different parameters (overlap significance)	109
3.5	GSC performance with different parameters (F1 score)	110
3.6	GSC performance when applied to gene sets of different sizes	111
3.7	GSO performance when applied to gene sets of different sizes	112
3.8	Overlap of associations identified by the GSC and text mining methods	113
3.9	Overlap of associations identified by the GSO and text mining methods	113
3.10	Associations identified between different disease and cell type classes	116
4.1	Size of the human variant data sets	146
4.2	Numbers of disease-associated phenotypes	149
4.3	Size of the mouse variant data set	151
4.4	Number of terms and relationships in the Uberpheno ontology	153
4.5	Evaluations completed and data sets used in this chapter	157
4.6	Network degrees of disease and non-disease genes	161
4.7	Numbers of articles reporting interactions involving each disease and non-disease gene	162

4.8	Numbers of interactions per article reporting interactions involving each disease and non-disease gene	163
4.9	The effect of parameter selection on loci definition	165
4.10	ALPACA performance with different parameters	167
4.11	Correlations between prioritisation scores and number of associated data	170
4.12	The effect of gene network degree on method performances	172
4.13	ALPACA performance using different context identification methods . .	176
4.14	RA gene set enrichment results.	180
A.1	Cell-type-specific PPI networks and the corresponding MeSH terms . .	227
A.2	Details of the data downloaded and used in this thesis	233

Abbreviations

1KGP	1000 Genomes Project
3C	Chromosome conformation capture
4C	Circularised chromosome conformation capture
5C	Carbon-copy chromosome conformation capture
AD	Autosomal dominant
ALPACA	Analysing Loci using Phenotypic And Cellular Associations
AR	Autosomal recessive
AUC	Area under the curve
AVEXIS	Avidity-Based Extracellular Interaction Screen
BBB	Blood brain barrier
BBID	Biological Biochemical Image Database
BH	Benjamini-Hochberg
BIND	Biomolecular Interaction Network Database
BioGRID	Biological General Repository for Interaction Datasets
BMI	Body mass index
bp	Base pair
CAD	Coronary artery disease
CAGE	Cap analysis of gene expression
CEU	Individuals with Northern and Western European ancestry from Utah, USA
cHi-C	Capture Hi-C
CL	Cell Ontology
CNSH	Central nervous system hypomyelination

CNV	Copy number variation
CVI	Common variable immunodeficiency
DAPPLE	Disease Association Protein-Protein Link Evaluator
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbSNP	The Single Nucleotide Polymorphism Database
DEG	Differentially-expressed gene
DEPICT	Data-Driven Expression Prioritised Integration for Complex Traits
DHS	DNase I hypersensitivity site
DIP	Database of Interacting Proteins
DNase I	Deoxyribonuclease I
DO	Disease Ontology
DPI	Decomposition peak identification
ENCODE	Encyclopaedia of DNA Elements
eQTL	Expression quantitative trait loci
FANTOM5	Functional Annotation of the Mammalian Genome 5
FDR	False discovery rate
FF	FANTOM5 ontology
FLD	Frontotemporal lobar degeneration
FN	False negative
FP	False positive
FWER	Family-wise error rate
GFC	Glomerulosclerosis, focal segmental
GI	Genetic interaction
GNF	Genomics Institute of the Novartis Research Foundation
GO	Gene Ontology
GRAIL	Gene Relationships Among Implicated Loci
GSC	Gene Set Compactness
GSO	Gene Set Overexpression
GWAS	Genome-wide association study

HD	Hirschsprung's disease
HGMD	Human Gene Mutation Database
HINT	High-Quality Interactomes
HPO	Human Phenotype Ontology
HPRD	Human Protein Reference Database
HT	High-throughput
HTLC1	Human T cell lymphotropic virus 1 infection
HUSA	Hemolytic uremic syndrome, atypical
ICA	Independent component analysis
IL2RA	Interleukin 2 receptor, alpha
IMPC	International Mouse Phenotyping Consortium
JPT+CHB	Individuals with Japanese ancestry from Tokyo, Japan and individuals with Han Chinese ancestry from Beijing, China
JSD	Jensen-Shannon distance
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KLD	Kullback-Leibler divergence
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
LQTS	Long QT syndrome
LT	Low-throughput
Mb	Megabase
MCID	Mitochondrial complex I deficiency
MDEPC	Monocyte derived endothelial progenitor cell
MDM	Mycobacterial disease, Mendelian
MeSH	Medical Subject Heading
MGD	Mouse Genomics Database
MICA	Most informative common ancestor
MID	Monocyte immature derived
MIM	Mendelian Inheritance in Man
MINT	Molecular Interaction Database
MLNS	Mucocutaneous lymph node syndrome

MP	Mammalian Phenotype Ontology
MS	Multiple sclerosis
NETBAG+	Network-Based Analysis of Genetic Associations +
NetWAS	Network-Wide Association Study
NIH	National Institute for Health
NK	Natural killer
NLP	Natural language processing
NSWL	Nephrotic syndrome with lesion of minimal change glomerulonephritis
OBO	Open Biomedical Ontologies
OMIM	Online Mendelian Inheritance in Man
OPEN	Objective Prioritisation for Enhanced Novelty
OWL	Web Ontology Language
PATO	Phenotypic Quality Ontology
PCC	Pearson's correlation coefficient
PEOWMDD	Progressive external ophthalmoplegia with mitochondrial DNA deletions
PharmGKB	Pharmacogenomics Knowledgebase
PhenIX	Phenotypic Interpretation of Exomes
PHIVE	Phenotypic Interpretation of Variants in Exomes
PICS	Probabilistic Identification of Causal SNPs
PIDC	Primary idiopathic dilated cardiomyopathy
PINBPA	Protein Interaction Network-Based Pathway Analysis
PolyPhen-2	Polymorphism Phenotyping v2
PPI	Protein-protein interaction
PrePPI	Predicting Protein-Protein Interactions
PRINCE	Prioritisation and Complex Elucidation
QT	Quantitative trait
RA	Rheumatoid arthritis
RGD	Rat Genome Database
RLE	Relative log expression

ROAST	Rotation Gene Set Testing
ROC	Receiver operating characteristic
RWR	Random walk with restart
RWRH	Random Walk with Restart on Heterogeneous Network
SAV	Single amino acid variant
SGA	Synthetic genetic array
SLE	Systemic lupus erythematosus
SMAD7	SMAD Family Member 7
SNOMED	Systematised Nomenclature of Medicine Reference Terminology
SNP	Single nucleotide polymorphism
SSP	Subacute sclerosing panencephalitis
STAT4	Signal Transducer and Activator of Transcription 4
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SuSPect	Disease-Susceptibility-based SAV Phenotype Prediction
SVM	Support vector machine
T2D	Type II diabetes
TAP-MS	Tandem affinity purification with mass spectrometry
TF	Transcription factor
TN	True negative
TP	True positive
TPM	Tags per million
TSS	Transcription start site
UCSC	University of California, Santa Cruz
UMLS	United Medical Language System
UniProt	Universal Protein Resource
US	Uveomeningoencephalitic syndrome
VCAM1	Vascular cell adhesion molecule 1
VEGAS	Versatile Gene-Based Association Study
WNT10B	Wingless-Type MMTV Integration Site Family, Member 10B

Y2H	Yeast two-hybrid
YRI	Individuals with Yoruba ancestry from Ibadan, Nigeria
ZP	Zebrafish Ontology

List of publications

Cornish A. J., Filippis I., David A., & Sternberg M. J. (2015). Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types. *Genome Medicine*, 7:95.

Chapter 1

Introduction

1.1 Outline of thesis

Genome-wide association studies (GWAS) have identified thousands of loci associated with various human traits, including disease susceptibilities (Li et al., 2016). For many of these loci, the causal variants and genes have yet to be identified. The aim of the work described in this thesis is to use an unbiased network-based approach to identify the genes in trait-associated loci most likely to be causal. Previously developed network-based methods often use generic networks, containing interactions that occur throughout the human body. In this work, I integrate protein-protein interaction (PPI) and gene expression data to generate cell-type-specific PPI networks and use these networks to identify the genes most likely to be causal.

In the introduction to this thesis, I discuss the methods that have been used to identify disease-causing variants and genes (Section 1.2). I then discuss how network-based methods have been used to study the relationships between genes (Section 1.3) and how various types of biological data have been used to predict which genes are likely to cause disease (Section 1.4). Finally, I discuss how methods have been developed to predict the tissues and cell types in which diseases are most likely to manifest and the insights gained from these predictions (Section 1.5).

In Chapter 3, I report the development of the Gene Set Compactness (GSC) and Gene Set Overexpression (GSO) methods, which identify disease-manifesting

cell types. The GSC method predicts these associations using cell-type-specific PPI networks. Both the GSC and GSO methods identify well characterised disease-associated cell types and associations that warrant further study.

In Chapter 4, I build upon the work of the previous chapter and describe the development and testing of ALPACA (Analysing Loci using Phenotypic And Cellular Associations). ALPACA uses cell-type-specific PPI networks and phenotype data from humans and mice to identify the genes in trait-associated loci most likely to be causal. ALPACA uses the GSO method to determine which cell-type-specific PPI network are best suited to studying each disease.

1.2 Disease variants

Identifying disease-causing variants and the genes they affect is important in understanding disease, its diagnosis and in identifying therapeutic targets. Since the 1980s, a number of methods have been developed to identify causal variants and genes. In this section, I describe these methods and the databases that have been established to store data relating to variants and their phenotypic effects, in humans (*Homo sapiens*), mice (*Mus musculus*) and other organisms.

1.2.1 Identifying disease variants

Genetic mapping has been used to identify variants and genes underlying many traits, including disease susceptibilities. Two main approaches have emerged over the last 40 years: linkage analysis and association analysis. These methods are hypothesis-free and rely upon identifying correlations between the trait of interest and variants in the genome. Linkage analysis maps genes to traits through the co-segregation of the trait and genetic markers across multiple generations. Association analysis identifies trait-associated variants by finding correlations between trait measurements and variant frequencies in a population. Both approaches have advantages and disadvantages and are best suited to studying certain types of traits.

Linkage analysis

Linkage analysis is a technique originally developed in *Drosophila melanogaster* in 1913 (Sturtevant, 1913). It follows the inheritance of traits and genetic markers across multiple generations. If a trait and a marker demonstrate a correlated segregation pattern, then it can be deduced that the gene responsible for the trait lies close to the marker. The inability to design crosses, small family sizes and an insufficient number of genetic markers limited the use of linkage analysis in humans until the 1980s, when Botstein et al. (1980) suggested that naturally occurring DNA polymorphisms could be used to study the inheritance of chromosomal regions. This development led to the localisation of genes associated with many Mendelian disorders, including Huntington's disease (Gusella et al., 1983).

While linkage analysis has been successful in identifying genes associated with Mendelian disorders, it has been less successful in studying complex diseases. This is partly due to locus heterogeneity, in which multiple loci influence the development of a disease. Whilst genes have been mapped to a number of Mendelian subtypes of complex diseases, including breast cancer (Hall et al., 1990) and type II diabetes (T2D) (Vionnet et al., 1992), these genes explain only a small fraction of disease incidence. For this reason, over the previous decade, GWAS have largely supplanted linkage analysis as the preferred method for studying the genetic basis of complex disease (Altshuler et al., 2008).

As the cost of genome sequencing has declined, linkage analysis has begun to be used alongside whole exome and whole genome sequencing to map Mendelian disorders (Ott et al., 2015). GWAS only interrogate common variants and are therefore unable to identify rare variants associated with disease. As the accuracies of sequencing technologies have improved, it has become possible to use sequencing technologies in conjunction with linkage analysis to identify rare disease variants. In this approach, variants can be excluded if they are not shared between affected individuals, or are present in unaffected individuals (Ott et al., 2015). Variants can be further filtered by removing common variants, such as the variants represented in the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium, 2015), or through the application of bioinformatics tools that predict variant pathogenicity,

such as PolyPhen-2 (Adzhubei et al., 2010) and SuSPect (Yates et al., 2014). In combination with sequencing, linkage analysis has been used to identify variants associated with a number of Mendelian disorders, including familial hypertension (Louis-Dit-Picard et al., 2012).

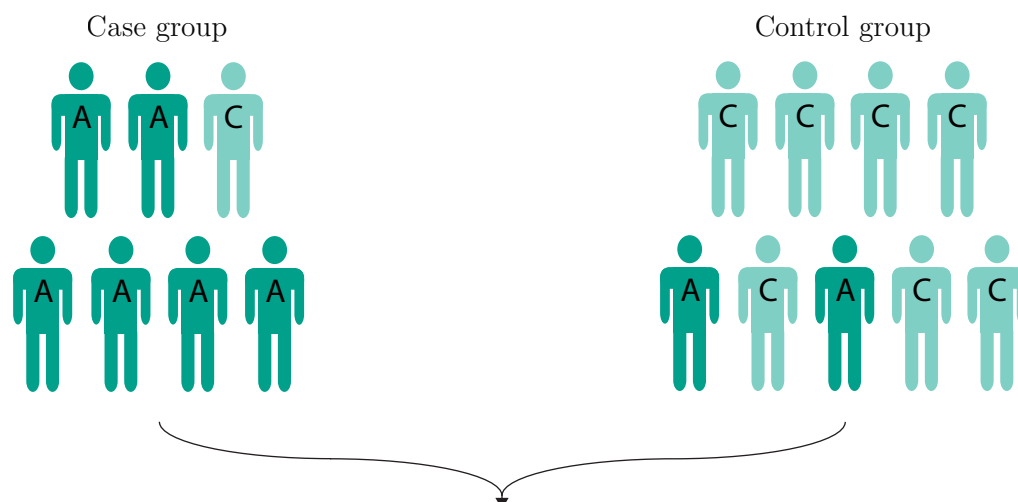
Association analysis

Association analysis identifies trait-associated variants by finding correlations between trait measurements and variant frequencies in a population. Over the previous decade, GWAS have identified thousands of trait-associated single nucleotide polymorphisms (SNPs). Traits studied include susceptibilities to complex disease, such as coronary artery disease (CAD) (The CARDIoGRAMplusC4D Consortium, 2015), multiple sclerosis (MS) (Sawcer et al., 2011) and T2D (DIAGRAM Consortium et al., 2014), and quantitative traits (QTs), such as blood lipid levels (Teslovich et al., 2010), body mass index (BMI) (Locke et al., 2015) and height (Wood et al., 2014).

Multiple GWAS methodologies exist to study different types of traits. Complex disease susceptibilities tend to be studied using a case-control setup (Figure 1.1). In this setup, the individuals studied are divided into a case group, containing individuals affected by the disease, and a control group, generally containing individuals without a history of the disease or a diagnosis at the start of the study (Sawcer et al., 2011; The CARDIoGRAMplusC4D Consortium, 2013).

Developments in microarray technology facilitated the large-scale genotyping of SNPs required by GWAS (Bush & Moore, 2012). In many GWAS, the number of SNPs genotyped is between 500,000 and 1,000,000 (Bush & Moore, 2012), far lower than the total number of SNPs that exist in the human genome (The 1000 Genomes Project Consortium, 2015). GWAS are however still able to capture the majority of variation that exists in the genome, as they exploit linkage disequilibrium (LD) between alleles. LD is the non-random association of alleles across multiple loci. It exists as a result of the shared ancestry of chromosomes. If a mutation occurs in a genome, then the new variant will exist only in the chromosome in which it occurred, which will be marked with a distinct combination of variants.

A) Case and control groups with genotype at a single position shown



B) Contingency table of genotype frequencies

	A	C
Cases	6	1
Controls	2	7

↓
Association significance p-value

C) Manhattan plot of association significances

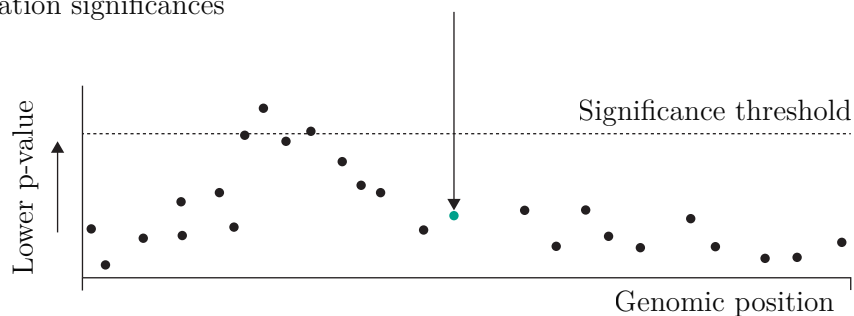


Figure 1.1: The case-control GWAS setup. A) Individuals are split into case and control groups. In this illustrative example, the genotype of each individual at a single position is shown. B) Methods such as contingency tables and logistic regression can be used to determine whether the frequency of an allele differs significantly between the two groups. In this example, an association significance p-value is produced using a contingency table. C) A Manhattan plot can be used to visualise the association significances of multiple SNPs in a genomic region or across the entire genome. A genome-wide significance threshold more stringent than the significance thresholds often employed in statistical testing is used to account for the testing of multiple hypotheses. In this example, the tested position (indicated by the arrow) does not pass the genome-wide significance threshold.

Over successive generations, recombination events and new mutations reduce the strength of association between the new variant and the surrounding variants. Due to the nature of recombination, LD between loci remains stronger the smaller the distance between the loci (The International HapMap Consortium, 2007). The SNPs genotyped in GWAS are therefore unlikely to themselves be the causal variants, but instead ‘tag’ larger regions of the genome associated with the trait of interest. A combination of correlated SNPs across multiple loci is known as a haplotype. The HapMap Project (The International HapMap Consortium, 2007) and the 1KGP (The 1000 Genomes Project Consortium, 2015) have studied haplotypes in multiple populations.

Methods such as contingency tables and logistic regression are used to determine whether the frequency of a SNP differs significantly between case and control groups (Bush & Moore, 2012). Once a trait-associated SNP (and therefore a trait-associated locus) has been identified, attempts can be made to identify the causal variant or variants through SNP imputation, the targeted sequencing of the locus, or the application of bioinformatics tools. In SNP imputation, SNPs not genotyped in the original study are inferred using the haplotypes of the individuals studied (Figure 1.2) (Bush & Moore, 2012). Bioinformatics tools often incorporate additional sources of biological data to identify the variants or genes in trait-associated loci most likely to be causal (see Section 1.4 for further details).

It is difficult to determine the translational success of GWAS as the time passed since many GWAS were completed is not sufficient to determine whether any targets identified will lead to approved therapeutics. There is some evidence however that trait-associated variants, genes and pathways identified in GWAS do represent useful therapeutic targets. For example, two genes (VCAM1 and IL2RA) identified as associated with MS in a GWAS are already targeted by MS treatments (Sawcer et al., 2011). More recently, SMAD7 antisense oligonucleotides (*mongersen*) were seen to produce significantly higher levels of clinical remission amongst Crohn’s disease patients than a placebo in a phase 2 clinical trial (Monteleone et al., 2015). SMAD7 is located near a Crohn’s-disease-associated SNP (Jostins et al., 2012), suggesting that GWAS can aid in drug discovery.

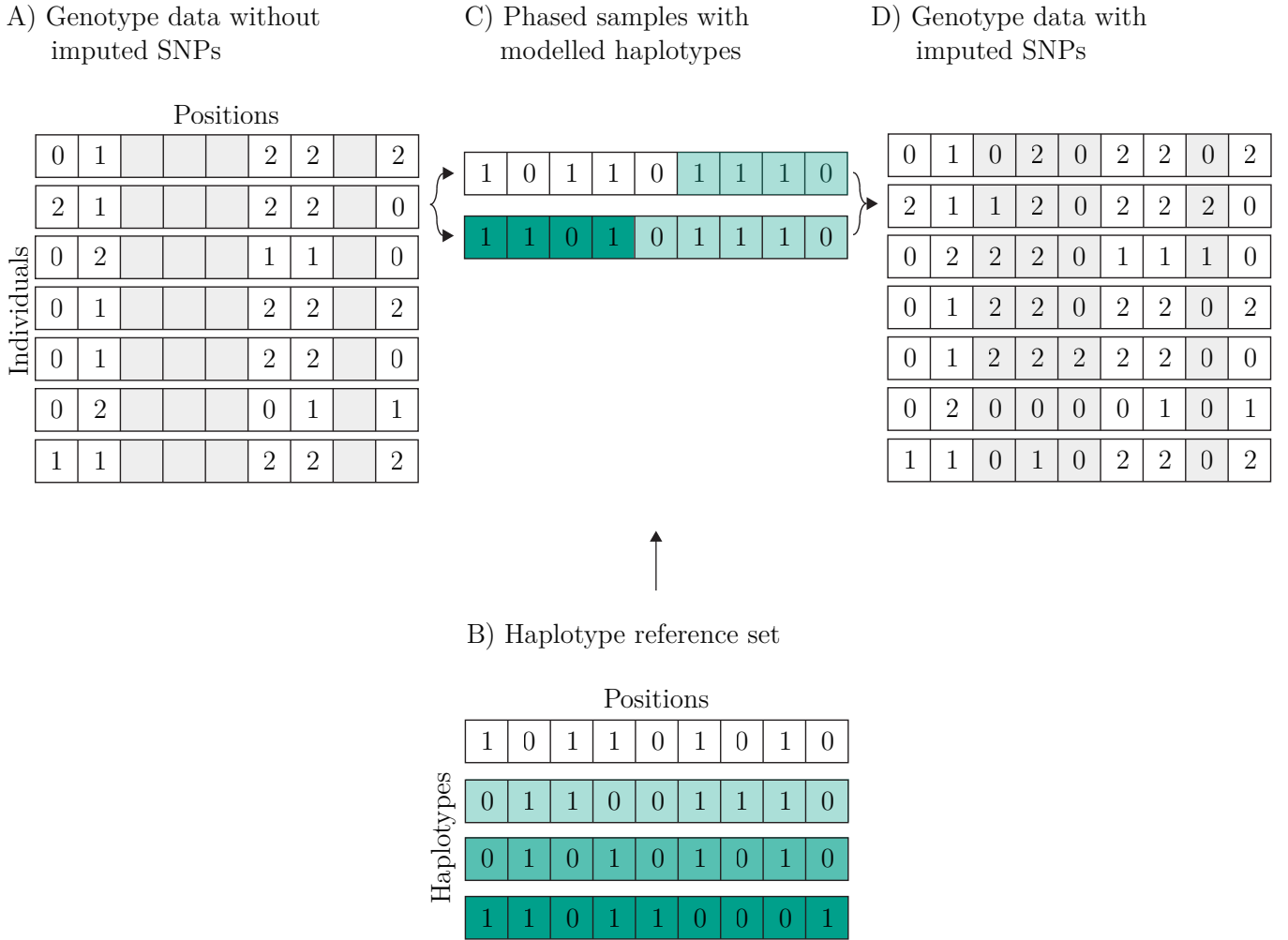


Figure 1.2: The imputation of SNPs in unrelated individuals in an illustrative example. A) The raw genotype data from seven individuals. Some positions have been genotyped, whilst others have not and are therefore empty. B) The haplotype reference set used to impute the SNPs, possibly drawn from the HapMap Project or the 1KGP. C) The haplotypes of a phased individual, determined by matching the genotyped positions to the haplotypes in the reference set. The different colour shades indicate which reference haplotypes the individual matches. D) The data from the seven individuals with SNPs imputed at the missing positions.

While GWAS have identified thousands of trait-associated variants, there are a number of limitations to the methodology that prevent the studies from fully explaining the heritability of many traits. It has been demonstrated that the greater the number of participants in a GWAS, the greater the power of the study to identify variants that pass genome-wide significance thresholds (Teslovich et al., 2010). Multiple studies have boosted their power to identify trait-associated variants by amalgamating data from individuals from multiple studies, resulting in meta-analyses containing up to 250,000 individuals (DIAGRAM Consortium et al., 2014; Wood et al., 2014; The CARDIoGRAMplusC4D Consortium, 2015). Practical constraints make further increases in study size difficult and therefore inhibit further increases in detection power. This increases the demand for tools that are able to combine the results of GWAS with other sources of data to identify candidate variants and genes for further study.

1.2.2 Human variant databases

Several databases have been established to catalog genetic variation in humans and make these data available for use by others. While some of these databases catalog only genetic variation, others associate genetic variation with phenotypic data. Databases that contain phenotypic data include those that focus on disease, such as ClinVar (Landrum et al., 2016), Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015) and UniProtKB/Swiss-Prot (Yip et al., 2008), and those that focus on drug interactions, such as the Pharmacogenomics Knowledgebase (PharmGKB) (Whirl-Carrillo et al., 2012). In this section I discuss three of the most-comprehensive freely available disease variant databases (Table 1.1), the data they contain and any curation procedures conducted.

ClinVar

The ClinVar database contains variants identified through research and clinical testing (Landrum et al., 2016). Variants represent sequence changes at single locations and combinations of changes across multiple locations. ClinVar contains variants with various levels of clinical significance, including pathogenic, likely

Resource	Disease variants	Diseases	Genes	Resource reference
ClinVar	75,512	5,538	3,064	Landrum et al. (2016)
HGMD Professional	109,521	-	-	Stenson et al. (2014)
HGMD Public	80,173	-	-	Stenson et al. (2014)
OMIM	-	6,443	3,424	Amberger et al. (2015)
UniProtKB/Swiss-Prot	25,890	2,969	2,243	Yip et al. (2008)

Table 1.1: The number of disease variants and the numbers of diseases and genes the disease variants map to in five human disease variant databases. Table A.2 contains details of when and from where these data were downloaded. Variants in ClinVar are considered disease variants if they are marked pathogenic or likely pathogenic. OMIM does not attempt to record all disease variants identified in genes and instead focuses on documenting disease at the level of the gene. For this reason, the number of disease variants in OMIM is not given. The numbers of disease variants in HGMD Professional and HGMD Public were obtained from the study by Peterson et al. (2013) and relate to the 11 April 2013 releases of the resources. These two resources are not available to download in full without a license and I was therefore unable to compute the numbers of disease variants currently contained in these resources and the numbers of diseases and genes these variants map to.

pathogenic, risk-related, benign and protective variants. Variants are associated with phenotypic data and data relating to the evidence supporting each variant. Variants are either submitted directly to ClinVar or obtained from other databases, such as OMIM, and there is therefore an overlap between the data contained in ClinVar and OMIM. Diseases referenced in ClinVar are recorded using a number of controlled vocabularies, including Mendelian Inheritance in Man (MIM) numbers (Amberger et al., 2015), Systematised Nomenclature of Medicine Reference Terminology (SNOMED) clinical terms (Stearns et al., 2001) and United Medical Language System (UMLS) terms (Bodenreider, 2004). This facilitates the integration of data from ClinVar with data from other resources.

Variants in ClinVar are categorised depending on whether they have been

submitted once or on multiple occasions, whether they have been reviewed by an expert panel or professional society and whether there are conflicting data associated with the submission. This is done with the aim of simplifying the evaluation of the medical importance of variants in ClinVar (Landrum et al., 2016). As of 3 March 2015, ClinVar contains 281,023 variants, of which 75,512 are marked pathogenic or likely pathogenic. However, of these 75,512 pathogenic or likely pathogenic variants, only 2,896 have so far been reviewed by an expert panel or professional society, and 64,475 have been submitted by only a single user. It is therefore difficult to determine the reliability of much of the data in ClinVar.

OMIM

OMIM documents the genetic basis of disease and contains free-text descriptions of these diseases (Amberger et al., 2015). Unlike ClinVar and UniProtKB/Swiss-Prot, OMIM focuses on documenting disease at the level of the gene, and for this reason it does not attempt to record every causal variant in each disease-associated gene. Instead, OMIM selects variants based on a number of criteria, including high population frequency, distinctive phenotype, historical significance and whether each variant was the first to be identified linking the gene to the disease (Amberger et al., 2011). As of 19 May 2015, OMIM contains 6,870 associations between 6,443 diseases and 3,424 genes. While OMIM traditionally focused on Mendelian disorders, it now also contains data on complex diseases.

OMIM relies on the curation of literature and is therefore vulnerable to study bias, in which better-studied genes are identified as being associated with a greater number of diseases and therefore overrepresented in the database (Köhler et al., 2008). Further work needs to be completed to determine the extent of this study bias and how it may affect bioinformatics tools that use data from OMIM.

UniProtKB/Swiss-Prot

The UniProtKB/Swiss-Prot database contains disease-associated and neutral single amino acid variants (SAVs) (Yip et al., 2008). The database forms the manually curated part of the Universal Protein Resource (UniProt). Putative variants are

identified in relevant publications by a team of curators, either manually or with the aid of automated text mining (Famiglietti et al., 2014). Both manually and automatically identified variants are then reviewed by a curation team. Users of the database can also suggest variants for inclusion. The review of variants includes manual validation of the protein sequence and a critical examination of the experiments and data that led to the identification of the variant and the associated disease (Poux et al., 2014).

While UniProtKB/Swiss-Prot is an expertly curated resource, the quality of its data still depends on the accuracy of the literature it curates. This dependency on literature accuracy is illustrated by the SIRT5 protein, which was originally described in UniProtKB/Swiss-Prot as activating the CPS1 protein through deacetylation. This record was later changed after it was shown that this mechanism was incorrect (Poux et al., 2014).

Currently, UniProtKB/Swiss-Prot contains only variants corresponding to missense mutations and small insertions and deletions (Famiglietti et al., 2014), meaning that some variant types are not represented (Peterson et al., 2013). A new data format is currently being developed by the maintainers of UniProtKB/Swiss-Prot that will allow for the inclusion of other variant types, such as frame-shift mutations (Famiglietti et al., 2014).

As of 2 March 2015, UniProtKB/Swiss-Prot contains 70,687 variants, of which 25,890 are marked as disease variants. 25,809 of these 25,890 disease variants are accompanied by a MIM number describing the disease. This use of a formal vocabulary to describe disease again facilitates the integration of data from UniProtKB/Swiss-Prot with data from other resources.

Other databases

Alongside the freely available disease variant databases, there are also databases that require a paid license. The Human Gene Mutation Database (HGMD) is one of the largest such databases (Stenson et al., 2014). Two versions of the HGMD are available: a freely available public version (HGMD Public) and the license-requiring full version (HGMD Professional). As of 11 April 2013, the Public and Professional

versions of HGMD contain 80,173 and 109,521 disease variants respectively (Peterson et al., 2013). The Public version of the database is accessible only through a web interface, meaning that it is difficult to integrate data from it with data from other resources. The Professional version of HGMD contains the majority of the disease variants contained in ClinVar, OMIM and UniProtKB/Swiss-Prot (Peterson et al., 2013) but its license requirement limits its use.

The number of variant databases that are currently available has led to the development of resources that combine variants reported in different databases. DisGeNET is one of the largest databases of this kind (Pinero et al., 2015). Like OMIM, DisGeNET documents the genetic basis of disease at the level of the gene. It combines disease-gene associations from UniProtKB/Swiss-Prot, associations identified through automated text mining and associations predicted using data from other organisms. Version 2.1 of DisGeNET contains 381,056 associations between 13,185 diseases and 16,666 genes.

1.2.3 Non-human variant databases

Databases also exist that catalog genetic variation in other organisms. These databases include FlyBase (dos Santos et al., 2015), the Mouse Genomics Database (MGD) (Bult et al., 2016) and the Rat Genome Database (RGD) (Shimoyama et al., 2015). Many of these databases also contain data on the phenotypes associated with each cataloged allele (Table 1.2). The development of controlled phenotype vocabularies, both within species (Köhler et al., 2014) and across species (Köhler et al., 2013), has allowed data from these databases to be used to study human disease (Chen et al., 2012; Robinson et al., 2014).

The MGD is one of the most comprehensive model organism databases (Bult et al., 2016). It contains data on spontaneous, engineered and conditional mouse mutations. Included in these data are the phenotypic effects of the mutations, which are recorded using terms from the Mammalian Phenotype Ontology (MP) (Smith et al., 2005). Incidental mutations found during mutational screens are also recorded, although the effects of these mutations are not always clear from the models. Data in the MGD comes from published literature and researcher submissions.

Resource	Alleles with phenotypic data	Genes with phenotypic data	Resource reference
FlyBase	80,122	24,962	dos Santos et al. (2015)
MGD	42,147	13,584	Bult et al. (2016)
RGD	-	1,414	Shimoyama et al. (2015)

Table 1.2: The number of alleles and genes with phenotypic data in three model organism variant databases. Table A.2 contains details of when and from where the data were downloaded. The RGD is not organised by allele and therefore the number of alleles with phenotypic data is not given for the RGD.

The MGD has been specifically designed to aid in the study of human disease (Bult et al., 2016). It therefore provides a list of human-mouse gene orthologs generated using data from HomoloGene (NCBI Resource Coordinators, 2016) that can be used to map phenotypes observed in mouse mutants to human genes. This has led to the development of methods that use data from the MGD to study human disease, including MouseFinder (Chen et al., 2012) and PHIVE (Robinson et al., 2014) (see Section 1.4.5 for further details).

The success of the MGD led to the formation of the International Mouse Phenotyping Consortium (IMPC), who are currently conducting a large-scale mouse phenotyping project (Brown & Moore, 2012). The aim of this project is to phenotype 20,000 protein-coding genes in mice by 2022. The project will produce phenotypic data for many genes for which there are currently no phenotypic data available in mammals.

1.3 Biological networks

Genes and their protein products do not function independently from one another. There has therefore been interest in incorporating interactions between genes and proteins to better understand and predict the phenotypic effects of genetic variants (Krauthammer et al., 2004; Köhler et al., 2008; Yates et al., 2014; Tasan et al.,

2015). The complete collection of interactions that occur between proteins in a cell is often referred to as the interactome (Sanchez et al., 1999). The network is a frequently used method of representing interactions between biological entities, including genes, proteins and regulatory molecules. In this section, I discuss some of the types of networks currently being used to study relationships between genotypes and phenotypes, the resources that contain network data and the work being done to transition from generic networks to networks specific to certain contexts, such as tissues and cell types.

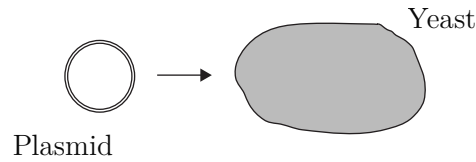
1.3.1 Types of biological network

Large-scale networks have been generated in a number of organisms, including yeast (Costanzo et al., 2010), flies (Guruharsha et al., 2011) and humans (Rolland et al., 2014). These networks often represent either physical interactions between proteins (Guruharsha et al., 2011; Rolland et al., 2014) or functional interactions between genes (Costanzo et al., 2010). Networks have also been generated by combining physical interactions between protein, functional interactions between genes and interactions between other molecules, such as transcription factors (TFs) and miRNAs (Himmelstein & Baranzini, 2015).

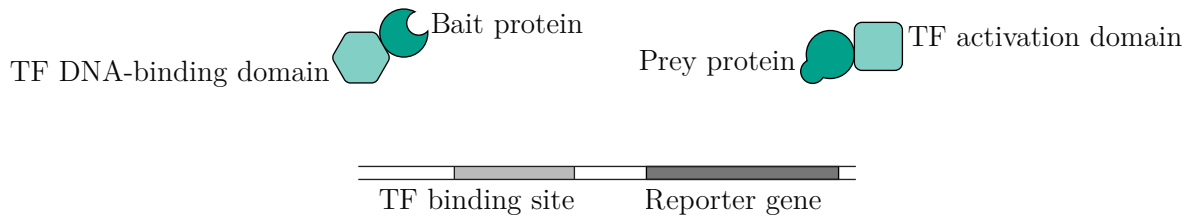
Physical interactions between proteins have been identified using two main approaches: hypothesis-free high-throughput (HT) screens of interactions, in which the presence of interactions between a large number of proteins are systematically tested for, and hypothesis-led low-throughput (LT) studies, in which the presence of interactions between a small number of proteins of interest are tested for (Califano et al., 2012).

Multiple technologies have been developed to test for PPIs. The yeast two-hybrid (Y2H) assay exploits the modular nature of eukaryotic TFs (Figure 1.3) (Fields & Song, 1989). In the assay, yeast are transfected with a plasmid containing two fusion genes: one containing the DNA-binding domain of a TF fused to the first protein of interest (the bait) and one containing the activation domain fragment of a TF fused to the second protein of interest (the prey). If a bait protein and a prey protein interact, then the two TF fragments co-localise, the reporter gene is expressed and

A) Yeast transfected with a plasmid containing two fusion genes and a reporter gene



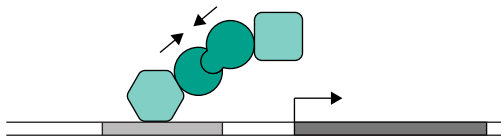
B) Fusion genes transcribed and translated



Interaction occurs

Interaction does not occur

C) Reporter gene transcribed



D) Reporter gene not transcribed

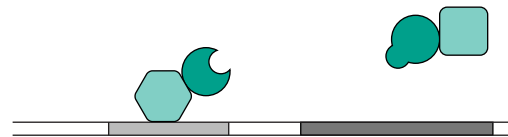
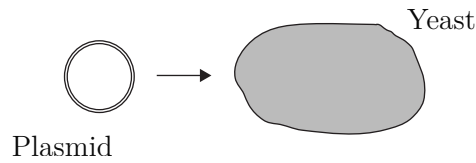
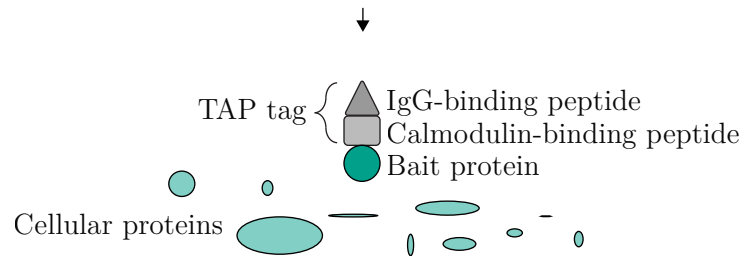


Figure 1.3: Identification of PPIs using the Y2H assay. A) Yeast is transfected with a plasmid containing two fusion genes and a reporter gene. B) The fusion genes are expressed. The protein product of the first fusion gene contains the DNA-binding domain of a TF fused to the first interacting protein (the bait). The protein product of the second fusion gene contains the TF activation domain bound to the second interacting protein (the prey). C) If the bait and prey proteins interact, then the activation domain is brought into contact with the TF binding site and the reporter gene is expressed. This provides evidence that the two proteins interact. D) If the bait and prey proteins do not interact, then the activation domain is not brought into contact with the TF-binding site and the reporter gene is not expressed. This does not provide evidence that the two proteins interact.

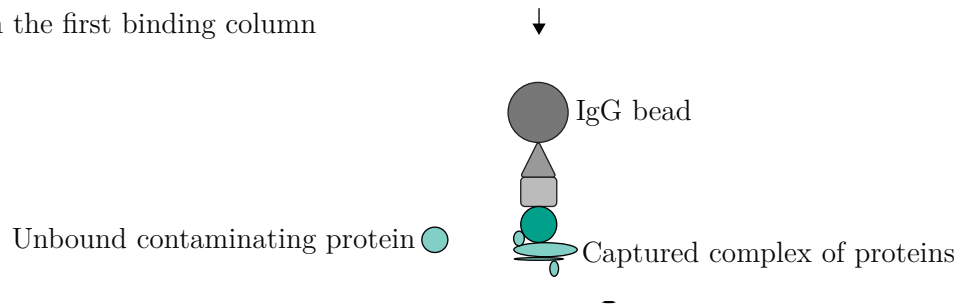
A) Yeast is transfected with a plasmid containing the fusion gene



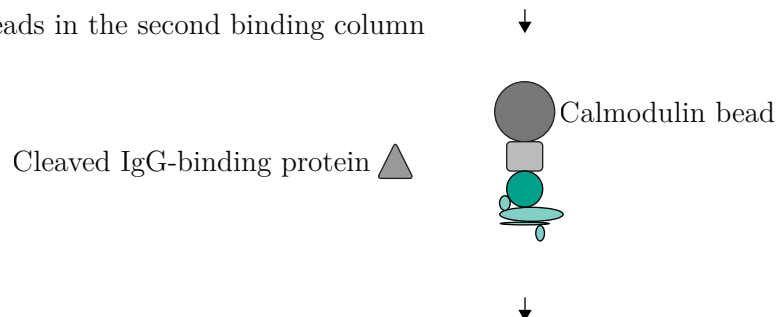
B) Fusion gene transcribed and translated



C) Addition of IgG beads in the first binding column



D) Addition of calmodulin beads in the second binding column



Separation of proteins followed by mass-spectrometry

Figure 1.4: Identification of PPIs using TAP-MS. A) Yeast is transfected with a plasmid containing a fusion gene, which itself contains an IgG-binding domain, a calmodulin-binding domain and the bait protein. B) The fusion gene is expressed in the yeast cells. C) Cells are lysed and their contents added to the first affinity column, which contains IgG beads. D) To remove contaminating proteins, the IgG-binding peptides are cleaved from the calmodulin-binding peptides and calmodulin beads are added to the solution. Finally, proteins are separated and mass spectrometry is used to determine which proteins form the complex.

the interaction can be identified. Another method, tandem affinity purification with mass spectrometry (TAP-MS), identifies protein complexes using a two-step purification process combined with mass spectrometry (Figure 1.4) (Rigaut et al., 1999). Unlike Y2H assays, TAP-MS does not identify direct binary interactions between proteins and therefore binary interactions need to be inferred using either the spoke or matrix model (Figure 1.5). The quality of the interactions identified by these two and other methods has been analysed and debated. One study suggested that Y2H assays and TAP-MS may represent complementary methods for identifying PPIs and that Y2H assays may identify a greater number of transient signalling interactions (Yu et al., 2008). Another study indicated that Y2H assays may produce a high number of false negatives (Braun et al., 2009).

Functional interactions between genes can be identified using methods such as synthetic genetic array (SGA) analysis, or predicted through data integration. In SGA analysis, the presence of a functional interaction between two genes is tested for by first mutating each gene individually (Tong et al., 2001). The effects of these mutations on cell fitness are then measured. Next, the genes are mutated together and the effect of this double mutation on cell fitness is measured. From the fitness effects of the single mutations, it is possible to predict the fitness effect of the double mutation, given that the pair of tested genes do not functionally interact. If the observed fitness effect of the double mutation deviates significantly from the predicted fitness effect, then it can be deduced that the two genes functionally interact.

While SGA analysis has successfully identified interactions in single-celled organisms, such as yeast (Costanzo et al., 2010), practical considerations have limited its application to multi-cellular organisms. This has led to the development of methods that infer functional interactions between genes by integrating multiple data types, including genomic, phylogenetic, proteomic, structural and transcriptomic data (Franke et al., 2006; Lee et al., 2011; Greene et al., 2015; Tasan et al., 2015). The edges in these networks often represent the probability that a functional interaction occurs between two genes.

All of the previously described methods have limitations. Methods that detect

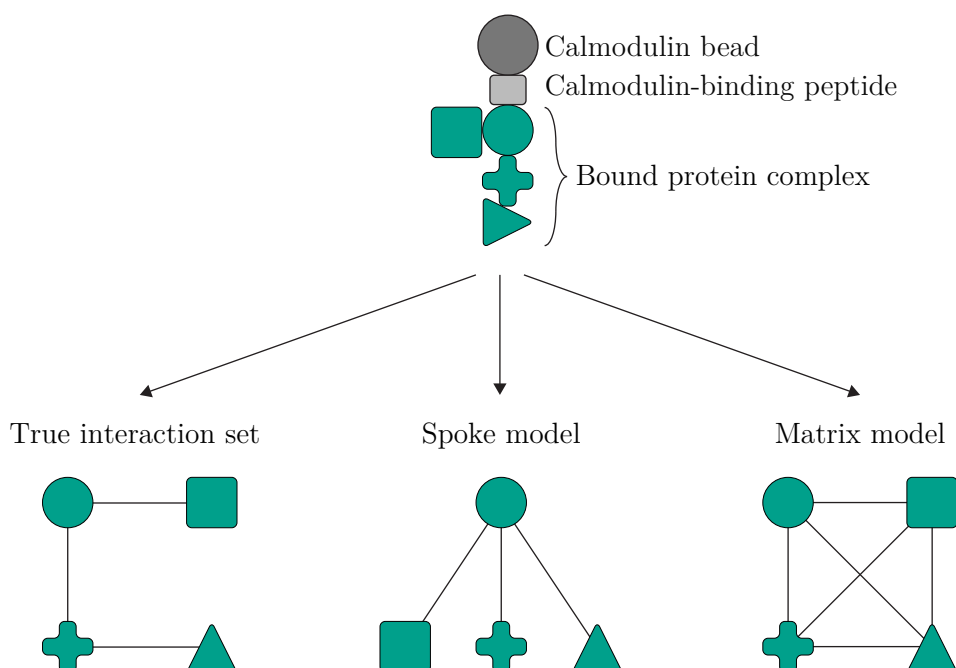


Figure 1.5: Inferring edges representing binary PPIs from data produced in TAP-MS using the spoke and matrix models. The green circle represents the bait protein and the other green shapes represent different proteins that form the complex. The bait protein is bound to a calmodulin-binding peptide, which is itself bound to a calmodulin bead. In the spoke model, interactions are included between the bait protein and all other proteins. In the matrix model, interactions are included between all pairs of proteins. If the complex contains more than two proteins, then the true set of interactions cannot be inferred using the data produced by the TAP-MS method.

physical interactions between proteins often require a certain set of conditions and therefore interactions that cannot take place under these conditions may not be identified (Braun et al., 2009). For example, interactions identified by Y2H assays need to be able to take place in yeast cells in order to be detected. Extracellular interactions may therefore not be identified. This has led to the development of new methods, such as Avidity-Based Extracellular Interaction Screens (AVEXIS), that test for extracellular PPIs (Bushell et al., 2008). It is likely that a combined set of methodologies will need to be used to build a complete map of the physical and functional interactions that occur between proteins and genes across biological

Resource	Genes	Interactions	Resource reference
BioGRID	15,098	151,883	Chatr-Aryamontri et al. (2015)
HI-II-14	4,210	13,374	Rolland et al. (2014)
IntAct	12,078	133,588	Orchard et al. (2014)
PrePPI	11,815	80,499	Garzo et al. (2013)
STRING	16,858	1,736,931	Szklarczyk et al. (2015)

Table 1.3: The number of human genes and the number of interactions between these genes in five of the most comprehensive PPI databases. Some PPI databases do not report different protein isoforms and I therefore mapped all interactors to Ensembl gene identifiers and report the number of these genes and interactions between these genes in the database. Table A.2 contains details of when and from where these data were downloaded. I used STRING version 10.0. Only interactions between two humans interactors are included. I removed all interactions in BioGRID and IntAct that were not marked as direct interactions, associations or direct associations. I considered all interactions in STRING with an experimental score greater than zero.

contexts.

1.3.2 PPI databases

As previously mentioned, PPIs have been identified in both HT and LT studies. The largest HT screen completed so far in humans used Y2H assays to identify 13,374 PPIs (Rolland et al., 2014). Unlike LT studies, HT screens do not focus on genes of interest and therefore the data generated in these studies are less likely to be biased towards certain genes (Rolland et al., 2014). Questions have been raised however about the accuracy of the interactions identified in some HT screens (Braun et al., 2009).

While LT studies may be biased towards genes of interest, they have identified tens of thousands of PPIs (Szklarczyk et al., 2015). A number of databases have been established to curate the PPIs reported in the literature and make them available for use by the scientific community (Table 1.3). The MIntAct Project recently added

data from a number of these databases to the IntAct database (Orchard et al., 2014). Outside of the MIntAct Project exists the BioGRID database, which curates PPIs, functional interactions, interactions between genes and chemicals and post-translational modifications (Chatr-Aryamontri et al., 2015).

There also exists a number of databases that do not themselves curate published literature, but instead collate interactions reported in other databases and integrate these data with other biological data. These databases include STRING (Szklarczyk et al., 2015) and PrePPI (Garzo et al., 2013). STRING contains experimentally determined PPIs reported in a set of databases that includes BIND (Bader et al., 2003), BioGRID (Chatr-Aryamontri et al., 2015), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), IntAct (Kerrien et al., 2012) and MINT (Licata et al., 2012). STRING uses a naive Bayesian approach to combine these data with genomic and transcriptomic data and interactions identified using automated text mining, to predict interactions and generate confidence scores for the experimentally determined interactions (von Mering et al., 2005). Using STRING, it is therefore possible to select only the highest-confidence experimentally determined PPIs. Much like STRING, PrePPI uses a Bayesian framework to integrate evolutionary, functional, structural and transcriptomic data to predict PPIs and generate confidence scores. PrePPI however collates a smaller set of databases. The experimentally determined interactions it considers have also not been updated since August 2011. PrePPI therefore contains fewer experimentally determined interactions than STRING.

1.3.3 Context-specific networks

It is understood that some genes are expressed only in certain contexts and may therefore only perform functions in these contexts (Velculescu et al., 1999). It follows that functional and physical interactions involving these context-specific genes may also be context-specific. However, many of the methods used to detect interactions are unable to determine whether the detected interactions are ubiquitous or context-specific. This has led to the development of methods that integrate multiple types of data to generate context-specific interaction networks. In the literature, tissues

and cell types are often referred to interchangeably. In this thesis, I therefore refer to tissues and cell types collectively as contexts.

A number of studies have generated context-specific PPI networks by integrating PPI data from HT and LT studies and gene expression data from the contexts of interest (Table 1.4). This has been done using two approaches: vertex removal and edge reweighting (Figure 1.6) (Magger et al., 2012). In the vertex removal approach, proteins and their associated interactions are removed if the corresponding genes are expressed below a certain threshold in the context (Bossi & Lehner, 2009; Lopes et al., 2011; Magger et al., 2012; Barshir et al., 2014; Liu et al., 2014b; Kotlyar et al., 2016). In the edge removal approach, edges in the network are reweighted based on the expression scores of the corresponding genes, so that the weights of edges connecting highly expressed genes are greater than the weights of edges connecting less-highly expressed genes (Magger et al., 2012). Both of these approaches rely on the hypothesis that interactions between proteins are less likely to occur in a context if the protein is not present in the context or present at lower levels.

Using these context-specific networks, a number of observations have been made about the functions that genes and proteins perform in different contexts. Housekeeping genes and proteins are defined as those that are expressed ubiquitously across tissues and cell types and are therefore believed to perform crucial functions in maintaining the mechanisms that support cellular life (Velculescu et al., 1999). Bossi & Lehner (2009) divided proteins into housekeeping and tissue-specific proteins, depending on whether they were expressed in the majority of samples from a panel of tissues. They noted that housekeeping proteins interacted with both housekeeping and tissue-specific proteins, suggesting that even housekeeping proteins are involved in tissue-specific functions. Barshir et al. (2014) analysed the expression of known disease-associated genes across tissue samples and found that whilst the majority of these diseases are manifested in only a limited number of tissues, many of the associated genes are housekeeping genes and therefore expressed in the majority of tissues. Barshir et al. next looked at the degree of these disease-associated genes across 16 tissue-specific PPI networks and observed that disease-associated genes tended to have a higher degree in networks specific to tissues in which the associated

Study	Interaction type	Data used	N. contexts
Bossi & Lehner (2009)	Physical	PPI, Tr	79
Lee et al. (2009)	Physical	PPI, Tr	79
Lopes et al. (2011)	Physical	PPI, Tr	84
Magger et al. (2012)	Physical	PPI, Tr	60
Guan et al. (2012)	Functional	F, Phe, PPI, Tr	107
Schaefer et al. (2013)	Physical	PPI, Tr	79
Lundby et al. (2014)	Physical	PPI	1
Barshir et al. (2014)	Physical	A, PPI, Tr	16
Li et al. (2014)	Physical	PPI, Tr	79
Greene et al. (2015)	Functional	Per, PPI, S, Tr	144
Kotlyar et al. (2016)	Physical	A, PPI, Tr	29

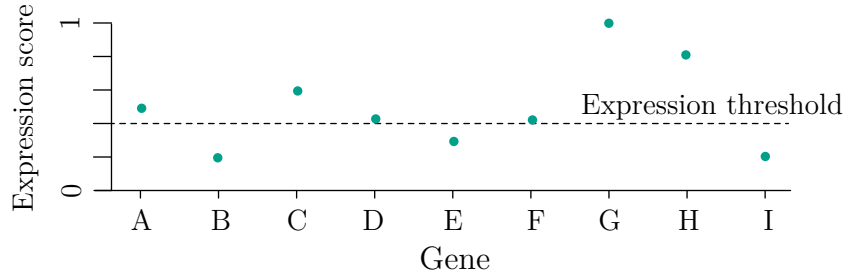
Table 1.4: Details of studies that have generated context-specific physical and functional interaction networks. Studies are ordered by publication date. Given is the type of interaction that the networks represent, the data used to generate the networks and the number of context-specific networks generated in each study. A: Protein abundance, F: Functional, Per: Chemical and genetic perturbation, Phe: Phenotypic, PPI: Protein-protein interaction, S: Sequence, Tr: Transcriptomic.

disease was manifested. They therefore concluded that disruption to tissue-specific interactions may lead to the development of the disease phenotype in the disease-manifesting tissues.

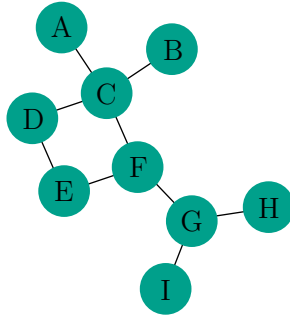
While the majority of PPIs identified in HT and LT studies are not associated with any specific context or contexts, a limited number of interactions known to occur in specific contexts have been identified. Lundby et al. (2014) used quantitative interaction proteomics and cardiac tissue from mice to identify 670 PPIs that occur in mouse cardiac tissue. The large number of distinct tissues and cell types present in mammals and the costs of this technology currently prohibit it from being used to generate proteome-wide context-specific PPI networks.

Guan et al. (2012) generated mouse functional interaction networks for 107

A) Gene expression data



B) Generic PPI network



C) Context-specific PPI networks

Vertex removal

Edge reweight

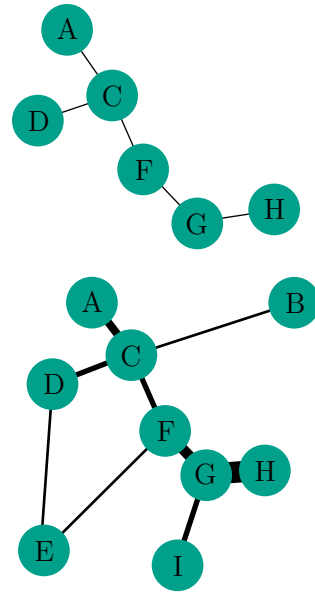


Figure 1.6: Generating context-specific PPI networks through the integration of gene expression and PPI data using the vertex removal and edge reweighting approaches. (A) Gene expression data. In this illustrative example, expression values have been normalised to range between 0 and 1. (B) Generic PPI network. (C) Context-specific PPI networks generated using the two approaches. In the vertex removal approach, vertices corresponding to genes expressed below a certain threshold (the dashed line) are removed from the network along with all associated edges. In the edge reweighting approach, edges are weighted based on the expression scores of the interacting genes, so that edges connecting highly expressed genes are given higher weights than edges connecting less-highly expressed genes. In the figure, edge thickness corresponds to the weight of the edge, so that higher-weight edges are thicker. The length of each edge is proportional to the reciprocal of its weight.

contexts. They used a number of different data types, including phenotypic, PPI and transcriptomic data. These data were integrated using a set of gold standard interactions specific to each context. Gold standard interactions were generated by identifying pairs of genes known to be expressed in each context and found in the same Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway (Kanehisa & Goto, 2000) or annotated with overlapping sets of Gene Ontology (GO) terms (The Gene Ontology Consortium, 2014). For each context, they used the corresponding set of gold standard interactions to score each data source by its relevance to the context. Some data sources were thought to represent the same information and therefore Guan et al. normalised the data sets to account for this. A Bayesian framework was used to integrate these data and generate a functional interaction network for each context. Using these networks, Guan et al. were able to investigate the different roles that housekeeping genes perform in different contexts. WNT10B is expressed in many tissues during development, but has different interacting partners in the muscle and bone networks, reflecting the distinct functions that the gene performs in muscle development and bone formation.

Greene et al. (2015) used the same approach to generate functional interaction networks for 144 human contexts. In total, they combined data from over 14,000 publications. Greene et al. also developed the NetWAS method, which uses these networks and data from GWAS to identify disease-associated genes (see Section 1.4.6 for further details).

1.4 Methods for identifying causal genes

Causal variants and genes are often difficult to identify in GWAS because of LD between variants and the short and long range regulation of genes. A single trait-associated SNP may implicate tens of potential causal genes (Raychaudhuri et al., 2009) and functional analysis of all of these genes may be both time consuming and expensive. Bioinformatics methods have therefore been developed to exploit additional sources of biological data to identify the genes most likely to be causal. As the quality and coverage of omics data sets continues to improve, new methods will need to be developed to use these data. In this section I discuss some of the

previously proposed gene prioritisation methods and the data they use (Table 1.5). Some of these methods prioritise genes in trait-associated loci identified in GWAS and others conduct genome-wide gene prioritisation.

Different methods use different approaches to study causal mechanisms. While some methods attempt to identify individual genes that cause a trait of interest, others attempt to improve their power by identifying networks of causal genes. Leiserson et al. (2013) named these two strategies ‘causal gene identification’ and ‘causal network identification’. A number of methods apply the causal network identification strategy to GWAS data. These methods include PINBPA, which uses a greedy search algorithm to identify sub-networks in PPI networks enriched with genes containing trait-associated SNPs (Baranzini et al., 2009), and NETBAG+, which uses a greedy search algorithm, copy number variation (CNV) data and phenotypic data to identify networks of functionally related genes that may underpin the trait of interest (Gilman et al., 2012). PINBPA and NETBAG+ have identified networks enriched with genes associated with autism, MS and schizophrenia (Baranzini et al., 2009; Gilman et al., 2012). While methods that use this causal network identification strategy have been seen to be successful, they prioritise sets of genes rather than individual genes and therefore require different benchmarking approaches to methods that use the causal gene identification strategy. In this thesis, I identify causal genes using the causal gene identification strategy and I therefore, in this section, focus on discussing methods that employ this strategy.

Gene prioritisation methods use a range of data types, including epigenetic, functional, genomic, PPI, phenotypic, proteomic, text-mined and transcriptomic data, from both humans and other organisms. In this section I also discuss the data types used by the most successful gene prioritisation methods. The list of methods covered in this section is not exhaustive and additional methods are discussed in the reviews by Bromberg (2013) and Leiserson et al. (2013).

1.4.1 Text mining

Biological information has been extracted from a number of different textual data sources, including published literature (Raychaudhuri et al., 2009; Pletscher-Frankild

Method name	Data used	Reference
Freudenberg2002	F, Phe	Freudenberg & Propping (2002)
Krauthammer2004	PPI	Krauthammer et al. (2004)
ENDEAVOUR	F, Pa, Pr, PPI, R, S, Te, Tr	Aerts et al. (2006)
Oti2006	PPI	Oti et al. (2006)
Prioritizer	F, GI, Pa, PPI, Tr	Franke et al. (2006)
Lage2007	Phe, PPI	Lage et al. (2007)
GeneWanderer	PPI	Köhler et al. (2008)
GRAIL	Te	Raychaudhuri et al. (2009)
RWRH	Phe, PPI	Li & Patra (2010)
PRINCE	Phe, PPI	Vanunu et al. (2010)
DAPPLE	PPI	Rossin et al. (2011)
HumanNet	GI, Phy, Pr, PPI, S, Te, Tr	Lee et al. (2011)
MouseFinder	Phe	Chen et al. (2012)
Magger2012	Phe, PPI, Te, Tr	Magger et al. (2012)
Jacquemin2013	Phe, PPI, Te, Tr	Jacquemin & Jiang (2013)
OPEN	Phy, Pr, R, Tr	Deo et al. (2014)
ExomeWalker	PPI	Smedley et al. (2014)
Li2014	PPI, R, Tr	Li et al. (2014)
Lundby2014	PPI	Lundby et al. (2014)
DEPICT	F, Pa, Phe, PPI, R, Tr	Pers et al. (2015)
PrixFixe	F, GI, Phy, Pr, PPI, R, Tr	Tasan et al. (2015)
NetWAS	F, GI, PPI, R, Tr	Greene et al. (2015)

Table 1.5: A non-exhaustive list of previously developed gene prioritisation methods. If the method was not named in the original study then I have named the method by concatenating the name of the first author of the study and the year of publication. Some of these methods prioritise genes in loci and others conduct genome-wide gene prioritisation. Each of these methods also use known disease-associated genes or data from GWAS. GI: Genetic interaction, F: Functional, Pa: Pathway, Phe: Phenotypic, Phy: Phylogenetic, Pr: Protein domain, PPI: Protein-protein interaction, R: Regulatory, S: Sequence, Te: Text mined, Tr: Transcriptomic.

et al., 2015; Hoehndorf et al., 2015a), clinical synopses (Lage et al., 2007) and patient medical records (Jensen et al., 2014a). This has been done through manual curation of these resources and through automated text mining. Many biological databases, such as the disease variant databases ClinVar (Landrum et al., 2016), OMIM (Amberger et al., 2015) and UniProtKB/Swiss-Prot (Yip et al., 2008) (discussed in Section 1.2.2) and the network databases BioGRID (Chatr-Aryamontri et al., 2015) and IntAct (Orchard et al., 2014) (discussed in Section 1.3.2), contain data gleaned from published literature, primarily through manual curation. The labour-intensive nature of manual curation (Famiglietti et al., 2014) has led however to the development of a number of methods that use automated text mining to identify relationships between various biological entities, including diseases and genes (Raychaudhuri et al., 2009; Pletscher-Frankild et al., 2015), variants (Pers et al., 2013), microRNAs (Xie et al., 2013) and associated phenotypes (Hoehndorf et al., 2015a).

Perez-Iratxeta et al. (2002) developed one of the first methods that systematically mapped diseases to genes using text mining. Their approach uses Medical Subject Heading (MeSH) terms, which describe the topics covered in each article indexed in PubMed, to link disease phenotypes to genes through known chemistry and gene functions. Gene Relationships Among Implicated Loci (GRAIL) (Raychaudhuri et al., 2009) built upon this approach by directly extracting terms from article titles and abstracts, instead of relying on MeSH terms, which are limited in number. Extracting terms directly from free text raises a number of problems however (Pletscher-Frankild et al., 2015). For example, the poor standardisation of some terms, such as diseases, tissues and gene names, makes it difficult to identify different terms that correspond to the same concept. This has led to methods such as DISEASES (Pletscher-Frankild et al., 2015) to employ dictionary-based approaches that use controlled vocabularies to identify sets of terms that map to the same concepts. The development of methods such as Aber-OWL (Hoehndorf et al., 2015b), which uses text mining to identify published literature relating to terms from various biomedical ontologies, will facilitate the use of text mining to characterise relationships between other biological entities.

Another problem with the previously described text mining methods is that they tend to rely on the simple co-occurrence of annotations or terms and therefore cannot infer the nature of the relationships between entities (Pletscher-Frankild et al., 2015). For example, co-occurrence-based methods are unable to distinguish between positive and negative relationships and will therefore identify both the sentences ‘IL2RA is associated with psoriasis’ and ‘IL2RA is not associated with psoriasis’ as evidence that IL2RA is associated with psoriasis. This has led to the development of methods that parse text using approaches such as natural language processing (NLP) to more accurately characterise relationships between entities (Jensen et al., 2014b; Lee et al., 2016). Methodological improvements in fields such as NLP and the improved standardisation of biomedical terms, including those proposed by the Open Biomedical Ontologies (OBO) Foundry (Smith et al., 2007), will further our ability to mine biological information from free text.

1.4.2 Physical interactions

Functionally related genes and proteins tend to form interconnected modules in PPI networks (Hartwell et al., 1999). It has been suggested that disruption to these functional modules may lead to distinct diseases (Barabási et al., 2011). This hypothesis is supported by the observation that proteins associated with the same disease are more likely to interact in PPI networks than proteins associated with different diseases (Goh et al., 2007; Bauer-Mehren et al., 2011). This tendency of proteins associated with the same disease to interact makes it possible to use PPI networks to prioritise disease-associated genes. These network-based methods score proteins that interact with known disease-associated proteins higher than proteins that do not interact with known disease-associated proteins; an approach known as ‘guilt-by-association’ (Tranchevent et al., 2011). Methods tend to use a one-to-one mapping between proteins and genes, as many PPI databases do not contain data for specific protein isoforms (Mishra et al., 2006; Jensen et al., 2009).

Krauthammer et al. (2004) developed one of the first methods that used the guilt-by-association principle to prioritise disease-associated genes. They first generated a PPI network through automated mining of published literature. Krauthammer et al.

next identified known disease-associated genes. They then measured the distance between the protein products of these genes and the protein products of other genes in the network using the shortest paths distance measure. Genes were then ranked by the distance between their protein products and the protein products of known disease-associated genes, so that genes positioned closer were ranked higher. Through manual examination of the results, Krauthammer et al. concluded that their method successfully identifies potential disease-associated genes.

The establishment of databases of known biological interactions and the release of HT PPI data sets led to the development of a number of PPI-based gene prioritisation methods. Oti et al. (2006) developed a method using interactions from a number of databases, including HPRD. Like Krauthammer et al. (2004), Oti et al. used the guilt-by-association principle to prioritise genes. Unlike Krauthammer et al. however, Oti et al. did not rank proteins using the shortest paths distance measure, but instead ranked proteins by the numbers of disease-associated proteins that they directly interacted with. This approach is known as the direct neighbourhood distance measure.

Alongside the previously mentioned shortest paths and direct neighbourhood distance measures, a number of other methods have been developed to measure the distances between vertices in networks. These distance measures are also sometimes referred to as label propagation algorithms, as they can be thought of as propagating a label, score or weight across a network from higher-weight vertices to lower-weight vertices. Distance measures can be divided into two classes: local distance measures, that consider only direct interactions, and global distance measures, that also consider indirect interactions. Local distance measures include the direct neighbourhood measure (Figure 1.7A) and global distance measures include the shortest paths measure (Figure 1.7B) and the random walk with restart (RWR) measure (Figure 1.7C).

Köhler et al. (2008) used leave-one-out cross-validation to compare the performance of these measures when used to prioritise disease-associated genes. The performance of the local direct neighbourhood measure was relatively poor, possibly due to the measure's inability to rank proteins that did not directly interact with

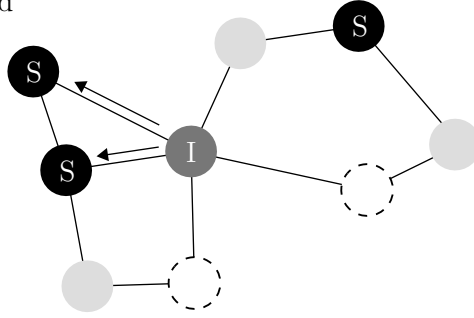
the protein product of any known disease-associated gene. The global RWR method performed best, leading Köhler et al. to develop the RWR-based GeneWanderer gene prioritisation method. Navlakha & Kingsford (2010) conducted similar analyses and also observed the best performance amongst random-walk-based distance measures. Random-walk-based approaches have been used in a number of other prioritisation methods, including the RWRH (Li & Patra, 2010) and ExomeWalker (Smedley et al., 2014) methods.

PPI networks have also been used to identify causal genes in trait-associated loci. DAPPLE identifies PPI sub-networks that connect genes in loci (Rossin et al., 2011). Rossin et al. applied DAPPLE to GWAS of height, lipid levels and T2D and noted that genes in height and lipid-level-associated loci were enriched with PPIs connecting them to genes in other associated loci. The genes that formed these loci-connecting sub-networks also tended to be over-expressed in disease-related tissues, providing evidence that these sub-networks explain the molecular mechanisms that link the genes and identify them as being causal.

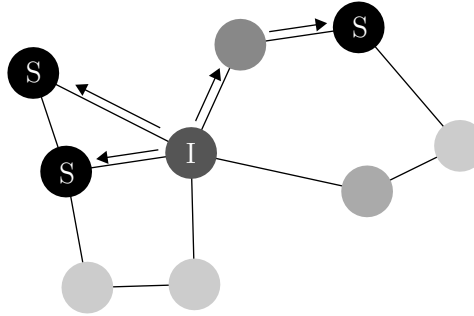
While it has been demonstrated that PPI networks can aid in the prioritisation of disease-associated genes, a number of limitations of the data, such as accuracy and completeness, are likely to limit the performance of PPI-based methods (Menche et al., 2015). Furthermore, for some proteins, no interactions have yet been identified (Chatr-Aryamontri et al., 2015), meaning that the corresponding genes cannot be prioritised using this approach.

There has also been criticism of the use of the guilt-by-association principle with PPI networks in general. Gillis & Pavlidis (2012) noted that the majority of the interactions in PPI and genetic interaction (GI) networks do not link functionally related proteins and genes. Instead, a small number of ‘critical’ edges link the functionally related components. Study bias may also be artificially inflating the performance of PPI-based gene prioritisation methods. Das & Yu (2012) used literature curation to generate the HINT PPI database. Studies that contributed to the database were categorised as HT or LT, depending on whether they reported 100 or more interactions. Das & Yu generated two PPI networks, one containing only interactions reported in HT studies and one containing only interactions reported

A) Direct neighbourhood



B) Shortest paths



C) Random walk with restart (RWR)

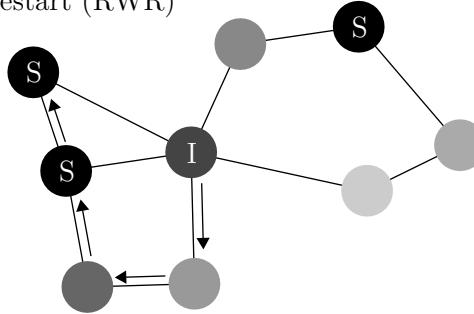


Figure 1.7: Three network distance measures that have been used to prioritise disease-associated genes. Seed vertices, which may be associated with the disease of interest, are marked with an *S*. The remaining vertices are coloured by their distance to the seed set using each measure. Darker vertices are located closer. A vertex of interest is marked with an *I* and arrows indicate paths used to measure the distance between the vertex of interest and the seed set. (A) In the direct neighbourhood measure, only direct interactions are considered and the measure is therefore unable to score the dashed vertices. (B) In the shortest paths measure, the shortest paths between the vertex of interest and the seed set are used as a measure of distance. (C) In the RWR measure, distance is measured by simulating random walks (see Section 2.6.1). For simplicity, only a single walk of length four is shown.

in LT studies. In the HT PPI network, there was no significant correlation between protein degree and the number of studies reporting interactions involving each protein. Conversely, in the LT PPI network, proteins with a large degree tended to be reported in more studies, demonstrating that some proteins tend to be revisited more often in LT studies.

1.4.3 Functional relationships

Networks of functional interactions have also been used to prioritise disease-associated genes. Franke et al. (2006) developed Prioritizer, which uses a functional interaction network to identify functionally related genes across multiple disease-associated loci. Franke et al. hypothesised that the disruption of these shared functions may lead to the development of the disease. A Bayesian classifier was used to integrate gene expression, functional and PPI data and generate a functional interaction network. To identify functionally related genes across trait-associated loci, the distances between each gene in each trait-associated locus and genes in all other trait-associated loci were measured using the shortest paths distance measure. These distances were then used to score the genes so that genes located nearer genes in other loci in the functional interaction network were scored higher than genes located further away. Using cross-validation and disease-gene associations from OMIM, Franke et al. demonstrated that Prioritizer successfully uses functional interaction networks to identify causal genes.

A functional interaction network was also used by Lee et al. (2011) to prioritise disease-associated genes. As well as gene expression and PPI data, Lee et al. used gene co-citation, domain co-occurrence, gene neighbour, phylogenetic and structural data to generate the functional interaction network. Lee et al. tested how well each individual data source predicted functional interactions between genes and found that all sources were able to identify interactions with varying performance. Performance was best when all data sources were integrated, demonstrating that integrating multiple data sources best identifies functionally related genes. Lee et al. also tested the performance of various label propagation algorithms when prioritising genes and found that global distance measures performed better than local distance

measures, supporting the observations made by Köhler et al. (2008) and Navlakha & Kingsford (2010) and discussed in the previous section.

More recently, Tasan et al. (2015) developed PrixFixe, which uses a functional interaction network to identify functionally related genes across trait-associated loci. PrixFixe takes trait-associated SNPs as input, which it uses to define trait-associated loci. Each locus may span multiple candidate genes. PrixFixe selects a single gene from each of these loci. The aim of PrixFixe is to find the set of genes, one from each locus, that is most densely connected in the functional interaction network. A genetic algorithm is used to identify this densely connected gene set. The algorithm starts with 5,000 randomly generated gene sets and runs over a number of ‘generations’, each of which contains a ‘mutation’ step and a ‘mating’ step. In the mutation step, each gene in each gene set is switched with a gene from the same locus at a 5% probability. Each gene set is then given a ‘fitness’ score, which is based on how densely connected the gene set is in the functional interaction network. In the mating step, 5,000 new gene sets are generated by mating pairs of gene sets. These offspring gene sets are generated by randomly selecting one gene from each locus from each parent. When selecting pairs of gene sets to mate, gene sets with higher fitness scores are preferentially chosen, so that the fitness of the gene sets increases over the generations. If the mean fitness of the 5,000 gene sets does not increase across a generation, then the genetic algorithm is terminated. Finally, each gene in each locus is scored by the number of gene sets they appear in. Tasan et al. applied PrixFixe to a number of GWAS and demonstrated that the method is able to identify gene sets enriched with known disease-associated genes and genes involved in disease-related pathways and processes. Like Franke et al. (2006) and Lee et al. (2011), Tasan et al. integrated multiple data sources to generate their network, which covers about 19,000 protein-coding genes. This illustrates an advantage of using functional interaction networks over PPI networks: functional interaction networks tend to cover a greater number of genes and are therefore able to prioritise a greater number of genes.

1.4.4 Phenotypic similarities

Phenotypic data have also been used to identify causal genes. These data are often used under the hypothesis that phenotypically similar diseases tend to be associated with functionally similar genes. Freudenber g & Propping (2002) developed one of the first methods that used phenotypic similarities between diseases to prioritise disease-associated genes. They first clustered diseases in OMIM using the attributes associated with each disease record, which included details about disease etiology and the affected tissues. They next pooled the genes known to be associated with each disease in each cluster and identified functions shared across these genes by measuring GO term enrichment (The Gene Ontology Consortium, 2014). To prioritise genes associated with a particular disease, Freudenber g & Propping then identified the disease cluster that was phenotypically most similar to the disease of interest. By noting the functions of genes associated with this disease cluster, Freudenber g & Propping were then able to prioritise genes with similar functions. Through cross validation, Freudenber g & Propping showed that their method ranked the true disease-causing gene in the top 3% of genes 33% of the time, demonstrating the potential of using phenotypic data to study the genetic basis of disease. An advantage of this and similar phenotype-based gene prioritisation methods is that no genes have to have been previously identified as being associated with the disease of interest in order to prioritise genes.

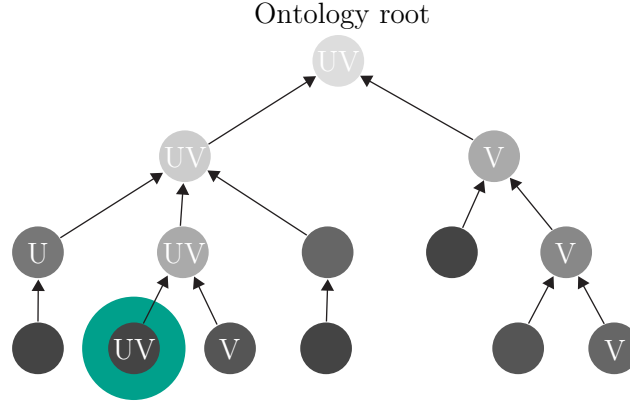
Lage et al. (2007) combined PPI and phenotypic data to identify disease-associated protein complexes. Like Freudenber g & Propping (2002), Lage et al. used disease records in OMIM to measure the phenotypic similarity of diseases. However, unlike Freudenber g & Propping, Lage et al. measured phenotypic similarity by identifying occurrences of clinically relevant UMLS terms in the text and clinical synopsis parts of each disease record. The number of occurrences of each UMLS term in each record was then transformed into a phenotypic vector. The phenotypic similarity of two diseases could then be quantified by measuring the similarity of the two phenotypic vectors, using the cosine of the angles between the vectors after normalisation. For a particular disease of interest, Lage et al. prioritised protein complexes by identifying complexes containing proteins associated with diseases

phenotypically similar to the disease of interest. The use of text mining allowed for the exploitation of the text and clinical synopsis parts of the disease records in OMIM, which are often richer than the record attributes. However, the size of this text and the number of medical terms that it contains varies between OMIM records, with the records of better-studied diseases tending to be more detailed (Wang et al., 2010). This variability has been seen to introduce bias into phenotypic similarity scores computed using this method and result in better-studied diseases being scored as phenotypically more similar (Wang et al., 2010). Furthermore, OMIM has traditionally focused on Mendelian disorders and therefore alternative data sources may be required to accurately quantify the phenotypic similarity of complex diseases.

PRINCE combines a network-based approach and phenotypic similarity scores computed by van Driel et al. (2006) to prioritise disease-associated genes (Vanunu et al., 2010). Like Lage et al. (2007), van Driel et al. measured the phenotypic similarity of diseases by mining the text of disease records in OMIM. Disease-gene associations were extracted from the GeneCards database (Safran et al., 2010). In PRINCE, genes are first scored by the phenotypic similarity between the diseases they are known to be associated with and the disease of interest. If a gene is not known to be associated with any disease, then it is given a score of zero. These scores are then applied to a PPI network. PRINCE next uses a label propagation algorithm to propagate these scores across the network, so that proteins that directly or indirectly interact with high-scoring proteins are also scored highly. Vanunu et al. compared the performance of PRINCE against GeneWanderer (Köhler et al., 2008), which also applies a label propagation algorithm to a PPI network to prioritise genes, but does not use phenotypic data. PRINCE significantly outperformed GeneWanderer in cross-validation, demonstrating the advantage of using phenotypic data to prioritise disease-associated genes.

The recently developed PhenIX method uses phenotypic data to identify genetic variants that may cause a disease of interest in an individual (Zemojtel et al., 2014). Unlike the previously described methods however, PhenIX uses an ontology-based approach to measure the phenotypic similarity of diseases, instead of text mining.

A) Higher MICA semantic similarity score



B) Lower MICA semantic similarity score

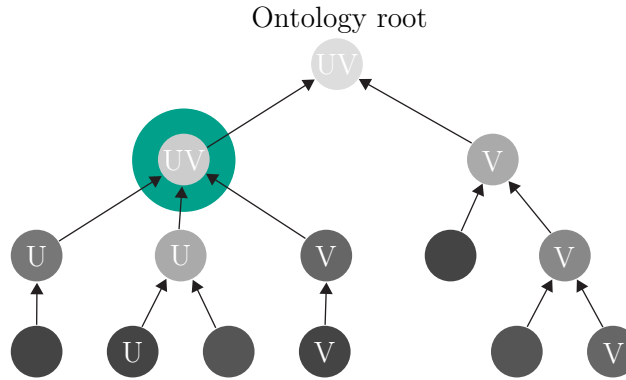


Figure 1.8: Measuring the similarity of two sets of disease-associated phenotypes using the MICA semantic similarity metric. Vertices represent terms in a phenotype ontology. Terms from each set are annotated with a U and a V respectively. Terms that are contained in both sets are annotated with both a U and a V. Because of the true path rule, all ancestors of a disease-associated term are also said to be associated with the disease. Terms are weighted by their IC. The darker the vertex, the higher the IC of the term. The larger green circles highlight the most informative common ancestor in each example. The most informative common ancestor in (A) has a higher IC than the most informative common ancestor in (B). The two diseases in (A) would therefore be identified as being phenotypically more similar than the two diseases in (B).

To compute the similarity of sets of phenotypes, PhenIX uses the Human Phenotype Ontology (HPO), which contains terms for more than 10,000 phenotypic abnormalities that occur in humans (Köhler et al., 2014). HPO terms have been systematically mapped to many diseases (Köhler et al., 2014). To compare the similarity of two sets of phenotypes, all terms in the HPO are first scored by how ‘informative’ they are. A phenotypic term is considered more informative if it is associated with fewer diseases and therefore more specific. An IC score is produced for each HPO term by computing the negative logarithm of the proportion of the diseases in the PhenIX data set associated with each term (Köhler et al., 2009). This means that the fewer diseases associated with a term, the greater its IC score. If a disease is associated with a term in the HPO, then it follows under the true path rule, that the disease is also associated with all ancestors of the term (Pesquita et al., 2009). This means that IC scores decrease towards the root of the ontology. To compare the phenotypic similarity of two sets of phenotypes, all ancestral terms are first identified and added to the sets. The intersection of these two sets is then identified and the greatest IC score amongst the terms in the intersection is used to score the phenotypic similarity of the sets. Therefore, if two diseases are both associated with a highly specific phenotypic term, then they are given a higher phenotypic similarity score. This method is known as the MICA semantic similarity metric (Figure 1.8).

Using ontologies to measure the phenotypic similarity of diseases avoids some of the problems of text mining, such as vocabulary synonyms and the over-use of certain medical terms. They may therefore more accurately measure phenotypic similarity (Wang et al., 2010; Schofield et al., 2010).

1.4.5 Data from model organisms

Model organisms have long been used to study human disease (Götz et al., 1995; Withers et al., 1998). The development of databases of genotype and phenotype data from model organisms has more recently facilitated the systematic and large-scale use of model organism data. As described in Section 1.2.3, the MGD is one of the largest such databases and has been used in a number of methods to study

human disease (Bult et al., 2016).

Mouse models of human disease have been traditionally generated using sequence similarity data. If a gene is known to be associated with a human disease, then it may be possible to generate a mouse model of the disease by mutating the gene. This approach has been used to generate mouse models of diseases such as Alzheimer’s disease (Götz et al., 1995) and T2D (Withers et al., 1998). While this approach has been successful, it relies upon knowledge of the genetic basis of the disease. A number of methods have therefore been developed that take the reverse approach and instead search for potential disease-associated genes by identifying mouse alleles known to produce similar phenotypes.

Early attempts to systematically use data from model organisms to study human disease were hampered by the lack of formal vocabularies describing phenotypic abnormalities across species. Phenotypes tended to be described using free-text, making systematic comparisons difficult. Washington et al. (2009) therefore developed the Uberon anatomical ontology, by integrating a number of species-specific and cross-species ontologies. Washington et al. annotated 11 gene-linked human diseases reported in OMIM with anatomical terms and used these terms to compare the diseases to alleles reported in a number of model organism databases. By quantifying the similarity of these sets of anatomical terms using an ontology-based approach similar to MICA (Zemojtel et al., 2014), Washington et al. were able to identify genes in model organisms orthologous to known human disease-associated genes. This demonstrated that model organism data could be used to prioritise diseases-associated genes in humans.

Robinson et al. (2008) used the text and clinical synopsis parts of disease records in OMIM to annotate human diseases with HPO terms. Chen et al. (2012) later developed MouseFinder, which uses these annotations and an ontology-based similarity metric to identify mouse genes associated with phenotypes similar to a disease of interest. The method can then prioritise the human orthologs of these genes. Chen et al. used disease-associated genes from OMIM to demonstrate that MouseFinder is able to prioritise disease-associated genes in humans. The focus of OMIM on Mendelian disorders may however limit the power of MouseFinder to

identify genes associated with complex diseases.

While mouse data have been successfully integrated with human data and used to prioritise disease-associated genes, there are a number of limitations to these data. As of 19 March 2015, there are 9,900 human genes with mouse orthologs with phenotypic data in the MGD. This means that many human genes cannot currently be studied using mouse data alone. Furthermore, the amount of phenotypic data associated with mouse alleles in the MGD is not uniform. Some alleles are associated with tens of terms and some alleles are associated with only a single term (Bult et al., 2016). The completion of the IMPCs large-scale mouse phenotyping project should help resolve the uneven coverage of the mouse genome and phenome in the MGD (see Section 1.2.3 for further details).

Differences in mouse and human genomics and physiology also limit the use of mouse data when studying some human diseases. For example, the genomic response to inflammation differs between mice and humans, which may explain why many drug candidates for inflammatory diseases tested in mice perform poorly in human trials (Seok et al., 2013). However, as the amount of genotypic and phenotypic data in model organism databases continues to increase, these resources will become increasingly useful for studying human disease.

1.4.6 Context-specific data

Many network-based gene prioritisation methods (such as those discussed in Sections 1.4.2 and 1.4.3) use generic networks that contain interactions that occur throughout the body. Diseases however tend to manifest in a limited number of tissues and cell types. There has therefore been interest in developing gene prioritisation methods that use interactions specific to the tissues and cell types in which a disease manifests.

Magger et al. (2012) integrated gene expression and PPI data to generate context-specific PPI networks (see Section 1.3.3 for further details). After generating these networks, Magger et al. incorporated them into PRINCE (Vanunu et al., 2010), which originally used generic PPI networks to prioritise genes, to determine whether PRINCE performs better when run using context-specific networks. Magger et al. conducted cross validation using sets of disease-associated genes from the GeneCards

database (Safran et al., 2010) to compare the performance of PRINCE using generic and context-specific PPI networks. For each disease, a suitable context was identified using the disease-context association data generated by Lage et al. (2008) by mining the PubMed database (see Section 1.5.1 for further details). The performance of PRINCE was significantly better when run using the context-specific PPI networks. This demonstrates that using data specific to a disease-associated context can aid in the prioritisation of disease-associated genes. Although PRINCE is available to download, Magger et al. have not made their context-specific version of PRINCE available for use. As previously mentioned, PRINCE was developed using data from OMIM and it is therefore unclear whether it would perform as well on complex diseases as it does on Mendelian disorders. Magger et al. used gene expression data from heterogeneous tissue samples to generate their context-specific PPI networks. Further work needs to be completed to determine whether the use of data from less heterogeneous samples would improve method performance.

The context-specific PPI networks generated by Magger et al. (2012) were also used by Jacquemin & Jiang (2013) to prioritise disease-associated protein complexes. Jacquemin & Jiang failed however to demonstrate that the performance of their method was significantly better when run using context-specific PPI networks over generic PPI networks.

Li et al. (2014) also prioritised disease-associated genes using context-specific PPI networks. They generated these networks by integrating gene expression, methylation and PPI data. Genes known to be aberrantly methylated in a disease of interest were used as seed genes. Li et al. used the PageRank label propagation algorithm to prioritise genes that interacted with many aberrantly methylated genes (Brin & Page, 1998). Like Magger et al. (2012), Li et al. used the disease-context association data set generated by Lage et al. (2008) to identify suitable disease-associated contexts. Like Jacquemin & Jiang (2013), Li et al. also failed to demonstrate that using context-specific PPI networks over generic PPI networks improved method performance, making it difficult to determine the added value of their context-specific networks.

Lundby et al. (2014) used data from a GWAS of long QT syndrome (LQTS) and

tissue-specific PPIs to identify genes associated with LQTS. LQTS is a Mendelian disorder that results in the prolongation of the QT interval. It is a major risk factor in sudden cardiac death. Lundby et al. started with five proteins previously identified as being associated with LQTS. As described in Section 1.3.3, they used quantitative interaction proteomics and cardiac tissue from mice to identify proteins with which the five LQTS-associated proteins interact. In total 670 proteins were identified as interacting with the five proteins. Lundby et al. demonstrated that LQTS-associated loci identified in the GWAS were enriched with these 670 proteins, indicating that the cardiac-tissue-specific PPIs may be useful in identifying the causal genes. To validate their results, Lundby et al. analysed the functions of three of the genes seen to interact with the LQTS-associated proteins by knocking out orthologs of the genes in zebrafish (*Danio rerio*). One of these gene knockouts affected cardiac repolarisation in the zebrafish, indicating that the gene may play a role in LQTS. Whilst these results are promising, Lundby et al. did not test whether similar results could be produced using PPIs not specific to cardiac tissue. Currently, few tissue-specific PPIs are present in the curated interaction databases. This means that in order to apply this method to other diseases and tissues, new PPIs would need to be identified using quantitative interaction proteomics, which may not be feasible because of the costs involved.

NetWAS uses context-specific functional interaction networks and data from GWAS to prioritise disease-associated genes (Greene et al., 2015). As described in Section 1.3.3, Greene et al. used a Bayesian approach to integrate data from more than 14,000 publications to produce human functional interaction networks for 144 contexts. In NetWAS, SNP-wise summary statistics from a GWAS are converted to gene-wise association scores using the VEGAS method (Liu et al., 2010). VEGAS identifies genes enriched with trait-associated SNPs, whilst taking into account LD between these SNPs. VEGAS produces a p-value for each gene, describing the probability of observing a distribution of p-values amongst the SNPs in the gene at least as strong as that observed by chance, given the LD structure of the SNPs. Genes with $p < 0.01$ are chosen as positive examples and 10,000 genes with $p > 0.01$ are chosen as negative examples. A support vector machine (SVM)

is then trained using these positive and negative examples. The features used by the SVM are the weights of the edges between each gene and each labelled example in a functional interaction network specific to a disease-associated context. Genes are reranked using the distance from the SVM-generated hyperplane. Greene et al. benchmarked NetWAS by comparing the position of known hypertension-associated genes from OMIM in the gene rankings output by VEGAS and NetWAS when applied to a hypertension GWAS. In this analysis, NetWAS was run using a kidney-specific network, as the kidney is known to play an important role in regulating blood pressure (Greene et al., 2015). The known hypertension-associated genes were ranked higher in the ranking output by NetWAS, suggesting that context-specific functional interaction networks can aid in prioritising disease-associated genes. Furthermore, Greene et al. recorded the positions of the known hypertension-associated genes when NetWAS was run using each of the other 143 context-specific networks and a generic network. The kidney-specific network produced one of the best rankings, suggesting that using a network specific to a disease-manifesting context can improve the performance of the method. Unlike previously described methods, NetWAS relies on manual selection of suitable disease-associated contexts, making unbiased benchmarking of the method difficult. Greene et al. only measured the performance of NetWAS with and without context-specific networks in the single hypertension example. Further diseases should be tested to determine whether the context-specific functional interaction networks outperform generic functional interaction networks for all diseases, or whether the context-specific functional interaction networks are only useful for studying some diseases.

1.5 Methods for mapping diseases to contexts

There has been interest in systematically mapping diseases to the tissues and cell types in which they manifest. One reason for this is that whilst the tissues and cell types involved in the development of many diseases are well understood, for other diseases the disease-manifesting tissues and cell types have not yet been identified or are still being debated. The identification of these tissues and cell types would improve our understanding of the etiology of these diseases.

The emergence of gene prioritisation methods that use data specific to the tissues and cell types that manifest disease also drive the requirement for systematic maps between diseases and contexts. As described in Section 1.4.6, a number of gene prioritisation methods use the systematic map generated by Lage et al. (2008) through mining the PubMed database. Lage et al. analysed only 73 contexts and therefore methods that use data from different contexts cannot rely solely on the map generated in this study. The lack of larger maps has led to methods such as NetWAS to require users to specify the context to be used (Greene et al., 2015). This makes the use of the method and the unbiased assessment of its performance more difficult.

Since it was observed that regulatory regions are enriched in disease-associated SNPs identified in GWAS (Maurano et al., 2012), there has been interest in using regulatory regions to identify causal SNPs and provide explanations as to how causal SNPs are functionally related to diseases. Regulatory regions are known to be highly cell-type-specific (Heintzman et al., 2009) and therefore it is necessary to identify the tissues or cell types manifesting a disease to use regulatory regions to study SNPs associated with the disease. The availability of systematic maps between diseases and contexts will therefore aid these analyses.

In this section I discuss the studies that have systematically mapped diseases to contexts, along with the methods and data they use. In these studies, diseases are generally mapped to tissues or cell types. There is emerging evidence however that these methods could also be used to map diseases to other contexts, such as stages of differentiation (Seumois et al., 2014).

1.5.1 Text mining

Lage et al. (2008) generated the first large-scale systematic map between diseases, tissues and cell types. Associations were identified by counting the number of articles indexed in PubMed that mentioned particular diseases and contexts. This was done using MeSH terms (NCBI Resource Coordinators, 2016). MeSH is a controlled vocabulary developed to record the topics mentioned in each article indexed in PubMed.

Study	Context-specific data	Phenotypes analysed	Contexts analysed
Lage et al. (2008)	Article MeSH annotations	1,054 diseases	73 tissues and cell types
Hu et al. (2011)	Gene expression	3 diseases	79 tissues and cell types
Maurano et al. (2012)	Epigenetic (DHSs)	2 diseases and 1 QT	115 tissues and cell types
Magger et al. (2012)	Gene expression and PPI	Unspecified	60 tissues and cell types
Gerasimova et al. (2013)	Epigenetic (H3K4me1)	1 disease	8 tissues and cell types
Trynka et al. (2013)	Epigenetic (H3K4me3)	5 diseases and 3 QTs	34 tissues and cell types
Börnigen et al. (2013)	Gene expression and PPI	165 diseases	36 tissues
Swindell et al. (2014)	Gene expression	1 disease	10 cell types
Slowikowski et al. (2014)	Gene expression	2 diseases and 2 QTs	79 tissues and cell types
Seumois et al. (2014)	Epigenetic (H3K4me2)	15 diseases	10 tissues and cell types
Farh et al. (2015)	Epigenetic (H3K27ac)	39 diseases	33 tissues and cell types

Table 1.6: Details of studies that have systematically mapped diseases and QTs to contexts. Each of these studies is discussed in Section 1.5. Studies are ordered by publication date. If epigenetic data were used to generate the mapping, then the chromatin marker used to identify regulatory DNA is also given.

Lage et al. (2008) first manually mapped 1,054 diseases contained in OMIM and 73 contexts from the Genomics Institute of the Novartis Research Foundation (GNF) tissue atlas (Su et al., 2004) to MeSH terms. PubMed was then queried with these MeSH terms to count the number of articles that mentioned each disease and context individually and the number of articles that mentioned each disease/context pair. If the number of articles that mentioned a disease/context pair was large, relative to the number of articles that individually mentioned the disease and the context, then the disease and the context were said to be associated (Jackson et al., 1989). Lage et al. randomly selected 190 of the associations found using this approach and attempted to manually identify molecular mechanisms supporting each association. They were able to demonstrate that above an association score of 8%, this text mining method achieved a precision of 0.80.

As described in Section 1.4.1, co-occurrence-based text mining has a number

of inherent limitations. Associations can also only be identified if they are already contained in the literature, meaning that text mining is unable to identify novel associations. These issues have led to the development of alternative methods (discussed below) that use epigenetic, gene expression and PPI data to map diseases to contexts.

1.5.2 Gene expression data

Once Lage et al. (2008) had used text mining to systematically map diseases to contexts, they used the map to study whether disease-associated genes and protein complexes are over-expressed in the tissues and cell types associated with each disease. They noted that both genes and protein complexes associated with non-cancerous diseases were significantly over-expressed in the contexts identified by the text mining method as being associated with each disease. This was to be expected, as it is intuitive that disruption to a gene that is highly expressed in a tissue or cell type, and therefore likely functional in the tissue or cell type, will lead to disease in that tissue or cell type.

Interestingly, Lage et al. (2008) demonstrated that whilst genes associated with non-cancerous diseases tend to be over-expressed in tissues related to each disease, cancer-associated genes tend not to be over-expressed in the tissues in which the tumours originate. A later study by Muir & Nunney (2015) contradicted this result however. Muir & Nunney analysed the expression of 23 cancer-associated genes across samples from 12 tissue and found that the expression of the genes was highest in tissues related to the cancers. This disparity between the results of Lage et al. and Muir & Nunney may be explained by the fact that whilst Lage et al. systematically mapped cancers to tissues, Muir & Nunney manually generated this mapping. The disparity could also be due to the smaller number of genes considered by Muir & Nunney. Nevertheless, the results produced by Lage et al. indicate that gene expression data could be used to systematically map at least non-cancerous diseases to contexts.

Hu et al. (2011) built upon the observations of Lage et al. (2008) and developed a method that uses gene expression data and disease-associated loci identified in

GWAS to identify likely disease-manifesting tissues and cell types. To test an association between a disease and a context, genes in each disease-associated locus are scored using the normalised expression values of the genes in the context. Each disease-associated locus is then scored using the highest gene score amongst the genes in the locus, adjusted for the number of genes considered. Significance scores for the disease-context associations are produced by comparing the observed mean locus scores against mean locus scores generated using randomly selected sets of SNPs.

Hu et al. (2011) first applied their method to three autoimmune diseases (Crohn’s diseases, systemic lupus erythematosus (SLE) and rheumatoid arthritis) and 79 tissues and cell types using human gene expression data generated by Su et al. (2004). As was to be expected, associations were identified between the three autoimmune diseases and tissues and cell types of the immune system. To further narrow-down the disease-manifesting cell types, Hu et al. used mouse gene expression data from a range of immune cell types and identified associations between the diseases and specific immune cells, including an association between SLE and transitional B cells, and an association between Crohn’s disease and epithelium-associated stimulated dendritic cells. This use of mouse data, rather than human data, is likely to have limited the identification of disease-manifesting cell types, as gene expression differs between humans and mice (Su et al., 2004). This compromise was probably made because of the lack of high-quality human cell type gene expression data available at the time. Slowikowski et al. (2014) later applied the method developed by Hu et al. to four additional phenotypes using the same gene expression data set and identified an association between blood cell count and bone marrow CD71+ early erythroid cells, demonstrating that the method can be successful when applied to non-immune cell types. While these results indicate that gene expression data can be used to identify associations between diseases and contexts, the method of Hu et al. can only be applied to disease-associated loci identified in GWAS and not to genes identified using other methods, such as whole genome sequencing.

A similar approach was taken by Swindell et al. (2014) to test for associations between psoriasis and ten different cell types. Swindell et al. identified associations

using both genes located near psoriasis-associated SNPs identified in GWAS and genes differentially expressed between samples of skin, with and without psoriatic lesions, from individuals with psoriasis. They noted that the genes located near the psoriasis-associated SNPs tended to be over-expressed in immune cell types, while the differentially-expressed genes (DEGs) tended to be over-expressed in both skin and immune cell types, including macrophages. A disadvantage of using DEGs to identify psoriasis-associated cell types is that the DEGs identified are likely to be highly dependent on the cell types present in the different skin samples. Macrophages are known to infiltrate psoriatic lesions (Nestle et al., 2009) and this may explain why many of the DEGs identified are known to be highly expressed in macrophages.

1.5.3 Gene expression and PPI data

Some studies have combined gene expression and PPI data to systematically map diseases to contexts. Magger et al. (2012) integrated gene expression and PPI data to generate 60 context-specific PPI networks. As described in Section 1.4.6, Magger et al. then ran PRINCE (Vanunu et al., 2010) using these networks and noted that PRINCE performed better when run using a PPI network specific to a tissue or cell type associated with each disease. Using this observation, Magger et al. developed a method to systematically map diseases to contexts. Given a disease and a set of associated genes, Magger et al. ‘left out’ a gene from the associated gene set. They then applied PRINCE to the remaining set of associated genes, using each of the context-specific PPI networks in turn, producing a ranked list of prioritised genes for each context. Magger et al. next noted the position of the ‘left out’ gene in each of the ranked lists. Using the observation that PRINCE tends to perform better when run using PPI networks specific to disease-associated contexts, Magger et al. inferred the context most likely to be associated with each disease by identifying the context-specific PPI network that resulted in the ‘left out’ gene being ranked highest. Magger et al. validated this approach by comparing their results to the text mined disease-context associations generated by Lage et al. (2008). They noted that in 53% of cases, the context ranked highest by Lage et al. was also ranked highest by their method, demonstrating that gene expression and PPI data can be

combined to identify disease-context associations.

Börnigen et al. (2013) used a different approach to integrate gene expression and PPI data and identify disease-context associations. Their method, which they call TissueRanker, compares the co-expression of disease-associated protein complexes across contexts, under the hypothesis that if the genes involved in a protein complex are co-expressed in a certain tissue, then that protein complex is more likely to be functionally active in the tissue. Using disease-gene associations from OMIM, Börnigen et al. used their method to identify associations between 165 diseases and 36 contexts. Like Magger et al. (2012), Börnigen et al. validated TissueRanker using the disease-context associations generated by Lage et al. (2008). Börnigen et al. further compared the performance of TissueRanker against disease-context associations generated using only the expression levels of the associated genes. They demonstrated that considering entire protein complexes, rather than individual proteins, improved performance significantly. They also compared the performance of TissueRanker when applied to different tissues. Performance was highest for adipose and placental tissue and lowest for blood and lymphatic tissue. Lower performance may be due to greater heterogeneity in some of the tissue samples, with some tissues containing a larger number of different cell types, each with its own distinct gene expression profile. Further work will need to be completed to determine whether using expression data from homogenous samples containing a single cell type, rather than heterogenous tissue samples containing multiple cell types, improves the performance of TissueRanker and similar methods.

1.5.4 Epigenetic data

In addition to gene expression and PPI data, epigenetic data have also been used to map diseases to contexts. In 2012, Maurano et al. performed deoxyribonuclease I (DNase I) mapping to identify DNase I hypersensitivity sites (DHSs) in 115 tissues and cell types. DHSs are known to be markers of transcriptionally active regulatory DNA in cells (Thurman et al., 2012) and therefore this DHS-mapping identified transcriptionally active regulatory regions in each of the tissues and cell types. Maurano et al. measured the enrichment of SNPs identified in GWAS as being

associated with three phenotypes (cardiac conduction measured using QRS duration, Crohn’s disease and MS) in the transcriptionally active regulatory regions identified in each of the tissues and cell types. Enrichment was greatest in the regulatory regions active in tissues and cell types related to the phenotypes. For example, SNPs associated with Crohn’s disease were most greatly enriched in T helper 17 cells and SNPs associated with MS in CD3+ T-cells. This demonstrates that active regulatory regions offer an alternative way of systematically mapping diseases to contexts.

Trynka et al. (2013) mapped 15 different chromatin marks across 14 different cell types from the Encyclopaedia of DNA Elements (ENCODE) project (The ENCODE Project Consortium, 2011) to further study the enrichment of disease-associated SNPs in regulatory regions. Regions marked by 4 of the 15 different chromatin marks (DHS, H3K4me3, H3K79me2 and H3K9ac) were enriched with disease-associated SNPs. Each of these 4 marks is known to identify regulatory DNA and therefore these results support the conclusions of Maurano et al. (2012). Trynka et al. further mapped the most significantly enriched chromatin mark (H3K4me3) in 34 tissues and cell types from the National Institute for Health (NIH) Epigenomics Project (Bernstein et al., 2010). They compared the enrichment of SNPs identified in GWAS as being associated with eight different phenotypes across the 34 tissues and cell types and identified a number of associations, including an association between plasma low-density lipoprotein (LDL) concentration and liver tissue, and an association between T2D and pancreatic islet cells. While these associations reveal little new about the cellular mechanisms that underpin these phenotypes, the work of Trynka et al. highlights which chromatin marks are most useful when analysing the context-specific impact of variants.

A couple of studies have used epigenome-based disease-context mapping to identify cell types associated specifically with asthma. Gerasimova et al. (2013) compared the enrichment of 2,510 SNPs in tight LD ($r^2 > 0.8$) with asthma-associated SNPs identified in a GWAS across transcriptional enhancers in eight different tissues and cell types (identified using H3K4me1 marks). Significant enrichment was identified in CD4+ T cell enhancers, a cell population known to

be involved in asthma (Erb & Le Gros, 1996), as well as in liver and adipose cell enhancers, cell populations not generally thought to be involved in the disease. Enrichment in these cell types increased further when ubiquitously active enhancers were excluded and only cell-type-specific enhancers considered, demonstrating that the differences between cell types are highly informative.

Seumois et al. (2014) further analysed the association between asthma and T cells. They mapped transcriptional enhancers active in both naive and memory CD4+ T cells (identified using H3K4me2 marks) in order to determine whether asthma-associated SNPs are likely to affect T cells at a particular stage in their differentiation. Memory CD4+ T cells were divided into two subtypes: T_H1 and T_H2 cells. Seumois et al. identified enrichment of asthma-associated SNPs in the enhancers gained by T_H2 cells, indicating that the asthma-associated SNPs may influence disease susceptibility by affecting the differentiation of this cell population. Furthermore, Seumois et al. analysed an additional 14 immune-related diseases and 8 contexts and noted that HIV and SLE-associated SNPs were enriched in enhancers lost in T_H2 cells, suggesting that SNPs associated with these diseases may affect the cell population earlier in their differentiation. Overall, the results produced by Seumois et al. indicate that epigenome-based disease-context mapping may be able to provide information on the stages of cell type differentiation most likely to be affected by disease-associated variants.

Finally, Farh et al. (2015) built upon the previously described methods to generate the largest epigenome-based disease-context mapping available to date. Farh et al. developed the PICS method, which uses the patterns of observed SNP-trait association and the haplotype structure of the associated loci identified in GWAS to predict the variants most likely to be causal. Using PICS, Farh et al. predicted causal variants for 39 diseases and measured the enrichment of these variants in regulatory elements mapped in 33 tissues and cell types (identified using H3K27ac marks). Farh et al. identified a number of associations, including associations between autoimmune diseases and cell types of the immune system, and associations between neurodegenerative diseases and brain tissues. Farh et al. also compared associations identified using non-coding SNPs and epigenetic data to

associations identified using protein-coding SNPs and expression data and concluded that using non-coding SNPs and epigenetic data was more informative. Farh et al. only analysed the expression of coding SNPs in three diseases however and therefore further work should be completed to compare the performance of these two approaches.

1.6 Scope of thesis

The studies discussed in Section 1.5 demonstrate that multiple types of data can be combined and used to systematically map diseases to contexts. In Chapter 2, I detail the methods I use in this thesis to evaluate method performance, normalise and cluster gene expression profiles and measure distances across networks. In Chapter 3, I describe the development of the GSC and GSO methods, which use gene expression and PPI data to identify disease-manifesting cell types. In Chapter 4, I use the GSO method to identify the cell-type-specific PPI networks best suited to prioritising disease-causing genes.

Multiple methods have used networks to prioritise disease-causing genes. The studies detailed in Section 1.4 suggest that the performance of these methods could be improved by incorporating context-specific interactions. Currently, a lack of context-specific interaction data and systematic maps between many diseases, tissues and cell types makes it difficult to prioritise disease-causing genes using context-specific networks. In this thesis, I demonstrate that cell-type-specific PPI networks generated by integrating gene expression and PPI data can aid in identifying disease-causing genes.

Chapter 2

Materials and methods

In this section I outline the statistics I use to evaluate the performance of the methods described in this thesis, the methods used to normalise gene expression data and the methods used to compute distances and propagate scores across networks.

2.1 Performance evaluation

I use the following statistics to evaluate the performance of the methods described in this thesis. The measures are based on the number of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs) output by a method.

2.1.1 Performance statistics

The precision of a method (also known as the positive predictive value) is the proportion of cases classified as positive by the method that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.1)$$

The sensitivity of a method (also known as the recall) is the proportion of truly positive cases that are classified as positive by the method:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2.2)$$

The specificity of a method is the proportion of truly negatives cases that are classified as negative by the method:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2.3)$$

The F1-score (also known as the F-score) is the harmonic mean of the precision and the recall of a method. The F1-score weights the precision and recall equally, thereby providing a single statistic that describes the performance of a method:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \quad (2.4)$$

2.1.2 ROC curves

I use receiver operating characteristic (ROC) curve plots to visualise the performance of the methods. The curves are generated by plotting the sensitivity of the classifier against $1 - \text{specificity}$ of the classifier at a number of threshold settings.

ROC curves can be compared by computing the area under the curve (AUC). The AUC represents the probability that a randomly chosen positive case will be ranked higher by the classifier than a randomly chosen negative case. If a classifier has perfect discriminating power, then the AUC of the corresponding ROC curve would equal 1. If a classifier does not perform better than expected by chance, then the expected AUC of the corresponding ROC curve would be 0.5.

DeLong's method is a non-parametric test used to compare ROC curves and their AUCs (DeLong et al., 1988). In DeLong's method, the null hypothesis is that the difference between two AUCs is equal to 0. DeLong's method takes into account that ROC curves are often generated by applying classifiers to the same data set, resulting in correlated curves. The test uses generalised U -statistics to estimate a covariance matrix and thereby compare correlated ROC curves.

In this thesis I use the `pROC` R-package to generate ROC curves, estimate AUCs and apply DeLong's method (Robin et al., 2011).

2.2 Statistical tests

2.2.1 Fisher's exact test

Fisher's exact test is used to test for independence in contingency tables. It is often used to analyse categorical data, in which objects are classified in multiple ways. In this thesis I only apply Fisher's exact test to 2×2 contingency tables. The test can however be applied to larger tables.

To apply Fisher's exact test, let T be a 2×2 contingency table:

$$T = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad (2.5)$$

The probability of observing any individual set of values, under the null hypothesis of independence, follows the hypergeometric distribution:

$$P(A = a, B = b, C = c, D = d) = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! (a+b+c+d)!} \quad (2.6)$$

To compute a p-value representing the probability, under the null hypothesis of independence, that an observation at least as extreme as that observed occurs, it is necessary to sum probabilities calculated using the hypergeometric distribution. In the one-tailed version of Fisher's exact test, probabilities are calculated for the observed contingency table and all other possible tables that are equally as extreme, or more extreme, than that observed (with the same or a smaller value of a). In the two-tailed version of the test, probabilities are also calculated for the tables that are equally as extreme, or more extreme, but in the other direction. The sum of these probabilities is used as a measure of statistical significance.

2.2.2 The Mann-Whitney U test

The Mann-Whitney U test (also known as the Mann-Whitney-Wilcoxon test) is a non-parametric test used to determine whether two samples are drawn from the same population. Unlike Student's t -test, it does not rely on the populations being

normally distributed.

To compute the U -statistic, both samples are first combined and ranked. Let R_1 be the sum of the ranks of the values from sample 1 and n_1 and n_2 be the sizes of samples 1 and 2. Then:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad (2.7)$$

This U -statistic is computed twice, with each sample being considered as sample 1 in each case. The smaller of these U -statistics is then used to determine the level of statistical significance at which the null hypothesis should be rejected.

2.3 Adjusting for multiple testing

2.3.1 Bonferroni correction

Bonferroni correction is a method of reducing the type-I error rate when conducting multiples tests (Dunn, 1961). It is a method of maintaining the family-wise error rate (FWER), which is the probability of making one or more false discoveries. If the desired significance level for a whole family of tests is α , and k is the number of hypotheses, then using the method each hypothesis is tested at a significance level of α/k .

2.3.2 Benjamini-Hochberg procedure

The Benjamini-Hochberg (BH) procedure is a method for controlling the false discovery rate (FDR) when conducting multiple tests (Benjamini & Hochberg, 1995). The BH procedure is considered less conservative than Bonferroni correction (Benjamini & Hochberg, 1995).

To control the FDR using the BH procedure, let $H_1 \dots H_m$ be m tested null hypotheses and $p_1 \dots p_m$ be the corresponding p-values. P-values are first ordered from the smallest to the largest, so that $p_{(1)} \dots p_{(m)}$, where $p_{(1)}$ is the smallest p-value and $H_{(1)} \dots H_{(m)}$ are the corresponding null hypotheses. To control at a FDR of α , the highest value of k is found so that $p_{(k)} \leq (k \times \alpha)/m$. Null hypotheses

$H_{(1)} \dots H_{(k)}$ are then rejected. The q-value is the minimum FDR at which the null hypothesis can be rejected.

2.4 Hierarchical clustering

In this thesis I cluster gene expression profiles using complete-linkage hierarchical clustering using the method implemented in the `stats` R-package (Ihaka & Gentleman, 1996). To cluster a set of elements using complete-linkage hierarchical clustering, each element is started in its own cluster. When applying the method to gene expression profiles, each element is a single profile. At each step, the two closest clusters are combined. The Jensen-Shannon distance (JSD) between the furthest elements in pairs of clusters is used as a measure of distance between clusters (Endres & Schindelin, 2003). Clusters are combined until all elements are contained in a single cluster. The process of clustering elements can be visualised using a dendrogram, where the distances along the branches indicate the distances at which clusters were combined.

To compute the JSD between gene expression profiles, I first normalise the gene expression profiles so that their values sum to one. I then give genes with expression values of zero a pseudo-expression value of 1×10^{-25} , as the inclusion of expression values equal to zero would produce infinite distances. I can then compute the JSD J between samples:

$$J(P, Q) = \sqrt{\frac{D(P||M) + D(Q||M)}{2}} \quad (2.8)$$

where P and Q are vectors representing the normalised gene expression profiles, $M = (P + Q)/2$ and D is the Kullback-Leibler divergence:

$$D(P||Q) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right) \quad (2.9)$$

2.5 Normalising read counts

I use the relative expression of genes across the samples profiled by The FANTOM Consortium (2014) to identify disease-associated contexts and generate context-specific PPI networks. These read counts are dependant on the depths to which the samples were sequenced. To compare read counts across samples, it is therefore necessary to adjust for these differences in sequencing depth. In this thesis I use the relative log expression (RLE) method (Anders & Huber, 2010), implemented in the `edgeR` R-package (Robinson et al., 2010), to normalise read counts to account for differences in sequencing depth.

The easiest way to make the read counts of genes comparable across samples is to divide the read counts by a sizing factor, based on the total number of reads in the sample, which is also known as the library size. Small numbers of highly expressed genes can however have a strong effect on library size and therefore can result in this method producing poor relative gene expression value estimates. It was for this reason that Anders & Huber (2010) developed the RLE method, which instead uses as sizing factors the median ratio of the read counts of each gene to the geometric mean of the read counts across all samples:

$$\hat{s}_j = \operatorname{median}_i \frac{k_{i,j}}{(\prod_{v=1}^m k_{iv})^{1/m}} \quad (2.10)$$

where $\hat{s}_j$ is the sizing factor of sample j and $k_{i,j}$ is the read count of gene i in sample j . This method ensures that small numbers of highly expressed genes do not over influence the sizing factors used to normalise the read counts.

2.6 Random-walk-based network algorithms

Random-walk-based methods can be used to measure distances between vertices in networks and propagate scores across networks. They work by simulating the movement of a random walker across the network. This random walker travels from vertex to vertex, along connecting edges. Unlike the shortest paths method (Cormen et al., 2009), which measures the distance between pairs of vertices using a single

path, random-walk-based methods consider all possible paths (Köhler et al., 2008). I use a RWR-based method in Chapter 3 to measure the distance between pairs of vertices and a RWR-based method in Chapter 4 to propagate scores across networks.

2.6.1 Measuring distances

To measure the distance from vertex i to vertex j , a random walker is started at vertex i . At each time step t , the random walker can either move to one of the vertices directly connected (adjacent) to its current vertex, or revert back to its starting vertex i with probability r . If the walker does not revert to its starting vertex, but instead moves to an adjacent vertex, then which vertex the walker moves to is controlled by a probability distribution based on the number of adjacent vertices and the weights of the connecting edges. If the network edges are not weighted, then the probability that the walker moves to each adjacent vertex is uniform. If the network edges are weighted, then the probability that the walker moves to each adjacent vertex is proportional to the weight of the connecting edge.

In Chapter 3, I use an approach similar to the iterative method described by Köhler et al. (2008) to compute distances. To measure the distance between vertex i and vertex j in graph G , let n be the number of vertices in G , A be the column-normalised adjacency matrix of G and P be a probability matrix with dimensions $n \times n$. $P_{i,j}^t$ is the probability that a walker starting from vertex i is located at vertex j at time t . P^0 is the initial probability distribution and an identity matrix. The probabilities are computed iteratively:

$$P^{t+1} = (1 - r)AP^t + rP^0 \quad (2.11)$$

Iterations are completed until the difference between the probability distributions between time steps (P^t and P^{t+1}), measured using the Manhattan distance, falls below a certain cutoff value. In Chapter 3, an iteration cutoff of $|S| \times 10^{-5}$ is used, where $|S|$ is the size of the gene set tested.

2.6.2 Propagating scores

In Chapter 4, I use an approach similar to the propagation algorithm used by Vanunu et al. (2010) to propagate scores across networks. Let G again be the graph, n be the number of vertices in G and A' be the normalised adjacency matrix of G . In A' , each edge is weighted by the sum of the edge weights of each interacting gene. Let D be a diagonal matrix where D_{ii} is the sum of the i th row of the adjacency matrix A . Then let $A'_{ij} = A_{ij}/\sqrt{D_{ii}D_{jj}}$. Let Y^t be the distribution of scores across vertices at time t and be a vector of length n . Y^0 is the initial distribution of scores and is normalised to sum to one. The distribution of scores is computed iteratively:

$$Y^{t+1} = (1 - r)A'Y^t + rY^0 \quad (2.12)$$

In Chapter 4, the scores represent phenotypic relevance scores. Scores are propagated from different initial distributions and compared to generate measures of significance. To make these sets of scores comparable, a fixed number of iterations are completed. As justified in Section 4.3.3, I use ten iterations to propagate the scores.

Chapter 3

Identifying associations between diseases and cell types

In this chapter I describe the generation of 73 cell-type-specific PPI networks through the integration of gene expression and PPI data. I also report the development of the GSC and GSO methods, which identify associations between diseases and contexts. The GSC method predicts associations by comparing the clustering of disease-associated genes across context-specific PPI networks. The GSO method predicts associations by identifying contexts in which disease-associated genes are over-expressed. I analyse the performance of these two methods using associations independently identified by mining the PubMed database. The GSC method predicts a number of previously identified disease-cell-type associations, as well as associations that warrant further study. This work has been published by Cornish et al. (2015).

The text mining work described in this chapter was completed with Dr. Ioannis Filippis (see Section 3.2.5 for further details).

3.1 Introduction

3.1.1 Motivation

There are many methods in bioinformatics that use networks to study disease (Moreau & Tranchevent, 2012; Lage, 2014). However, the majority of these methods use generic networks, containing all interactions known to occur between genes or proteins throughout the body, rather than networks specific to the context of the disease being studied. These contexts include the tissues and cell types in which the disease manifests. This is partly due to a lack of information about the contexts under which interactions between genes and proteins take place (Lage, 2014). The majority of methods that identify interactions are unable to determine whether an interaction takes place only in a certain tissue or cell type. It may even be the case that some proteins identified as interacting rarely or never interact *in vivo*, either because the proteins never come into contact, or because the required interaction conditions are never satisfied (Gonzalez & Kann, 2012).

Magger et al. (2012) demonstrated that the performance of network-based methods may be affected by their failure to use context-specific networks. There has therefore been an interest in generating context-specific networks through the integration of data (Bossi & Lehner, 2009; Lopes et al., 2011; Magger et al., 2012; Guan et al., 2012; Barshir et al., 2014; Liu et al., 2014b; Greene et al., 2015). However, there are currently few cell-type-specific networks available, partly due to the lack of high-quality gene expression data available for many cell types. In 2014, the Functional Annotation of the Mammalian Genome 5 (FANTOM5) project released high-quality gene expression data for a large number of primary cell type samples (The FANTOM Consortium, 2014), creating the opportunity to generate cell-type-specific networks through the integration of these and other data.

The use of context-specific data in network-based methods has also been limited by the lack of large-scale maps between diseases and many contexts. Some methods that currently use context-specific networks rely on a limited number of text mined associations (Magger et al., 2012; Li et al., 2014). Others require users to manually select the context to use (Greene et al., 2015). The availability of larger disease-

context maps would simplify the use of these methods and aid in their development and testing. Furthermore, the systematic mapping of diseases to cell types may result in the identification of novel disease-cell-type associations that improve our understanding of disease etiology.

3.1.2 Generating context-specific networks

As described in Section 1.3.3, there are currently a number of methods that integrate context-specific data to generate context-specific networks. Data types used by these methods include epigenetic, gene expression, PPI and chemical and genetic perturbation data.

Vertex removal and edge reweighting methods (Figure 1.6) have both been used to generate context-specific PPI networks. In the vertex removal method, proteins are removed from the network if they are expressed below a certain threshold. This may make this method less suitable for certain tasks, such as genome-wide gene prioritisation, as the removal of proteins limits the number of genes that can be prioritised. In the edge reweighting method, no proteins are removed entirely from the network and edges are instead up or down weighted depending on the expression levels of the corresponding genes. This method therefore avoids the previously mentioned issue.

Functional interaction and PPI networks represent different types of biological relationship between genes and proteins. While these network types have been used in a number of different tasks, it is not yet clear whether certain network types are best suited to certain tasks. The lack of cell-type-specific PPI networks prompted me to generate the 73 cell-type-specific networks described in this chapter.

3.1.3 Mapping diseases to contexts

As described in Section 1.5, diseases have previously been systematically mapped to the tissues and cell types in which they are most likely to manifest using multiple methods and data sources. Lage et al. (2008) produced one of the earliest mappings between diseases and contexts by mining the PubMed database for articles that co-mention diseases and contexts.

More recently, epigenetic and gene expression data have been used to associate diseases with tissues and a limited number of cell types. Hu et al. (2011) mapped diseases to immune cell types by quantifying the enrichment of cell-type-specific genes in disease-associated loci. Maurano et al. (2012) noted that genomic regions functioning as enhancers of transcription were enriched with disease-associated SNPs identified in GWAS. Furthermore, enrichment was strongest in enhancers specific to tissues and cell types related to each disease, allowing Trynka et al. (2013), Gerasimova et al. (2013) and Seumois et al. (2014) to identify a number of disease-context associations. Farh et al. (2015) extended this analysis further, identifying associations between 39 diseases and 33 contexts.

Gene expression and PPI data have also been integrated to associate diseases with contexts. Magger et al. (2012) noted that PRINCE (Vanunu et al., 2010) tends to perform better when run using PPI networks specific to contexts related to each disease. Magger et al. used this observation to systematically map diseases to contexts. Börnigen et al. (2013) used PPI data to generate protein complexes. They observed that proteins in disease-associated protein complexes were often co-expressed in tissues related to the disease and used this information to map diseases to tissues.

While these methods have successfully mapped diseases to tissues and a limited number of cell types, a lack of high-quality cell-type-specific data has previously prohibited the large-scale mapping of diseases to cell types. The release in 2014 of high-quality gene expression data from a range of cell types by The FANTOM Consortium facilitates this mapping. In this chapter I systematically map diseases to the cell types in which they are most likely to manifest using these data.

3.2 Materials and methods

3.2.1 Gene expression data

In this chapter I describe the generation of cell-type-specific PPI networks through the integration of gene expression data from the FANTOM5 project (The FANTOM Consortium, 2014) and PPI data. The FANTOM Consortium generated these data

Processing step	Number of samples	Number of profiles	Minimum value	Maximum value
CAGE peaks	573	-	0	2,297,740
Hierarchical organisation of samples	362	-	0	2,297,740
Sample normalisation	362	-	0	573,441
Quality control	324	-	0	573,441
Generating gene expression profiles	-	74	0	476,071
Percentile normalisation	-	74	0	1
Numbers of highly expressed genes	-	73	0	1

Table 3.1: The effect of each of the processing steps on the gene expression data. Each of the processing steps in this table is reflected in Section 3.2.1. For each step, the number of samples or gene expression profiles remaining at the end of the step and the minimum and maximum values in these samples/profiles are given.

by profiling gene expression in a large panel of human primary cell samples using cap analysis of gene expression (CAGE) with single-molecule cDNA sequencers. In this section I describe the normalisation of these data, the quality control procedures I applied and how I combined the samples to generate gene expression profiles for 73 cell types. Table 3.1 contains details of the number of samples and gene expression profiles in the data set at the end of each of these processing steps, along with the range of values they span.

CAGE peaks

Using CAGE technology, The FANTOM Consortium (2014) identified transcription start sites (TSSs) and quantified their expression (Kanamori-Katayama & Itoh, 2011). TSSs were identified by The FANTOM Consortium using decomposition peak identification (DPI). DPI identifies TSSs by first generating CAGE tag clusters that contain the contiguous regions of CAGE signal in individual biological states. It next uses independent component analysis (ICA) to decompose these CAGE tag clusters and identify individual peaks (Hyvärinen & Oja, 2000). Finally, peaks that overlap

are merged to produce the peaks reported in the study. The FANTOM Consortium assigned each peak located within 500bp of the 5' end of a gene transcript in the *hg19* human genome reference assembly to that gene. I downloaded the CAGE peak raw read counts for the primary cell samples from the FANTOM5 project website (Table A.2 contains details of when and from where these and other data used in this thesis were downloaded).

Hierarchical organisation of samples

Many of the human primary cell samples profiled in the FANTOM5 Project are drawn from the same populations of cells (The FANTOM Consortium, 2014). The GSC and GSO methods described later in this chapter each require a single gene expression profile for each cell type they consider. I therefore combined samples corresponding to the same cell type to produce individual expression profiles.

How to best define a 'cell type' and identify populations of cells that represent distinct cell types is an unresolved problem (Carniol et al., 2016). One approach is to hierarchically organise cell types in an ontology (Meehan et al., 2011). This allows cell types to be defined at different resolutions. In order to combine the samples from the FANTOM5 Project, I therefore decided to first organise the samples into a hierarchical structure. I did this using information provided by The FANTOM Consortium (2014) and Andersson et al. (2014).

Andersson et al. (2014) organised 362 of the samples into groups based on the function and the morphology of the cells profiled. Each of these groups contains between one and 32 samples. Andersson et al. were unable to assign each of the remaining samples to a single group. I therefore discarded these samples. Andersson et al. describe these broad sample groups as 'facets'.

Each of the FANTOM5 project samples is accompanied by a sample name, which contains a description of the cell type and a donor identifier. If these donor identifiers are removed then samples can be grouped by the sample name. These sample-name-based groups are smaller than the facets defined by Andersson et al. (2014). The samples in each of the sample-name-based groups are also all contained in the same facet. I therefore refer to these finer sample-name-based groups as 'sub-facets'. Each

sub-facet maps to exactly one facet and therefore the sub-facet and facet groupings can be considered a two-level hierarchy of sample organisation (Figure 3.1B). I use this two-level hierarchy of samples to combine the samples.

As previously mentioned, it was necessary to produce a single gene expression profile for each cell type. It was not obvious however whether it would be best to combine samples at the level of the sub-facet, the level of the facet or at a mixture of the levels. I did not want to combine samples with distinct gene expression profiles, as this may have resulted in the production of hybrid profiles that did not represent gene expression in any cell type. I also did not want to produce many similar gene expression profiles, as this may have had a deleterious effect on the performance of the GSC and GSO methods. I therefore decided to combine the samples at a mixture of the levels, depending on how similar the gene expression profiles of the samples in each sub-facet and facet were. This procedure is described later in this section.

Sample normalisation

I used the `biomaRt` R-package to assign genes lacking an Ensembl gene identifier such an identifier (Durinck et al., 2009). I discarded peaks that could not be assigned an Ensembl gene identifier. For each gene, I summed the read counts of the associated peaks to generate a gene-wise read count, as described by Sardar et al. (2014). Having a single score for each gene allows for the integration of the expression and PPI data (see Section 3.2.3 for further details). I then normalised the gene-wise read counts using the RLE method (Figure 3.1C, as described in Section 2.5). The values output by the RLE method represent normalised tags per million (TPM).

Quality control

I conducted quality control to remove any spurious samples. I did this by measuring the similarity of the gene expression profiles of the samples in each sub-facet. If a sample had a gene expression profile that differed strongly from the other samples in the sub-facet, then I discarded the sample. Some of the sub-facets contained only a single sample. In these cases, it was not possible to conduct quality control and

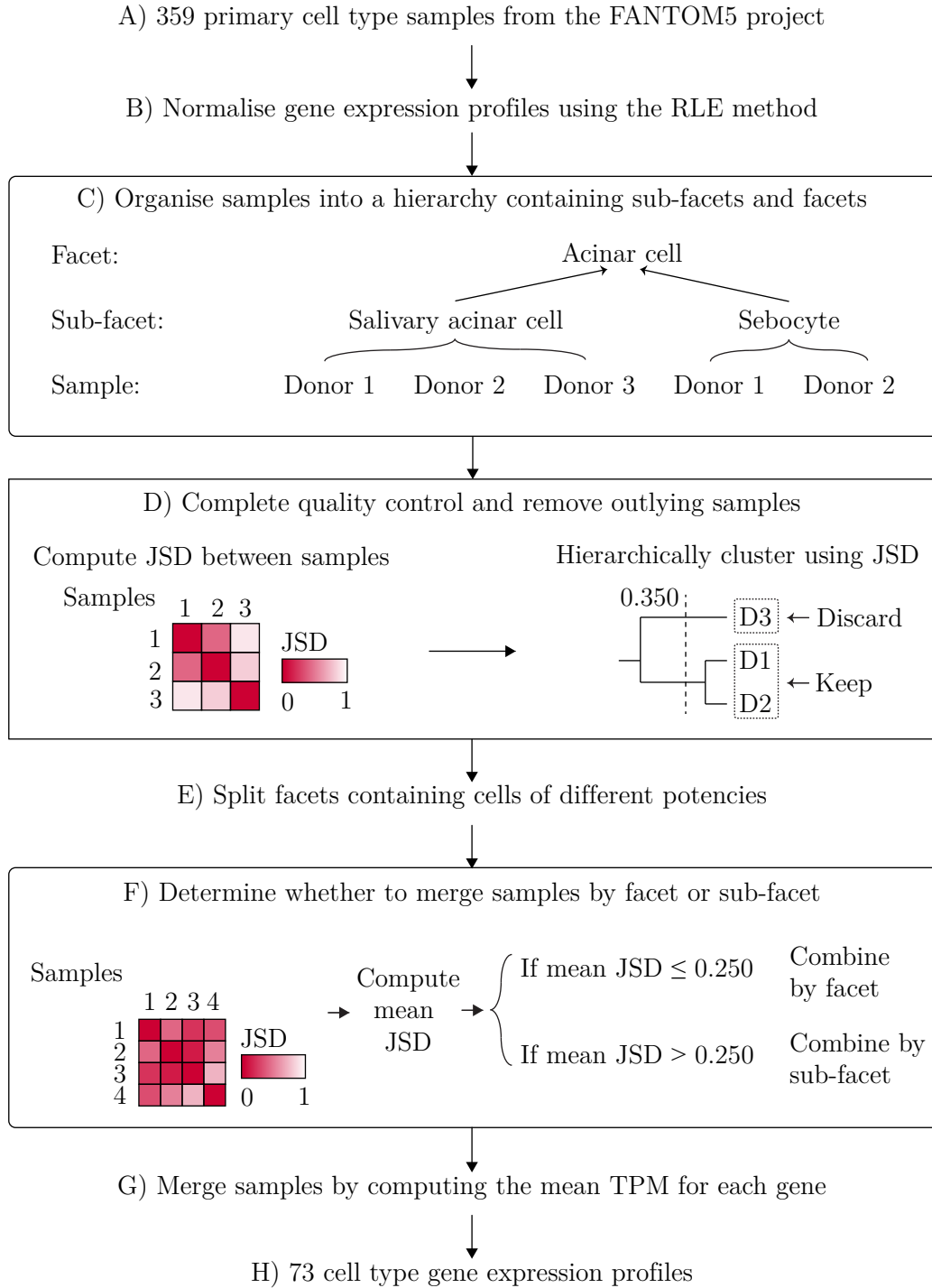


Figure 3.1: The pipeline used to normalise, filter and combine the FANTOM5 project sample gene expression profiles to produce gene expression profiles for 73 cell types. As justified in Section 3.2.1, the 3 hepatocyte samples are discarded, and therefore 359, rather than 362, samples are used.

Tree cut value	N. samples discarded	N. facets discarded
0.500	0	0
0.450	1	0
0.400	4	0
0.350	17	0
0.300	41	2
0.250	90	8
0.200	145	16
0.150	268	42
0.100	327	58
0.050	341	64

Table 3.2: The number of samples that would be discarded during quality control using different tree cut values. Also given is the number of complete facets that would be discarded. I consider a facet discarded if all samples in the facet are discarded. At the start of this quality control procedure, there were 341 samples and 64 facets remaining.

I therefore discarded these sub-facets and samples, resulting in the loss of 21/362 (5.8%) of the samples. I next clustered samples using complete-linkage hierarchical clustering and the JSD between the samples (Figure 3.1D, see Section 2.4 for details). The JSD metric was also used by Andersson et al. (2014) to cluster FANTOM5 project samples. I cut the resulting tree at a height of 0.350 to split the samples in each sub-facet into discreet clusters. For each of the sub-facets, I discarded samples not contained in the largest cluster. If no clusters contained more than one sample, then I discarded all samples in the sub-facet. Using smaller tree cut values would increase the number of samples removed. I decided to cut the tree at a height of 0.350, as this resulted in the removal of the most spurious samples, whilst avoiding the complete loss of any single facet (Table 3.2). This procedure resulted in the loss of 17/341 (5.0%) of the remaining samples.

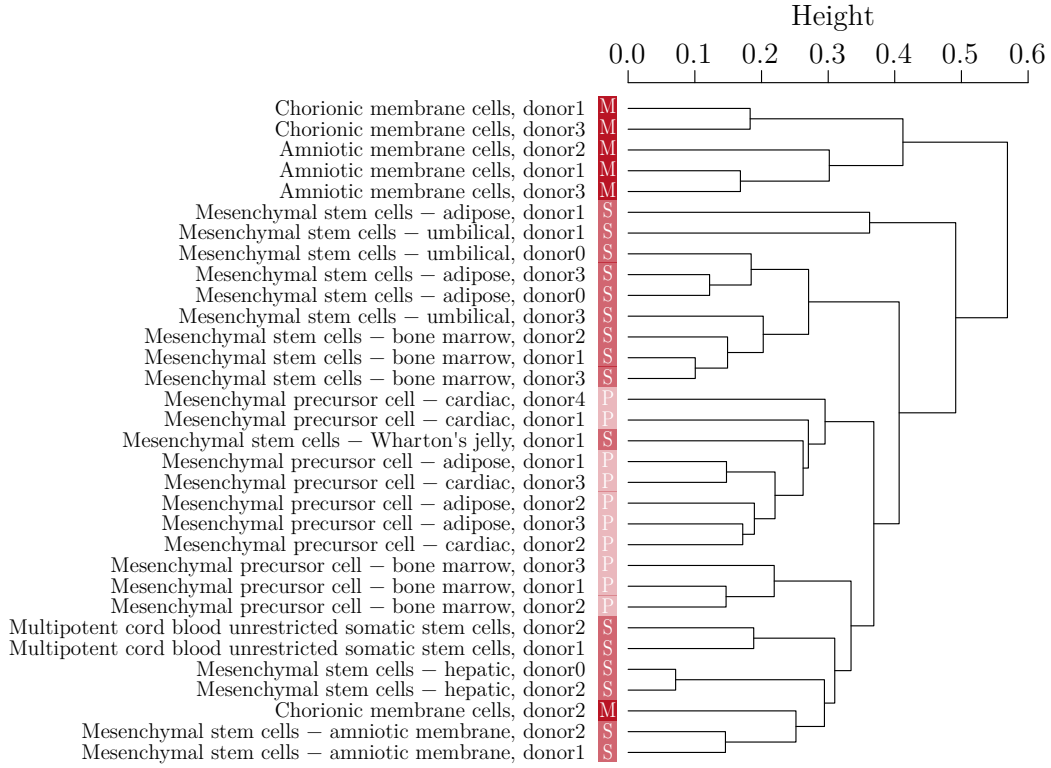
Generating gene expression profiles

To produce gene expression profiles that accurately represent individual cell types, it was necessary to combine groups of samples corresponding to the same cell type. Samples can either be combined at the level of the sub-facet, the facet or using a mixture of these groupings.

Some of the facets defined by Andersson et al. contain samples of different potencies. It is known that cells of different potencies have different gene expression profiles (Kulterer et al., 2007). This can be seen when samples in these multi-potency facets are clustered using their gene expression profiles, complete-linkage hierarchical clustering and the JSD between the samples (Figure 3.2). I therefore split facets containing samples of different potencies into separate groups (Figure 3.1E). I split the ‘mesenchymal cell’ facet into three facets (a ‘mesenchymal stem cell’ facet, a ‘mesenchymal precursor cell’ facet and a ‘mesenchymal somatic cell’ facet) and the ‘monocyte’ facet into two facets (a ‘CD14+ monocyte derived endothelial progenitor cell’ facet and a ‘monocyte’ facet). I assigned sub-facets to the new facets based on the potency of the cell type, as indicated in the name of the sub-facets.

As previously explained, it is necessary to combine samples and generate individual gene expression profiles representing distinct cell types. To do this I computed the mean JSD between all pairs of samples in each facet to measure the similarity of the gene expression profiles (Figure 3.1F). This allowed me to identify facets in which the samples have more and less similar gene expression profiles. A smaller mean JSD indicates that the samples in a facet have more-similar gene expression profiles. To ensure that sub-facets containing samples with similar expression profiles were not overrepresented in the final set of expression profiles, I combined samples at the level of the facet if the mean JSD was less than 0.250 and combined samples at the level of the sub-facet if the mean JSD was greater than 0.250. I chose a threshold of 0.250 after manual inspection of the sub-facets contained in each facet. The cell type facets contain different numbers of samples and sub-facets (Figure 3.3). The ‘mesenchymal stem cell’ facet contains seven sub-facets that represent stem cell populations found in a diverse set of tissues, including adipose, bone marrow and hepatic tissue. Conversely, the ‘vascular associated

A) Clustering of samples in the ‘mesenchymal cell’ facet



B) Clustering of samples in the ‘monocyte’ facet

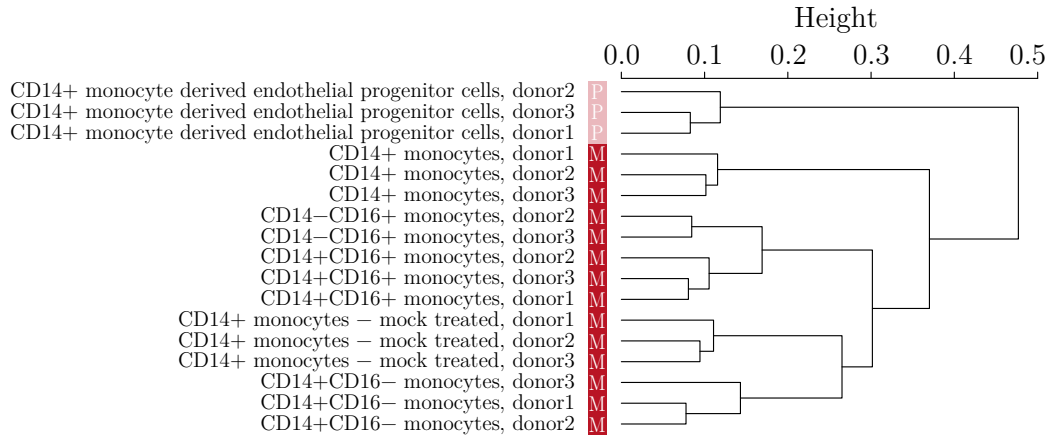


Figure 3.2: The clustering of the gene expression profiles in the (A) ‘mesenchymal cell’ and (B) ‘monocyte’ facets. Clustering was completed using complete-linkage hierarchical clustering and the JSD between samples. Samples tend to cluster with samples of similar potencies. M: mature cell, P: precursor or progenitor cell, S: stem cell.

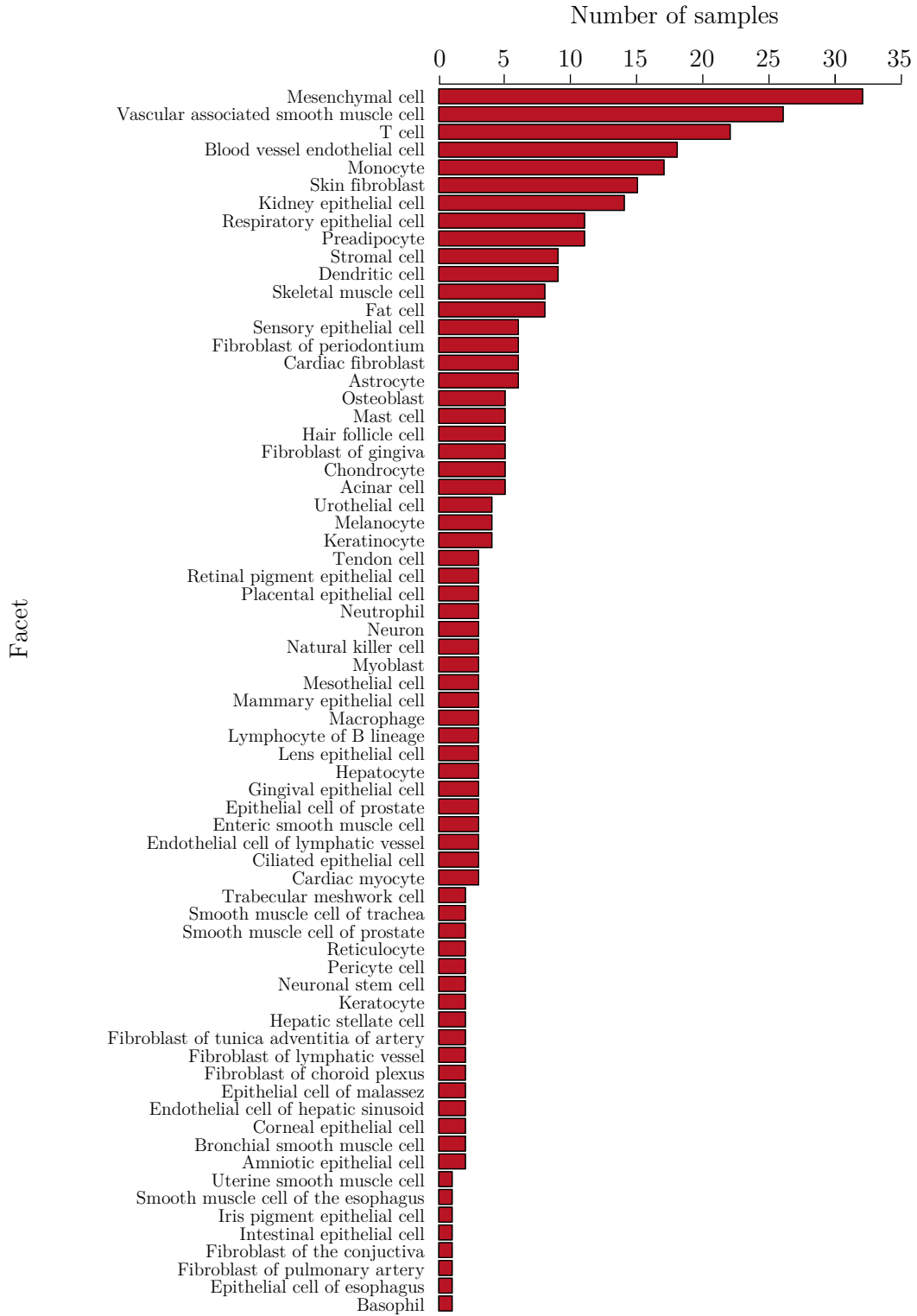


Figure 3.3: The number of FANTOM5 project samples in each cell type facet defined by Andersson et al. (2014). All samples mapped to facets by Andersson et al. are included.

smooth muscle cell’ facet contains ten sub-facets, many of which are functionally similar. I hypothesised that disruption to cells from the different sub-facets contained in the ‘mesenchymal stem cell’ facet may produce different diseases. The mean JSD between samples in the ‘mesenchymal stem cell facet’ is 0.250. I therefore chose a threshold of 0.250 to ensure that the samples contained in this facet were combined at the level of the sub-facet, whilst samples in facets with more-similar expression profiles were combined at the level of the facet. I combined samples by computing the mean TPM (Figure 3.1G), producing gene expression profiles for 74 cell types.

Percentile normalisation

When generating context-specific networks, I use gene expression data as a proxy for protein abundance data, as high-quality protein abundance data is currently only available for a limited number of cell types (Kim et al., 2014; Wilhelm et al., 2014). Barshir et al. (2014) demonstrated that the correlation between mRNA transcript levels and protein abundance for genes in a single sample is low. Wilhelm et al. (2014) conducted similar analyses, but noted that the correlation between mRNA transcript levels and protein abundance for a single gene across multiple samples is higher. This supports the hypothesis that mRNA transcript translation rates are the dominant factor in determining protein abundance across cellular conditions (Schwanhäusser et al., 2011). Therefore, while relative gene expression may be a poor proxy for relative protein abundance when comparing multiple genes in a single sample, relative gene expression may be a better proxy for relative protein abundance when comparing a single gene across multiple samples. For this reason, I use the relative expression of genes across samples, rather than the relative expression of genes within samples, to generate cell-type-specific PPI networks and identify disease-cell-type associations.

To compute the relative expression value of each gene, I divided the expression value of each gene by the mean expression value of the gene across all gene expression

profiles:

$$e'_{i,l} = \frac{ke_{i,l}}{\sum_{j=1}^k e_{i,j}} \quad (3.1)$$

where k is the number of gene expression profiles, $e_{i,l}$ is the expression value of gene i in profile l and $e'_{i,l}$ is the relative expression value of gene i in profile l .

Many cell types contain a small number of highly expressed genes. For example, the expression of HBB in reticulocytes is 176,857 times higher than the median expression value of the gene across all cell types. If the raw relative expression values were used to generate cell-type-specific PPI networks, then these highly expressed genes would produce a small number of edges of very high weight. This would make it difficult to compute distances between vertex pairs using the RWR method, as these high-weight edges would strongly influence the movement of the random walker. I therefore percentile-normalised the relative gene expression scores. For each expression profile, I ranked and transformed the relative gene expression values so that they ranged uniformly from zero to one, where zero represents the most relatively under-expressed gene in the profile and one represents the most relatively over-expressed gene in the profile. This percentile normalisation limits the range of expression scores, preventing the generation of context-specific PPI networks with very high edge weights.

Numbers of highly expressed genes

After division by the mean expression value, many of the genes in the hepatocyte expression profile had large relative expression scores. The hepatocyte expression profile contained 364 genes with a relative expression score greater than 57 (in the top 0.1% of scores), compared to a median number of 6 across all cell types. Application of two alternative sample normalisation methods (the trimmed mean of M-values method (Robinson et al., 2010) and the upper quartile normalisation method (Bullard et al., 2010)) did not resolve this difference. Inclusion of the hepatocyte gene expression profile in the data set used by the GSC and GSO methods may have a deleterious impact on the performance of the methods, as

many incorrect associations could be identified between diseases and hepatocytes. I therefore removed the hepatocyte samples from the gene expression data set and recomputed the relative gene expression scores. This reduced the number of cell types in the final data set from 74 to 73.

3.2.2 Protein-protein interaction data

I obtained PPI data from the STRING database (Franceschini et al., 2013). I chose to use STRING because it contains a large number of expertly curated experimentally validated PPIs. It also provides confidence scores for each of these interactions. This means that it is possible to use data from STRING to select only the highest-confidence experimentally validated PPIs.

I downloaded all experimentally validated interactions between human proteins from version 9.1 of the STRING database. I mapped each protein to an Ensembl gene identifier using the `biomaRt` R-package (Durinck et al., 2009). If an Ensembl gene identifier could not be found for a protein, then I removed it from the network along with all of the corresponding interactions. To ensure that all interactions were of high confidence, I discarded interactions with confidence scores less than 0.8 (see Section 3.3.3 for justification). The final data set contains 32,275 interactions between 7,332 proteins.

3.2.3 Generating context-specific PPI networks

I generated context-specific PPI networks by integrating the percentile-normalised gene expression profiles generated using data from the FANTOM5 project and the PPI data from STRING. Let $G = (V, E)$, where V is a set of n vertices representing genes and $E \subseteq V \times V$ is a set of m edges representing physical interactions between the protein products of the genes. To generate a PPI network specific to a certain context, each edge in E is weighted using the product of the percentile-normalised expression scores of the corresponding genes:

$$w_{i,j,l} = x_{i,l}x_{j,l} \tag{3.2}$$

where $x_{i,l}$ is the percentile-normalised expression score of gene i in context l and $w_{i,j,l}$ is the weight of the edge connecting gene i and gene j .

If two genes are relatively over-expressed in a context compared to other contexts, then the weight of the edge connecting the genes will be greater. This represents a greater probability that an interaction takes place between the protein products of the genes in the context. This method is similar to the ‘edge reweight’ method described by Magger et al. (2012) (Figure 1.6). The method described here is different in two ways however, in that it (1) uses percentile-normalised gene expression profiles and (2) does not apply an arbitrary cutoff to the gene expression data.

3.2.4 Disease-gene association data

The GSC and GSO methods identify disease-associated contexts using known disease-associated genes. I downloaded disease-gene associations from version 2.1 of the DisGeNET database (Pinero et al., 2015). I chose to use DisGeNET because it combines multiple databases of disease-gene associations. As described in Section 1.2.2, DisGeNET contains associations identified through literature curation, automated text mining and associations predicted using model organisms. DisGeNET represents one of the largest databases of disease-gene associations currently available (Pinero et al., 2015). Version 2.1 of the database contains 13,185 diseases associated with one or more genes.

I applied a number of filtering steps to the DisGeNET data set to select the highest quality disease-gene associations relevant to the analyses described in this chapter. First, I removed associations identified using automated text mining and predicted using animal models, as I deemed these less likely to be of high quality. I also removed associations related to the genetic response to environmental chemicals, as these are less relevant to the study of disease. The removal of these associations reduced the number of diseases in the data set, associated with one or more genes, to 3,856. Each of the diseases in the DisGeNET database is assigned to one or more broader disease classifications. It has previously been observed that unlike other diseases, cancer-associated genes are not over-expressed in the tissues in which the

cancer is located (Lage et al., 2008). I therefore removed diseases belonging to the ‘neoplasms’ class from the data set. I also removed diseases belonging to the ‘congenital, hereditary, and neonatal diseases and abnormalities’ class, because of the lack of fetal gene expression data in the FANTOM5 project data set. This reduced the number of diseases associated with one or more genes to 2,143. Next, I removed associations if they were supported by only a single evidence source, as I deemed these less likely to be of high quality, reducing the number of diseases associated with one or more genes to 1,898. Many of the disease names in DisGeNET are followed by a number that represents a disease sub-type. I pooled these disease sub-types to increase the number of genes associated with each disease, reducing the number of diseases to 1,679. I also removed genes with no expression data in the FANTOM5 project data set, reducing the number of diseases to 1,557. Some of the entries in DisGeNET (such as ‘DNA damage’) represent molecular processes rather than diseases. I therefore removed diseases not mapping to MeSH terms from the disease MeSH tree (tree C and F03 in the 2015 MeSH tree structure). This produced a data set containing 1,544 diseases associated with one or more genes. I mapped genes to Ensembl gene identifiers using the `biomaRt` R-package (Durinck et al., 2009).

3.2.5 Mapping diseases to contexts

In this section I outline three methods for systematically mapping diseases to contexts. In this thesis, these methods are used exclusively to identify associations between diseases and cell types. They could however also be used to map diseases to other contexts, such as tissues.

Gene Set Overexpression

It has been shown that disease-associated genes tend to be over-expressed in the tissues and cell types in which the disease manifests (Lage et al., 2008). Here described is the GSO method, an approach that I developed to systematically map diseases to contexts by identifying gene expression profiles in which sets of disease-associated genes are over-expressed. Hu et al. (2011) used a similar method with GWAS data to map diseases to a limited number of cell types. Unlike the method

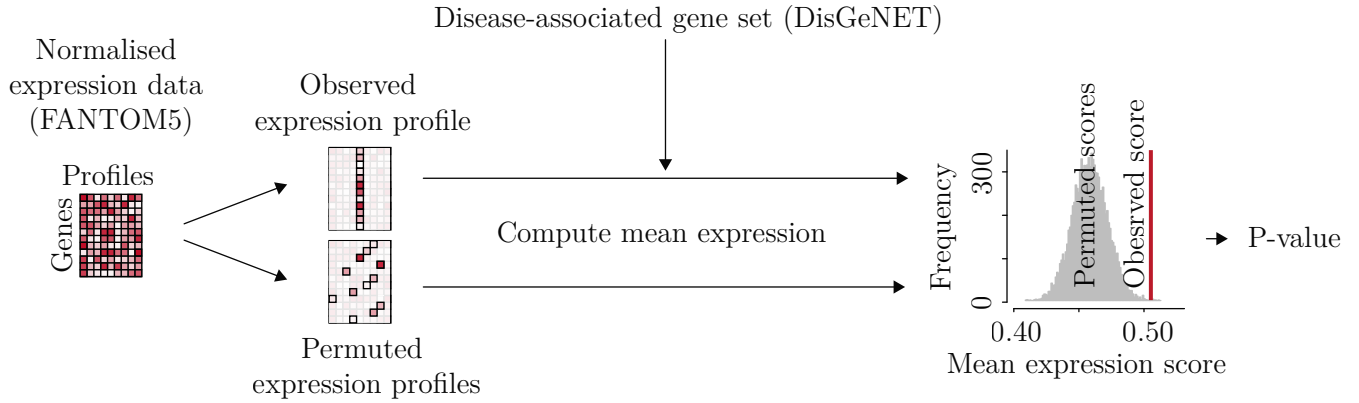


Figure 3.4: Overview of the GSO method. The mean percentile-normalised gene expression score of the disease-associated gene set is compared between the observed and permuted expression profiles. An empirical p -value, describing the probability that a mean expression score at least as high as that observed occurs by chance is generated by taking the proportion of permuted expression profiles in which the mean expression score of the gene set is equal to or greater than the mean expression score of the gene set in the observed expression profile.

developed by Hu et al., the GSO method is not limited to GWAS data and can be applied to sets of disease-associated genes identified using different approaches.

The GSO method uses a permutation-based approach to identify contexts in which a set of disease-associated genes is significantly more over-expressed than expected by chance (Figure 3.4). If a set of disease-associated genes is identified as being over-expressed in a given context, then the disease is said to be associated with the context.

To determine whether disease D is associated with context l , let S be the set of genes known to be associated with D . u permuted gene expression profiles are then generated. Each permuted expression profile is generated by sampling, for each gene, a cell type at random from the 73 cell types considered and then taking the percentile-normalised gene expression score of the gene from that cell type. This means that like the observed expression profile, the permuted expression profiles consist of percentile-normalised gene expression scores.

Genes in S with no available expression score are given scores of 0.5 (the median

percentile-normalised gene expression score). An empirical p-value, describing the probability that a mean gene expression score at least as high as that observed occurs by chance, is produced by taking the proportion of permuted expression profiles in which the mean expression score of S is equal to or greater than the mean expression score of S in the expression profile observed for l . A minimum p-value of $1/u$ is applied to ensure that no p-values equal 0. I ran the GSO method using 10,000 permutations to produce the results described in this thesis.

Gene Set Compactness

Sets of disease-associated genes are often enriched in certain cellular pathways, processes and protein complexes, the disruption of which may lead to the development of the disease (Baranzini et al., 2009). Many of these cellular components are represented in PPI networks, meaning that sets of disease-associated genes tend to cluster in PPI networks (Bauer-Mehren et al., 2011). As described in Chapter 1, there is currently little information about the contexts in which interactions between proteins take place, meaning that the PPI networks generally used in bioinformatics contain pathways, processes and protein complexes active and present in different contexts. If context-specific PPI networks were available, then we may observe that sets of disease-associated genes cluster even more significantly in PPI networks specific to contexts related to each disease, as the cellular components whose disruption leads to the development of the disease would be more likely to be represented in these context-specific networks.

Here described is the GSC method, a novel approach that systematically maps diseases to contexts by comparing the clustering of sets of disease-associated genes across context-specific PPI networks. If a set of disease-associated genes clusters more significantly than expected by chance in a PPI network specific to a certain context, given the data used to generate the networks, then the disease is said to be associated with the context.

The GSC method is based on the compactness function, which is defined as the mean distance between pairs of vertices in a graph (Glaab et al., 2010). The

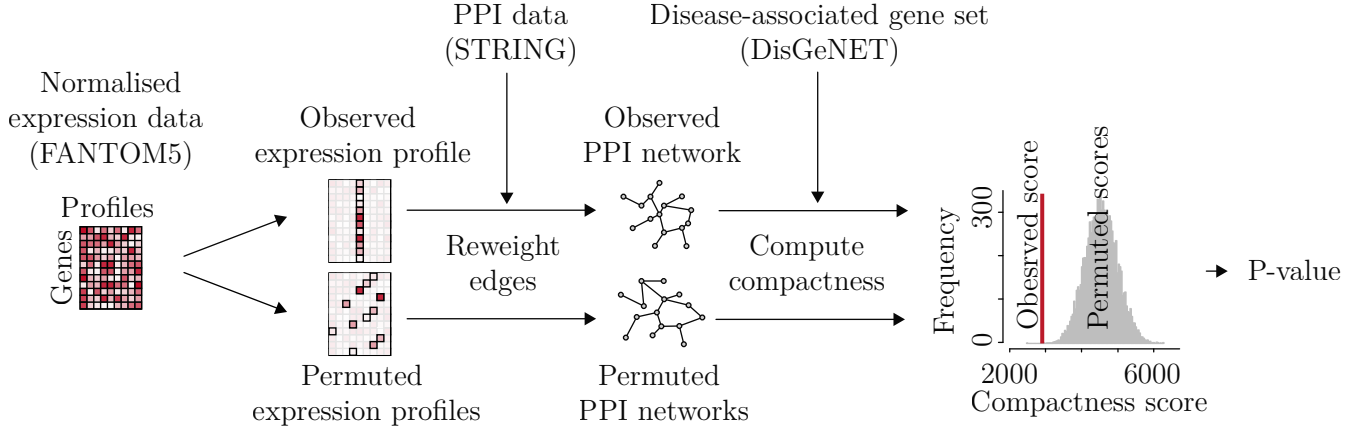


Figure 3.5: Overview of the GSC method. The compactness score of the disease-associated gene set is computed in the observed context-specific PPI network and u permuted PPI networks. An empirical p -value describing the probability that a compactness score at least as low as that observed occurs by chance, given the data used to generate the networks, is generated by taking the proportion of permuted networks in which the compactness score of the disease-associated gene set is less than or equal to the compactness score of the gene set in the observed context-specific PPI network.

compactness score C of vertex set P in graph G is

$$C(P, G) = \frac{\sum_{i, j \in P} d^G(i, j)}{|P|^2} \quad (3.3)$$

where $d^G(i, j)$ is the distance between vertex i and vertex j in graph G and $|P|$ is the size of vertex set P . Distances can be computed using different methods, including the shortest paths (Cormen et al., 2009) and RWR (Köhler et al., 2008) methods. A smaller compactness score for P indicates that the vertices in P are located closer to one another in G . If the compactness score of a gene set is smaller than expected by chance, then the gene set can be said to cluster in the network (Cornish & Markowetz, 2014).

To identify disease-context associations using the GSC method, context-specific PPI networks are first generated using the method described in Section 3.2.3. The

compactness score of the set of disease-associated genes S is then computed in each of these observed context-specific PPI networks. The RWR method is used to compute distances between vertex pairs (see Section 2.6.1 for details). For each vertex, all vertices are ranked by their distance to the vertex, so that the vertex on which the random walker is most likely to be located is ranked first. These ranks are used as the vertex pair distances in Equation 3.3.

To determine whether the compactness score of S in a context-specific network is smaller than expected by chance, given the data used to generate the network, the gene expression profiles used to generate the observed networks are permuted, using the same approach as the GSO method. These permuted expression profiles are then used to generate u permuted networks. A empirical p-value is then produced for each context by taking the proportion of permuted PPI networks in which the compactness score of S is less than or equal to the compactness score of S in the observed context-specific PPI network. A minimum p-value of $1/u$ is applied to ensure that no p-values equal 0. I ran the GSC method using 10,000 permutations to produce the results described in this thesis.

Text mining

The work described in this section was carried out by Dr. Ioannis Filippis (I.F.) and myself. I.F. and I designed the research, I.F. and I completed the mapping of diseases and cell types to MeSH terms, I.F. wrote the text mining software and I analysed the results.

We used text mining and the PubMed database to generate an independent set of disease-cell-type associations (Figure 3.6). This was done by first mapping each disease in the DisGeNET data set and each cell type in the FANTOM5 project data set to one or more MeSH terms. MeSH is a controlled vocabulary that describes the topics mentioned in each article indexed in PubMed (NCBI Resource Coordinators, 2016). In this method, MeSH terms are used to query the PubMed database and identify articles that mention each disease and cell type. If the number of articles that mention both a disease and a cell type is greater than expected by chance, given the number of articles individually mentioning the terms, then the disease and the

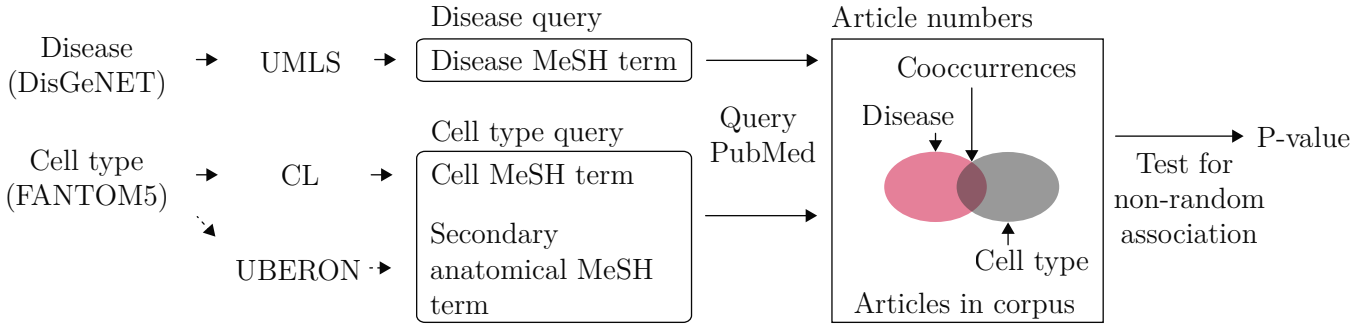


Figure 3.6: Overview of the text mining method. Each disease and cell type is mapped to one or two MeSH terms using cross-referenced controlled vocabularies. These MeSH terms are then used to identify articles indexed in PubMed individually and co-mentioning the disease and the cell type. Fisher’s exact test is used to determine whether the number of articles co-mentioning the terms is greater than expected by chance, given the number of articles that mention the terms individually, under the null hypothesis that the disease is not associated with the cell type.

cell type can be said to be associated (Korbel et al., 2005).

We used the UMLS controlled vocabulary (Bodenreider, 2004) to map diseases in the DisGeNET data set to MeSH terms. In the DisGeNET database, each disease is mapped to a UMLS term. We used the UMLS Metathesaurus to map each of these terms to a MeSH term (Bodenreider, 2004). We mapped diseases mapped to UMLS terms not present in the UMLS Metathesaurus to MeSH terms by querying the MeSH database (<http://www.ncbi.nlm.nih.gov/mesh>) with the disease name.

We mapped each cell type represented as a facet or sub-facet in the FANTOM5 project data set to one or two MeSH terms using multiple cross-referenced controlled vocabularies. While the disease-related sections of the MeSH vocabulary are sufficiently comprehensive to map the majority of DisGeNET diseases to a unique MeSH term, the cell-type-related sections of the MeSH vocabulary are less extensive, meaning that it was not possible to assign each cell type a single unique MeSH term. We therefore assigned some cell types two MeSH terms: one from the cell type MeSH tree (A11) and one from another anatomical MeSH tree (sub-trees of A that are not A11), in order to differentiate between the cell types. When querying PubMed for articles mentioning these cell types, we identified articles annotated with both terms.

We used the FANTOM5 ontology (FF) provided by The FANTOM Consortium to map the cell types to MeSH terms. This ontology describes the relationships between each of the FANTOM5 project samples. It also contains cross-references to terms from other controlled vocabularies, including the Cell Ontology (CL), a cross-species cell type ontology (Meehan et al., 2011), and Uberon, a cross-species anatomical ontology (Mungall et al., 2012). Using the CL cross-referencing, we mapped each FANTOM5 project cell type to a CL term. If the FF term corresponding to a facet or sub-facet was not cross-referenced with a CL term, then ancestral terms in the FF ontology, up to a distance of two, were considered. If this consideration of ancestral FF terms still did not identify a suitable CL term, then we manually chose a representative CL term from the 8 July 2014 release of the CL. If the consideration of ancestral FF terms identified multiple CL terms, then we manually chose the most suitable CL term. CL terms were chosen manually by searching for the occurrence of words in the facet and sub-facet names in the names and definitions of CL terms. Representative MeSH terms were finally identified by querying the MeSH database with the mapped CL terms.

If the CL term mapped to a cell type did not uniquely identify the cell type in the data set, then we identified an additional anatomical MeSH term. We identified anatomical MeSH terms using the 21 June 2014 release of Uberon. For each of the cell types, a representative Uberon term was identified using the same approach used to identify CL terms. However, due to the smaller number of Uberon terms cross-referenced in the FF ontology, ancestral terms up to a distance of five were considered. We identified anatomical MeSH terms by querying the MeSH database with the mapped Uberon terms.

Once each disease and cell type had been mapped to one or two MeSH terms, we used the Entrez Programming Utilities service (NCBI Resource Coordinators, 2016) to count the number of articles in the PubMed database annotated with the terms. The PubMed database was queried for each disease and cell type individually and each disease/cell-type pair. We used Fisher’s exact test to determine whether the number of articles mentioning both a disease and a cell type was greater than expected by chance, given the number of articles individually mentioning the disease

and the cell type (Cheung et al., 2012). For each disease/cell-type pair, the p-value was produced using a contingency table containing the number of articles co-mentioning the disease and the cell type, the number of articles mentioning the disease but not the cell type, the number of articles mentioning the cell type but not the disease, and the number of articles in the PubMed database that mention neither the disease nor the cell type. This mining of the PubMed database was completed on 23 April 2015.

Due to the ontological structure of the MeSH vocabulary, some diseases and cell types mapped to MeSH terms that are either ancestors or descendants of MeSH terms mapped to other diseases and cell types. These relationships were not considered when comparing the results of the different disease-context association identification methods.

3.3 Results

3.3.1 Network topology features

Using the method described in Section 3.2.3, I integrated PPI and gene expression data to generate 73 cell-type-specific PPI networks. Each of these cell-type-specific networks was generated using a different gene expression profile and therefore the topological features of the networks differ. In this section I discuss the topological features of the cell-type-specific PPI networks and compare these features to the features of the unweighted generic PPI network generated using the same PPI data.

Edge weights in the cell-type-specific networks are computed using the product of the percentile-normalised gene expression scores. These scores range between 0 and 1 and therefore the edge weights also range between 0 and 1. Figure 3.7 shows the combined distribution of edge weights from the 73 cell-type-specific networks. Across all networks, the median edge weight is 0.337. Between networks however, the median edge weights vary. The highest median edge weight occurs in ‘Mesenchymal Stem Cells - Adipose’ (0.433) and the lowest median edge weight occurs in ‘Neurons’ (0.299). This difference between networks occurs because although the expression scores for each cell type have been percentile normalised to range between 0 and

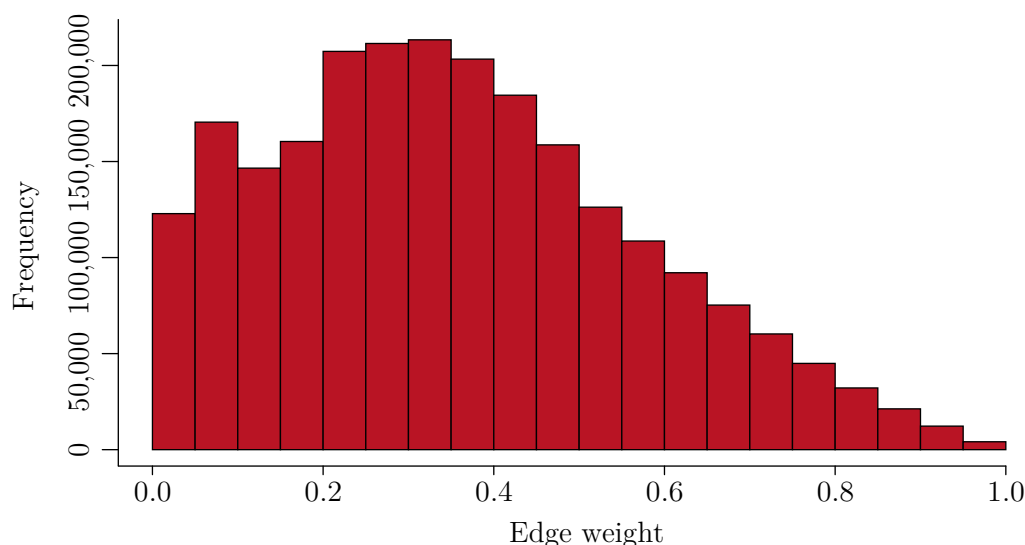


Figure 3.7: Histogram showing the frequency of edges with different weights across the 73 cell-type-specific PPI networks. I combined the edge weights from all 73 networks to generate this histogram.

1, genes with higher degrees in the networks will exert a greater influence on the distribution of edge weights, resulting in networks with different median edge weights.

I next compared the centrality of genes in the generic and cell-type-specific networks. A number of different network centrality measures are available, including vertex degree and betweenness (White & Smyth, 2003). The fact that the edge reweighting method used to generate the cell-type-specific networks does not result in the removal of any edges means that each gene has equal degree across the generic and cell-type-specific networks. Vertex degree is therefore not a useful measure for comparing the networks. For this reason I used vertex betweenness to compare the centrality of genes across the networks.

I computed the betweenness of each vertex in the generic and each of the cell-type-specific networks. If a network has a small number of hub vertices of high impact then we may expect the network to have a small number of vertices with higher betweenness scores, and a much larger number of vertices with lower betweenness scores. To determine whether this is the case in the generic and cell-type-specific networks, I computed ten quantiles for the distribution of

Quantile	Generic	Cell-type-specific (mean)
0%	0	0
10%	0	0
20%	0	0
30%	0.737	0
40%	102	0
50%	565	1.60
60%	2,450	283
70%	7,330	7,330
80%	12,041	14,666
90%	27,004	38,137
100%	11,555,460	8,912,198

Table 3.3: Vertex betweenness scores in the generic and cell-type-specific PPI networks. For each network I first computed the betweenness score of each vertex and then computed 10 quantiles from the resulting distribution of scores. For the cell-type-specific networks, I then computed the mean of each quantile across the networks.

betweenness scores in each network (Table 3.3). In both the generic and cell-type-specific networks, there are small numbers of vertices with high betweenness scores, indicating that each of the networks contains hub vertices. In the cell-type-specific networks, the betweenness scores of vertices in the 80% and 90% quantiles are greater than the betweenness scores of vertices in the same quantiles in the generic network, suggesting that the cell-type-specific networks may contain hub vertices with greater impact on network topology.

3.3.2 Disease-associated cell-type-specific sub-networks

Figure 3.8 shows sub-networks from two cell-type-specific PPI networks, containing genes associated with psoriasis and arthritis and their interacting partners. The two

cell-type-specific PPI networks were chosen manually to illustrate the enrichment of high-weight edges in PPI networks specific to cell types related to the diseases. To determine whether these sub-networks are enriched with high-weight edges, I permuted the expression profiles of the chosen cell types and used these permuted expression profiles to generate permuted sub-networks. Each of the permuted sub-networks contained the same set of genes as the observed sub-network. To measure enrichment, I compared the number of high-weight edges in the observed and permuted sub-networks. I define high-weight edges as those in the top 1% of edge weights (a weight greater than 0.900). I generated empirical p-values by measuring the proportion of permuted sub-networks with a greater number of high-weight edges than the observed sub-networks. Both the monocyte-specific psoriasis sub-network and the arthritis-specific neutrophil sub-network are enriched with high-weight edges (both $p < 0.0001$). To improve visual interpretation, genes with more than 15 interacting partners were not included in Figure 3.8. These genes were however included in the enrichment analysis.

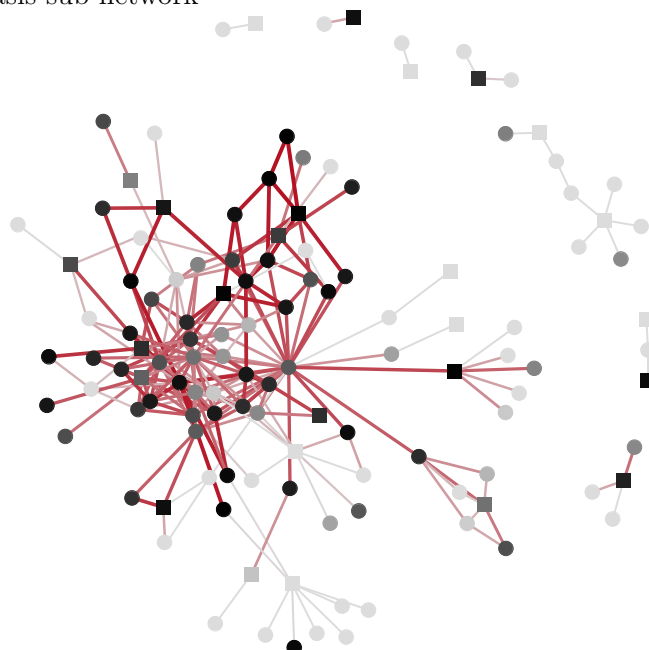
The two sub-networks indicate that the monocyte and neutrophil-specific PPI networks may be useful for studying psoriasis and arthritis. As previously mentioned, many methods in bioinformatics that use networks to study disease use a generic network, rather than a network specific to a context related to the disease being studied. The enrichment of high-weight edges connecting the disease-associated genes in Figure 3.8 indicates that interactions between disease-associated genes may be stronger in PPI networks specific to cell types related to the disease. These PPI networks may therefore be useful in tasks such as gene prioritisation.

3.3.3 Parameter selection

The GSC method uses two parameters: the confidence score cutoff applied to the STRING data set (see Section 3.2.2 for details) and the random walk restart probability r (see Section 2.6.1 for details). In this section I test how these two parameters affect the performance of the GSC method.

To measure how parameter choice affects GSC method performance, I compared the associations identified by the GSC method to the associations identified by the

A) Monocyte-specific psoriasis sub-network



B) Neutrophil-specific arthritis sub-network

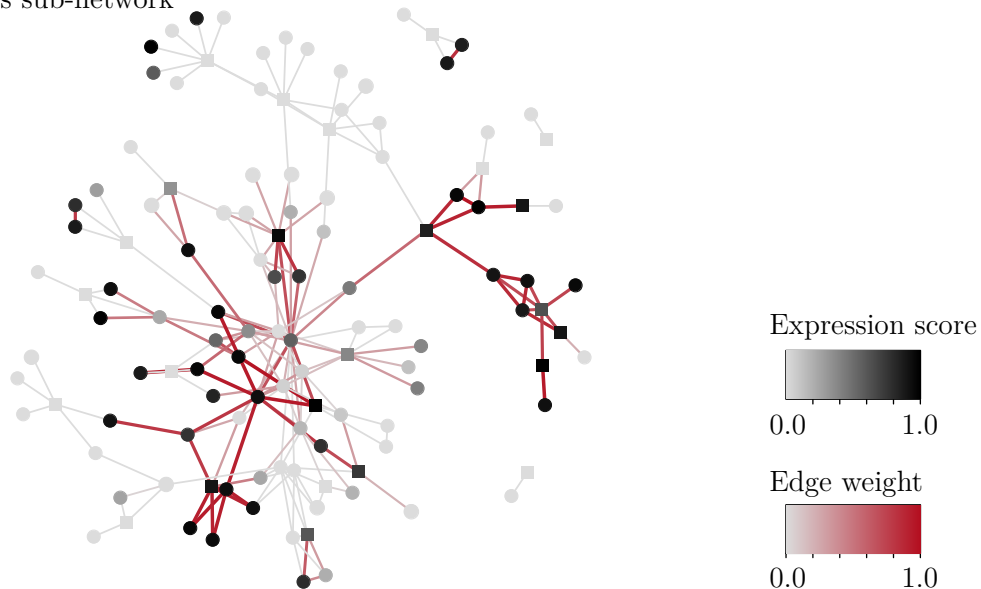


Figure 3.8: Sub-networks from cell-type-specific PPI networks containing disease-associated genes (squares) and their interacting partners (circles). Genes with greater normalised gene expression scores are darker in colour. Edges with greater weights are darker in colour and thicker. To improve visual interpretation of the sub-networks, genes with more than 15 interacting partners were excluded. Both sub-networks are enriched with high-weight edges (both $p < 0.0001$, without excluding genes with more than 15 interacting partners).

text mining method. The GSC method works by measuring the distances between pairs of vertices in context-specific PPI networks. The GSC method therefore cannot be applied to diseases with only a single known associated gene. I applied the GSC method to the 503 diseases in the filtered DisGeNET data set with two or more associated genes and the 73 cell types in the FANTOM5 project data set. I used text mining to generate an independent set of associations between these 503 diseases and 73 cell types. I ran the GSC method using STRING confidence score cutoffs of 0.0, 0.2, 0.4, 0.6 and 0.8 and values of r equal to 0.1, 0.3, 0.5, 0.7 and 0.9. I applied the BH procedure (see Section 2.3.2 for details) to the output sets of p-values to correct for multiple testing. I then applied a FDR cutoff of 10% to the computed q-values to produce sets of disease-cell-type associations (a FDR cutoff of 10% is used throughout the remainder of this chapter, unless otherwise specified).

Table 3.4 shows the significance of the overlap between the associations identified by the GSC method and the text mining method for each combination of tested parameters. Table 3.5 shows the F1-score (see Section 2.1 for details) for the GSC method, computed using associations identified by the text mining method as true positives. Running the GSC method using different confidence score cutoffs had little effect on performance. I therefore use interactions from STRING with a confidence score of 0.8 or greater throughout the remainder of this chapter.

Using different values of r also had little effect on method performance. A number of methods that use the RWR method and biological networks to identify protein complexes (Macropol et al., 2009), predict gene-phenotype associations (Li et al., 2010) and prioritise disease-associated genes (Smedley et al., 2014) successfully use restart probabilities of 0.7. I therefore set r equal to 0.7 to complete the remaining analyses in this chapter.

3.3.4 Effect of gene set size on method performance

The number of genes associated with each disease in the DisGeNET data set varies. To determine whether there is a minimum number of disease-associated genes required by the GSC and GSO methods to successfully identify disease-cell-type associations, I organised diseases into sets of diseases with similar numbers of

Restart r	STRING interaction confidence score cutoff				
	0.0	0.2	0.4	0.6	0.8
0.1	9×10^{-238}	4×10^{-243}	2×10^{-245}	7×10^{-231}	1×10^{-236}
0.3	3×10^{-247}	4×10^{-233}	3×10^{-243}	1×10^{-244}	4×10^{-224}
0.5	7×10^{-239}	8×10^{-231}	7×10^{-245}	1×10^{-231}	2×10^{-211}
0.7	4×10^{-238}	3×10^{-240}	7×10^{-227}	3×10^{-228}	3×10^{-202}
0.9	4×10^{-229}	6×10^{-230}	7×10^{-221}	4×10^{-215}	4×10^{-192}

Table 3.4: The performance of the GSC method when run using interactions from STRING passing different confidence score cutoffs and using different restart probabilities r . The performance of the GSC method was measured by computing the significance of the overlap between the associations identified by the GSC method and the associations identified by the text mining method. I measured overlap significance using Fisher’s exact test. I generated sets of disease-cell-type associations by applying a FDR cutoff of 10% to the results of the methods.

associated genes. I then applied the GSC and GSO methods to these sets of diseases and measured their performance using the associations identified by the text mining method as true positives.

Each of the generated disease sets contains at least 20 diseases. To generate the sets, I first sorted diseases by their number of associated genes. I next added diseases to a disease set, starting from the disease with the greatest number of associated genes. When the number of diseases in the disease set reached 20, I started a new set and added the following diseases to this new set. Diseases with the same number of associated genes were always added to the same disease set.

As previously mentioned, the GSC method works by measuring the distance between pairs of genes in a network and therefore cannot be applied to diseases with only a single known associated gene. The permutation-based approach used by the GSO method to compute p-values means that if a disease is associated with only a single gene, then it is not possible for the method to identify associations that pass a FDR of 10%. To measure the effect of gene set size on method performance,

Restart r	STRING interaction confidence score cutoff				
	0.0	0.2	0.4	0.6	0.8
0.1	0.237	0.239	0.240	0.232	0.245
0.3	0.239	0.232	0.238	0.237	0.233
0.5	0.234	0.229	0.237	0.228	0.225
0.7	0.231	0.232	0.225	0.226	0.218
0.9	0.221	0.223	0.219	0.216	0.220

Table 3.5: The performance (F1-score) of the GSC method when run using interactions from STRING passing different confidence score cutoffs and using different restart probabilities r . I measured the performance of the GSC method using the associations identified by the text mining method as true positives. I generated sets of disease-cell-type associations by applying a FDR cutoff of 10% to the results of the methods.

I therefore applied both the GSC and GSO methods to the 503 diseases in the DisGeNET data set associated with two or more genes.

As the number of genes associated with each disease increases, the performance of the GSC (Table 3.6) and GSO (Table 3.7) methods improve. The overlap between the associations identified by the GSC and GSO methods and the text mining method is significant for sets of diseases with three or more associated genes ($p < 0.05$, after Bonferroni correction). For this reason, I completed the remaining analyses described in this chapter using the 352 diseases in the DisGeNET data set associated with three or more genes.

3.3.5 Comparison of the associations identified by the methods

The GSC and GSO methods identified 752 and 599 associations between the 352 diseases and the 73 cell types at a 10% FDR. The overlaps between these sets of disease-cell-type associations and the associations identified by the text mining method are significant at FDRs of 5%, 10% and 20% (Tables 3.8 and 3.9), providing validation for the GSC and GSO methods.

Number of genes	Number of diseases	Precision	Recall	F1-score	Overlap significance	Overlap significance adjusted
54 - 237	20	0.365	0.347	0.356	1×10^{-19}	2×10^{-18}
31 - 53	21	0.593	0.437	0.503	7×10^{-45}	9×10^{-44}
22 - 30	20	0.481	0.262	0.339	2×10^{-19}	2×10^{-18}
16 - 21	25	0.521	0.330	0.404	6×10^{-29}	8×10^{-28}
12 - 15	22	0.489	0.193	0.277	8×10^{-15}	9×10^{-14}
9 - 11	30	0.481	0.176	0.258	3×10^{-17}	3×10^{-16}
7 - 8	28	0.391	0.060	0.104	2×10^{-05}	2×10^{-04}
6	30	0.643	0.099	0.172	9×10^{-14}	1×10^{-12}
5	31	0.412	0.053	0.095	2×10^{-05}	3×10^{-04}
4	48	0.395	0.071	0.121	7×10^{-10}	8×10^{-09}
3	77	0.456	0.063	0.110	3×10^{-15}	3×10^{-14}
2	151	0.097	0.006	0.011	2×10^{-01}	1×10^{-00}
1	1041	-	-	-	-	-

Table 3.6: The performance of the GSC method when applied to sets of diseases with different numbers of associated genes. The performance of the method was measured using the associations identified by the text mining method as true positives. The GSC method cannot be applied to diseases with only a single known associated gene and therefore the performance statistics for this row are not available. Sets of disease-cell-type associations were produced by applying a FDR cutoff of 10% to the results of the methods. Overlap significance was computed using Fisher’s exact test and p-values adjusted for multiple testing using Bonferroni correction.

Number of genes	Number of diseases	Precision	Recall	F1-score	Overlap significance	Overlap significance adjusted
54 - 237	20	0.429	0.322	0.370	4×10^{-21}	5×10^{-20}
31 - 53	21	0.634	0.353	0.454	9×10^{-38}	1×10^{-36}
22 - 30	20	0.468	0.206	0.286	6×10^{-15}	8×10^{-14}
16 - 21	25	0.433	0.226	0.297	4×10^{-17}	5×10^{-16}
12 - 15	22	0.406	0.114	0.178	7×10^{-08}	8×10^{-07}
9 - 11	30	0.476	0.141	0.217	7×10^{-14}	8×10^{-13}
7 - 8	28	0.500	0.067	0.118	3×10^{-07}	4×10^{-06}
6	30	0.630	0.094	0.163	8×10^{-13}	9×10^{-12}
5	31	0.643	0.069	0.124	9×10^{-09}	1×10^{-07}
4	48	0.567	0.071	0.127	4×10^{-13}	5×10^{-12}
3	77	0.229	0.027	0.048	6×10^{-04}	7×10^{-03}
2	151	0.114	0.010	0.018	5×10^{-02}	6×10^{-01}
1	1041	-	-	-	-	-

Table 3.7: The performance of the GSO method when applied to sets of diseases with different numbers of associated genes. The performance of the GSO method was measured using the associations identified by the text mining method as true positives. Due to the permutation-based approach used by the GSO method to compute p-values, no associations involving diseases with only a single known associated gene can pass a 10% FDR cutoff. Performance statistics for this row are not given for this reason. Sets of disease-cell-type associations were produced by applying a FDR cutoff of 10% to the results of the methods. Overlap significance was computed using Fisher's exact test and p-values adjusted for multiple testing using Bonferroni correction.

	FDR		
	5%	10%	20%
Identified by the GSC method	547	752	1031
Identified by the text mining method	1847	1913	1992
Identified by the GSC and text mining methods	278	355	436
Overlap significance	9×10^{-172}	5×10^{-202}	7×10^{-219}

Table 3.8: The number of disease-cell-type associations identified by the GSC method, by the text mining method and by both methods at various FDRs. Sets of disease-cell-type associations were produced by applying FDR cutoffs to the results of the methods. Overlap significance was measured using Fisher’s exact test.

	FDR		
	5%	10%	20%
Identified by the GSO method	431	599	913
Identified by the text mining method	1847	1913	1992
Identified by the GSO and text mining methods	223	292	377
Overlap significance	4×10^{-139}	1×10^{-169}	6×10^{-183}

Table 3.9: The number of disease-cell-type associations identified by the GSO method, by the text mining method and by both methods at various FDRs. Sets of disease-cell-type associations were produced by applying FDR cutoffs to the results of the methods. Overlap significance was measured using Fisher’s exact test.

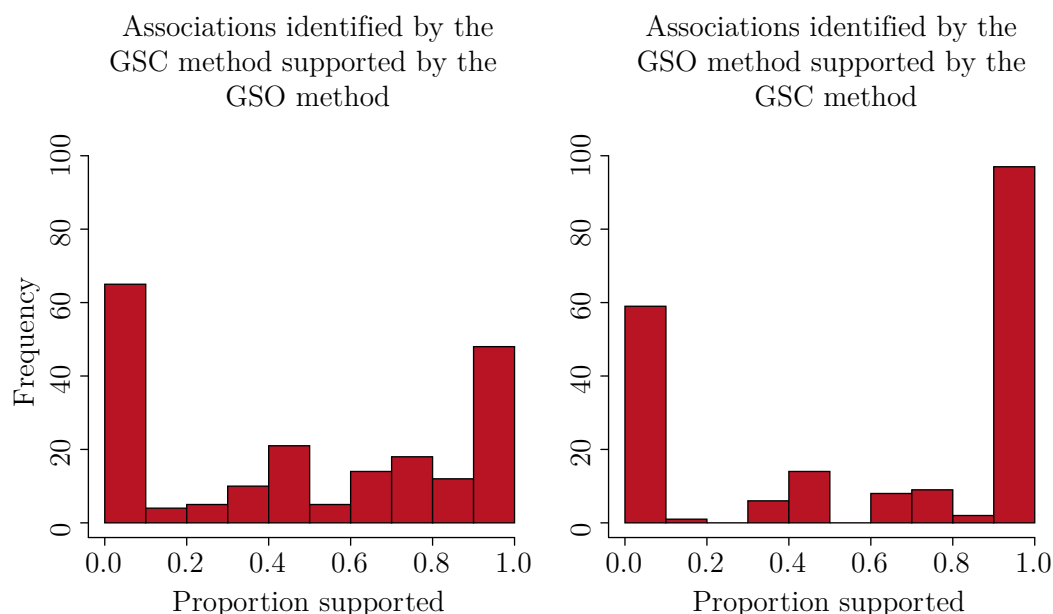
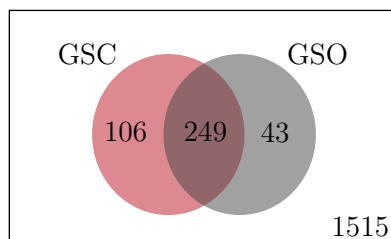


Figure 3.9: Histograms showing the support between the GSC and GSO methods. For each disease, the proportion of associations identified by one of the methods supported by the other was computed. Diseases were excluded if no cell types were identified as associated. A value of one indicates that all associations identified by the first method are supported by the second. A value of zero indicates that no associations identified by the first method are supported by the second. Sets of associations were produced by applying a FDR cutoff of 10% to the results of the methods.

Many of the associations identified by the GSC method are supported by the GSO method and *vice versa* (Figure 3.9). For 48.0% of the tested diseases, at least 50% of the associated cell types identified by the GSC method were also identified by the GSO method. Of the 355 associations identified by the GSC method, supported by the text mining method, 249 associations were also identified by the GSO method (Figure 3.10). However, 106 of the associations identified by the GSC method, supported by the text mining method, were not identified by the GSO method. This indicates that the GSC and GSO methods represent complementary approaches for identifying disease-cell-type associations.

As described in Section 3.2.5, I used the text mining method to independently identify associations between diseases and cell types. A disadvantage of using the

A) Associations supported by the text mining method



B) Associations not supported by the text mining method

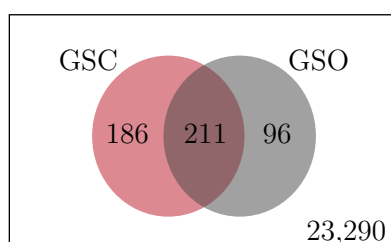


Figure 3.10: Associations identified by the GSC and GSO methods (A) supported and (B) not supported by the text mining method. Sets of associations were produced by applying a FDR cutoff of 10% to the results of the methods.

co-occurrence of disease and cell type MeSH terms in PubMed article annotations to identify associations is that this method does not take into account the relationships between the diseases and the cell types in the articles. This prevents the text mining method from distinguishing between those cell types directly affected by the disease-associated genes and those cell types indirectly affected later in the development of the disease. For example, psoriasis is an immune-related skin disorder caused by environmental factors, genetic factors and disruption to the immune system (Raychaudhuri et al., 2014). Because of the immunological basis of psoriasis, the text mining method identifies a number of immune cells as associated with the disease. Cell types present in the skin, including keratinocytes, are affected in the disease and therefore also mentioned in articles referring to psoriasis. The text mining method therefore identifies keratinocytes as being associated with psoriasis ($q < 1 \times 10^{-318}$). The text mining method is however unable to determine whether the psoriasis-associated genes directly affect keratinocytes, disrupting pathways and

Disease class	Cell type class	GSC	GSO	Text mining
Cardiovascular	Cardiovascular	6.0% (6/100)	6.2% (5/81)	28.6% (65/227)
Cardiovascular	Immune system	43.0% (43/100)	42.0% (34/81)	17.6% (40/227)
Immune system	Immune system	88.2% (149/169)	89.9% (116/129)	58.5% (185/316)
Mental disorders	Neural	57.6% (19/33)	52.0% (13/25)	72.7% (24/33)

Table 3.10: The proportion of cell types identified as associated with a disease class that are from certain cell type classes, by the GSC, GSO and text mining methods. I used the MeSH and CL terms mapped to each disease and cell type in Section 3.2.5 to identify diseases and cell types belonging to each class. If diseases were descendants of the following MeSH terms, then I mapped them to the corresponding disease classes: C14 (cardiovascular diseases), C20 (immune system diseases) and F03 (mental disorders). Likewise, if cell types were descendants of the following CL terms, then I mapped them to the corresponding cell type classes: CL:0002139 and CL:0002494 (cardiovascular cells), CL:0000738 (immune system cells) and CL:0002319 (neural cells).

processes active in these cell, or whether they instead disrupt pathways and processes active in immune cells, and keratinocytes are instead indirectly affected as a result of this disruption to immune cells.

The GSC, GSO and text mining methods identify different numbers of associations between diseases and cell types of certain classes (Table 3.10). For example, a greater proportion of the cell types identified as associated with immune system diseases by the GSC and GSO methods are cell types of the immune system, compared to the text mining method. Also, a greater proportion of the cell types identified as associated with cardiovascular diseases by the text mining method are cell types of the cardiovascular system, compared to the GSC and GSO methods. Many of the cell types identified by the GSC and GSO methods as being associated with cardiovascular disease are immune system cell types, possibly reflecting the role of the immune system in the development of some cardiovascular diseases (Frostedgard, 2013). These differences between disease classes may indicate that the

GSC and GSO methods are less effective when applied to certain disease classes. As previously mentioned, the text mining method is unable to differentiate between cell types directly affected by disease-associated genes and cell types indirectly affected later in disease development. This may explain some of the differences in the number of associations identified between the disease and cell type classes between methods.

3.3.6 Disease-cell-type associations identified by the GSC method

As previously mentioned, the GSC method identified 752 associations between the 352 diseases and the 73 cell types (Figure 3.11). Some of these diseases-cell-type associations are already well characterised, while others may warrant further study. In this section I detail some of the associations identified by the GSC method.

Well-characterised associations

The GSC method identified a number of well-characterised associations, including associations between neural cell types and mental disorders, such as neurons and bipolar disorder ($q < 0.007$). Neurons were also identified as being associated with substance abuse disorders, such as alcoholism ($q < 0.007$). A large number of associations were also identified between cell types of the immune system and autoimmune disorders, such as neutrophils and celiac disease ($q < 0.007$) and monocytes and ulcerative colitis ($q < 0.007$). Additionally, cell types of the immune system were identified as being associated with susceptibilities to a number of infectious diseases, including macrophages and susceptibility to malaria ($q < 0.013$). The GSC method also identified associations between a number of highly localised diseases and cell types, including lens epithelial cells and retinal diseases ($q < 0.007$).

Multiple sclerosis and mast cells

The GSC method identified mast cells as being associated with MS ($q < 0.027$). Mast cells are known to be involved in various allergic diseases, including asthma (Bradding et al., 2006). Whether mast cells are involved in the development of MS is not yet known, although there is some emerging evidence for an association. The blood brain barrier (BBB) is partly regulated by mast cells (Esposito et al., 2001)



Figure 3.11: Heat map of a subset of the disease-cell-type associations identified by the GSC method. The darker red the cell, the stronger the association. I adjusted p -values for multiple testing using the BH procedure. Each disease and cell type is involved in at least two associations at a 10% FDR. I organised diseases and cell types using complete-linkage hierarchical clustering, run using the $-\log_{10}$ of the q -values. Figure A.1 contains all associations between the 352 diseases and 73 cell types. MID: monocyte immature derived, MDEPC: monocyte derived endothelial progenitor cell.

and increased permeability of the BBB has been identified as one of the earliest changes to occur in MS (Minagar & Alexander, 2003). These observations have led to clinical trials of *masitinib*, a drug that acts as an inhibitor of mast cell activity, in MS patients. In a phase 2a clinical trial, *masitinib* was seen to produce small but non-significant improvements in MS patients (Vermersch et al., 2012). A phase 2b/3 clinical trial is currently underway (ClinicalTrials.gov identifier: NCT01433497).

Preeclampsia and adipocytes

One of the less well characterised associations identified by the GSC method is an association between preeclampsia and adipocytes ($q < 0.084$). Preeclampsia is a pregnancy complication characterised by the new onset of hypertension and proteinuria during the second half of pregnancy (Turner, 2010). It affects 5–8% of pregnancies. Multiple genes have been identified as being associated with preeclampsia (Pinero et al., 2015) but how these genes influence the development of the condition is unclear. Disruption to multiple cell populations, including endothelial cells, cells of the immune system and adipocytes, have been suggested to contribute to the development of preeclampsia.

During pregnancy, endothelial cells are key to the remodelling of the maternal vessels that provide nutrients and oxygen to the placenta and developing foetus (Powe et al., 2011). This process is disrupted in preeclampsia, possibly as a result of dysregulation in the endothelial cell population (Zhou et al., 1997).

Natural killer cells are known to produce signalling proteins that affect endothelial cell migration (Laresgoiti-Servitje et al., 2010). These proteins have been observed at higher concentrations in preeclampsia, leading to the suggestion that aberrant natural killer cell signalling may also contribute to the development of the condition.

Adipocytes produce adipokines, which have also been observed to affect endothelial cell function (Roberts et al., 2011). Obesity is associated with a three-fold increase in the risk of preeclampsia and dysregulated adipokine production (Roberts et al., 2011), supporting the hypothesis that adipocytes may be involved in the development of the condition. The association identified between preeclampsia

and adipocytes by the GSC method suggests that preeclampsia-associated genes may affect an individual's risk of developing the condition by disrupting the processes and pathways active in adipocytes.

Osteoarthritis and chondrocytes

Osteoarthritis is a disease of the joints that results in the degradation of the cartilage and bone (Glyn-Jones et al., 2015). It affects 14% of people over the age of 60 and is therefore considered an age-related disease (Glyn-Jones et al., 2015). While many people exhibit age-related changes in their joints, often as a result of general 'wear and tear', only some of these people display the symptoms characteristic of osteoarthritis (Loeser et al., 2012). There is therefore interest in identifying the factors that separate these groups of people, to better understand the etiology of the disease and aid in its diagnosis. Multiple variants have been identified that increase an individual's risk of developing the disease (Landrum et al., 2016), but how these variants influence disease susceptibility is not known.

Chondrocytes are the only cell type located in adult cartilage; the main tissue affected in osteoarthritis (Goldring & Goldring, 2007). These cells are responsible for repairing the damage that occurs to the tissue over time and it has therefore been suggested that osteoarthritis may occur as a result of disruption to chondrocyte function (Goldring & Goldring, 2007).

It has also been suggested that cells of the immune system may promote the development of the disease. Wang et al. (2011a) observed parts of the inflammatory complement system to be at higher concentrations in the synovial fluids of individuals with early stage osteoarthritis. To test the importance of these components, Wang et al. created mice models of osteoarthritis that were deficient in the components. These mice exhibited less cartilage loss than mice not deficient in the components, indicating that parts of the immune system may play a role in the development of osteoarthritis.

While the immune system may play a role in disease development, the GSC method identified chondrocytes as the cell type most-strongly associated with osteoarthritis ($q < 0.007$). This indicates that osteoarthritis-associated genes may

influence disease development by disrupting chondrocyte function.

3.3.7 Cell-type-based diseasomes

A diseasome represents a network of diseases connected by shared or similar disease features or mechanisms (Liu et al., 2014a). Diseasomes have been generated using known disease-associated genes (Goh et al., 2007), affected molecular pathways (Lee et al., 2008), co-morbidity data (Hidalgo et al., 2009), reported symptoms (Zhou et al., 2014) and through the integration of these data (Liu et al., 2014a). There has been an interest in mapping diseasomes as it has been demonstrated that identifying connections between diseases may aid in tasks such as drug repurposing (Liu et al., 2014a).

I generated cell-type-based diseasomes using the disease-cell-type associations identified by the GSC method (Figures 3.12, 3.13, 3.14 and 3.15). In the diseasomes, diseases were connected if they were identified by the GSC method as being associated with similar cell types. I first corrected p-values output by the GSC method for multiple testing using the BH procedure. To reduce the number of vertices in the diseasomes, I removed diseases associated with no cell types at a 5% FDR. For each pair of diseases, I computed Pearson’s correlation coefficient (PCC) using the $-\log_{10}$ of the q-values and set the diagonal of the resulting correlation matrix to 0. Each disease in the correlation matrix is represented as a vertex in the diseasome. I connected each disease to a set number of diseases with which it correlated most strongly. Vertex area is proportional to the number of cell types identified as associated.

Vertices in the diseasomes are coloured by disease class. I identified these disease classes using the MeSH terms mapped to each disease in Section 3.2.5. For each disease, I identified ancestral MeSH terms at the second level of the MeSH ontology, as these represent broader disease classifications. Some diseases have multiple ancestral MeSH terms at this level and it was therefore necessary to select between multiple classes for some diseases. In each of these cases, I used the ancestral MeSH term occurring most frequently amongst all diseases to classify each disease, as this reduced the number of classes represented in the diseasomes and therefore

makes them easier to interpret visually. I collected together diseases belonging to disease classes represented fewer than eight times in the diseasome in an ‘other’ class, to again make the diseasomes easier to interpret visually. I coloured edges connecting diseases of the same class with the class colour to illustrate the number of edges connecting diseases of the same class. I generated the diseasomes using the Fruchterman and Reingold layout algorithm (Fruchterman & Reingold, 1991) with default parameters, implemented in the **igraph** R-package (Csardi & Napusz, 2006).

The cell-type-based diseasomes could alternatively be generated by connecting all disease pairs that pass a certain correlation threshold. This approach however produces diseasomes in which some vertices have a large degree and others have a small degree or a degree of zero. This occurs because some sets of diseases are associated with very similar sets of cell types, for example, autoimmune diseases and the cell types of the immune system. These diseasomes would be difficult to interpret visually. For this reason, I decided to connect each disease to a fixed number of diseases.

Many diseases in the diseasome are connected to diseases of the same class, a feature seen in many previously generated diseasomes (Liu et al., 2014a). This was to be expected, as many of the disease classifications relate to specific anatomical structures that contain specific sets of cell types. Not all diseases however are connected to diseases of the same class, although many of these cases can be explained. For example, celiac disease is classified as a ‘nutritional and metabolic disease’, but interacts with three ‘immune system diseases’, reflecting the fact that celiac disease is an autoimmune disorder (Green & Cellier, 2007).

I generated multiple diseasomes by connecting each disease to the two (Figure 3.12), three (Figure 3.13), four (Figure 3.14) and five (Figure 3.15) diseases with which it correlated most strongly. Clustering of diseases of the same class can be observed in each of the diseasomes. The clustering of diseases in the diseasomes makes the labelling of vertices with disease names difficult. In Figure 3.13, the clustering of diseases can be observed and the vertices are far enough apart to allow disease name labelling. For this reason, I labelled vertices with disease names only in Figure 3.13.

Identifying novel associations between diseases may aid in understanding the mechanisms that underpin disease. Endometriosis is a complex gynaecological disorder, defined as the presence of endometrial glands and stroma outside of the uterus (Burney & Giudice, 2012). It is thought to affect 6–10% of women of reproductive age and can cause chronic pelvic pain and infertility. While a number of factors contributing to the development of the disease have been suggested, there is still much to learn about the etiology of endometriosis (Burney & Giudice, 2012).

In Figure 3.13, endometriosis interacts with two immune-related diseases: acquired immunodeficiency syndrome and bacterial vaginosis, suggesting that the immune system may play a role in the development of the condition. It has previously been noted that endometriosis is associated with increased risks of certain inflammatory and autoimmune disorders, including SLE and rheumatoid arthritis (Sinai et al., 2002), supporting the idea that shared mechanisms may underly the development of these diseases. Furthermore, abnormal inflammatory responses have been observed in women diagnosed with endometriosis (Podgaec et al., 2007). These abnormal responses include significantly higher interferon-gamma and interleukin-10 levels, compared to women without an endometriosis diagnosis (Podgaec et al., 2007). It is however not known whether this abnormal response contributes to the development of the condition or whether it is caused by the condition. The interaction between endometriosis and the two immune system diseases suggests that the immune system may play an active role in the development of the disease. Further work integrating disease-cell-type associations with additional data will need to be completed to identify novel links between endometriosis and other complex conditions and diseases.

3.4 Method implementation and data availability

I created an R-package called `DiseaseCellTypes` that contains implementations of the GSC and GSO methods, the method used to generate the cell-type-specific PPI networks and the method used to produce the cell-type-based diseasomes. The `DiseaseCellTypes` package also contains a vignette that describes how the results outlined in this chapter can be reproduced. The package also contains the

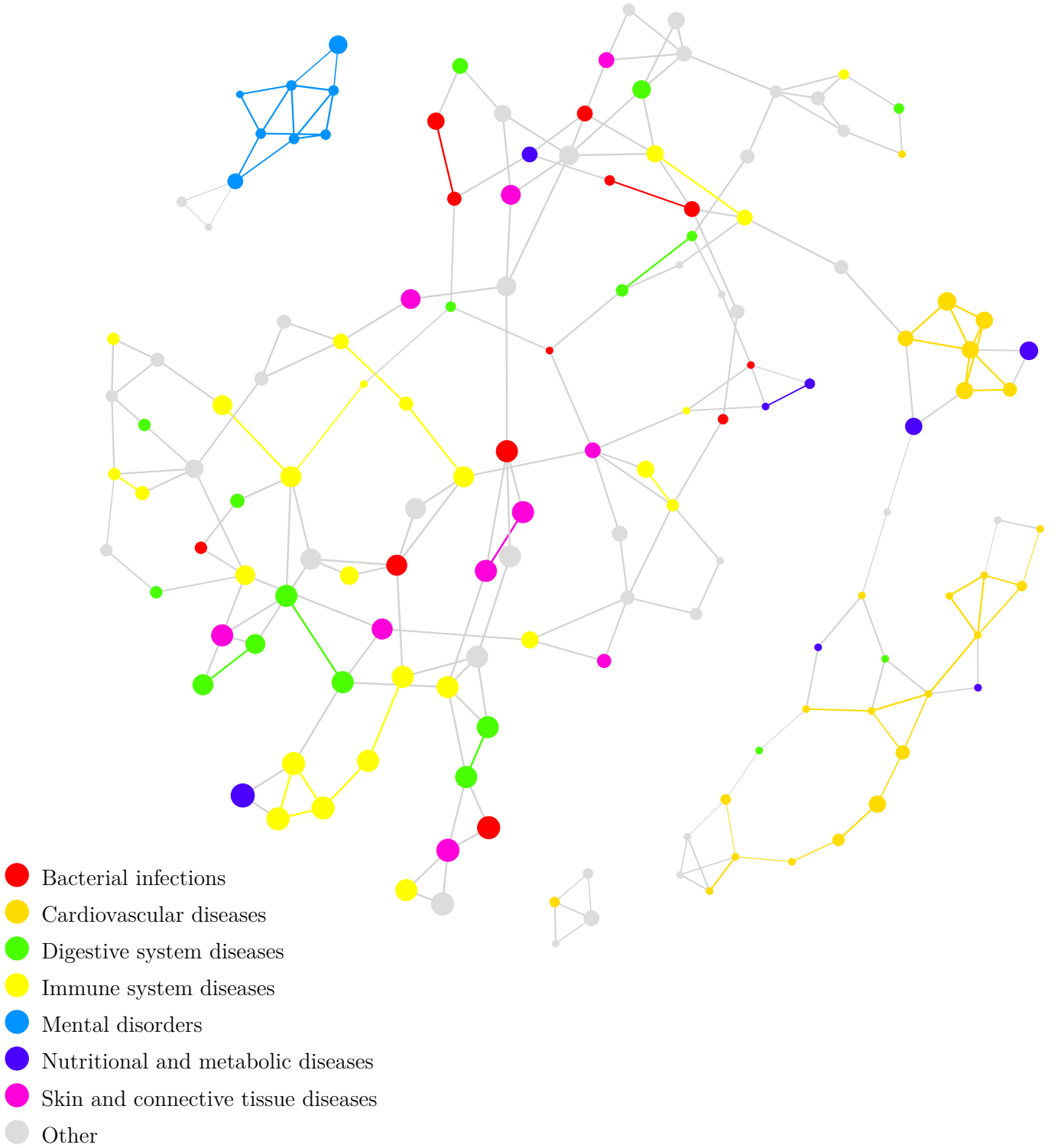


Figure 3.12: *Diseasome generated by connecting each disease to the two diseases with which it correlates most strongly, with respect to the cell types identified as associated by the GSC method. Vertices are coloured by the class of the disease. Edges that connect two diseases of the same class are also coloured by the disease class.*

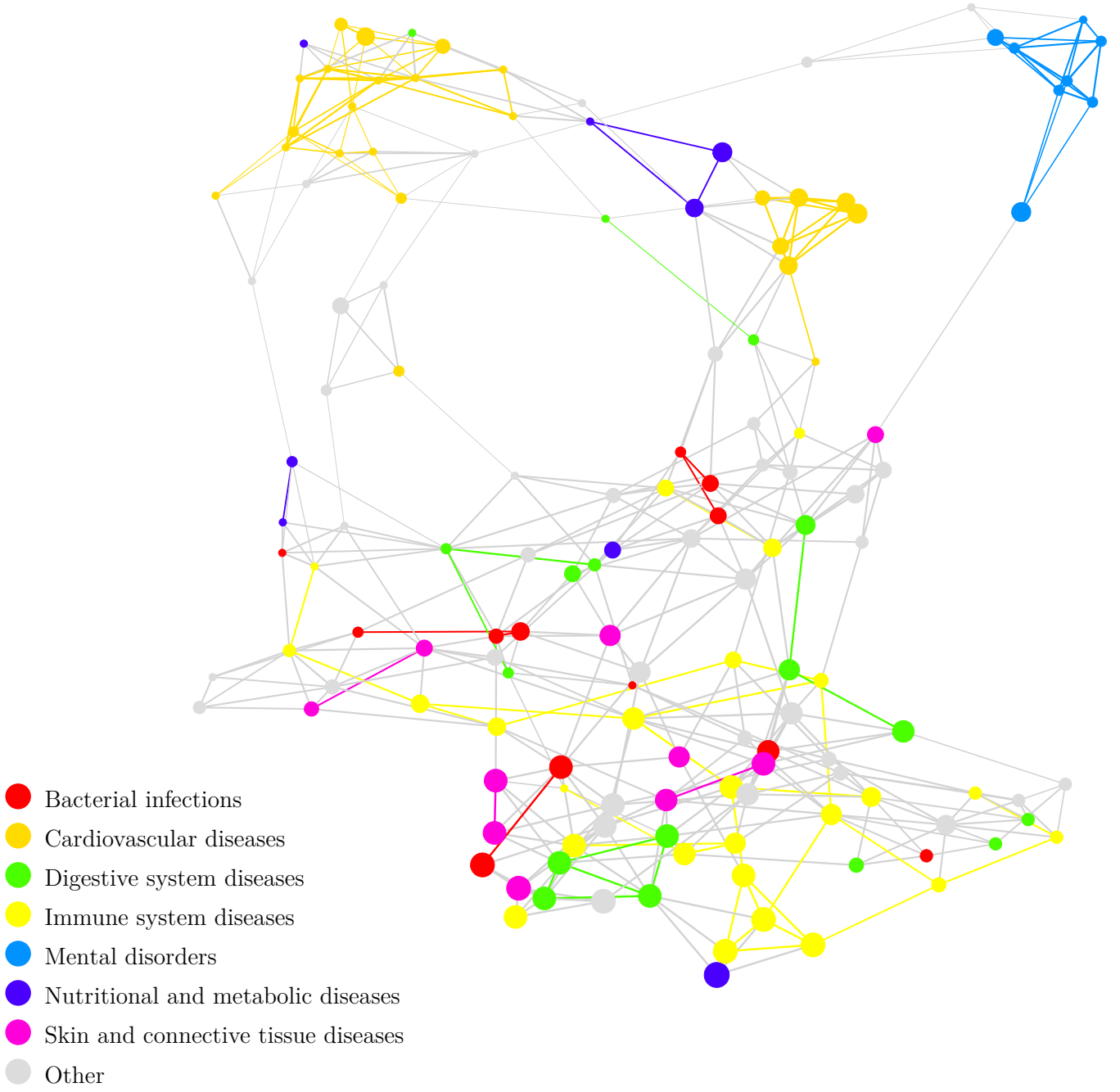


Figure 3.14: *Diseasome* generated by connecting each disease to the four diseases with which it correlates most strongly, with respect to the cell types identified as associated by the GSC method. Vertices are coloured by the class of the disease. Edges that connect two diseases of the same class are also coloured by the disease class.

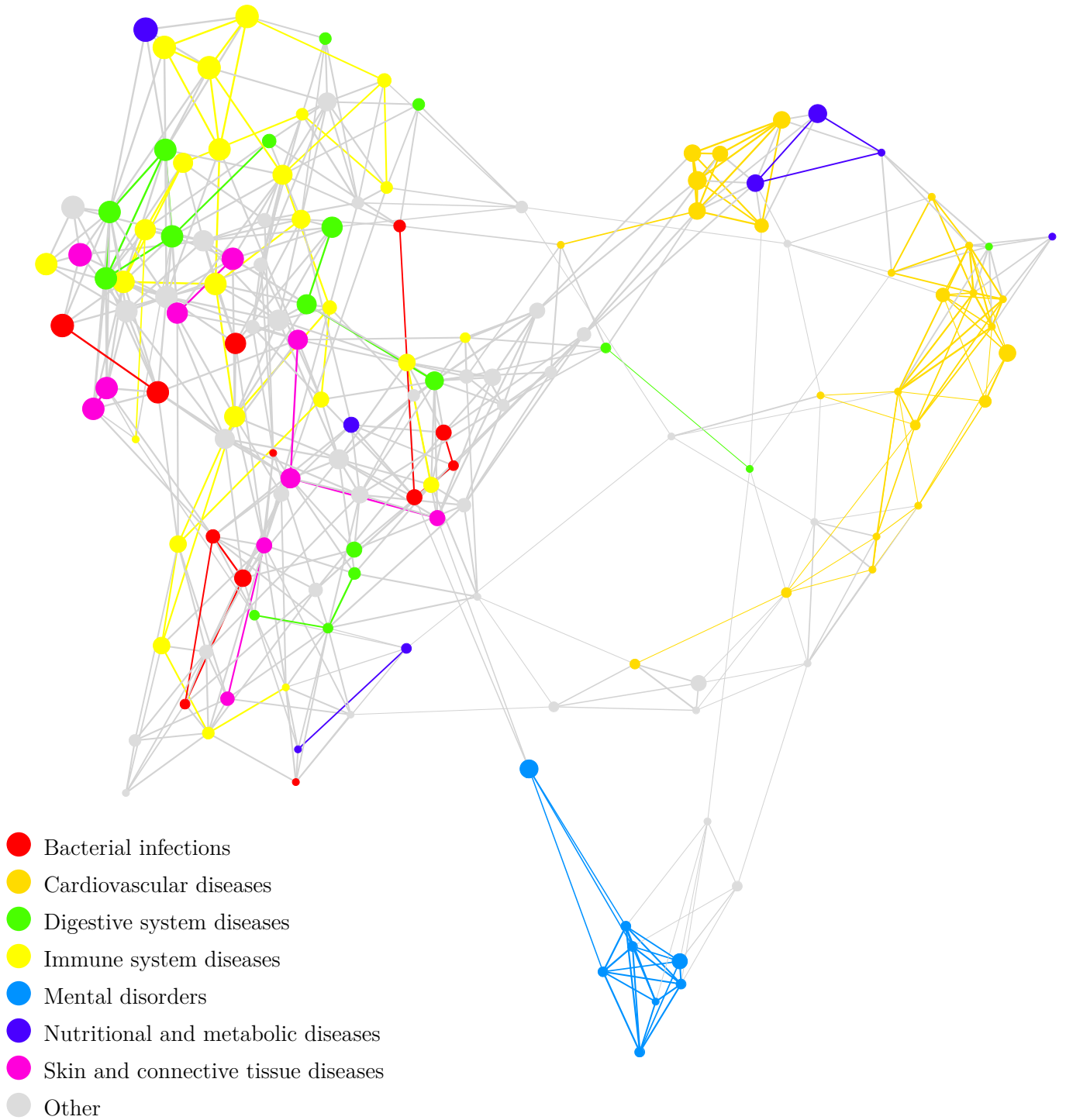


Figure 3.15: *Diseasome* generated by connecting each disease to the five diseases with which it correlates most strongly, with respect to the cell types identified as associated by the GSC method. Vertices are coloured by the class of the disease. Edges that connect two diseases of the same class are also coloured by the disease class.

data required to reproduce the results. The `DiseaseCellTypes` package can be downloaded from <http://alexjcornish.github.io/DiseaseCellTypes> and the 73 cell-type-specific PPI networks can be downloaded from http://alexjcornish.github.io/Cell_Type_Interactomes. While previous studies have integrated known disease-associated genes, gene expression data and PPI data to identify associations between diseases, tissues and cell types (Hu et al., 2011; Börnigen et al., 2013), the `DiseaseCellTypes` package is the first package to use this approach to allow users to identify associations between diseases and contexts using their own data.

3.5 Discussion

The GSC method identified at least one cell type associated with 202 of the 352 diseases with 3 or more associated genes at a 10% FDR. There are a number of reasons that may explain why the GSC method identified no cell types as associated with the remaining 150 diseases.

One explanation may be that for some of these diseases, the true disease-manifesting cell type is not represented in the 73 gene expression profiles generated using data from the FANTOM5 project. For example, no cells from the prostate are represented in the 73 profiles. This may explain why no cell types are identified as being associated with ‘prostatic hyperplasia’.

The lack of a single or a small number of disease-manifesting cell types may also explain why the GSC method does not identify cell types associated with some diseases. Primary ovarian insufficiency is a multi-factorial disease and defined as ovarian function failure before the age of 40 (Nelson, 2009). The GSC method identifies no cell types as being associated with primary ovarian insufficiency, possibly due to its multi-factorial nature.

The primary cell samples used by The FANTOM Consortium represented healthy cells (Andersson et al., 2014). Gene expression however is known to change when a cell is exposed to different environmental conditions, such as stressful conditions (de Nadal et al., 2011), reflecting the activation of different sets of pathways and processes in the cell. If a certain set of genes, some of which may be disease-

associated, only perform a function in a cell at a certain time, such as during exposure to stressful conditions, then the GSC method may fail to identify an association between the gene set and the cell type. Condition-specific functioning of genes may therefore explain why the GSC method does not identify cell types associated with some diseases.

While the overlaps between the disease-cell-type associations identified by the GSC, GSO and text mining methods were significant (Tables 3.8 and 3.9), the lack of a single unique MeSH term for some of the 73 cell types may affect the quality of the text-mined disease-cell-type associations. As described in Section 3.2.5, a lack of unique MeSH terms means that it was necessary to assign some cell types both a cell type MeSH term and a secondary anatomical MeSH term, in order to distinguish them from other cell types. However, just because an article is annotated with two MeSH terms related to a cell type, it does not necessarily follow that the article mentions the specific cell type. For example, an article annotated with the ‘amnion’ and ‘epithelial cells’ MeSH terms does not necessarily mention amniotic epithelial cells. The MeSH vocabulary increased in size from 236,253 terms in September 2012 (NCBI Resource Coordinators, 2013) to 253,057 terms in September 2014 (NCBI Resource Coordinators, 2015). As the size of the vocabulary continues to increase, the number of cell type MeSH terms will also likely increase, facilitating improved mapping and possibly higher-quality text mined associations.

Identifying disease-cell-type associations using text mining of article text, rather than associated MeSH terms, is made difficult by the large number of cell type name synonyms and acronyms (Neves et al., 2013). As the automated identification of specific cell types in article text improves, this approach may prove a useful alternative to MeSH-term-based text mining.

As demonstrated in Section 3.3.4, the greater the number of genes known to be associated with a disease, the better the performance of the GSC and GSO methods. The low number of genes known to be associated with many diseases limits the number of diseases that the methods can currently be applied to. As genome sequencing costs continue to decline (Hayden, 2014), the number of known disease-associated genes will continue to increase, allowing the methods to be applied to

new diseases, some of which may be rarer and less well studied than the 352 diseases considered in these analyses.

In the future, improvements could be made to how the permuted gene expression profiles used by the GSO and GSC methods are generated. Currently, each permuted gene expression profile is generated by sampling, for each gene, a cell type at random and then taking the percentile-normalised gene expression score of the gene from that cell type. This means that the correlations between gene expression scores that are likely to exist in the observed gene expression profiles are not preserved in the permuted gene expression profiles. This may lead to the GSO and GSC methods generating inflated significance estimates, as correlations between the observed gene expression scores may increase the variation in the scores produced by these methods.

The problem of preserving gene correlations also arises in gene set testing. Rotation Gene Set Testing (ROAST) is a gene set test that uses rotation, a Monte Carlo approach, instead of permutation to avoid incorrectly assuming the independence of genes (Wu et al., 2010). A similar rotation-based approach could be incorporated into the GSO and GSC methods to similarly avoid the current assumption of independence.

3.6 Conclusions

In this chapter I integrated gene expression data from the FANTOM5 project and PPI data from STRING to generate 73 cell-type-specific PPI networks. Using these networks, I demonstrated that sets of disease-associated genes tend to cluster more in PPI networks specific to cell types related to a disease. This indicates that these PPI networks may be useful in tasks such as prioritising new disease-associated genes and identifying pathways affected by these genes.

In this chapter I also described the systematic mapping of diseases to the cell types in which they are most likely to manifest. The GSC method uses the observation that disease-associated genes tend to cluster more in disease-related cell-type-specific PPI networks to generate this mapping. Using text mining, we independently validated many of the associations identified by the GSC method. Many of the other associations identified represent interesting candidates for future study.

As the amount of available cell type and tissue specific epigenetic, transcriptomic and PPI data continues to increase, it will be necessary to integrate these data to better understand the effects of genomic variation on human health.

Chapter 4

Prioritising genes in trait-associated loci

This chapter describes the development and benchmarking of ALPACA, a method that uses context-specific PPI networks and phenotype data from humans and mice to prioritise genes in trait-associated loci. ALPACA (Figure 4.1) builds upon the work described in the previous chapter and uses the GSO method to identify the context-specific PPI network best suited to prioritising genes associated with each disease. I use cross-validation to evaluate the performance of ALPACA and PRINCE and demonstrate that ALPACA outperforms PRINCE and avoids being biased towards better-studied genes.

4.1 Introduction

4.1.1 Motivation

While GWAS have identified thousands of trait-associated SNPs (Li et al., 2016) moving from these SNPs to the genes they affect is a challenging task. As described in Section 1.2.1, SNPs identified in GWAS are rarely themselves causal and instead tag regions of the genome that contain the causal variant or variants. These regions can sometimes contain many genes. Identifying which, if any, of these candidate genes are causal is crucial in understanding the genotype-to-phenotype relationship

and therefore the etiology of the disease. Conducting functional analyses of all candidate genes is often not possible due to the expenses involved. It would therefore be useful to be able to identify if any of the genes are likely to be causal using additional data.

Causal variants can affect genes in a number of ways. Variants can be located in the protein-coding region of a gene and produce an amino acid change, affect transcript splicing or introduce a premature stop codon (Landrum et al., 2016). Variants can lie close to a gene, either upstream or downstream, and affect the expression of the gene by disrupting a *cis*-acting regulatory mechanism (Veyrieras et al., 2008). Variants can also be positioned distantly from the causal gene, either on the same chromosome or on a different chromosome, and affect gene expression by disrupting a *trans*-acting regulatory mechanism (Mifsud et al., 2015). The lack of data on *trans*-acting regulatory mechanisms currently prohibits the systematic mapping of disease-causing variants to many of the genes they affect (Mifsud et al., 2015). For this reason, ALPACA is focused on identifying causal genes located close to causal variants.

4.1.2 Data sources

Databases of genotype-to-phenotype relationships in humans and other organisms are important resources for studying disease. Methods such as PRINCE demonstrate that disease-associated genes can be prioritised using genes known to cause diseases phenotypically similar to the disease of interest (Vanunu et al., 2010). Furthermore, methods such as MouseFinder demonstrate that this approach can be extended to data obtained from other organisms (Chen et al., 2012). While some methods use the phenotypic similarity of human diseases (Freudenberg & Propping, 2002; Lage et al., 2007; Vanunu et al., 2010; Li & Patra, 2010; Zemojtel et al., 2014) or the phenotypic similarity of human diseases and mouse models (Oellrich et al., 2012; Chen et al., 2012; Robinson et al., 2014) to prioritise disease-associated genes, few methods combine both of these data sources (Hoehndorf et al., 2011) and none combine both with network-based analysis. As the amount of genotype and phenotype data available for different organisms continues to increase, methods that integrate these

data will become increasingly important.

While the majority of gene prioritisation methods use generic physical or functional interaction networks to prioritise disease-associated genes, a number of methods have recently been developed that use networks specific to the tissues and cell types that manifest each disease (Magger et al., 2012; Jacquemin & Jiang, 2013; Li et al., 2014; Greene et al., 2015). ALPACA prioritises genes using cell-type-specific PPI networks generated by integrating PPI and gene expression data. Previously developed methods that use context-specific networks either require a user to manually select a suitable context (Greene et al., 2015) or use text-mined disease-context associations (Magger et al., 2012; Jacquemin & Jiang, 2013; Li et al., 2014). ALPACA, on the other hand, uses the GSO method to identify disease-associated contexts. In Section 4.3.8, I demonstrate that ALPACA performs better when run using PPI networks specific to contexts identified by the GSO method.

4.1.3 Generating association scores

A criticism of many gene prioritisation methods is that the scores they produce for each gene are difficult to interpret, possibly limiting their adoption (Moreau & Tranchevent, 2012). Genes are often assigned scores that can be used to rank genes but cannot be used as a measure of confidence of association between any single gene and the trait of interest. Of the methods described in Section 1.4 that prioritise genes by propagating scores across networks, only Prioritizer generates scores that represent the likelihood of observing results at least as extreme as that observed given that no association exists (Franke et al., 2006). The remaining methods generate scores that can only be used to rank genes (Krauthammer et al., 2004; Oti et al., 2006; Köhler et al., 2008; Li & Patra, 2010; Vanunu et al., 2010; Lee et al., 2011; Smedley et al., 2014).

An additional criticism of some network-based gene prioritisation methods is that they may be biased towards better-studied genes (Wang et al., 2011b; Oti et al., 2011). If a gene is better studied then it is more likely to be known to be involved in a greater number of PPIs (Das & Yu, 2012) and associated with a greater number of diseases (Köhler et al., 2008). Differences in the number of PPIs each

gene is involved in may bias network-based methods towards better-studied genes, as genes involved in greater numbers of PPIs are more likely to interact with genes known to be associated with the disease of interest. Similarly, if a gene is known to be associated with a greater number of diseases, then there is a greater chance that at least one of the diseases is phenotypically similar to the disease of interest. This means that methods that score genes by their association to phenotypically similar diseases, such as PRINCE (Vanunu et al., 2010), may be especially vulnerable to study bias.

Prioritizer produces values representing the likelihood of observing scores at least as extreme as that observed, given that no association exists, by taking a different approach to other gene prioritisation methods (Franke et al., 2006). As described in Section 1.4.3, Prioritizer identifies genes in disease-associated loci that interact with genes in other disease-associated loci in a functional interaction network. Functional interactions may occur between these genes by chance and therefore Prioritizer permutes the loci and re-scores the genes to estimate the likelihood that the number of observed interactions occurs by chance.

Zhang et al. (2011) developed a network-based method that uses permutation to identify genes involved in signalling pathways. Unlike Franke et al. (2006), Zhang et al. generate p-values by repeatedly redistributing the initial scores applied to the network and re-propagating them. For each gene, a p-value is generated by comparing the observed score of the gene against the score of the gene in each permutation. This allows the method to account for differences in network topology between genes. ALPACA builds upon this permutation-based approach and uses it to produce meaningful scores for each gene and avoid being biased towards better-studied genes (see Section 4.3.5).

4.2 Materials and methods

In this section I first outline the ALPACA gene prioritisation method. I then describe the method used by ALPACA to define trait-associated loci and identify the genes that these loci contain. This is followed by a description of how I integrated human disease variant data from ClinVar, OMIM and UniProtKB/Swiss-Prot, disease-

phenotype-term mappings and mouse phenotype data from the MGD to define sets of phenotypes associated with each human protein-coding gene. I next detail how ALPACA uses the simGIC semantic similarity metric to measure the phenotypic relevance of each human protein-coding gene to the trait of interest. Finally, I describe how I have evaluated the performance of ALPACA and compared it against the PRINCE gene prioritisation method.

4.2.1 The ALPACA method

ALPACA uses context-specific PPI networks and genotype and phenotype data from humans and mice to identify genes in trait-associated loci that may contain causal variants or be affected by causal variants. The first step of the method is to define the trait-associated loci and identify the candidate genes that these loci contain (this process is described in Section 4.2.2). Trait-associated SNPs passing a genome-wide significance threshold (such as the often used $p < 5 \times 10^{-8}$ threshold) are used to define these loci. Input SNPs should be in the format of reference SNP numbers from The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001). SNPs passing this significance threshold may not all represent distinct loci and may instead be in partial LD and correspond to the same association signal. To define lead SNPs in each loci I use the method of Wood et al. (2014) and select the most significantly associated SNP if multiple SNPs lie within 1Mb of each other. The population from which individuals in the GWAS were drawn (either individuals with Northern and Western European ancestry from Utah, USA (CEU), individuals with Japanese ancestry from Tokyo, Japan and individuals with Han Chinese ancestry from Beijing, China (JPT+CHB) or individuals with Yoruba ancestry from Ibadan, Nigeria (YRI)) and the name of the panel used to impute SNPs (either the HapMap Project or the 1KGP) must also be specified to correctly estimate LD. By default, loci are defined and the candidate genes contained in these loci identified using an LD cutoff of $r^2 = 0.3$ and with gene boundary extensions of 50kb (this choice of parameters is explained and justified in Section 4.3.2).

Human protein-coding genes are next scored by their phenotypic relevance to the trait of interest. This is done by first defining sets of phenotypes known to

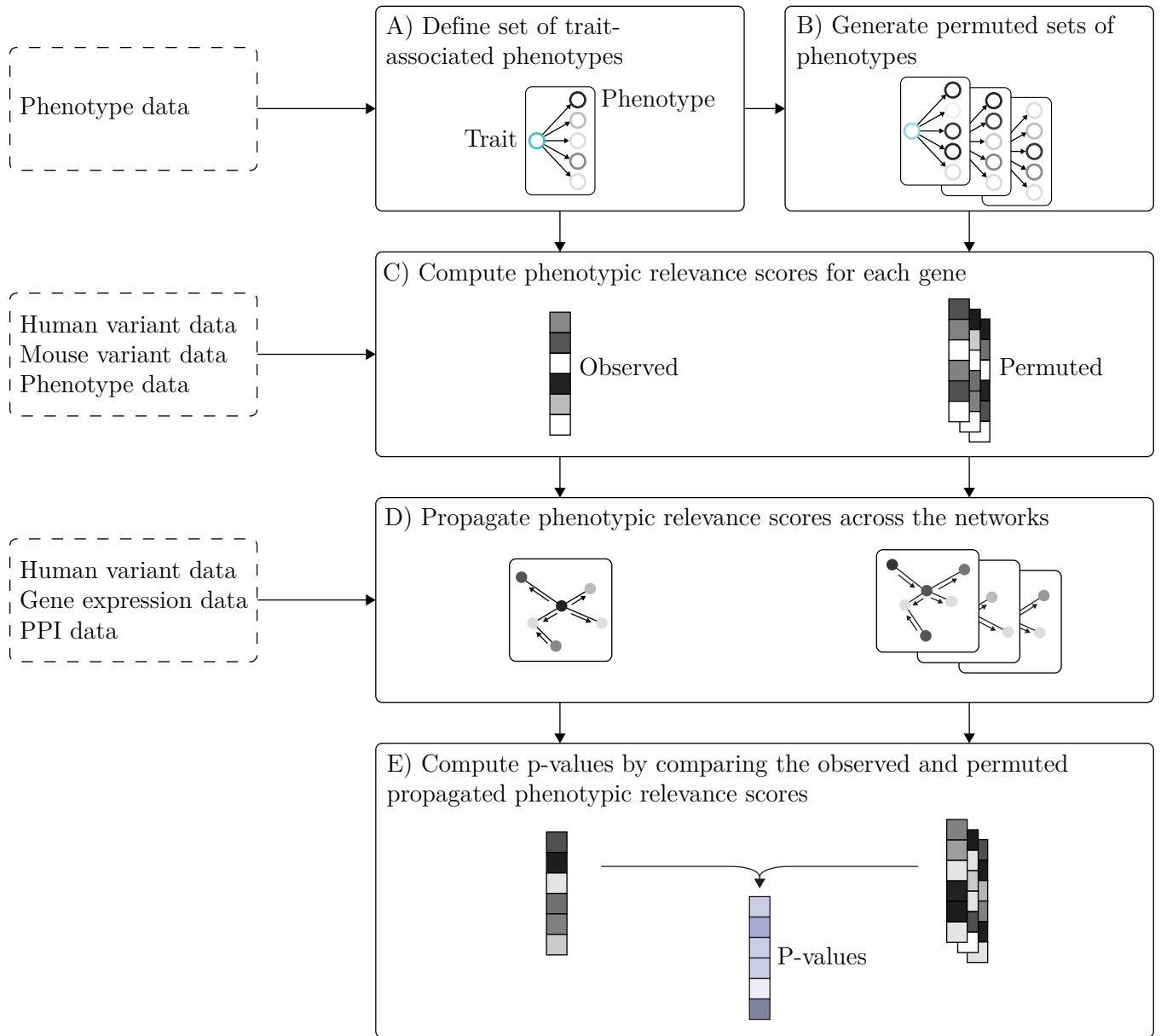


Figure 4.1: Overview of the ALPACA gene prioritisation method. The data used by the method are specified in dashed boxes to the left. A) A set of phenotypes associated with the trait of interest is defined. B) Sets of permuted phenotypes are generated. In this figure, three permutations are being used. C) Genes are scored by their phenotypic relevance to the trait of interest using the *simGIC* semantic similarity metric. Each box represents a gene and the darker the box, the greater the phenotypic relevance score. D) Each set of phenotypic relevance scores is applied to a network and propagated. The GSO method is used to identify a suitable cell-type-specific PPI network. If no network is identified, then a generic PPI network is used. E) A *p*-value is produced for each gene by taking the proportion of permutations in which the propagated phenotypic relevance score is greater than or equal to the observed phenotypic relevance score.

be associated with each gene. In ALPACA, terms from the HPO and the MP are used to denote phenotypes. The sets of gene-associated phenotypes are generated by integrating human disease variant data from ClinVar, OMIM and UniProtKB/Swiss-Prot (as described in Section 4.2.3), disease-phenotype-term mappings generated by Hoehndorf et al. (2015a) and the HPO (as described in Section 4.2.4) and mouse phenotype data from the MGD (as described in Section 4.2.5). It is also necessary to describe the trait of interest using terms from the HPO and the MP (Figure 4.1A). This set of trait-associated phenotypes can either be defined by the user, or if the trait of interest is a disease susceptibility, taken from the predefined sets of disease-phenotype-term mappings (as described in Section 4.2.4). ALPACA uses the simGIC semantic similarity metric to measure the similarity of each set of gene-associated phenotypes and the set of trait-associated phenotypes and thereby score genes by their phenotypic relevance to the trait of interest (as described in Section 4.2.6, Figure 4.1C). Let J be a vector and denote the phenotypic relevance score of each human protein-coding gene.

The scores in J are next propagated across a PPI network (Figure 4.1D). When possible, ALPACA uses a PPI network specific to a context associated with the trait of interest (the generation of these networks is described in Section 4.2.7). This trait-associated context can either be defined by the user, or if the trait of interest is a disease susceptibility, identified automatically using the GSO method (as described in Section 3.2.5). ALPACA uses gene expression data from the FANTOM5 project (as described in Section 3.2.1) and is therefore able to use 73 different cell-type-specific PPI networks. P-values generated by the GSO method describing the significance of association are adjusted for multiple testing using Bonferroni correction. A context is said to be associated with a trait at $p < 0.05$. If multiple contexts are identified as associated then the most significantly associated context is chosen. If multiple contexts have the same p-value then the context in which the mean normalised gene expression of the disease-associated gene set is highest is chosen. A generic PPI network is used to propagate scores if no trait-associated context is identified.

ALPACA uses the RWR network propagation algorithm to propagate the

phenotypic relevance scores across the PPI network (as described in Section 2.6.2). I chose to use this algorithm to propagate the scores because it has repeatedly been shown to be one of the most successful network propagation algorithms when prioritising genes (Köhler et al., 2008; Navlakha & Kingsford, 2010). Using this algorithm, genes that interact with genes of high phenotypic relevance are scored higher than genes that do not interact with genes of high phenotypic relevance. By default, the RWR algorithm is run with a restart probability of $r = 0.5$ and 10 iterations (this choice of parameters is justified in Section 4.3.3). Let K be a vector and denote these propagated phenotypic relevance scores.

As previously described, network-based gene prioritisation methods can be biased towards better-studied genes and produce scores that are difficult to interpret. ALPACA therefore uses a permutation-based approach to avoid this bias and produce more meaningful scores. In this permutation-based approach, the observed scores in K are compared against scores generated using permuted sets of phenotypes. These permuted sets of phenotypes are generated using the predefined sets of disease-phenotype-term mappings (described in Section 4.2.4, Figure 4.1B). To generate each permuted set of phenotypes, a disease from the set of diseases to which phenotypes have been mapped is sampled. A set of phenotypes, equal in size to the set of phenotypes associated with the trait of interest, is then sampled from the set of phenotypes mapped to the sampled disease. If the trait of interest is associated with a greater number of phenotypes than the set of phenotypes mapped to the sampled disease, then all phenotypes associated with the sampled disease are used. Phenotypic relevance scores are then computed using this permuted set of phenotypes and propagated across a PPI network. If a context-specific network was used to propagate the observed phenotypic relevance scores, then a randomly chosen context-specific PPI network is used to propagate the permuted phenotypic relevance scores. If a generic network was used to propagate the observed phenotypic relevance scores, then a generic network is also used to propagate the permuted phenotypic relevance scores. For each gene, a p-value can then be produced describing the probability of observing a propagated phenotypic relevance score at least as high as that observed by chance, given the data used by ALPACA and that the gene is not

associated with the disease. This is done by taking the proportion of the permuted phenotype sets in which the propagated phenotypic relevance score of the gene is equal to or greater than the phenotypic relevance score generated using the observed set of phenotypes (Figure 4.1E). A minimum p-value of $1/n_{sim}$ is applied to ensure that no p-values equal 0, where n_{sim} is the number of permutations completed. In this chapter ALPACA is run using 1,000 permutations. These p-values are used to prioritise the candidate genes in the trait-associated loci.

Illustrative example

In this section I will provide a brief illustrative example of the ALPACA computational procedure. To do this, I will describe the application of ALPACA to three loci (rs10168266, rs2233434, rs3125734) identified by Myouzen et al. (2012) as being associated with rheumatoid arthritis (RA) and passing a genome-wide significance threshold of $p < 5 \times 10^{-8}$. I ran ALPACA using the default parameters described in this section. As the participants in the study were of Japanese ancestry, I used the JPT+CHB reference panel to estimate LD between SNPs.

ALPACA identifies two genes in the rs10168266 locus, nine genes in the rs2233434 locus and one gene in the rs3125734 locus. ALPACA next scores each of these genes by their phenotypic relevance to RA. Phenotype terms describing RA are taken from the set of disease-associated phenotypes and the simGIC semantic similarity metric is used to compare the similarity of this set of phenotypes to the sets of phenotypes associated with each gene. Four of the twelve genes have no known associated phenotypes and therefore are assigned a score of zero. The highest scoring gene is STAT4 in the rs10168266 locus (with a score of 0.066). The GSO method and the eight known RA-associated genes in the combined ClinVar/OMIM/UniProtKB/Swiss-Prot data set are next used to identify a disease-associated context. The GSO method identifies ‘Monocyte’ as the most strongly associated context and therefore the phenotypic relevance scores are propagated across the monocyte-specific PPI network. This step allows three of the four genes with no known associated phenotypes to be scored. The remaining gene (ENSG00000272442) cannot be scored as it is not represented in the PPI data set. STAT4 is still the highest scoring gene

(with a score of 0.000241), although the difference between this gene score and the other gene scores is now smaller. 10,000 permuted phenotype sets are next generated and the same procedure applied to each of these sets. P-values are generated for each of the genes by comparing the observed propagated phenotypic relevance scores to the permuted propagated phenotypic relevance scores. STAT4 is the gene with the smallest p-value (0.021).

4.2.2 Defining trait-associated loci

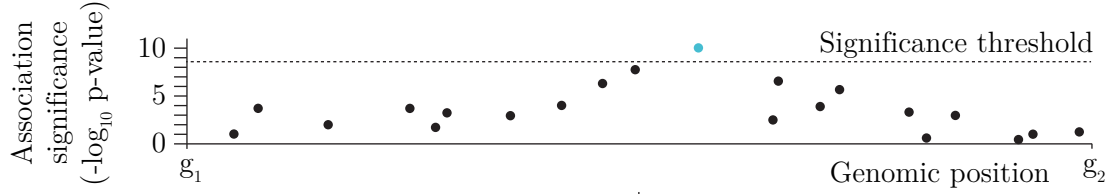
In this section I describe the method used by ALPACA to define the trait-associated loci and identify the genes these loci contain. This method is similar to the approach taken by comparable gene prioritisation methods, including GRAIL (Raychaudhuri et al., 2009), DAPPLE (Rossin et al., 2011), OPEN (Deo et al., 2014), DEPICT (Pers et al., 2015) and PrixFixe (Tasan et al., 2015).

As previously described, trait-associated SNPs identified in GWAS are rarely themselves causal. These SNPs instead tag regions of the genome in LD that contain the causal variant or variants. These regions can contain tens of genes, of which zero or more may contain or be affected by the causal variant or variants. In order to identify the genes that are affected, it is first necessary to define the region tagged by the SNP.

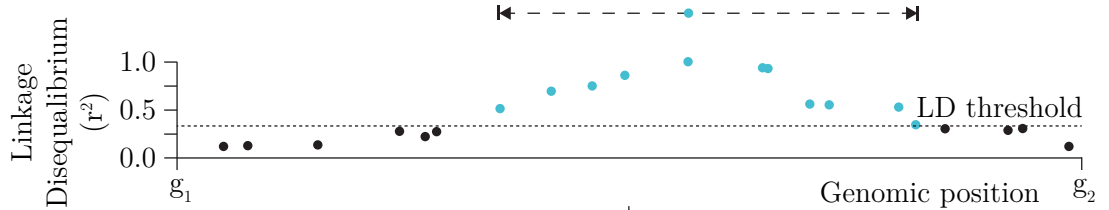
Defining regions of SNPs in LD

Trait-associated SNPs are defined as those SNPs that pass a genome-wide significance threshold (Figure 4.2A). To define a trait-associated region, I first identify SNPs in LD with each trait-associated SNP. Patterns of LD vary between populations and it is therefore important to take into account the population studied in the GWAS when estimating LD (Bush & Moore, 2012). Different GWAS also use different reference panels to impute SNPs. The majority of GWAS use data produced by the HapMap Project (The International HapMap Consortium, 2007) or the 1KGP (The 1000 Genomes Project Consortium, 2015). These two projects have genotyped different sets of SNPs and it is therefore also important to use data from the correct project when defining associated regions (Buchanan et al., 2012).

A) Identify SNPs that pass a genome-wide significance threshold



B) Identify region of SNPs in LD



C) Identify candidate genes

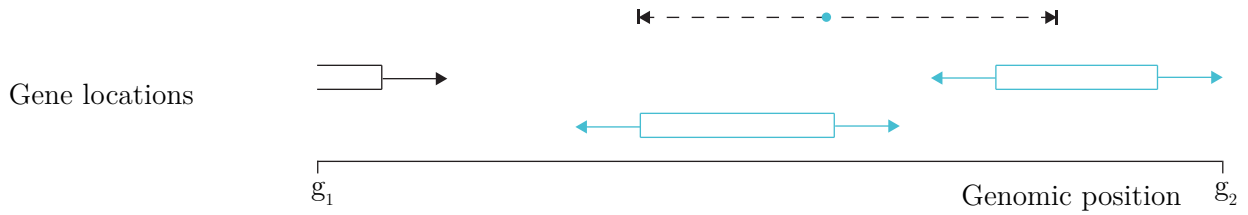


Figure 4.2: Defining trait-associated loci and identifying the genes these loci contain.

A) SNPs that pass a genome-wide significance threshold are identified. The finely dashed line represents a significance threshold, the circles represent SNPs and the blue circle represents the only SNP that passes the significance threshold. SNPs are located in a genomic region spanning from g_1 to g_2 . B) SNPs in LD with the lead SNP are identified and a region spanning these SNPs defined. The finely dashed line represents an LD cutoff, blue circles represent SNPs in LD with the lead SNP and the broadly dashed line represents the region. C) Genes in the associated region are identified. Rectangles indicate gene locations and the arrows represent the boundary extensions added to incorporate cis-regulatory elements. Genes coloured blue are at least partly contained in the associated region and are therefore considered candidate genes. I justify the parameters that I use to define trait-associated regions and identify the genes that these regions contain in Section 4.3.2.

I obtained HapMap Project and 1KGP data from LocusZoom (Pruim et al., 2010). LocusZoom is a tool for visualising trait-associated loci. It displays the LD between SNPs and therefore contains HapMap Project and 1KGP data in a format from which it is easy to estimate LD. I chose to use data from the October 2008 phase II release of the HapMap Project (The International HapMap Consortium, 2007) and the June 2010 pilot phase release of the 1KGP (The 1000 Genomes Project Consortium, 2010) as these data sets both contain three populations commonly studied in GWAS: CEU, JPT+CHB, and YRI.

I used PLINK to estimate LD between pairs of SNPs (Purcell et al., 2007). PLINK estimates LD by first estimating haplotype frequencies using an expectation maximisation algorithm (Qin et al., 2002). Correlation coefficients can then be computed from these haplotype frequencies. These correlation coefficients are used as a measure of LD. It is not computationally feasible or useful to compute LD between all pairs of SNPs as the LD between SNPs decreases with genomic distance. I therefore computed, for each SNP, the correlation coefficient (r^2) between the SNP and the 1,000 closest SNPs located within 1Mb.

Regions of the genome in LD with each trait-associated SNP can be defined using these SNP pair estimates. For each trait-associated SNP, I identify all SNPs in LD ($r^2 > 0.3$, see Section 4.3.2 for justification of this parameter) and generate a region spanning these SNPs (Figure 4.2B). Overlapping regions are merged. I downloaded SNP locations for the *hg19* genome assembly from version 75 of the Ensembl database (Yates et al., 2016).

Extending regions to recombination hotspots

Myers et al. (2005) estimated that 80% of recombination occurs within only 10–20% of sequence. These regions of high-recombination are known as ‘recombination hotspots’. Some gene prioritisation methods refine their definition of trait-associated loci by extending each region in each direction to the closest recombination hotspot (Raychaudhuri et al., 2009; Rossin et al., 2011; Deo et al., 2014; Himmelstein & Baranzini, 2015).

To determine whether extending trait-associated regions to the closest

recombination hotspots aids in loci definition, I downloaded recombination hotspot location data for the phase II release of the HapMap Project and the pilot phase release of the 1KGP from the HapMap Project website and the 1KGP `ftp` server. The HapMap Project and the 1KGP both identified recombination hotspots using the LDhat method to estimate recombination rates and the LDhot likelihood ratio test to call hotspots (McVean et al., 2004). I mapped recombination hotspot locations to the *hg19* genome assembly using the University of California, Santa Cruz (UCSC) Genome Browser LiftOver tool (Speir et al., 2016).

In Section 4.3.2, I demonstrate that extending regions to the closest recombination hotspots does not aid in defining loci enriched with trait-associated genes. In the final method I therefore do not extend each region to the closest recombination hotspots when defining trait-associated loci.

Identifying candidate genes

Finally, genes contained in the trait-associated loci are identified (Figure 4.2C). A gene is said to be contained in a loci if it at least partly overlaps the associated region. If one gene is contained in multiple trait-associated loci then the loci are merged. I downloaded gene locations in the *hg19* genome assembly from version 75 of the Ensembl database (Yates et al., 2016). In the Ensembl database, the start and end positions of each protein-coding gene correspond to the smallest and largest start and end position of all transcripts associated with the gene. I downloaded only the positions of protein-coding genes as ALPACA uses PPI data to prioritise genes and is therefore only able to prioritise protein-coding genes.

Some gene prioritisation methods extend the boundaries of each gene a set distance upstream and downstream in order to incorporate *cis*-regulatory elements (Rossin et al., 2011; Deo et al., 2014; Tasan et al., 2015). In Section 4.3.2, I demonstrate that extending gene boundaries 50kb upstream and 50kb downstream results in greater enrichment of trait-associated genes in trait-associated loci. In this chapter I therefore extend gene boundaries 50kb upstream and 50kb downstream when identifying candidate genes.

4.2.3 Human disease variant data

In this section I describe the integration of disease variant data from ClinVar (Landrum et al., 2016), OMIM (Amberger et al., 2015) and UniProtKB/Swiss-Prot (Yip et al., 2008) to generate the sets of disease-associated genes used in this chapter. This data set is used to define the sets of gene-associated phenotypes, identify disease-associated contexts and generate the artificial disease loci used to evaluate ALPACA. I chose not to incorporate data from DisGeNET (which is used in Chapter 3) as DisGeNET contains many associations reported in GWAS for which no causal variant has yet been identified (Pinero et al., 2015). Not including these associations reduces the risk of introducing knowledge bias when evaluating the performance of ALPACA.

I chose to use data from ClinVar, OMIM and UniProtKB/Swiss-Prot as these databases represent three of the largest freely available disease variant databases (Peterson et al., 2013). It was necessary to integrate these three databases as each database contains disease variants not reported in the other databases (Peterson et al., 2013). I chose not to use data from HGMD as the Public version of HGMD cannot be downloaded in full and the Professional version requires a license.

The ClinVar variant summary file (`variant_summary`) was downloaded from the ClinVar `ftp` server. I discarded variants not marked as pathogenic or likely pathogenic. As described in Section 1.2.2, ClinVar assigns a review level to each variant describing its medical importance. To maximise the number of variants represented in the final data set, I decided to not discard variants at any review level. Variants in ClinVar are assigned to a gene if the gene contains the variant.

The OMIM morbid map file (`morbidmap`) was downloaded from the OMIM website. The morbid map contains the cytogenetic locations of each disease in OMIM and details of whether the molecular basis of the disorder is known. I discarded those entries in which the molecular basis of the disease was unknown.

UniProtKB/Swiss-Prot data (`humsavar`) was downloaded from the UniProt `ftp` server. Variants in UniProtKB/Swiss-Prot are assigned to a gene if they are located in the protein-coding region of the gene. I discarded the variants not assigned to a gene.

	Data type	Number	Of total
<i>ClinVar</i>			
	All variants	281,023	100.0%
	Pathogenic/likely pathogenic variants	75,488	26.9%
	Pathogenic/likely pathogenic variants with Ensembl gene ID	65,726	23.4%
	Pathogenic/likely pathogenic variants with Ensembl gene ID and DO term	46,749	16.6%
<i>OMIM</i>			
	All entries	6,870	100.0%
	With identified molecular mechanism	5,482	79.8%
	With identified molecular mechanism and Ensembl gene ID	5,473	79.7%
	With identified molecular mechanism, Ensembl gene ID and DO term	4,334	63.1%
<i>UniProtKB/Swiss-Prot</i>			
	All variants	70,687	100.0%
	Disease-associated variants	25,809	36.5%
	Disease-associated variants with Ensembl gene ID	25,747	36.4%
	Disease-associated variants with Ensembl gene ID and DO term	23,679	33.5%

Table 4.1: The numbers of variants and entries in the three human disease variant databases with different types of associated data and the percentage of all variants and entries these numbers represent. I combined these three data sets to produce the final disease-gene association data set used in this chapter.

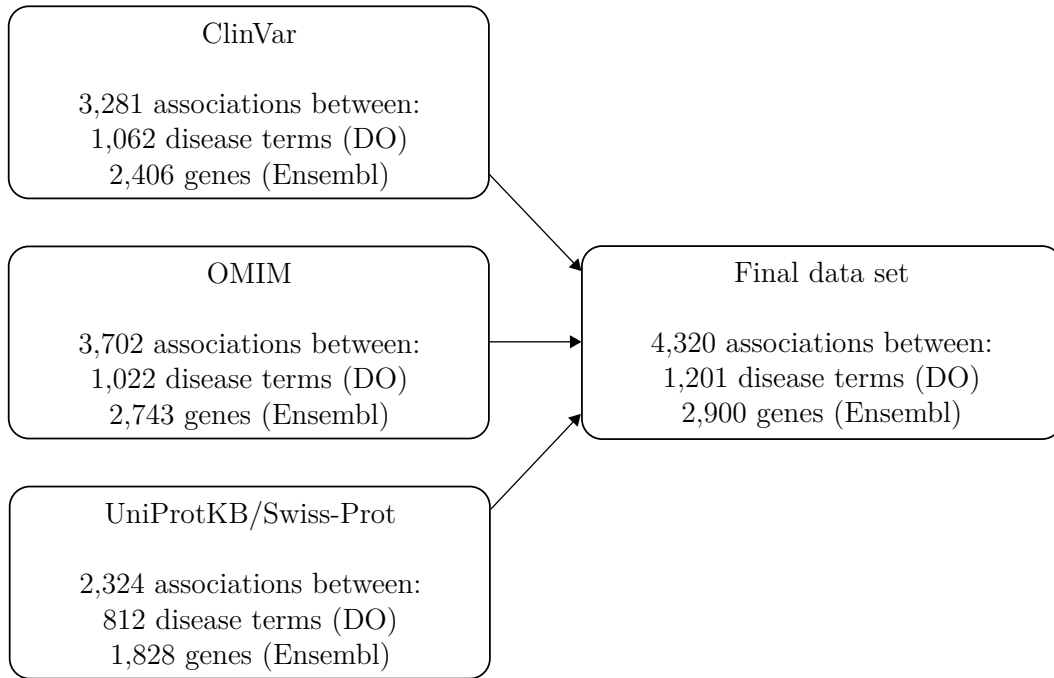


Figure 4.3: Integrating disease-gene associations from ClinVar, OMIM and UniProtKB/Swiss-Prot. A gene is said to be associated with a disease if a variant located in the gene in ClinVar or UniProtKB/Swiss-Prot is mapped to the disease or if OMIM reports such an association. I have mapped diseases referenced in the three databases to disease terms from the DO and genes to Ensembl gene identifiers.

ClinVar, OMIM and UniProtKB/Swiss-Prot report the diseases mapped to each variant and gene using different controlled vocabularies. To integrate the data it was therefore necessary to map the terms from each vocabulary to a single vocabulary. I decided to map the terms to terms from the DO as the DO represents one of the largest disease-focused controlled vocabularies currently available and contains terms for both Mendelian disorders and complex diseases (Kibbe et al., 2015). The DO also contains a large number of cross-references to other controlled vocabularies, facilitating the mapping of terms from other vocabularies to the DO.

ClinVar, OMIM and UniProtKB/Swiss-Prot all use MIM numbers to describe disease (Amberger et al., 2015). ClinVar also uses terms from a number of other vocabularies, including SNOMED clinical terms (Stearns et al., 2001) and UMLS terms (Bodenreider, 2004). The DO provides cross-referencing between MIM

numbers, SNOMED clinical terms, UMLS terms and DO terms. I used these cross-references to map diseases to DO terms (Table 4.1).

Not all MIM numbers can be directly mapped to DO terms using the cross-references provided by the DO, as some MIM numbers refer to disease descriptions that are more specific than any disease in the DO. In these cases I used the MeSH ontology to identify the most suitable DO term. To do this, I first used the UMLS Metathesaurus to map each MIM number to a MeSH term (Bodenreider, 2004). The DO provides cross-referencing between MeSH terms and DO terms. If a MIM number mapped to a MeSH term that was not cross-referenced with a DO term, then I considered the ancestors of the MeSH term. This was done repeatedly until a MeSH term cross-referenced with a DO term was identified. I was then able to map the MIM number to a DO term using this ancestral MeSH term.

I mapped genes referenced in ClinVar, OMIM and UniProtKB/Swiss-Prot to Ensembl gene identifiers using the `biomaRt` R-package (Durinck et al., 2009). I discarded variants and genes if I could not map them to both a DO term and an Ensembl gene identifier. In the final data set, a gene is said to be associated with a disease if a variant located in the gene in ClinVar or UniProtKB/Swiss-Prot is mapped to the disease or if OMIM reports an association between the gene and the disease. The final data set contains 4,320 associations and spans 1,201 diseases and 2,900 genes (Figure 4.3).

4.2.4 Human disease phenotype terms

In this section I describe the generation of the disease-phenotype-term mapping data set used in this chapter. This data set is used to define the sets of gene-associated and trait-associated phenotypes. I generate the data set by integrating data from Hoehndorf et al. (2015a) and the HPO (Table 4.2).

Hoehndorf et al. (2015a) mapped terms from the HPO and the MP to diseases from the DO using automated text mining. They obtained and parsed article titles and abstracts using the Aber-OWL: PubMed framework to identify occurrences of terms from the HPO, the MP and the DO (Hoehndorf et al., 2015b). If the number of articles co-mentioning a disease term and a phenotype term was large compared to

	Method	Associations
<i>Hoehndorf et al. (2015a)</i>		
	Text mining	124,213
<i>HPO</i>		
Clinical features sections of OMIM records		61,974
Published medical literature		4,062
Disease reviews		32,421
Clinical experience of individuals		25

Table 4.2: The numbers of disease-phenotype-term mappings generated by Hoehndorf et al. (2015a) and the HPO using different methods. Hoehndorf et al. mapped phenotype terms from the HPO and the MP to disease terms from the DO. The HPO mapped phenotype terms from the HPO to MIM numbers.

the number of articles individually mentioning the terms, then the phenotype could be said to be associated with the disease. Hoehndorf et al. used these phenotypic associations to predict disease-gene associations and noted that performance was best when the top 21 phenotypes associated with each disease were considered. They therefore generated sets of 21 phenotypes associated with each disease.

As part of the development of the HPO, terms from the HPO were mapped to MIM numbers by parsing the clinical features sections of OMIM records, by parsing published medical literature, by collecting annotations from clinical reviews of diseases and manually from the clinical experience of individuals (Köhler et al., 2014).

I downloaded the disease-phenotype-term mappings generated by Hoehndorf et al. (2015a) and the HPO. As previously mentioned, Hoehndorf et al. mapped phenotype terms to disease terms from the DO, while the HPO mapped phenotype terms to MIM numbers. I used the cross-referencing provided by the DO to map these MIM numbers to disease terms from the DO. It was then possible to combine the two data sets (Figure 4.4). The final data set contains 163,540 unique mappings

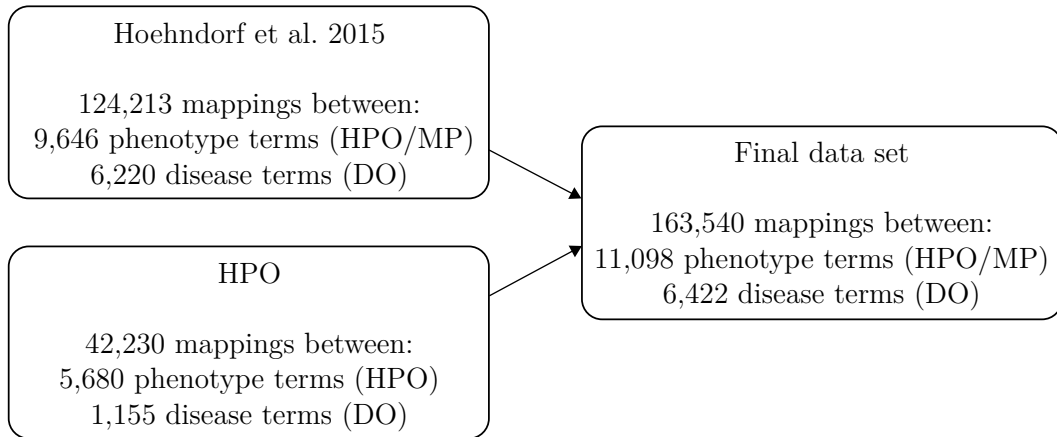


Figure 4.4: Integrating disease-phenotype-term mappings generated by Hoehndorf et al. (2015a) and the HPO. The HPO mapped phenotype terms to diseases using multiple methods. It was not possible to map all MIM numbers in the HPO data set to DO terms. The number of mappings from the HPO given in this figure is therefore lower than the total number of mappings from the HPO given in Table 4.2.

between phenotype terms from the HPO and the MP and disease terms from the DO.

4.2.5 Mouse phenotype data

In this section I describe the parsing of the mouse phenotype data used in this chapter. This data set is used to define the sets of gene-associated phenotypes.

I obtained mouse phenotype data from the MGD (Bult et al., 2016). The MGD is the largest database of its kind and, as of 14 November 2015, contains phenotype data for 45,288 mouse alleles. It also contains data on orthologous human genes, facilitating the integration of data from the MGD with human data to study human disease (Chen et al., 2012; Robinson et al., 2014).

To integrate data from the MGD with human variant data it was necessary to map mouse genes to their human orthologs. I downloaded lists containing mouse alleles (`MGI_PhenotypicAllele`) and mouse-human gene orthologs (`HMD_HumanPhenotype`) from the MGD ftp server. This list of mouse-human gene orthologs was generated using data from HomoloGene (NCBI Resource Coordinators, 2016), which uses sequence similarity to detect homology between

	Data type	Number	Of total
	All alleles	45,288	100.0%
	Alleles with reported mouse gene	37,838	83.5%
	Alleles with orthologous human gene	31,924	70.5%
Alleles with orthologous human gene with an Ensembl gene ID		31,894	70.4%

Table 4.3: The numbers of alleles in the mouse variant data set with various types of associated data. An orthologous human gene can only be identified for an allele if the allele is associated with a mouse gene.

genes. I mapped each MGD allele to its human ortholog using these lists. I next used the **biomaRt** R-package to assign human Ensembl gene identifiers to each human ortholog (Durinck et al., 2009). If I could not map a mouse allele to a human Ensembl gene identifier then I discarded the allele. Some mouse alleles mapped to multiple human genes and in these cases I duplicated the mouse allele entry. The final data set contains 31,992 associations between 31,894 mouse alleles and 10,431 human genes (Table 4.3).

I also downloaded the phenotypes reported to be associated with each allele (**MG1_GenePheno**) from the MGD **ftp** server. This phenotype data is based on observations made by the submitters of each allele to the MGD and was recorded using terms from the MP (Smith et al., 2005). The MGD contains a number of conditional mutants in which a gene of interest is knocked out at a specific developmental stage or in a particular tissue (Blake et al., 2011). I have decided not to include phenotypes associated with conditional mutations in my data set as the phenotypic effects of these mutations are likely to be highly biased towards the anatomical structures in which they were triggered.

4.2.6 Measuring phenotype similarity

As described in Section 4.2.1, ALPACA scores genes by their phenotypic relevance to the trait of interest. In this section I detail how the sets of disease-associated genes (described in Section 4.2.3), the disease-phenotype mappings (described in Section

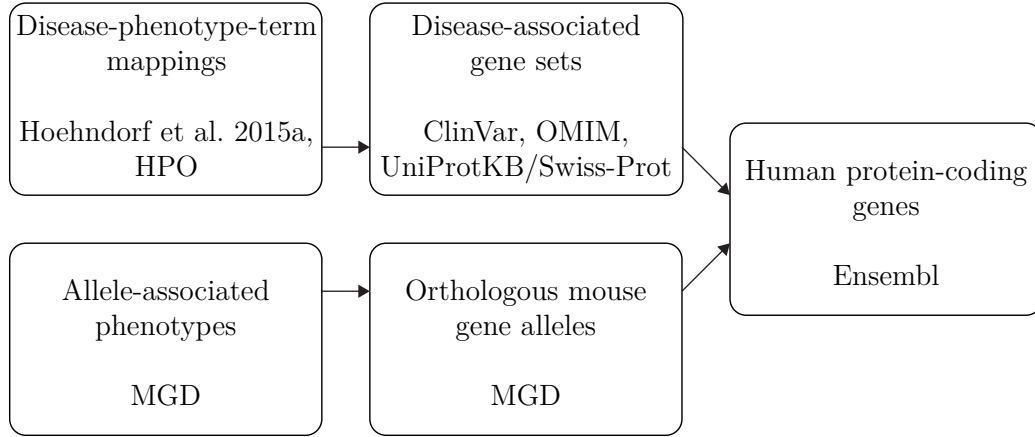


Figure 4.5: Integrating data to generate the sets of gene-associated phenotypes used in this chapter. Given are the data sets used to map phenotypes to genes and the resources from which these data were obtained.

4.2.4) and the mouse phenotype data (described in Section 4.2.5) were integrated and used to define sets of gene-associated phenotypes (Figure 4.5). I then explain how the simGIC semantic similarity metric (Pesquita et al., 2008) is used to measure the similarity of sets of gene-associated and trait-associated phenotypes.

To define sets of phenotypes associated with each human protein-coding gene, I first used the disease-associated gene sets to identify human diseases associated with each gene. Using these associations, I assigned the sets of disease-phenotype-term mappings to each gene. For each human gene, I also identified alleles of orthologous mouse genes reported in the MGD. I then assigned the phenotype terms associated with these alleles to the human genes. The set of phenotype terms associated with each human gene therefore represents the phenotypes observed to be associated with variants of the gene in human, and variants of orthologous genes in mice. The final data set contains 460,921 associations between phenotype terms from the HPO and the MP and human genes. 8,791 human genes were assigned at least one phenotype term.

To apply the simGIC semantic similarity metric to terms from the HPO and the MP, it is necessary to use an ontology that contains terms from both of these ontologies. Uberpheno is such an ontology (Köhler et al., 2013). Uberpheno was generated by combining terms from the HPO, the MP and a generated Zebrafish

Ontology	Number of terms	Number of relationships
HPO	11,009	17,454
MP	10,011	13,297
ZP	12,438	18,462
Total	33,458	49,213

Table 4.4: The number of terms in the Uberpheno ontology from the HPO, the MP and the ZP and the number of relationships associated with these terms. The number of relationships associated with a term is defined as the number of ‘is_a’ relationships originating from the term.

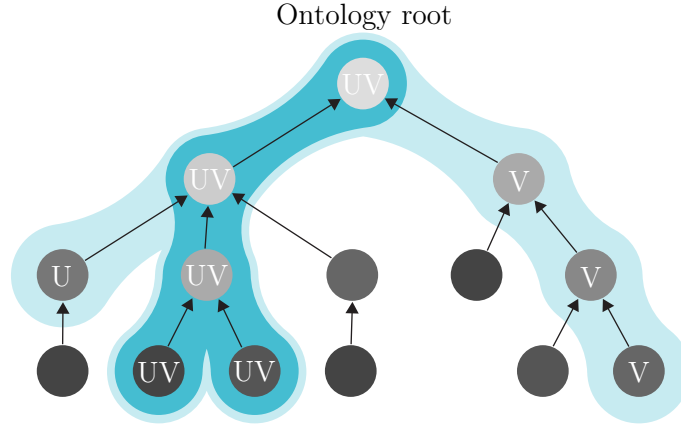
Ontology (ZP) (Table 4.4) using logical definitions associated with each phenotype term and a reasoning process that identified relationships between the terms. I downloaded the Uberpheno ontology files from the Uberpheno web page.

To apply the simGIC metric it is first necessary to estimate the IC of each term in Uberpheno. Here the IC of each term in Uberpheno is the negative logarithm of the proportion of diseases and alleles associated with the term. Following the ‘true path’ rule, ancestors of terms associated with a disease or an allele are also associated with the disease or the allele (Pesquita et al., 2009) and therefore the IC of a term is always greater than or equal to the IC of its ancestors. For two sets of phenotypes, the simGIC similarity score is the Jaccard similarity coefficient weighted by the IC of each term (Figure 4.6). If $I(c)$ is the IC of term c and E_1 and E_2 are two sets of phenotypes, then their simGIC similarity score equals:

$$\text{simGIC}(E_1, E_2) = \frac{\sum_{x \in E_1 \cap E_2} I(x)}{\sum_{y \in E_1 \cup E_2} I(y)} \quad (4.1)$$

The simGIC semantic similarity metric is one of a number of metrics that have been developed to measure the similarity of sets of ontology terms. Other methods include the MICA metric (described in Section 1.4.4) and the simUI metric (Pesquita

A) Higher simGIC similarity score



B) Lower simGIC similarity score

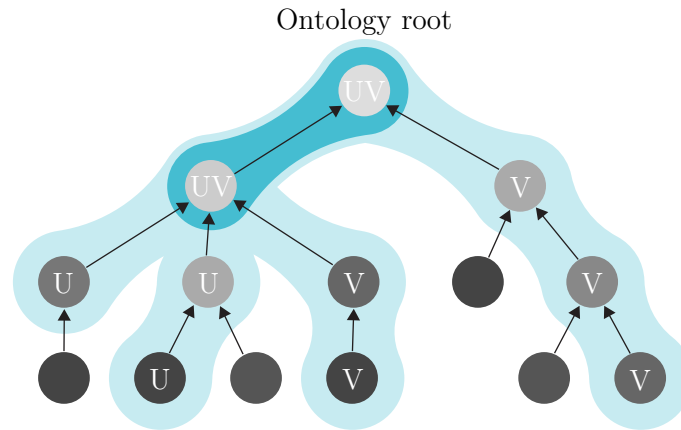


Figure 4.6: Measuring the similarity of a set of disease-associated phenotypes and a set of gene-associated phenotypes using the simGIC semantic similarity metric. Vertices represent terms in the ontology. Terms annotated with a *U* represent phenotypes known to be associated with the disease and terms annotated with a *V* represent phenotypes known to be associated with the gene. The darker the vertex, the greater the IC of the term. The simGIC similarity score is equal to the sum of the IC of terms from the intersection of the sets (the darker blue shape) divided by the sum of the IC of the terms from the union of the sets (the lighter blue shape). Therefore, the greater the IC of the overlap between the disease-associated phenotypes (*U*) and the gene-associated phenotypes (*V*), the greater the simGIC score.

et al., 2008). Pesquita et al. (2008) compared the performance of these metrics using the GO and found that the simGIC metric was best able to use GO annotations to identify genes that shared sequence similarity. For this reason I use the simGIC metric in ALPACA to measure the similarity of sets of phenotype terms.

4.2.7 Protein-protein interaction data

In this section I describe the PPI data used in this chapter and the generation of the context-specific PPI networks.

I downloaded human PPI data (`9606.protein.aliases.v10.txt`) from version 10.0 of the STRING database (Szklarczyk et al., 2015). The context-specific PPI networks generated using these data are used to prioritise genes. It is therefore desirable that the networks cover the greatest number of genes possible in order to maximise the number of genes that can be prioritised. I therefore downloaded all experimentally validated interactions and, unlike in Chapter 3, did not filter by confidence score. I mapped each protein to an Ensembl gene identifier using the STRING alias file (`9606.protein.aliases.v10.txt`). I removed both duplicate and looping interactions from this data set. The final data set contains 1,736,931 interactions between 16,858 genes.

I generated the context-specific PPI networks used by ALPACA using this PPI data and gene expression data from the FANTOM5 project (as described in Section 3.2.1). In the previous chapter, edges in the context-specific PPI networks are weighted using the product of the percentile-normalised expression scores of the interacting genes. In these networks, if either gene has a low expression score, or is not observed as being expressed, then the weight of the connecting edge is low. As observed in the previous chapter, many disease-associated genes are not observed as being expressed in the identified associated contexts. To ensure that ALPACA is still able to prioritise these genes I therefore use a different method to weight the edges of the context-specific PPI networks used in this chapter:

$$w_{i,j,l} = \max(x_{i,l}, x_{j,l}) \quad (4.2)$$

where $x_{i,l}$ is the percentile-normalised expression score of gene i in context l and $w_{i,j,l}$ is the weight of the edge connecting gene i and gene j in the PPI network specific to context l .

In Section 4.3.5, I compare the number of interactions that genes known and not-known to be disease-associated are involved in in five PPI resources. These five resources are BioGRID (Chatr-Aryamontri et al., 2015), HI-II-14 (Rolland et al., 2014), IntAct (Orchard et al., 2014), PrePPI (Garzo et al., 2013) and STRING (Szklarczyk et al., 2015). I removed all interactions from the BioGRID and IntAct data sets that were not marked as direct interactions (MI:0407), associations (MI:0914) or physical associations (MI:0915) (Hermjakob et al., 2004). Some of these resources do not record interactions between different protein isoforms. I therefore mapped interacting proteins to their respective genes and considered all interactions at the gene-level. I used the `biomaRt` R-package (Durinck et al., 2009) to assign Ensembl gene identifiers to proteins in BioGRID, HI-II-14, IntAct and PrePPI. I mapped proteins in STRING to Ensembl gene identifiers as previously described. I removed both duplicate and looping interactions from these data sets.

4.2.8 Evaluating method performance

Artificial disease loci

I evaluated the performance of ALPACA and PRINCE using cross-validation, in which I measured how successfully the methods are able to identify true disease-associated genes in disease-associated loci (Table 4.5). It is not always clear which, if any, genes in trait-associated loci identified in GWAS are causal. I therefore generated artificial disease-associated loci using known disease-associated genes and used these artificial disease loci to evaluate the performance of the methods. Similar approaches have previously been used to evaluate the performance of gene prioritisation methods (Köhler et al., 2008; Vanunu et al., 2010; Piro et al., 2013).

Some methods that have evaluated method performance using artificial loci have generated loci by selecting the N genes closest to a disease-associated gene, where N can range from 50 to 400 (Piro et al., 2013). This set of genes is then used as a candidate set to which the method is applied. These loci contain many more genes

Evaluation	Data sets used	Section
Study bias in PPI databases	BioGRID, HI-II-14, IntAct, PrePPI and STRING PPI resources	4.3.1
Calibrating trait-associated loci	Height-associated SNPs and putative height-associated genes	4.3.2
Network propagation parameter selection	Artificial retinitis pigmentosa disease loci	4.3.3
Type-1 error rate analysis	Simulated null disease loci	4.3.4
Effect of study bias on gene prioritisation	Simulated null disease loci	4.3.5
Comparison of method performance	Artificial GWAS disease loci	4.3.6
Performance using data from multiple species	Artificial GWAS disease loci	4.3.7
Performance using context-specific networks	Artificial GWAS disease loci	4.3.8
Comparison of edge weighting methods	Artificial GWAS disease loci	4.3.9
Case study	RA-associated SNPs	4.3.10

Table 4.5: The evaluations completed in this chapter and the data sets used in these evaluations. Also given is the reference number of the section in which the evaluation is described.

than the loci often identified in GWAS and I therefore took a different approach to generate the artificial loci used in these analyses. For a disease-associated gene, I identified the SNP positioned closest to the centre position of the gene (the start and end positions of a gene correspond to the smallest and largest start and end positions of all transcripts associated with the gene). I then used the method described in Section 4.2.2 to define a trait-associated locus and identify the genes it contains. This results in a locus that generally contains the disease-associated gene and a variable number of other genes, depending on the LD structure of the region surrounding the disease-associated gene. I used data from the CEU panel from the October 2008 phase II release of the HapMap Project to estimate LD when generating artificial loci.

I generated the artificial loci using disease-gene associations from the data set described in Section 4.2.3. To assess the performance of ALPACA when applied to GWAS, it was necessary to use only genes associated with diseases that are studied in GWAS. GWASdb is a database of manually curated and quality controlled trait-associated SNPs identified in GWAS (Li et al., 2016). The database provides details of each of the studies it contains, including DO terms describing the diseases studied. I removed those studies not mapped to a DO term or mapped to multiple DO terms. The DO terms mapped to the remaining studies represent a list of diseases that have been studied in GWAS. I removed associations from the disease-gene association data set that involved diseases not represented in this list of diseases. I then generated artificial disease loci using the remaining associations. I discarded those loci that did not contain the known disease-associated gene from which they were generated. 645 artificial disease loci were generated using this method.

Cross-validation

I used cross-validation and the artificial disease loci to evaluate method performance. The disease-associated genes with which the artificial loci were generated were considered positive examples. The remaining genes in the loci were considered negative examples. If an artificial locus contained another gene associated with the disease with which the locus was generated, then I removed this gene from the

candidate gene set. The identification of an association between a disease and a gene may subsequently affect the identification of additional diseases associated with the gene and the study of the gene in mice. Before applying ALPACA and PRINCE to each artificial locus, I therefore removed the associations between the positive gene and all diseases and mouse alleles from the data used by the methods, in order to reduce the risk of knowledge bias affecting method performance evaluation. If ALPACA is unable to score a gene, then I assigned the gene a p-value of 1. I use the scores given to the positive and negative examples in each locus to evaluate method performance.

Implementation of the PRINCE method

Although an implementation of the PRINCE method is available to download (Gottlieb et al., 2011), I decided to reimplement the method as this allowed me to run PRINCE using the same PPI and disease-gene association data as used by ALPACA. This ensures that any differences in method performance are not due to the data used. I applied ALPACA and PRINCE to the same sets of genes when comparing the performance of the methods. If PRINCE was unable to score a gene, then I assigned the gene a score of zero. Unlike ALPACA, PRINCE uses MIM numbers to describe diseases. I used the cross-referencing provided by the DO to identify suitable MIM numbers for each DO term represented in the set of artificial disease loci used to evaluate method performance. If multiple MIM numbers mapped to a single DO term, then I chose the number that occurred first numerically to represent the disease.

4.3 Results

4.3.1 Study bias in PPI databases

As previously discussed, network-based gene prioritisation methods may be biased towards better-studied genes (Wang et al., 2011b; Oti et al., 2011). Genes known to be disease-associated may represent some of these better-studied genes. It has also been suggested that better-studied genes may also be known to be involved in

greater numbers of interactions (Das & Yu, 2012). In this section I test whether known disease-associated genes are involved in greater numbers of interactions than genes not known to be disease-associated in a number of PPI resources.

I refer to genes known to be disease-associated as ‘disease genes’ and genes not known to be disease-associated as ‘non-disease genes’ for brevity. It may be the case however that some of these non-disease genes will be identified in the future as being disease-associated. If a gene is associated with at least one disease in the disease-gene-association data set described in Section 4.2.3, then I consider the gene a disease gene. All other genes are considered non-disease genes.

I analysed five PPI resources to determine whether disease genes are involved in greater numbers of interactions than non-disease genes. Four of these resources (BioGRID, IntAct, PrePPI and STRING) are literature-based PPI databases and contains PPIs reported in a large number of studies. These studies are a mixture of LT and HT studies. The other resource (HI-II-14) is a single HT PPI screen. As described in Section 1.3.1, LT studies are hypothesis-driven and may therefore be biased towards certain genes. HT screens on the other hand are not hypothesis-driven and should therefore not be biased towards certain genes.

It has been suggested that genes involved in greater numbers of interactions in the body may have a greater propensity to be disease-associated (Yates & Sternberg, 2013). One hypothesis for why this may be is that genes involved in many interactions may be less able to tolerate mutations as a greater amount of their sequence is required for function (Yates & Sternberg, 2013). If disease genes are truly involved in greater numbers of interactions, then we would expect to observe disease genes as being involved in greater numbers of interactions than non-disease genes in both the four literature-based PPI databases and the HT PPI screen. I generated PPI networks using data from each of the five resources to test this (as described in Section 4.2.7).

Table 4.6 shows the mean degree of disease and non-disease genes in the five networks. I considered only the 3,010 genes represented in all five resources to ensure that the results are comparable. In BioGRID, IntAct and PrePPI, disease genes do have a greater mean degree than non-disease genes and the distributions

Resource	Mean degree		Different
	Disease genes	Non-disease genes	
BioGRID	41.7	33.8	$p < 0.034$
HI-II-14	5.37	6.68	$p < 0.022$
IntAct	22.0	19.1	$p < 0.002$
PrePPI	22.0	16.4	$p < 7 \times 10^{-8}$
STRING	269	247	$p < 0.228$

Table 4.6: The network degrees of disease and non-disease genes in networks generated using data from five PPI resources. I considered only the 3,010 genes represented in all five resources. Section 4.2.7 contains details of the five resources and which interactions were considered. I used the Mann-Whitney U test to test whether the distributions of values were drawn from the same populations.

of degrees is significantly different in these two categories ($p < 0.05$). In STRING there is no significant difference between the distributions of disease and non-disease gene degrees. The distributions are significantly different however when all genes represented in STRING are considered ($p < 2 \times 10^{-10}$). Disease genes conversely do not have a higher mean degree in the HI-II-14 network.

The fact that disease genes do not have a higher mean degree in the HT PPI screen suggests that disease genes may not truly be involved in greater numbers of interactions than non-disease genes. The tendency of disease genes to have higher degrees in networks generated using data from literature-based PPI databases may instead occur as a result of these genes being better studied. On the other hand, it has been suggested that the technologies used to complete HT PPI screens may produce large numbers of false positives and false negatives (Braun et al., 2009). Large numbers of false positives and false negatives may therefore alternatively explain why disease genes are not observed as having higher degrees in the HI-II-14 network.

To further investigate why disease genes are reported as being involved in greater numbers of interactions than non-disease genes in literature-based PPI databases,

Resource	Mean number of articles		Different
	Disease genes	Non-disease genes	
BioGRID	16.6	11.4	$p < 3 \times 10^{-16}$
IntAct	5.36	4.12	$p < 3 \times 10^{-16}$

Table 4.7: The mean numbers of articles reporting interactions involving each disease and non-disease gene in BioGRID and IntAct. For each gene I identified the articles that reported at least one interaction involving the gene. I considered only the 11,370 genes represented in both resources. I used the Mann-Whitney U test to test whether the distributions of values were drawn from the same populations.

I compared the number of studies reporting interactions involving each gene to determine whether disease genes are better studied. Both BioGRID and IntAct provide the PubMed manuscript identifier of the study reporting each interaction they contain. It is therefore possible to identify the studies that have analysed interactions involving each gene. It may be the case that other studies have attempted to identify interactions involving the protein products of the genes but failed to find any interactions. It is not possible to account for these cases as negative interactions are generally not reported.

Some interactions are not accompanied by PubMed manuscript identifiers in IntAct and I therefore removed these interactions. For each disease and non-disease gene, I computed the numbers of articles that reported at least one interaction involving the gene. Table 4.7 shows that disease genes tend to be involved in interactions reported across a greater number of studies, in both BioGRID and IntAct, indicating that disease genes may be better studied.

I next attempted to determine whether the tendency of disease genes to be involved in greater numbers of interactions than non-disease genes, in literature-based PPI databases, occurs as a result of the fact that interactions involving disease genes tend to be reported in greater numbers of studies. To test this I computed the number of interactions that involved each disease and non-disease gene in both BioGRID and IntAct. As I was interested in the number of interactions reported

Resource	Mean number of interactions per article		Different
	Disease genes	Non-disease genes	
BioGRID	2.86	3.11	$p < 0.789$
IntAct	4.48	5.41	$p < 0.126$

Table 4.8: The mean numbers of interactions involving each disease and non-disease gene per article reporting interactions involving each gene. For each gene I identified the articles that reported at least one interaction involving the gene. I then divided the total number of interactions involving the gene by the number of articles reporting at least one interaction involving the gene, to give the number of interactions involving the gene per article reporting interactions involving the gene. I considered only the 11,370 genes represented in both resources. I used the Mann-Whitney U test to test whether the distributions of values were drawn from the same populations.

by each study, I did not remove duplicate interactions. For each gene, I divided the number of interactions by the number of articles that reported at least one interaction involving the gene, to produce a value that represents the mean number of interactions involving the gene reported per study reporting interactions involving the gene. If disease genes are truly involved in greater numbers of interactions than non-disease genes, then we may expect that the number of interactions identified per study would be greater for disease genes than non-disease genes. This is not observed in the data from BioGRID and IntAct however (Table 4.8). This result further suggests that the differences in the number of interactions involving disease and non-disease genes in literature-based PPI databases may occur as a result of the fact that disease genes tend to be the subject of greater numbers of studies than non-disease genes.

4.3.2 Calibrating trait-associated loci

In order to identify causal genes in trait-associated loci it is first necessary to define the loci and identify the genes they contain. Previously developed methods that do

this tend to use the same broad approach. They first define associated regions by identifying all SNPs in LD with each trait-associated SNP. Some methods then extend these regions in each direction to the closest recombination hotspots (Raychaudhuri et al., 2009; Rossin et al., 2011; Deo et al., 2014; Himmelstein & Baranzini, 2015). The methods finally identify the genes in these regions. When identifying these genes, some methods extend gene boundaries to incorporate *cis*-regulatory regions (Raychaudhuri et al., 2009; Rossin et al., 2011; Deo et al., 2014; Tasan et al., 2015) and others do not (Himmelstein & Baranzini, 2015; Pers et al., 2015). There are therefore three main parameters in this broad approach: 1) the r^2 cutoff used to select SNPs in LD with each trait-associated SNP, 2) whether or not to extend regions in each direction to the closest recombination hotspots and 3) whether or not to add flanking regions to gene boundaries to incorporate *cis*-regulatory regions. Many of the articles describing the development of comparable gene-prioritisation methods do not report the testing of the effect of any of these parameters on the definition of trait-associated loci, or on the subsequent performance of the methods applied to them (Raychaudhuri et al., 2009; Rossin et al., 2011; Deo et al., 2014; Himmelstein & Baranzini, 2015). The developers of DEPICT (Pers et al., 2015) did however test how each of these three parameters affected the performance of DEPICT. The developers of PrixFixe (Tasan et al., 2015) tested the effect of using different r^2 cutoffs on method performance, but did not test the other parameters. In this section I test how these three parameters affect the definition of trait-associated loci and the genes they contain.

To conduct these tests I used the observation made by Lui et al. (2012) that genes in height-associated loci are enriched with genes functional in the growth plate. The growth plate is the proliferative region of bone and plays a crucial role in bone elongation and therefore height (Lui et al., 2012). Lui et al. characterised genes functional in the growth plate by identifying genes expressed at higher levels in the growth plate compared to other tissues and by identifying genes spatially and temporally regulated in the growth plate. They completed this analysis in mice and rats and mapped identified genes to their human orthologs. In total, Lui et al. identified genes mapping to 427 human genes that satisfied at least two of the three

r^2	Without hotspot extension		With hotspot extension	
	No flank	50kb flank	No flank	50kb flank
0.1	0.101	0.266	0.430	0.366
0.2	0.124	0.100	0.314	0.132
0.3	0.014	0.007	0.042	0.026
0.4	0.004	0.009	0.020	0.019
0.5	0.001	0.008	0.028	0.052
0.6	2×10^{-4}	0.002	0.015	0.027
0.7	8×10^{-5}	0.001	0.014	0.036
0.8	3×10^{-4}	0.001	0.008	0.024
0.9	2×10^{-4}	4×10^{-4}	0.006	0.026

Table 4.9: The enrichment of height-associated genes in loci generated using different parameter sets. Loci were generated using different r^2 cutoffs, with and without extending regions in each direction to the closest recombination hotspots and with and without adding flanking regions of 50kb to gene boundaries. I measured the enrichment of putative height-associated genes in the defined candidate gene sets using the one-tailed version of Fisher’s exact test.

previously described conditions. They noted that height-associated loci identified in a height GWAS meta-analysis are enriched with these genes (Lango Allen et al., 2010).

I used this set of 427 putative height-associated genes to test the effect of the previously outlined parameters on the definition of trait-associated loci. I obtained 697 height-associated SNPs from a more recent and larger GWAS meta-analysis of height (Wood et al., 2014). I generated trait-associated loci using these SNPs, r^2 cutoffs of between 0.1 and 0.9, with and without extending regions to the closest recombination hotspots, and with and without adding flanking regions of 50kb to gene boundaries to incorporate *cis*-regulatory regions. I chose a value of 50kb as Veyrieras et al. (2008) demonstrated that the majority of *cis*-acting expression quantitative trait loci (eQTL) lie within 50kb of their target gene. I measured

the enrichment of putative height-associated genes in these loci using the one-tailed version of Fisher’s exact test. If trait-associated loci are correctly defined, then we would expect them to be enriched with genes associated with the trait. It is also desirable however to maximise the size of the loci, to maximise the likelihood that causal genes are contained amongst the candidate genes.

When regions are not extended to recombination hotspots and gene boundaries are not extended, the use of r^2 cutoffs of 0.3 or greater identifies sets of candidate genes enriched with putative height-associated genes ($p < 0.05$, Table 4.9). Extending regions to the closest recombination hotspots does not increase this enrichment. Adding flanking regions of 50kb to gene boundaries does increase this enrichment however. Throughout the remainder of this chapter, trait-associated loci are therefore defined using an r^2 cutoff of 0.3, without extending regions to their closest recombination hotspots and with gene boundary extensions of 50kb.

These results differ to the results of Pers et al. (2015). Pers et al. also used the putative height-associated genes identified by Lui et al. (2012) and the height-associated SNPs identified by Wood et al. (2014) to calibrate their definition of trait-associated loci. Pers et al. however did not measure the enrichment of putative height-associated genes in the loci, but instead measured the performance of their method (DEPICT) when run using different loci-defining parameter sets. They noted that DEPICT performed best when an r^2 cutoff of 0.5 was used. They also noted that extending regions to the closest recombination hotspots and adding flanking regions of 50kb to gene boundaries both reduced the performance of DEPICT. The differences between my results and the results produced by Pers et al. are likely due to this difference in loci calibration methodology and the performance of DEPICT when applied to gene sets of different sizes.

4.3.3 Network propagation parameter selection

The RWR propagation algorithm used by ALPACA to propagate scores across PPI networks uses two parameters: 1) the restart probability r and 2) the fixed number of iterations to complete. In this section I test how these parameters affect the performance of ALPACA.

Restart r	Number of iterations				
	10	20	30	40	50
0.1	0.763	0.764	0.763	0.764	0.763
0.3	0.770	0.768	0.770	0.772	0.770
0.5	0.769	0.770	0.772	0.770	0.771
0.7	0.769	0.767	0.769	0.770	0.766
0.9	0.768	0.766	0.770	0.770	0.768

Table 4.10: The performance of ALPACA when run using different RWR parameter sets. Two parameters are tested: 1) the restart probability r and 2) the fixed number of iterations to complete. ALPACA was applied to 56 artificial disease loci generated using genes known to be associated with retinitis pigmentosa. Scores represent AUCs.

I used cross-validation and artificial disease loci to test how these parameters affect method performance. Later in this chapter I use artificial disease loci generated using genes associated with diseases studied in GWAS to evaluate the performance of ALPACA. In order to keep the training and testing of ALPACA separate, I therefore use a different data set to test how the RWR algorithm parameters affect method performance. I chose to generate artificial disease loci using genes known to be associated with ‘retinitis pigmentosa’ (DOID:10584), as this is the disease known to be associated with the greatest number of genes not in the data set used to test the performance of ALPACA. I generated 56 artificial disease loci using genes associated with this disease. I then ran cross-validation using each of the loci with five different values of r (0.1, 0.3, 0.5, 0.7, 0.9) and five different numbers of iterations (10, 20, 30, 40, 50).

The choice of parameters had little effect on method performance (Table 4.10). The RWR algorithm has previously been observed to perform well when run with r equal to 0.5 (Vanunu et al., 2010). I therefore ran ALPACA with r equal to 0.5 in the remaining analyses in this chapter. The greater the number of iterations completed, the greater the time required by ALPACA to run. I therefore run ALPACA using 10 iterations.

4.3.4 Type-1 error rate analysis

I conducted simulations to determine whether the p-values generated by ALPACA approximately estimate type-1 error rates. I simulated 100 null GWAS by sampling 100 diseases from the set of diseases for which phenotype data are available. For each study, I simulated disease-associated SNPs by sampling 23 SNPs from the set of SNPs for which there are data in the October 2008 phase II release of the HapMap Project. I chose to sample sets of 23 SNPs as 23 is the mean number of trait-associated SNPs passing a genome-wide significance threshold reported by studies in the GWASdb data set.

These GWAS are simulated and therefore it can be expected that few or none of the genes in the loci are truly associated with the disease. If ALPACA does correctly estimate type-1 error rates then we would expect that the p-values generated by ALPACA when applied to these loci would be uniformly distributed.

I applied ALPACA to each of these simulated null GWAS and estimated LD using the CEU panel from the October 2008 phase II release of the HapMap Project. I discarded those loci that did contain genes known to be associated with each disease. It may be the case that some of the loci in these simulated null GWAS do contain genes associated with each disease that have not yet been identified. If this were the case then the type-1 error rate would appear to be inflated. It can be seen in Figure 4.7 that the p-values generated by ALPACA are roughly uniformly distributed, indicating that ALPACA approximately estimates type-1 error rates.

4.3.5 Effect of study bias on gene prioritisation

In this section I test how study bias may affect the performance of ALPACA and PRINCE. If a gene prioritisation method is biased towards genes involved in greater numbers of interactions, and disease-associated genes are involved in greater numbers of interactions in the interaction data used by the method as a result of study bias, then estimates of the performance of the method may be inflated in cross-validation. This may occur as the method may rank the disease genes higher as a result of them being better studied. A bias towards genes involved in greater numbers of interactions may also make a method less useful, as it could mean that the method

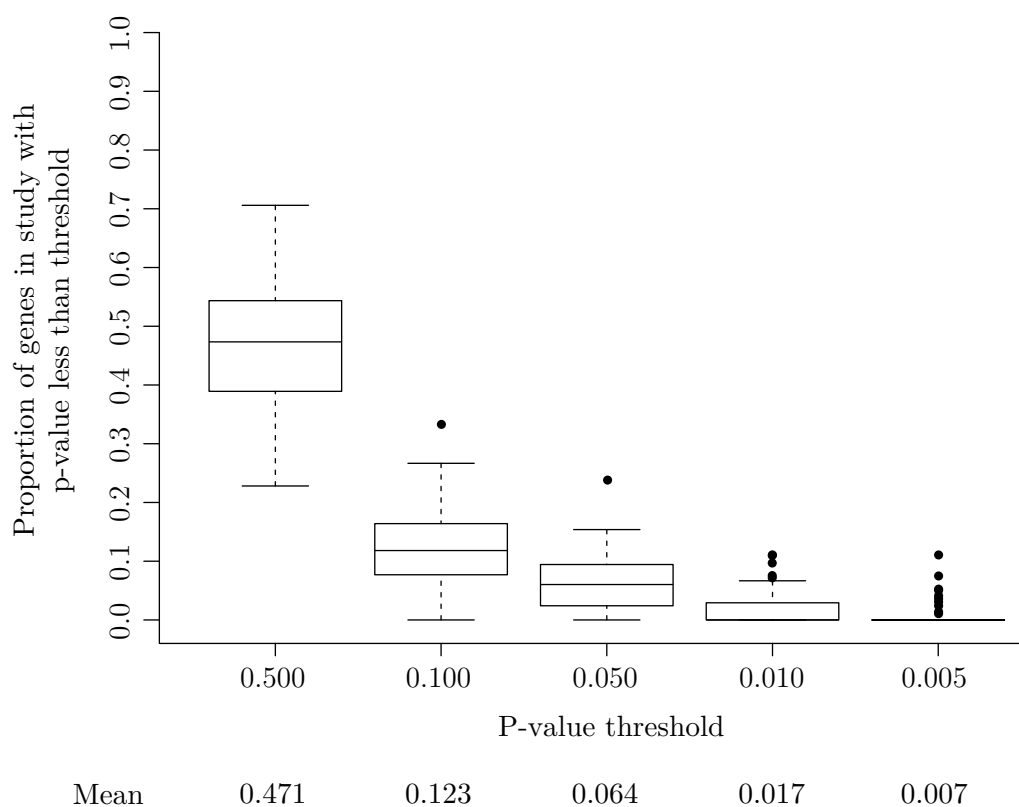


Figure 4.7: The p-values generated by ALPACA when applied to 100 simulated null GWAS. One disease and 23 associated SNPs were sampled to generate each null GWAS. For each p-value threshold, the proportion of genes in the loci in each study with a p-value less than or equal to the threshold is plotted. Each box shows the median proportion and the interquartile range. The box whiskers extend to the most extreme data points, up to a limit of 1.5 times the interquartile range. The mean proportion of genes with p-values less than the threshold is also given at the bottom.

Method	Correlation with gene network degree	Correlation with number of associations
ALPACA K	0.468	0.589
ALPACA p-values	-0.049	0.004
PRINCE	0.611	0.381

Table 4.11: Correlations between the gene scores output by ALPACA and PRINCE and the network degree of each gene and the number of associated diseases and mouse alleles. The observed propagated phenotypic relevance scores K and the p-values generated by ALPACA are both given. As PRINCE uses only associations between genes and human diseases, only these associations were considered when measuring the correlation between PRINCE scores and the number of associations. I used Spearman’s rank correlation coefficient to measure the correlations.

is less able to identify new disease genes for which fewer data may be available. It is therefore desirable to develop a gene prioritisation method that is not biased towards better-studied genes.

To determine whether ALPACA and PRINCE are biased towards better-studied genes, I applied both methods to the simulated null GWAS generated in Section 4.3.4. To reduce the risk of study bias, ALPACA compares the observed propagated phenotypic relevance scores K against scores generated using permuted sets of phenotypes. To determine whether this affects whether ALPACA is biased towards better-studied genes, I compared K and the p-values generated by ALPACA against the degree of each gene in the interaction network used and the number of diseases and mouse alleles associated with each gene (Table 4.11). Generating p-values by comparing observed and permuted scores reduced the correlation between gene score and both gene degree (from 0.468 to -0.049 , measured using Spearman’s rank correlation coefficient) and the number of associated diseases and mouse alleles (from 0.589 to 0.004). The permutation-based-approach used by ALPACA therefore lowers its bias towards better-studied genes.

I similarly compared the gene scores output by PRINCE against the degree of

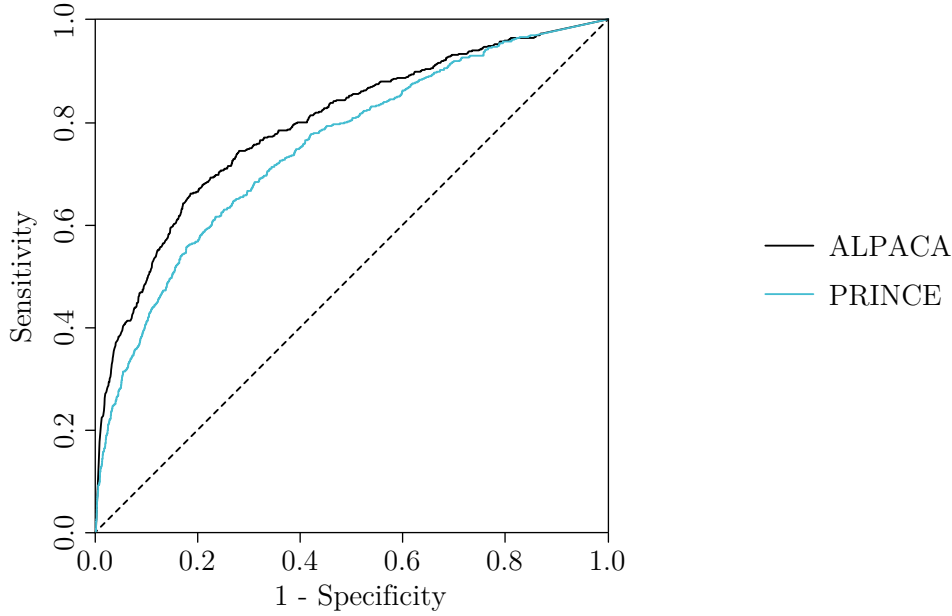


Figure 4.8: The performance of ALPACA and PRINCE in cross-validation. ALPACA achieves an AUC of 0.794 and PRINCE achieves an AUC of 0.750. The difference between these AUCs is significantly different from zero ($p < 2 \times 10^{-4}$, measured using the two-sided version of DeLong’s method).

each gene and the number of associated diseases. There is a correlation between gene scores and both gene degree (0.611) and the number of associated diseases (0.381), indicating that PRINCE is biased towards better-studied genes.

4.3.6 Comparison of method performance

I used cross-validation and artificial disease loci to compare the performance of ALPACA and PRINCE. PRINCE can only be applied to diseases with a MIM number. To compare the two methods I therefore used only the artificial disease loci generated using genes associated with diseases that could be mapped to MIM numbers using the cross-referencing provided by the DO. ALPACA significantly outperforms PRINCE in this cross-validation ($p < 2 \times 10^{-4}$, measured using the two-sided version of DeLong’s method, Figure 4.8)

As demonstrated in Section 4.3.5, PRINCE is biased towards genes involved in greater numbers of interactions. The permutation-based approach used by ALPACA

Degree of disease genes	Number of loci	ALPACA AUC	PRINCE AUC	Different
1 - 6	63	0.799	0.573	$p < 3 \times 10^{-10}$
7 - 17	77	0.787	0.677	$p < 6 \times 10^{-4}$
18 - 39	68	0.874	0.725	$p < 4 \times 10^{-6}$
40 - 80	72	0.842	0.783	$p < 0.068$
81 - 173	70	0.823	0.820	$p < 0.936$
174 - 333	70	0.799	0.823	$p < 0.406$
334 - 850	70	0.763	0.838	$p < 0.023$
851 - 3720	71	0.826	0.883	$p < 0.044$

Table 4.12: The performance of ALPACA and PRINCE when prioritising disease genes with different network degrees. I split the artificial disease loci into sets based on the network degree of each disease gene. The number of loci in each set is given. I used cross-validation to evaluate performance and give the AUC that each method achieved for each set. I used the two-sided version of DeLong’s method to determine whether the differences in AUCs are significantly different from zero.

means that it is less biased towards these genes. To determine how this difference affects the performance of the two methods, I split the artificial disease loci into sets based on the network degree of the disease genes. I then tested the performance of the methods when applied to these sets in cross-validation (Table 4.12).

To generate the loci sets I computed the degree of each disease gene in the network used by ALPACA and PRINCE. I then sorted the loci by this number. The degree of these genes ranged from one to 3,720. I identified seven quantiles in this sorted distribution of gene degrees and grouped the loci using these quantiles. Loci containing disease genes with the same degree were always added to the same set and therefore the number of loci in each set is not equal (Table 4.12).

ALPACA performs consistently when prioritising disease genes with different network degrees (Table 4.12). The method achieves an AUC of 0.799 when prioritising genes with degrees of between one and six and an AUC of 0.826 when

prioritising genes with degrees of between 851 and 3,720. The performance of PRINCE is more variable however. PRINCE achieves an AUC of only 0.573 when prioritising genes with degrees of between one and six and a much higher AUC of 0.883 when prioritising genes with degrees of between 851 and 3,720. ALPACA performs significantly better than PRINCE when prioritising genes with degrees of less than 39 and PRINCE performs significantly better than ALPACA when prioritising genes with degrees greater than 334 ($p < 0.05$).

This variability in the performance of PRINCE again demonstrates that it is biased towards better-studied genes. The permutation-based approach used by ALPACA ensures that it is not similarly biased. ALPACA may therefore be more suitable for identifying new disease genes for which fewer data may be available.

4.3.7 Performance using data from multiple species

To test whether the performance of ALPACA is improved by including both human and mouse data, I completed cross-validation using both human and mouse data, using only human data and using only mouse data (Figure 4.9). The performance of ALPACA is significantly better when it is run using both human and mouse data, compared to when it is run using only human data ($p < 2 \times 10^{-6}$, measured using the two-sided version of DeLong's method) and only mouse data ($p < 4 \times 10^{-6}$). This demonstrates that ALPACA successfully integrates and uses data from multiple species.

As previously mentioned, the ability to engineer mutations in mice means that phenotype data are available for a greater number of mouse genes than human genes. In the phenotype data set used by ALPACA, mouse phenotype data are available for 7,964 human genes and human phenotype data are available for only 2,899 human genes. Despite this, there was no significant difference in the performance of ALPACA when run using only mouse data and only human data. This may reflect the fact that whilst human phenotype data are available for fewer genes, they are more relevant to the study of human disease.

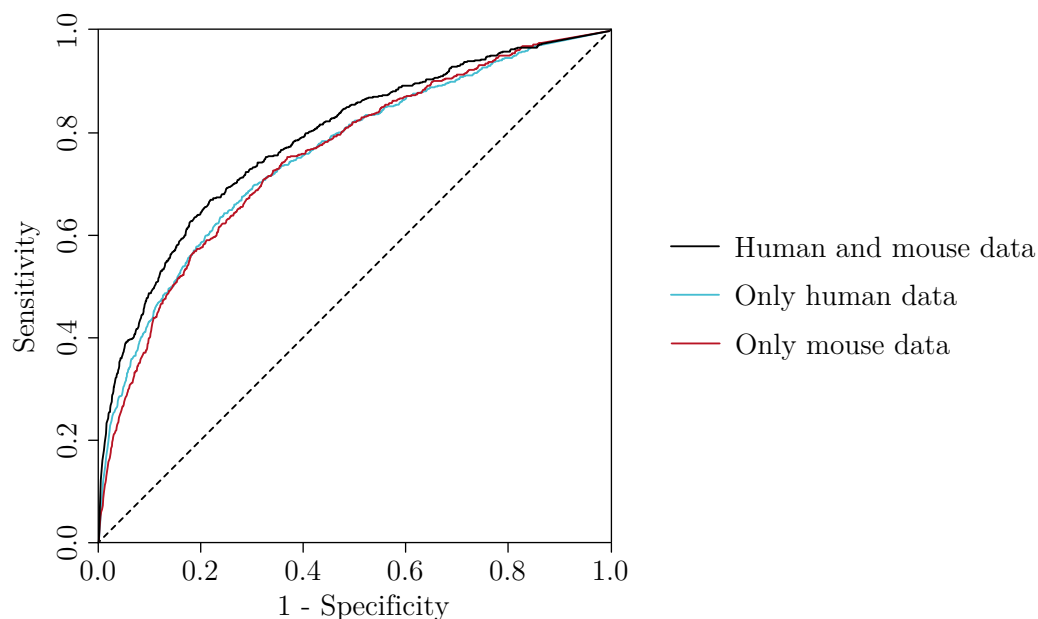


Figure 4.9: The performance of ALPACA when run using both human and mouse data, only human data and only mouse data. Performance was measured using cross-validation and artificial disease loci. The AUC of ALPACA is 0.788 when run using both human and mouse data, 0.757 when using only human data and 0.754 when using only mouse data. When ALPACA is run using both human and mouse data it performs significantly better than when run using only human data ($p < 2 \times 10^{-6}$) and only mouse data ($p < 4 \times 10^{-6}$). I used the two-sided version of DeLong's method to determine whether the differences in AUCs are significantly different from zero.

4.3.8 Performance using context-specific networks

ALPACA uses the GSO method to identify disease-associated contexts. If the GSO method identifies a disease-associated context, then a PPI network specific to this context is used to propagate scores. If the GSO method does not identify a disease-associated context, then a generic PPI network is used to propagate scores. In this section I test whether using context-specific PPI networks, instead of generic PPI networks, improves the performance of ALPACA. In Chapter 3, text mining was also used to identify disease-associated contexts. In this section I therefore also test whether similar results are observed when ALPACA is run using networks specific to contexts identified by the text mining method. As seen in Chapter 3, there is a large overlap in the associations identified by the GSO and GSC methods. The GSC method requires longer to run than the GSO method. For these reasons I have not evaluated the performance of ALPACA when run using the GSC method to identify disease-associated contexts.

I used cross-validation and the 645 artificial disease loci to assess the performance of ALPACA using the GSO and text mining methods to select context-specific PPI networks and using only generic PPI networks. The text mining was completed using the method described in Section 3.2.5 on 16 November 2015. I used the cross referencing provided by the DO to map the DO terms mapped to the artificial disease loci to MeSH terms. If this method failed to identify a MeSH term, then I mapped the DO term to a MeSH term by querying the MeSH database (<http://www.ncbi.nlm.nih.gov/mesh>) with the disease name. I could then use these MeSH terms and the MeSH terms mapped to each cell type in Section 3.2.5 to complete the text mining. Disease-context associations identified by the text mining method were used in ALPACA in the same way as the disease-context associations identified by the GSO method.

ALPACA performs significantly better when run using PPI networks specific to contexts identified by the GSO method ($p < 4 \times 10^{-5}$, Table 4.13). It does not perform significantly different however when run using PPI networks specific to contexts identified by the text mining method ($p < 0.100$).

The GSO and text mining methods identify disease-associated contexts in 253/

Context identification method	AUC	Different from none
<i>All cases</i>		
None	0.775	
GSO	0.789	$p < 4 \times 10^{-5}$
Text mining	0.766	$p < 0.100$
<i>Cases in which a context was identified by the GSO method</i>		
None	0.803	
GSO	0.838	$p < 2 \times 10^{-5}$
Text mining	0.798	$p < 0.523$
<i>Cases in which a context was identified by the text mining method</i>		
None	0.780	
GSO	0.796	$p < 2 \times 10^{-5}$
Text mining	0.769	$p < 0.955$

Table 4.13: The performance of ALPACA when context-specific PPI networks are not used (None) and when the GSO and text mining methods are used to identify suitable context-specific PPI networks. The GSO and text mining methods were not able to identify disease-associated contexts in every case. For this reason, the performance of ALPACA is given for all cases, for only the cases where the GSO method identified a disease-associated context, and for only the cases where the text mining method identified a disease-associated context. I used the two-sided version of DeLong’s method to determine whether the performance of ALPACA when run using context-specific PPI networks is significantly different to the performance of ALPACA when run using only generic PPI networks. ALPACA performs significantly better when context-specific PPI networks are used and identified by the GSO method, but not the text mining method.

645 (39.2%) and 523/645 (81.1%) of cases respectively. If no disease-associated context is identified then ALPACA uses a generic PPI network to propagate scores. In these cases, the results produced by ALPACA are the same as when no context-specific PPI networks are considered. For this reason, I split Table 4.13 to show the performance of ALPACA when all cases are considered, when only the 253 cases in which the GSO method identifies a disease-associated context are considered, and when only the 523 cases in which the text mining method identifies a disease-associated context are considered.

These results demonstrate that the performance of ALPACA is improved by using PPI networks specific to contexts identified by the GSO method. Methods that use context-specific networks to prioritise genes tend to use text mining to identify suitable contexts (Magger et al., 2012; Li et al., 2014). These results suggest that the performance of these methods may be improved by using the GSO method to identify suitable contexts.

4.3.9 Comparison of edge weighting methods

In Chapter 3, the cell-type-specific PPI networks were generated by weighting each edge with the product of the percentile-normalised gene expression scores of the interacting genes. Conversely in this chapter, the cell-type-specific PPI networks are generated by weighting each edge with the maximum of the percentile-normalised gene expression scores. I did this to try to ensure that genes that were not expressed highly in a given context could still be ranked highly by ALPACA if they interacted with many other highly expressed genes.

In this section, I test how this difference affects the performance of ALPACA. I again completed cross-validation using the artificial disease loci. I ran cross-validation twice, once using the networks generated using the product of the scores (as described in Section 3.2.3) and once using the maximum of the scores (as described in Section 4.2.7).

As previously mentioned, the GSO method does not identify a disease-associated context for all diseases. When no disease-associated context is identified, then the generic PPI network is used to prioritise genes. Edges are not weighted in

this generic PPI network and therefore when the GSO method does not identify a disease-associated context, it does not make a difference whether the cell-type-specific PPI networks were generated using the product or the maximum of the scores. This means that when all disease loci are considered (those for which the GSO method could and couldn't identify a context), then there is little difference in the performance of ALPACA in cross-validation using the product and the maximum of the scores (AUCs of 0.789 and 0.788 respectively). The GSO method identifies disease-associated contexts for 253/645 of the disease loci. When only these 253 disease loci for which the GSO method could identify a context are considered, then there is still little difference in the performance of ALPACA in cross-validation using the product and the maximum of the scores (AUCs of 0.838 and 0.836 respectively). There is no significant difference between these AUCs (at a significance level of $p < 0.05$ measured using the two-sided version of DeLong's method).

This analysis therefore provides no evidence that using the maximum of the percentile-normalised gene expression scores, instead of the product of the percentile-normalised gene expression scores, to weight the edges in the cell-type-specific PPI networks improves performance.

4.3.10 Case study

In Section 4.2.1, I applied ALPACA to RA to illustrate the computational procedure. In this section I describe a case study of RA using ALPACA and discuss the gene prioritisation results.

As described in Section 4.2.1, Myouzen et al. (2012) identified three loci as being associated with RA. I applied ALPACA to these loci using the JPT+CHB reference panel to estimate LD. These loci contain a total of twelve genes. ALPACA gives STAT4 the lowest p-value ($p < 0.021$). STAT4 does not however pass a significance threshold of $q < 0.1$ after correction for multiple testing using the BH procedure. Despite this, STAT4 may represent the most interesting candidate for future study as it is already known to be associated with the inflammatory disease SLE. Additionally, a number of abnormal immune system phenotypes have been observed in mice mutants of the orthologous gene, including increased susceptibility

to autoimmune diabetes and abnormal natural killer (NK) cell physiology.

I next completed genome-wide gene prioritisation of genes associated with RA using ALPACA to determine whether ALPACA is able to prioritise genes outside of the three loci. I did this again using the default ALPACA parameters. I removed genes known to be associated with RA and corrected for multiple testing using the BH procedure. 17 genes are identified at a FDR of 10%. None of these 17 genes is known to be associated with any human disease (in the data set used by ALPACA) but six have phenotype data available for their orthologous mouse genes. In some of these mouse mutants, altered RA risk has been observed (MGI:2182132), and in others, immune system abnormalities have been observed (MGI:2182133, MGI:3055564, MGI:5491497). The similarities between these phenotypes and RA are likely to partially explain why these genes are ranked highly. The remaining genes are ranked highly because of their proximity to the protein products of known RA-associated genes in the PPI network.

DAVID is a resource that facilitates a number of tasks, including measuring the enrichment of gene sets in lists of genes (Huang et al., 2009). I used DAVID to determine whether the genes ranked highly by ALPACA are enriched with genes involved in certain cellular pathways and processes. Huang et al. recommend applying DAVID to lists of between 100 and 2,000 genes, as this increases the statistical power of DAVID. As previously mentioned, only 17 genes passed a FDR cutoff of 10%, and I therefore instead consider the larger set of 686 genes that pass a significance cutoff of $p < 0.05$ (without correction for multiple testing) when completing this enrichment analysis. 662 of these 686 genes were mapped to a gene identifier by DAVID and could therefore be used in the enrichment analysis. DAVID can be used to measure the enrichment of gene sets of various types. These gene sets include functional sets, pathways and sets of known disease-associated genes. I ran DAVID using gene sets from the biological process, cellular component and molecular function GO categories (The Gene Ontology Consortium, 2014) and from the BBID (Becker et al., 2000), BioCarta (Nishimura, 2001) and KEGG (Kanehisa & Goto, 2000) pathway databases. As the background gene set I used all human protein-coding genes to which ALPACA was able to assign a p-value.

Term	Category	P-value	Q-value
Skeletal system development	GO biological process	2.8×10^{-11}	6.1×10^{-8}
Immune response	GO biological process	2.0×10^{-10}	2.2×10^{-7}
Skeletal system morphogenesis	GO biological process	3.5×10^{-8}	2.6×10^{-5}
Natural killer cell mediated cytotoxicity	KEGG pathway	4.2×10^{-7}	5.1×10^{-5}
Carbohydrate binding	GO molecular function	5.0×10^{-7}	3.0×10^{-4}

Table 4.14: Results of the gene set enrichment completed using the list of genes with $p < 0.05$ and DAVID. Shown are the five most significantly enriched gene sets, along with the enrichment p -values and q -values (computed using the BH procedure).

Table 4.14 shows the five most significantly enriched gene sets. Two of these gene sets are immune system related. This is of little surprise as RA is considered an autoimmune disease (McInnes & Schett, 2012). The enrichment of genes in the ‘natural killer cell mediated cytotoxicity’ KEGG pathway is of interest however, as there is emerging evidence that NK cells may play a role in the pathogenesis of autoimmune diseases, such as RA (Fogel et al., 2013). NK cells play an important immunoregulatory role and it has therefore been suggested that disruption to NK cells could lead to aberrant activation of other immune cell types (Fogel et al., 2013). NK cells have also been observed to accumulate in the synovium of patients with RA (Dalbeth & Callan, 2002), however whether this is a factor that contributes to the development of the disease or a result of the disease is unclear. More specifically, disruption to NK cell mediated cytotoxicity is already known to cause disease. Hemophagocytic lymphohistiocytosis is an immune dysregulatory syndrome caused by the failure of cells such as NK cells to kill persistently activated and infected cells, as a result of the disruption of NK cell mediated cytotoxicity (Risma & Jordan, 2012). Despite this, complete loss of NK cells in individuals tends to be associated with increased risk of viral infection (Orange, 2002), rather than increased risk of developing autoimmune diseases. Whether there is a link between RA and NK cells is therefore far from clear. It is also important to note that the gene list is enriched with genes in the ‘natural killer cell mediated cytotoxicity’ pathway, despite the fact

that the GSO method identified monocytes, rather than NK cells, as the cell type most strongly associated with RA.

Why the gene list is enriched with genes from the three other gene sets in Table 4.14 is unclear. A genome-wide DNA methylation study previously identified genes involved in skeletal system morphogenesis as being associated with osteoarthritis (Aref-Eshghi et al., 2015). Similarly, carbohydrate-binding proteins such as Galectin-1 are thought to play a role in inflammation in osteoarthritis (Toegel et al., 2016). Why the gene list of prioritised RA genes is enriched with genes from these gene sets is unclear however. It may be the case that, due to the phenotypic similarity of RA and osteoarthritis, some of the highly ranked genes are known to be associated with osteoarthritis, leading to the enrichment of genes from functional gene sets and pathways related to osteoarthritis.

4.4 Discussion

The study bias that exists in literature-based PPI databases may affect both the development of PPI-based gene prioritisation methods and the evaluation of their performance. The performance of many methods (including ALPACA in this thesis) is evaluated using cross-validation. This approach attempts to determine how well a method is able to predict new disease-gene associations by taking a known association, removing from the data used by the method all data that exists as a result of this association being known, and then testing whether the method is able to predict this ‘masked’ association. The difficulty in this approach comes in removing all data that exists as a result of the association being known. As observed in Section 4.3.1, genes known to be disease-associated tend to be the subject of a greater number of studies identifying PPIs, and this may explain why disease genes tend to be involved in greater numbers of interactions than non-disease genes in literature-based PPI databases. For any individual disease gene, it is however not possible to determine which of the interactions it is known to be involved in have been identified as a result of the gene being identified as being disease-associated. If a gene is involved in a greater number of interactions as a result of it being identified as being disease-associated, and a gene prioritisation method (such as PRINCE) is

biased towards genes involved in greater numbers of interactions, then estimates of the performance of the method may therefore be inflated in cross-validation. It may not be possible to truly evaluate the performance of many PPI-based gene prioritisation methods until a map of the interactome has been realised in a single high-quality unbiased HT screen. So far, HT screens have identified only a small fraction of the PPIs thought to take place in the human body (Stumpf & Thorne, 2008; Rolland et al., 2014) and are therefore not yet suitable for method performance evaluation.

Some gene prioritisation methods require the disease of interest to be defined before the method can be run. ExomeWalker (Smedley et al., 2014) and GeneWanderer (Köhler et al., 2008) use previously identified disease-associated genes from OMIM to prioritise genes and therefore require the disease of interest to be represented in OMIM. PRINCE (Vanunu et al., 2010) and RWRH (Li & Patra, 2010) also use OMIM to identify phenotypically similar diseases and therefore also require the disease of interest to be represented in OMIM. Other methods, such as DAPPLE (Rossin et al., 2011), NetWAS (Greene et al., 2015), Prioritizer (Franke et al., 2006) and PrixFixe (Tasan et al., 2015), do not require the disease of interest to be defined and therefore can be applied to any disease. ALPACA requires the disease of interest to be defined in order to identify genes associated with phenotypically similar diseases. As ALPACA uses sets of phenotypes to describe disease, it is however not limited to those diseases represented in any single vocabulary or database. In this thesis, I have applied ALPACA only to diseases represented in the DO to which phenotype terms have been mapped. I chose to do this as it allowed me to test the performance of ALPACA in a systematic and unbiased manner. There is no reason however why a user could not apply ALPACA to a set of phenotypes they have chosen to represent their own disease of interest. PhenIX takes a similar approach, allowing its users to prioritise disease variants using a set of phenotypes they have chosen (Zemojtel et al., 2014). Zemojtel et al. suggest that this is advantageous as it allows users to prioritise disease variants for patients with no clear diagnosis.

ALPACA currently considers all gene-associated phenotypes equally, no matter whether the phenotype was observed in humans or mice. As described in Section

4.3.7, ALPACA performs similarly when run using only phenotype data from humans and only phenotype data from mice. This is despite the fact that in the data sets used, mouse phenotype data are available for 7,964 human genes and human phenotype data only 2,899 human genes. This suggests, as may have been expected, that phenotype data from humans are more useful for identifying associations between human genes and human diseases than phenotype data from mice. It may therefore be logical that ALPACA be developed further to weight phenotype data from humans and mice differently. Phenotype weighting would become increasingly important if ALPACA were expanded to use data from additional species, as the more distantly related the species, the less useful phenotype data from the species may be.

4.5 Conclusions

In this chapter I have demonstrated that genes known to be disease-associated are the subject of greater numbers of studies identifying PPIs. Based on this observation, I developed ALPACA, which uses a permutation-based approach to prioritise causal genes in trait-associated loci identified in GWAS and avoid being biased towards better-studied genes. I used cross-validation to evaluate the performance of ALPACA and compare its performance to that of the PRINCE gene prioritisation method. ALPACA both outperforms PRINCE and, unlike PRINCE, is not biased towards genes involved in greater numbers of interactions or for which more phenotype data are available. This suggests that ALPACA may be better at identifying new disease genes, for which fewer data may be available. The permutation-based approach used by ALPACA also allows the method to produce more meaningful scores that provide an estimate of the probability of observing gene scores at least as extreme as that observed, given the data used and that the gene is not associated with the disease.

ALPACA builds on the work described in Chapter 3 and uses cell-type-specific PPI networks to prioritise genes. ALPACA uses the GSO method to determine which cell-type-specific PPI network is best suited to prioritising genes associated with each disease. I have demonstrated that using cell-type-specific PPI networks

identified by the GSO method improves the performance of ALPACA over using only generic PPI networks. Using cell-type-specific PPI networks identified by the text mining method does not improve method performance however. Other gene prioritisation methods that use context-specific networks to prioritise genes require the user of the method to specify the context to be used or use text mining to identify suitable contexts. The work described in this chapter demonstrates that the performance of these methods may be improved by using the GSO method to identify suitable contexts.

GWAS have identified thousands of trait-associated SNPs (Li et al., 2016). For many of these SNPs however, the effected genes have yet to be identified. Network-based gene prioritisation methods have successfully been used to prioritise disease-associated genes and therefore offer a way of identifying the genes in disease-associated loci most likely to be causal. As more genotypic, transcriptomic, interactomic and phenotypic data become available for additional species, it will become increasingly important to integrate these data to better understand relationships between genotype and phenotype.

Chapter 5

Discussion and future work

In this chapter I will discuss some of the difficulties in using a network-based approach to study disease. I will also explain how the availability of additional data in the near future will aid in the identification of disease genes and in understanding the phenotypic effects of these genes. Finally, I will outline some work that could be done to improve the performance of ALPACA and extend its capabilities.

5.1 Discussion

5.1.1 Moving towards a dynamic picture of the interactome

The interactome has been successfully used in a number of tasks, such as in studying gene and protein function (Warde-Farley et al., 2010), in identifying disease-causing variants and genes (Krauthammer et al., 2004; Vanunu et al., 2010; Rossin et al., 2011; Yates et al., 2014) and in characterising relationships between diseases (Menche et al., 2015). The performance and utility of these and similar interactome-based methods is likely to improve as we develop a better understanding of the interactome and its dynamic nature.

Tens of thousands of PPIs have been identified and collected in literature-based PPI databases (Orchard et al., 2014; Chatr-Aryamontri et al., 2015). The accuracy of the data stored in these databases has been debated however (Sprinzak et al., 2003). As shown in Section 4.3.1, these databases may also be vulnerable to study bias, which may influence the performance of the methods that rely on these data.

Large-scale HT screens offer the most promising way of generating unbiased maps of the interactome. Rolland et al. (2014) generated the largest HT screen of human PPIs currently available. This screen searched less than half of the PPI ‘search space’ however, meaning that many other possible interactions still need to be tested. Attempts are underway to finish the testing of all possible PPIs and complete a draft version of the human interactome (Rolland et al., 2014). Completion of this draft may allow us to answer questions about how the position of proteins in the interactome relates to their function and any disease associations. As demonstrated in Section 4.3.1, it is currently difficult to determine whether the protein products of disease-associated genes are involved in greater numbers of interactions than genes that are not disease-associated. A complete map of the interactome may allow us to answer this question. Furthermore, a complete and unbiased map of the interactome would allow us to more accurately assess the performance of interactome-based gene prioritisation methods.

The availability of a draft version of the human interactome will not however provide us with a complete understanding of how the interactome differs between biological conditions, such as tissues, cell types and developmental stages. As discussed in Section 1.3.3 and demonstrated in Chapter 3, data can be integrated to study this dynamic nature of the interactome. Dynamic maps of the interactome would also aid us in understanding the condition-specific functioning of genes. In yeast, SGA technology has already been used to build networks of functional interactions between genes in cells exposed to different conditions (Bandyopadhyay et al., 2010; Srivas et al., 2013). Using these networks, it is possible to study how the functional context of these networks changes to allow cells to respond to the conditions to which they are exposed (Cornish & Markowitz, 2014). New technologies will need to be developed before it is possible to map similar networks in more complex organisms. Data integration methods currently represent the most promising way of studying the dynamic nature of the human interactome.

5.1.2 Understanding the context-specific effects of disease genes

One of the limitations of ALPACA is that it is currently unable to identify contexts associated with the majority of diseases represented in the GWASdb database. As demonstrated in Section 4.3.8, the performance of ALPACA improves when a trait-associated context can be identified by the GSO method. It therefore follows that the overall performance of ALPACA may improve if additional trait-associated contexts could be identified.

ALPACA could fail to identify a trait-associated context for two reasons: 1) a truly trait-associated context does not exist in the set of 73 tested contexts and 2) a truly trait-associated context exists in the set of 73 tested contexts but the GSO method fails to identify it.

The first of these two options is likely to be true for many traits. It is estimated that there are at least 400 different cell types present in the human body (Meehan et al., 2011), although new rarer cell types are continuing to be identified (Lyubimova et al., 2015). Only a minority of the cell types present in the human body are therefore represented in the data set used by ALPACA. A lack of high quality gene expression data for many cell types currently makes it difficult to increase the number of cell types that can be considered by ALPACA.

Improvements in transcriptomic technologies will increase the number of cell types for which high quality gene expression data are available. To produce gene expression data for a cell type, it has previously been necessary to isolate and purify tens of thousands of cells of that type (Kanamori-Katayama & Itoh, 2011). Many cell types are difficult to extract in this quantity, limiting the number of cell types that could be profiled. Advances in single-cell transcriptomics have allowed for the profiling of individual cells (Klein et al., 2015; Macosko et al., 2015). These methods will therefore increase the number of cell types for which high quality gene expression data are available.

Single-cell transcriptomics will also allow for the profiling of gene expression in cells at specific developmental stages or under certain conditions. Seumois et al. (2014) demonstrated that some genetic lesions may influence cell types at certain stages of their differentiation, leading to the development of specific diseases. There

are currently limited gene expression data for specific cell types at multiple stages of differentiation and it is therefore not possible to determine whether the performance of ALPACA could be improved by being able to consider cell types at different stages of differentiation. Similarly, single-cell transcriptomics may make it possible to profile cell types during foetal development. Congenital diseases develop before birth and it may therefore not be possible to identify contexts associated with these diseases if only gene expression data from adult cells are used.

Some traits may also not be associated with any single cell type. The genetic lesions that cause the disease may instead affect multiple cell types or entire tissues. Additional disease-context associations may therefore be identified if ALPACA was able to consider sets of cell types or whole tissues. It has been demonstrated that tissue-specific networks can be successfully used to prioritise disease genes (Magger et al., 2012) and therefore it seems likely that the performance of ALPACA could be improved by considering both individual cell types and whole tissues.

The ability to identify disease-associated contexts may also be improved with the identification of additional disease-associated genes. As shown in Section 3.3.4, as the number of known disease-associated genes increases, so does the ability of the GSO method to identify disease-associated contexts. As the number of known disease-associated genes continues to increase, it is therefore likely that the ability of ALPACA to identify disease-associated contexts will improve.

As described in Section 1.5, disease-associated contexts have been identified using text mining, gene expression data, protein abundance data, PPI data and epigenetic data. Expanding ALPACA so that it can use additional data sources may also improve performance. Protein abundance data are currently available for only a few cell types (Kim et al., 2014; Wilhelm et al., 2014). As more of these data become available they could also be used to identify disease-associated contexts and generate context-specific networks (Barshir et al., 2014; Kotlyar et al., 2016). The results of Seumois et al. (2014) and Farh et al. (2015) suggest that epigenetic data can be used to identify disease-context associations that cannot be identified using gene expression data alone. Methods should therefore be developed to integrate gene expression and epigenetic data to more accurately identify disease-associated

contexts.

5.1.3 Identifying causal genes using trans-acting regulatory elements

One of the limitations of ALPACA is its inability to detect genes that may be affected by causal variants via *trans*-acting regulatory mechanisms. It has been estimated that as many as 88% of trait-associated SNPs identified in GWAS are located in non-coding regions (Hindorff et al., 2009). Many of these SNPs may influence the trait they are associated with by affecting the regulation of one or more genes. When identifying genes in trait-associated loci, ALPACA extends gene boundaries in order to incorporate *cis*-acting regulatory elements. Incorporating *trans*-acting elements is more difficult however, due to the scarcity of data linking these elements to the genes they influence.

A couple of methods have been developed to identify *cis* and *trans*-acting regulatory elements. eQTL are regions of the genome containing variants that affect the expression of one or more genes. They can be mapped using approaches analogous to linkage and association analysis (Albert & Kruglyak, 2015), but with the expression of a gene as the trait of interest. Whilst eQTL have begun to be integrated with the results of GWAS to locate causal variants (He et al., 2013; Chung et al., 2014), the relatively small number of *trans*-acting eQTL identified so far (Westra et al., 2013) prohibits the large-scale use of eQTL data to link trait-associated loci to the genes they may influence across large distances.

Chromatin conformation capture technologies (such as 3C, 4C, 5C and Hi-C) use cross-linking of DNA and proteins, followed by sequencing, to identify regions of chromatin that are in close contact with each other (Dekker et al., 2013). This approach represents an alternative method of linking trait-associated SNPs in regulatory regions to the genes that they may affect. The development of the capture Hi-C (cHi-C) methodology has facilitated the targeted identification of interactions between disease-associated loci and various regulatory elements. Jäger et al. (2015) used cHi-C to search for interactions between 14 colorectal cancer risk loci and other genomic regions to study how these loci may affect the regulation of gene

expression. They identified interactions between these loci and genomic regions tens of megabases away, indicating possible *trans*-acting regulatory mechanisms. While further work needs to be completed to determine whether these interactions are functional, ALPACA would be unable to identify these genes as being affected by the causal variants in these loci.

As the quality and quantity of genome-scale regulatory data improves, it will be important to incorporate these data into gene prioritisation methods to identify the genes affected by causal variants. Mifsud et al. (2015) used cHi-C to generate the largest map of physical interactions involving gene promoters available to date. They generated maps in two different human blood cell types, identifying more than 1.6 million significant interactions. By comparing the accessibility of chromatin interacting with the promoters of highly and lowly expressed genes, Mifsud et al. were able to demonstrate that at least some of the interactions identified are functional. Furthermore, they demonstrated that the promoter-interacting regions were enriched with disease-associated SNPs, suggesting that this and similar data will be useful in interpreting disease-associated SNPs in the future.

5.2 Future work

5.2.1 Using data from additional species

While ALPACA currently uses genotype and phenotype data from only humans and mice, it could in the future be extended to use data from other organisms. Currently, databases that link genotype to phenotype exist for only a few species. Many of the species-specific databases that do exist, such as the RGD (Shimoyama et al., 2015) and FlyBase (dos Santos et al., 2015), report phenotypes using a mixture of free-text and terms from controlled vocabularies, making data integration difficult. Some databases, such as the MGD, now report all phenotypes using only terms from a small number of controlled vocabularies (Bult et al., 2016). It was for this reason that I was able to use data from the MGD in ALPACA. As the number of databases taking this approach increases, the integration of data from additional species will become feasible.

Improvements in cross-species phenotype vocabularies and ontologies would also make it easier to integrate data from additional species. Uberpheno (which is used in ALPACA) contains terms for only three species: humans, mice and zebrafish (Köhler et al., 2013). There currently exists a large number of disparate ontologies describing relationships between various biological entities, including cell types (Meehan et al., 2011), diseases (Kibbe et al., 2015) and gene functions (The Gene Ontology Consortium, 2014). Over recent years there has been a concerted effort by organisations such as the OBO Foundry to increase the interoperability of these and other ontologies (Smith et al., 2007). This is being done by ensuring that each ontology defines each of its terms using terms from even more elementary ontologies. These elementary ontologies include the Phenotypic Quality Ontology (PATO), which describes phenotypes and traits in a species-agnostic manner (Gkoutos et al., 2005). Once terms in multiple ontologies have been defined using terms from these elementary ontologies, automated reasoning can be used to identify relationships between terms across ontologies, facilitating the integration of the ontologies. It is this work by the OBO that led to the development of Uberpheno. As a greater number of ontologies adopt these standards, ALPACA could be extended to incorporate the ontologies, allowing it to use data from additional species.

5.2.2 Applying ALPACA to different network types

As described in Section 1.3.3, multiple methods have been used to generate context-specific networks. These methods include integrating PPI, gene expression and protein abundance data to remove or reweight interactions more or less likely to occur in given contexts (Bossi & Lehner, 2009; Lee et al., 2009; Lopes et al., 2011; Magger et al., 2012; Schaefer et al., 2013; Barshir et al., 2014; Liu et al., 2014b; Kotlyar et al., 2016) and using a Bayesian framework to integrate multiple sources of data and identify pairs of genes that are likely to share a function (Guan et al., 2012; Greene et al., 2015).

Removing proteins from a PPI network limits the use of the network in gene prioritisation, as it reduces the number of genes represented in the network. For this reason, I decided not to apply ALPACA to networks generated using this vertex

removal method. It would however be useful to know how ALPACA performs when run using functional interaction networks generated using a Bayesian framework. These networks can be generated to cover a greater number of genes than many PPI networks (Greene et al., 2015) and they are therefore ideal for gene prioritisation. Greene et al. (2015) make their 144 context-specific functional interaction networks available to download. I was however unable to use these networks in ALPACA as many of the cell types represented in the FANTOM5 data set are not represented in the 144 contexts considered by Greene et al.. Furthermore, many of the 144 contexts considered by Greene et al. are not represented in the FANTOM5 data set. This means that it would not be possible to use the GSO method and the FANTOM5 gene expression data set to identify which of the 144 networks should be used by ALPACA to prioritise genes. To assess the performance of ALPACA run using functional interactions networks generated using Bayesian integration, it may be necessary to generate new functional interaction networks for each of the cell types in the FANTOM5 data set. Determining which network generation method produces networks most suited to gene prioritisation will be important in the development of future gene prioritisation methods.

5.2.3 Prioritising disease variants

ALPACA was developed to prioritise genes in disease-associated loci. The method could however be extended to prioritise disease variants. ExomeWalker (Smedley et al., 2014), eXtasy (Sifrim et al., 2013), PhenIX (Zemojtel et al., 2014) and SuSPectP (Yates et al., under review) all combine gene-level prioritisation with variant analysis, completed using methods such as PolyPhen-2 (Adzhubei et al., 2010) and SuSPect (Yates et al., 2014), to identify variants that are most likely to be pathogenic. Integration of gene-level prioritisation with variant analysis produces significant improvements in performance (Smedley et al., 2014; Zemojtel et al., 2014). The gene-level scores produced by ALPACA could also be combined with variant-level scores to prioritise disease variants. SuSPectP uses PRINCE, and ExomeWalker uses an approach similar to PRINCE, to generate the gene-level scores. These methods may therefore be biased towards variants in better-

studied genes. The use of ALPACA to generate the gene-level scores may avoid the introduction of this bias.

As the costs of genome sequencing continue to decline (Hayden, 2014) whole-exome and whole-genome sequencing will be used more frequently to identify both common and rare disease variants. The integration of unbiased gene prioritisation methods, such as ALPACA, with variant analysis methods may aid in the identification of new disease variants.

5.2.4 Making ALPACA available for use

In order for a gene prioritisation method to be used, it needs to be made available to the wider scientific community. Widely adopted methods tend to be made available either as a downloadable application (Vanunu et al., 2010; Pers et al., 2015; Tasan et al., 2015) or through a web server (Raychaudhuri et al., 2009; Rossin et al., 2011; Smedley et al., 2014), although some of the most widely adopted methods are available in both formats (Tranchevent et al., 2008). Both formats have advantages: if a method is available to download then it may be easier for a user to use their own data, whilst having a method available through a web server may make it easier to use, especially if the user is not familiar with command line utilities. Making methods available as downloadable applications is quicker and I therefore plan to make ALPACA available to download in the near future, possibly through a bioinformatics software repository such as Bioconductor (Gentleman et al., 2004). If ALPACA is to gain wider adoption however, it may be advantageous to also make it available for use through a web server.

Chapter 6

Conclusions

In this thesis I have generated one of the largest collections of cell-type-specific PPI networks currently available, by integrating PPI data from STRING (Szklarczyk et al., 2015) and gene expression data from the FANTOM5 project (The FANTOM Consortium, 2014). I have demonstrated that sets of disease-associated genes cluster more significantly in PPI networks specific to cell types related to the disease. This observation suggested that these networks may be better suited to prioritising disease-associated genes. For this reason, I developed ALPACA, which uses cell-type-specific PPI networks to identify genes in disease-associated loci most likely to be causal. In cross-validation, ALPACA performs better when run using PPI networks specific to contexts identified by the GSO method over generic PPI networks. This suggests that data specific to contexts identified by the GSO method may be more useful for studying certain diseases.

Tens of thousands of interactions have been identified as occurring between proteins (Szklarczyk et al., 2015). We are however only just beginning to develop an understanding of how the interactome varies between cell types and tissues and how these differences relate to the distinct functions that genes perform in these different contexts. In the near future, it will not be feasible to use HT screens to map the interactome in multiple contexts. It is therefore necessary to use data integration methods, such as those described in this thesis, to generate networks specific to certain contexts. While I have demonstrated that using context-specific PPI networks over generic PPI networks improves the performance of ALPACA,

there is still a lot more room for improvement. Using additional data and more sophisticated integration methods may improve the performance of ALPACA and similar methods and allow for the generation of context-specific networks better suited to studying disease.

ALPACA uses data from both humans and mice and is the first network-based gene prioritisation method that uses data from multiple species to score genes. Ontology-based approaches represent one of the most promising ways of comparing the similarity of sets of phenotypes and therefore ALPACA was developed to use an ontology-based approach to identify genes associated with sets of phenotypes similar to the phenotypes associated with the trait of interest. Systematic integration of phenotype data from multiple species requires that these data are recorded and made available in standardised and computer-readable forms. Databases such as ClinVar, OMIM, UniProtKB/Swiss-Prot and the MGD provide data in such a format and the Uberpheno cross-species ontology allows these data to be integrated. The success of ALPACA demonstrates that by using a cross-species ontology, genotype and phenotype data from multiple species can be used with a network-based approach to prioritise disease genes.

Methods such as linkage and association analysis have identified thousands of loci associated with various human traits (Landrum et al., 2016; Li et al., 2016). The next step is to understand how these loci and the genes and variants they contain relate to their associated traits. The GSC and GSO methods described in this thesis offer a way of predicting the cell types affected by sets of disease-associated genes. The methods identify both well characterised associations and associations that warrant further study. ALPACA uses the predictions made by the GSO method to help identify which genes in these trait-associated loci are most likely to be causal.

Bibliography

- Adzhubei I. A., Schmidt S., Peshkin L., Ramensky V. E., Gerasimova A., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.
- Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., et al. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.
- Albert F. W., & Kruglyak L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212.
- Altshuler D., Daly M. J., & Lander E. S. (2008). Genetic Mapping in Human Disease. *Science*, 322(5903):881–888.
- Amberger J., Bocchini C., & Hamosh A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Human Mutation*, 32(5):564–567.
- Amberger J. S., Bocchini C. A., Schiettecatte F., Scott A. F., & Hamosh A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43:789–798.
- Anders S., & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.
- Andersson R., Gebhard C., Miguel-Escalada I., Hoof I., Bornholdt J., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.
- Aref-Eshghi E., Zhang Y., Liu M., Harper P. E., Martin G., et al. (2015). Genome-wide DNA methylation study of hip and knee cartilage reveals embryonic organ

- and skeletal system morphogenesis as major pathways involved in osteoarthritis. *BMC Musculoskeletal Disorders*, 16(1):287.
- Bader G. D., Betel D., & Hogue C. W. V. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250.
- Bandyopadhyay S., Mehta M., Kuo D., Sung M.-K., Chuang R., et al. (2010). Rewiring of genetic networks in response to DNA damage. *Science*, 330:1385–1390.
- Barabási A.-L., Gulbahce N., & Loscalzo J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.
- Baranzini S. E., Galwey N. W., Wang J., Khankhanian P., Lindberg R., et al. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human Molecular Genetics*, 18(11):2078–2090.
- Barshir R., Shwartz O., Smoly I. Y., & Yeger-Lotem E. (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLOS Computational Biology*, 10(6):e1003632.
- Bauer-Mehren A., Bundschuh M., Rautschka M., Mayer M. A., Sanz F., et al. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLOS One*, 6(6):e20284.
- Becker K. G., White S. L., Muller J., & Engel J. (2000). BBID: the biological biochemical image database. *Bioinformatics*, 16(8):745–746.
- Benjamini Y., & Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.
- Bernstein B. E., Stamatoyannopoulos J. A., Costello J. F., Ren B., Milosavljevic A., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048.

- Blake J. A., Bult C. J., Kadin J. A., Richardson J. E., & Eppig J. T. (2011). The mouse genome database (MGD): Premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Research*, 39:D842–D848.
- Bodenreider O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270.
- Börnigen D., Pers T. H., Thorrez L., Huttenhower C., Moreau Y., et al. (2013). Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic Acids Research*, 41(18):e171.
- Bossi A., & Lehner B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5:260.
- Botstein D., White R. L., Skolnick M., & Davis R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331.
- Bradding P., Walls A. F., & Holgate S. T. (2006). The role of the mast cell in the pathophysiology of asthma. *Journal of Allergy and Clinical Immunology*, 117(6):1277–1284.
- Braun P., Tasan M., Dreze M., Barrios-Rodiles M., Lemmens I., et al. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods*, 6(1):91–97.
- Brin S., & Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117.
- Bromberg Y. (2013). Chapter 15: Disease Gene Prioritization. *PLOS Computational Biology*, 9(4):e1002902.
- Brown S. D. M., & Moore M. W. (2012). The International Mouse Phenotyping Consortium: Past and future perspectives on mouse phenotyping. *Mammalian Genome*, 23(9):632–640.

- Buchanan C. C., Torstenson E. S., Bush W. S., & Ritchie M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association*, 19:289–294.
- Bullard J. H., Purdom E., Hansen K. D., & Dudoit S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94.
- Bult C. J., Eppig J. T., Blake J. A., Kadin J. A., Richardson J. E., et al. (2016). Mouse genome database 2016. *Nucleic Acids Research*, 44:D840–D847.
- Burney R. O., & Giudice L. C. (2012). Pathogenesis and pathophysiology of endometriosis. *Fertility and Sterility*, 98(3):511–519.
- Bush W. S., & Moore J. H. (2012). Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):e1002822.
- Bushell K. M., Söllner C., Schuster-Boeckler B., Bateman A., & Wright G. J. (2008). Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Research*, 18(4):622–630.
- Califano A., Butte A. J., Friend S., Ideker T., & Schadt E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genetics*, 44(8):841–847.
- Carniol K., Lahav G., Suel G., & Troyanskaya O. (2016). On a Quest for Principles, Big Data in Hand. *Cell*, 165(5):1038–1040.
- Chatr-Aryamontri A., Breitkreutz B.-J., Oughtred R., Boucher L., Heinicke S., et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43:D470–D478.
- Chen C. K., Mungall C. J., Gkoutos G. V., Doelken S. C., Kohler S., et al. (2012). Mousefinder: Candidate disease genes from mouse phenotype data. *Human Mutation*, 33(5):858–866.

- Cheung W. A., Ouellette B. F., & Wasserman W. W. (2012). Inferring novel gene-disease associations using Medical Subject Heading Over-representation Profiles. *Genome Medicine*, 4(9):75.
- Chung D., Yang C., Li C., Gelernter J., & Zhao H. (2014). GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLOS Genetics*, 10(11):e1004787.
- Cormen T. H., Leiserson C. E., Rivest R. L., & Stein C. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 3rd edition, (2009).
- Cornish A. J., & Markowetz F. (2014). SANTA: Quantifying the Functional Content of Molecular Networks. *PLOS Computational Biology*, 10(9):e1003808.
- Cornish A. J., Filippis I., David A., & Sternberg M. J. (2015). Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types. *Genome Medicine*, 7:95.
- Costanzo M., Baryshnikova A., Bellay J., Kim Y., Spear E. D., et al. (2010). The Genetic Landscape of a Cell. *Science*, 327(5964):425–431.
- Csardi G., & Nepusz T. (2006). The igraph software package for complex network research. *International Journal of Complex Systems in Science*, 1695(5):1–9.
- Dalbeth N., & Callan M. F. C. (2002). A subset of natural killer cells is greatly expanded within inflamed joints. *Arthritis and Rheumatism*, 46(7):1763–1772.
- Das J., & Yu H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92.
- de Nadal E., Ammerer G., & Posas F. (2011). Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12):833–845.
- Dekker J., Marti-Renom M. A., & Mirny L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403.

- DeLong E. R., DeLong D. M., & Clarke-Pearson D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845.
- Deo R. C., Musso G., Tasan M., Tang P., Poon A., et al. (2014). Prioritizing causal disease genes using unbiased genomic features. *Genome Biology*, 15(12):534.
- DIAGRAM Consortium, AGEN-T2D Consortium, SAT2D Consortium, MAT2D Consortium, & T2D-GENES Consortium. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–244.
- dos Santos G., Schroeder A. J., Goodman J. L., Strelets V. B., Crosby M. A., et al. (2015). FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43:D690–D697.
- Dunn O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64.
- Durinck S., Spellman P. T., Birney E., & Huber W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191.
- Endres D., & Schindelin J. (2003). A New Metric for Probability Distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.
- Erb K. J., & Le Gros G. (1996). The role of Th2 type CD4+ T cells and Th2 type CD8+ T cells in asthma. *Immunology and Cell Biology*, 74(2):206–208.
- Esposito P., Gheorghe D., Kandere K., Pang X., Connolly R., et al. (2001). Acute stress increases permeability of the blood-brain-barrier through activation of brain mast cells. *Brain Research*, 888(1):117–127.
- Famiglietti M. L., Estreicher A., Gos A., Bolleman J., Géhant S., et al. (2014). Genetic variations and diseases in UniProtKB/Swiss-Prot: The ins and outs of expert manual curation. *Human Mutation*, 35(8):927–935.

- Farh K. K.-H., Marson A., Zhu J., Kleinewietfeld M., Housley W. J., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.
- Fields S., & Song O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.
- Fogel L. A., Yokoyama W. M., & French A. R. (2013). Natural killer cells in human autoimmune disorders. *Arthritis Research & Therapy*, 15(4):216.
- Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41:D808–D815.
- Franke L., van Bakel H., Fokkens L., de Jong E. D., Egmont-Petersen M., et al. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics*, 78(6):1011–1025.
- Freudenberg J., & Propping P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18(2):S110–S115.
- Frostegard J. (2013). Immunity, atherosclerosis and cardiovascular disease. *BMC Medicine*, 11:117.
- Fruchterman T., & Reingold E. (1991). Graph Drawing by Force-directed Placement. *Software - Practice and Experience*, 21(11):1129–1164.
- Garzo I., Zhang Q. C., Petrey D., & Honig B. (2013). PrePPI : a structure-informed database of protein protein interactions. *Nucleic Acids Research*, 41:D828–D833.
- Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.
- Gerasimova A., Chavez L., Li B., Seumois G., Greenbaum J., et al. (2013). Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data. *PLOS One*, 8(1):e54359.

- Gillis J., & Pavlidis P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLOS Computational Biology*, 8(3):e1002444.
- Gilman S. R., Chang J., Xu B., Bawa T. S., Gogos J. A., et al. (2012). Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nature Neuroscience*, 15(12):1723–1728.
- Gkoutos G. V., Green E. C. J., Mallon A.-M., Hancock J. M., & Davidson D. (2005). Using ontologies to describe mouse phenotypes. *Genome Biology*, 6(1):R8.
- Glaab E., Baudot A., Krasnogor N., & Valencia A. (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics*, 11:597.
- Glyn-Jones S., Palmer A. J. R., Agricola R., Price A. J., Vincent T. L., et al. (2015). Osteoarthritis. *Lancet*, 6736(14):1–12.
- Goh K.-I., Cusick M. E., Valle D., Childs B., & Vidal M. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(201):8685–8690.
- Goldring M. B., & Goldring S. R. (2007). Osteoarthritis. *Journal of Cellular Physiology*, 213:626–634.
- Gonzalez M. W., & Kann M. G. (2012). Chapter 4: Protein interactions and disease. *PLOS Computational Biology*, 8(12):e1002819.
- Gottlieb A., Magger O., Berman I., Ruppin E., & Sharan R. (2011). PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics*, 27(23):3325–3326.
- Götz J., Probst A., Spillantini M. G., Schäfer T., Jakes R., et al. (1995). Somatodendritic localization and hyperphosphorylation of tau protein in transgenic mice expressing the longest human brain tau isoform. *The EMBO Journal*, 14(7):1304–1313.
- Green P., & Cellier C. (2007). Celiac disease. *The New England Journal of Medicine*, 357:1731–1743.

- Greene C. S., Krishnan A., Wong A. K., Ricciotti E., Zelaya R. A., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576.
- Guan Y., Gorenshiteyn D., Burmeister M., Wong A. K., Schimenti J. C., et al. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLOS Computational Biology*, 8(9):e1002694.
- Guruharsha K. G., Rual J.-F., Zhai B., Mintseris J., Vaidya P., et al. (2011). A protein complex network of *Drosophila melanogaster*. *Cell*, 147(3):690–703.
- Gusella J., Wexler N., Conneally P., Naylor S., Anderson M., et al. (1983). A polymorphic DNA marker genetically linked to Huntington’s disease. *Nature*, 306(17):234–238.
- Hall J. M., Lee M. K., Newman B., Morrow J. E., Anderson L. A., et al. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689.
- Hartwell L. H., Hopfield J. J., Leibler S., & Murray A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.
- Hayden E. C. (2014). The \$1,000 genome. *Nature*, 507:294–295.
- He X., Fuller C. K., Song Y., Meng Q., Zhang B., et al. (2013). Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *American Journal of Human Genetics*, 92(5):667–680.
- Heintzman N. D., Hon G. C., Hawkins R. D., Kheradpour P., Stark A., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.
- Hermjakob H., Montecchi-Palazzi L., Bader G., Wojcik J., Salwinski L., et al. (2004). The HUPO PSI’s molecular interaction format—a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183.

- Hidalgo C. A., Blumm N., Barabási A.-L., & Christakis N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, 5(4):e1000353.
- Himmelstein D. S., & Baranzini S. E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLOS Computational Biology*, 11(7):e1004259.
- Hindorff L. A., Sethupathy P., Junkins H. A., Ramos E. M., Mehta J. P., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–9367.
- Hoehndorf R., Schofield P. N., & Gkoutos G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119.
- Hoehndorf R., Schofield P. N., Gkoutos G. V., Division E., Street D., et al. (2015). Analysis of the human diseasome reveals phenotype modules across common, genetic, and infectious diseases. *Scientific Reports*, 5:10888.
- Hoehndorf R., Slater L., Schofield P. N., & Gkoutos G. V. (2015). Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16(26).
- Hu X., Kim H., Stahl E., Plenge R., Daly M., et al. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *American Journal of Human Genetics*, 89(4):496–506.
- Huang D. W., Sherman B. T., & Lempicki R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.
- Hyvärinen A., & Oja E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

- Ihaka R., & Gentleman R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.
- Jackson D., Somers K., & Harvey H. (1989). Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence? *The Americal Naturalist*, 133(3):436–453.
- Jacquemin T., & Jiang R. (2013). Walking on a tissue-specific disease-protein-complex heterogeneous network for the discovery of disease-related protein complexes. *BioMed Research International*, 2013:1–29.
- Jäger R., Migliorini G., Henrion M., Kandaswamy R., Speedy H. E., et al. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, 6:6178.
- Jensen A. B., Moseley P. L., Oprea T. I., Ellesøe S. G., Eriksson R., et al. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022.
- Jensen K., Panagiotou G., & Kouskoumvekaki I. (2014). Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. *PLOS Computational Biology*, 10(1):e1003432.
- Jensen L. J., Kuhn M., Stark M., Chaffron S., Creevey C., et al. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37:D412–D416.
- Jostins L., Ripke S., Weersma R. K., Duerr R. H., McGovern D. P., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124.
- Kanamori-Katayama M., & Itoh M. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research*, 21:1150–1159.
- Kanehisa M., & Goto S. (2000). KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

- Kerrien S., Aranda B., Breuza L., Bridge A., Broackes-Carter F., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40: D841–D846.
- Keshava Prasad T. S., Goel R., Kandasamy K., Keerthikumar S., Kumar S., et al. (2009). Human Protein Reference Database-2009 update. *Nucleic Acids Research*, 37:D767–D772.
- Kibbe W. A., Arze C., Felix V., Mitra E., Bolton E., et al. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, 43: D1071–D1078.
- Kim M.-S., Pinto S. M., Getnet D., Nirujogi R. S., Manda S. S., et al. (2014). A draft map of the human proteome. *Nature*, 509:575–581.
- Klein A., Mazutis L., Akartuna I., Tallapragada N., Veres A., et al. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.
- Köhler S., Bauer S., Horn D., & Robinson P. (2008). Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82:949–958.
- Köhler S., Schulz M. H., Krawitz P., Bauer S., Doelken S., et al. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, 85(4):457–464.
- Köhler S., Doelken S. C., Ruef B. J., Bauer S., Washington N., et al. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, 2:30.
- Köhler S., Doelken S. C., Mungall C. J., Bauer S., Firth H. V., et al. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42:D966–D974.

- Korbel J. O., Doerks T., Jensen L. J., Perez-Iratxeta C., Kaczanowski S., et al. (2005). Systematic association of genes to phenotypes by genome and literature mining. *PLOS Biology*, 3(5):e134.
- Kotlyar M., Pastrello C., Sheahan N., & Jurisica I. (2016). Integrated Interactions Database: Tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, 44:D536–D541.
- Krauthammer M., Kaufmann C. A., Gilliam T. C., & Rzhetsky A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15148–15153.
- Kulterer B., Friedl G., Jandrositz A., Sanchez-Cabo F., Prokesch A., et al. (2007). Gene expression profiling of human mesenchymal stem cells derived from bone marrow during expansion and osteoblast differentiation. *BMC Genomics*, 8(1):70.
- Lage K. (2014). Protein-protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta*, 1842(10):1971–1980.
- Lage K., Karlberg E. O., Størling Z. M., Olason P. I., Pedersen A. G., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316.
- Lage K., Hansen N. T., Karlberg E. O., Eklund A. C., Roque F. S., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20870–20875.
- Landrum M. J., Lee J. M., Benson M., Brown G., Chao C., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44:D862–D868.
- Lango Allen H., Estrada K., Lettre G., Berndt S. I., Weedon M. N., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

- Laresgoiti-Servitje E., Gómez-lópez N., & Olson D. M. (2010). An immunological insight into the origins of pre-eclampsia. *Human Reproduction Update*, 16(5): 510–524.
- Lee I., Lehner B., Crombie C., Wong W., Fraser A. G., et al. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nature Genetics*, 40(2):181–188.
- Lee I., Blom U. M., Wang P. I., Shim J. E., & Marcotte E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21:1109–1121.
- Lee K., Lee S., Park S., Kim S., Kim S., et al. (2016). BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. *Database*, pages 1–13.
- Lee S.-A., Chan C.-H., Chen T.-C., Yang C.-Y., Huang K.-C., et al. (2009). POINeT: protein interactome with sub-network analysis and hub prioritization. *BMC Bioinformatics*, 10:114.
- Leiserson M. D. M., Eldridge J. V., Ramachandran S., & Raphael B. J. (2013). Network analysis of GWAS data. *Current Opinion in Genetics & Development*, 23:602–610.
- Li M., Zhang J., Liu Q., Wang J., & Wu F. X. (2014). Prediction of disease-related genes based on weighted tissue-specific networks by using DNA methylation. *BMC Medical Genomics*, 7:S4.
- Li M. J., Liu Z., Wang P., Wong M. P., Nelson M. R., et al. (2016). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Research*, 44:D869–D876.
- Li Y., & Patra J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224.
- Li Z., Santangelo T., Cubonova L., Reeve J., & Kelman Z. (2010). Affinity

- purification of an archaeal DNA replication protein network. *MBio*, 1(5):e00221–10.
- Licata L., Briganti L., Peluso D., Perfetto L., Iannuccelli M., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40: D857–D861.
- Liu C.-C., Tseng Y.-T., Li W., Wu C.-Y., Mayzus I., et al. (2014). DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Research*, 42:W137–W146.
- Liu J. Z., McRae A. F., Nyholt D. R., Medland S. E., Wray N. R., et al. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics*, 87:139–145.
- Liu W., Wang J., Wang T., & Xie H. (2014). Construction and Analyses of Human Large-Scale Tissue Specific Networks. *PLOS One*, 9(12):e115074.
- Locke A. E., Kahali B., Berndt S. I., Justice A. E., Pers T. H., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518: 197–206.
- Loeser R. F., Goldring S. R., Scanzello C. R., & Goldring M. B. (2012). Osteoarthritis: A disease of the joint as an organ. *Arthritis and Rheumatism*, 64(6):1697–1707.
- Lopes T. J. S., Schaefer M., Shoemaker J., Matsuoka Y., Fontaine J.-F., et al. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27(17):2414–2421.
- Louis-Dit-Picard H., Barc J., Trujillano D., Miserey-Lenkei S., Bouatia-Naji N., et al. (2012). KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nature Genetics*, 44(4):456–460.
- Lui J. C., Nilsson O., Chan Y., Palmer C. D., Andrade A. C., et al. (2012). Synthesizing genome-wide association studies and expression microarray reveals

- novel genes that act in the human growth plate to modulate height. *Human Molecular Genetics*, 21(23):5193–5201.
- Lundby A., Rossin E. J., Steffensen A. B., Acha M. R., Newton-Cheh C., et al. (2014). Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. *Nature Methods*, 11(8):868–874.
- Lyubimova A., Kester L., Wiebrands K., Basak O., Sasaki N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.
- Macosko E., Basu A., Satija R., Nemesh J., Shekhar K., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.
- Macropol K., Can T., & Singh A. K. (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10:283.
- Magger O., Waldman Y. Y., Ruppin E., & Sharan R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLOS Computational Biology*, 8(9):e1002690.
- Maurano M. T., Humbert R., Rynes E., Thurman R. E., Haugen E., et al. (2012). Systematic Localization of Common Disease-Associate Variation in Regulatory DNA. *Science*, 337(6099):1190–1195.
- McInnes I. B., & Schett G. (2012). The Pathogenesis of Rheumatoid Arthritis. *The New England journal of Medicine*, 365(23):2205–2219.
- McVean G. A. T., Myers S. R., Hunt S., Deloukas P., Bentley D. R., et al. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584.
- Meehan T. F., Masci A. M., Abdulla A., Cowell L. G., Blake J. A., et al. (2011). Logical development of the cell ontology. *BMC Bioinformatics*, 12:6.

- Menche J., Sharma A., Kitsak M., Ghiassian S., Vidal M., et al. (2015). Uncovering disease-disease relationships through the human interactome. *Science*, 347(6224): 1257601.
- Mifsud B., Tavares-Cadete F., Young A. N., Sugar R., Schoenfelder S., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606.
- Minagar A., & Alexander J. S. (2003). Blood-brain barrier disruption in multiple sclerosis. *Multiple Sclerosis*, 9(6):540–549.
- Mishra G. R., Suresh M., Kumaran K., Kannabiran N., Suresh S., et al. (2006). Human protein reference database-2006 update. *Nucleic Acids Research*, 34:D411–D414.
- Monteleone G., Neurath M. F., Ardizzone S., Di Sabatino A., Fantini M. C., et al. (2015). Mongersen, an Oral SMAD7 Antisense Oligonucleotide, and Crohn's Disease. *New England Journal of Medicine*, 372(12):1104–1113.
- Moreau Y., & Tranchevent L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536.
- Muir B., & Nunney L. (2015). The expression of tumour suppressors and proto-oncogenes in tissues susceptible to their hereditary cancers. *British Journal of Cancer*, 113(2):345–353.
- Mungall C. J., Torniai C., Gkoutos G. V., Lewis S. E., & Haendel M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1): R5.
- Myers S., Bottolo L., Freeman C., McVean G., & Donnelly P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.
- Myouzen K., Kochi Y., Okada Y., Terao C., Suzuki A., et al. (2012). Functional

- Variants in NFKBIE and RTKN2 Involved in Activation of the NF-KB Pathway Are Associated with Rheumatoid Arthritis in Japanese. *PLOS Genetics*, 8(9).
- Navlakha S., & Kingsford C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063.
- NCBI Resource Coordinators. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41:D8–D20.
- NCBI Resource Coordinators. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 43:D6–D17.
- NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44:D7–D19.
- Nelson L. (2009). Primary Ovarian Insufficiency. *The New England Journal of Medicine*, 360(6):606–614.
- Nestle F. O., Kaplan D. H., & Barker J. (2009). Psoriasis. *The New England Journal of Medicine*, 361(5):496–509.
- Neves M., Damaschun A., Mah N., Lekschas F., Seltmann S., et al. (2013). Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database*, 2013:bat020.
- Nishimura D. (2001). A view from the Web, BioCarta. *Biotech Software & Internet Report*, 2(3):117–120.
- Oellrich A., Hoehndorf R., Gkoutos G. V., & Rebholz-Schuhmann D. (2012). Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PLOS One*, 7(6):e38937.
- Orange J. S. (2002). Human natural killer cell deficiencies and susceptibility to infection. *Microbes and Infection*, 4(15):1545–1558.
- Orchard S., Ammari M., Aranda B., Breuza L., Briganti L., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42:D358–D363.

- Oti M., Snel B., Huynen M. A., & Brunner H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, 43(8):691–698.
- Oti M., Ballouz S., & Wouters M. A. (2011). In Silico Tools for Gene Discovery. *Methods in Molecular Biology*, 760:189–206.
- Ott J., Wang J., & Leal S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16:275–284.
- Perez-Iratxeta C., Bork P., & Andrade M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31:316–319.
- Pers T. H., Dworzyński P., Thomas C. E., Lage K., & Brunak S. (2013). MetaRanker 2.0: a web server for prioritization of genetic variation data. *Nucleic Acids Research*, 41:W104–W108.
- Pers T. H., Karjalainen J., Chan Y., Westra H.-J., Wood A., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, 6:5890.
- Pesquita C., Faria D., Bastos H., Ferreira A. E. N., Falcão A. O., et al. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9:S4.
- Pesquita C., Faria D., Falcão A. O., Lord P., & Couto F. M. (2009). Semantic similarity in biomedical ontologies. *PLOS Computational Biology*, 5(7):e1000443.
- Peterson T. A., Doughty E., & Kann M. G. (2013). Towards precision medicine: Advances in computational approaches for the analysis of human variants. *Journal of Molecular Biology*, 425(21):4047–4063.
- Pinero J., Queralt-Rosinach N., Bravo A., Deu-Pons J., Bauer-Mehren A., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015:bav028.
- Piro R. M., Molineris I., Di Cunto F., Eils R., & König R. (2013). Disease-gene discovery by integration of 3D gene expression and transcription factor binding affinities. *Bioinformatics*, 29(4):468–475.

- Pletscher-Frankild S., Palleja A., Tsafou K., Binder J. X., & Jensen L. J. (2015). DISEASES: Text mining and data integration of disease-gene associations. *Methods*, 74:83–89.
- Podgaec S., Abrao M. S., Dias J. A., Rizzo L. V., de Oliveira R. M., et al. (2007). Endometriosis: An inflammatory disease with a Th2 immune response component. *Human Reproduction*, 22(5):1373–1379.
- Poux S., Magrane M., Arighi C. N., Bridge A., O'Donovan C., et al. (2014). Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database*, 2014:bau016.
- Powe C. E., Levine R. J., & Karumanchi S. A. (2011). Preeclampsia, a disease of the maternal endothelium: The role of antiangiogenic factors and implications for later cardiovascular disease. *Circulation*, 123(24):2856–2869.
- Pruim R. J., Welch R. P., Sanna S., Teslovich T. M., Chines P. S., et al. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337.
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.
- Qin Z., Niu T., & Liu J. (2002). Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms. *American Journal of Human Genetics*, 71:1241–1247.
- Raychaudhuri S., Plenge R. M., Rossin E. J., Ng A. C. Y., Purcell S. M., et al. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLOS Genetics*, 5(6):e1000534.
- Raychaudhuri S. K., Mavarakis E., & Raychaudhuri S. P. (2014). Diagnosis and classification of psoriasis. *Autoimmunity Reviews*, 13:490–495.
- Rigaut G., Shevchenko A., Rutz B., Wilm M., Mann M., et al. (1999). A generic

- protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17:1030–1032.
- Risma K., & Jordan M. B. (2012). Hemophagocytic lymphohistiocytosis: updates and evolving concepts. *Current Opinion in Pediatrics*, 24(1):9–15.
- Roberts J. M., Bodnar L. M., Patrick T. E., & Powers R. W. (2011). The Role of Obesity in Preeclampsia. *Pregnancy Hypertension*, 1(1):6–16.
- Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.
- Robinson M. D., McCarthy D. J., & Smyth G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.
- Robinson P., Köhler S., Oellrich A., Wang K., Mungall C., et al. (2014). Improved exome prioritization of disease genes through cross species phenotype comparison. *Genome Research*, 24(2):340–348.
- Robinson P. N., Köhler S., Bauer S., Seelow D., Horn D., et al. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615.
- Rolland T., Tas M., Sahni N., Yi S., Lemmens I., et al. (2014). A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159:1212–1226.
- Rossin E. J., Lage K., Raychaudhuri S., Xavier R. J., Tatar D., et al. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLOS Genetics*, 7(1):e1001273.
- Safran M., Dalah I., Alexander J., Rosen N., Iny Stein T., et al. (2010). GeneCards Version 3: the human gene integrator. *Database*, 2010:baq020.
- Salwinski L., Miller C. S., Smith A. J., Pettit F. K., Bowie J. U., et al. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32: D449–D451.

- Sanchez C., Lachaize C., Janody F., Bellon B., Röder L., et al. (1999). Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an internet database. *Nucleic Acids Research*, 27(1):89–94.
- Sardar A. J., Oates M. E., Fang H., Forrest A. R. R., Kawaji H., et al. (2014). The evolution of human cells in terms of protein innovation. *Molecular Biology and Evolution*, 31(6):1364–1374.
- Sawcer S., Hellenthal G., Pirinen M., Spencer C. C. A., Patsopoulos N. A., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219.
- Schaefer M. H., Lopes T. J. S., Mah N., Shoemaker J. E., Matsuoka Y., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLOS Computational Biology*, 9(1):e1002860.
- Schofield P. N., Gkoutos G. V., Gruenberger M., Sundberg J. P., & Hancock J. M. (2010). Phenotype ontologies for mouse and man: bridging the semantic gap. *Disease Models & Mechanisms*, 3:281–289.
- Schwanhäusser B., Busse D., Li N., Dittmar G., Schuchhardt J., et al. (2011). Global quantification of mammalian gene expression control. *Nature*, 473:337–342.
- Seok J., Warren H. S., Cuenca A. G., Mindrinos M. N., Baker H. V., et al. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 110(9):3507–3512.
- Seumois G., Chavez L., Gerasimova A., Lienhard M., Omran N., et al. (2014). Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nature Immunology*, 15(8):777–788.
- Sherry S. T., Ward M. H., Kholodov M., Baker J., Phan L., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

- Shimoyama M., De Pons J., Hayman G. T., Laulederkind S. J. F., Liu W., et al. (2015). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Research*, 43:D743–D750.
- Sifrim A., Popovic D., Tranchevent L. C., Ardeshirdavani A., Sakai R., et al. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, 10(11): 1083–1084.
- Sinaii N., Cleary S. D., Ballweg M. L., Nieman L. K., & Stratton P. (2002). High rates of autoimmune and endocrine disorders, fibromyalgia, chronic fatigue syndrome and atopic diseases among women with endometriosis: a survey analysis. *Human Reproduction*, 17(10):2715–2724.
- Slowikowski K., Hu X., & Raychaudhuri S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics*, 30(17):2496–2497.
- Smedley D., Köhler S., Czeschik J. C., Amberger J., Bocchini C., et al. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, 30(22):3215–3222.
- Smith B., Ashburner M., Rosse C., Bard J., Bug W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.
- Smith C. L., Goldsmith C.-A. W., & Eppig J. T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6:R7.
- Speir M. L., Zweig A. S., Rosenbloom K. R., Raney B. J., Paten B., et al. (2016). The UCSC Genome Browser database: 2016 update. *Nucleic Acids Research*, 44: D717–D725.
- Sprinzak E., Sattath S., & Margalit H. (2003). How Reliable are Experimental Protein Protein Interaction Data? *Journal of Molecular Biology*, 327(5):919–923.

- Srivas R., Costelloe T., Carvunis A.-R., Sarkar S., Malta E., et al. (2013). A UV-induced genetic network links the RSC complex to nucleotide excision repair and shows dose-dependent rewiring. *Cell Reports*, 5(6):1714–1724.
- Stearns M. Q., Price C., Spackman K. A., & Wang Y. A. (2001). SNOMED clinical terms: overview of the development process and project status. *Proceedings / AMIA Annual Symposium*, 1:662–666.
- Stenson P. D., Mort M., Ball E. V., Shaw K., Phillips A. D., et al. (2014). The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9.
- Stumpf M., & Thorne T. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6959–6964.
- Sturtevant A. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59.
- Su A. I., Wiltshire T., Batalov S., Lapp H., Ching K. A., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067.
- Swindell W. R., Stuart P. E., Sarkar M. K., Voorhees J. J., Elder J. T., et al. (2014). Cellular dissection of psoriasis for transcriptome analyses and the post-GWAS era. *BMC Medical Genomics*, 7:27.
- Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43:D447–D452.
- Tasan M., Musso G., Hao T., Vidal M., Macrae C. A., et al. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods*, 12(2):154–159.

- Teslovich T. M. T., Musunuru K., Smith A. A. V., Edmondson A. C., Stylianou I. M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713.
- The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.
- The CARDIoGRAMplusC4D Consortium. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, 45(1):25–33.
- The CARDIoGRAMplusC4D Consortium. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130.
- The ENCODE Project Consortium. (2011). A user’s guide to the Encyclopedia of DNA elements (ENCODE). *PLOS Biology*, 9(4):e1001046.
- The FANTOM Consortium. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.
- The Gene Ontology Consortium. (2014). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43:1049–1056.
- The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.
- Thurman R. E., Rynes E., Humbert R., Vierstra J., Maurano M. T., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.
- Toegel S., Weinmann D., André S., Walzer S. M., Bilban M., et al. (2016). Galectin-1 Couples Glycobiology to Inflammation in Osteoarthritis through the Activation of an NF- κ B-Regulated Gene Network. *Journal of Immunology*, 196(4):1910–1921.

- Tong A. H., Evangelista M., Parsons A. B., Xu H., Bader G. D., et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368.
- Tranchevent L.-C., Barriot R., Yu S., Van Vooren S., Van Loo P., et al. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, 36:W377–W384.
- Tranchevent L.-C., Capdevila F. B., Nitsch D., De Moor B., De Causmaecker P., et al. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32.
- Trynka G., Sandor C., Han B., Xu H., Stranger B. E., et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.
- Turner J. A. (2010). Diagnosis and management of pre-eclampsia: An update. *International Journal of Women’s Health*, 2(1):327–337.
- van Driel M. A., Bruggeman J., Vriend G., Brunner H. G., & Leunissen J. A. M. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5):535–542.
- Vanunu O., Magger O., Ruppin E., Shlomi T., & Sharan R. (2010). Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 6(1):e1000641.
- Velculescu V., Madden S., Zhang L., Last A., Yu J., et al. (1999). Analysis of human transcriptomes. *Nature Genetics*, 23(4):387–388.
- Vermersch P., Benrabah R., Schmidt N., Zéphir H., Clavelou P., et al. (2012). Masitinib treatment in patients with progressive multiple sclerosis: a randomized pilot study. *BMC Neurology*, 12:36.
- Veyrieras J.-B., Kudaravalli S., Kim S. Y., Dermitzakis E. T., Gilad Y., et al. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLOS Genetics*, 4(10):e1000214.

- Vionnet N., Stoffel M., Takeda J., Yasuda K., Bell G. I., et al. (1992). Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. *Nature*, 356(6371):721–722.
- von Mering C., Jensen L. J., Snel B., Hooper S. D., Krupp M., et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33:D433–D437.
- Wang J., Zhou X., Zhu J., & Guo Z. (2010). Bias of phenotype similarity scores between diseases. *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, 3:8–11.
- Wang Q., Rozelle A. L., Lepus C. M., Scanzello C. R., Song J. J., et al. (2011). Identification of a central role for complement in osteoarthritis. *Nature Medicine*, 17(12):1674–1679.
- Wang X., Gulbahce N., & Yu H. (2011). Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, 10(5):280–293.
- Warde-Farley D., Donaldson S. L., Comes O., Zuberi K., Badrawi R., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38:W214–W220.
- Washington N. L., Haendel M. A., Mungall C. J., Ashburner M., Westerfield M., et al. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLOS Biology*, 7(11):e1000247.
- Westra H.-J., Peters M. J., Esko T., Yaghootkar H., Schurmann C., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243.
- Whirl-Carrillo M., McDonagh E. M., Hebert J. M., Gong L., Sangkuhl K., et al. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, 92(4):414–417.

- White S., & Smyth P. (2003). Algorithms for estimating relative importance in networks. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 266–275.
- Wilhelm M., Schlegl J., Hahne H., Moghaddas Gholami A., Lieberenz M., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509 (7502):582–587.
- Withers D. J., Gutierrez J. S., Towery H., Burks D. J., Ren J. M., et al. (1998). Disruption of IRS-2 causes type 2 diabetes in mice. *Nature*, 391(6670):900–904.
- Wood A. R., Esko T., Yang J., Vedantam S., Pers T. H., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46:1173–1186.
- Wu D., Lim E., Vaillant F., Asselin-Labat M. L., Visvader J. E., et al. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182.
- Xie B., Ding Q., Han H., & Wu D. (2013). miRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*, 29 (5):638–644.
- Yates A., Akanni W., Amode M. R., Barrell D., Billis K., et al. (2016). Ensembl 2016. *Nucleic Acids Research*, 44:D710–D716.
- Yates C. M., & Sternberg M. J. E. (2013). Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). *Journal of Molecular Biology*, 425(8):1274–1286.
- Yates C. M., Filippis I., Kelley L. A., & Sternberg M. J. E. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology*, 426(14):2692–2701.
- Yip Y. L., Famiglietti M., Gos A., Duek P. D., David F. P. A., et al. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human Mutation*, 29(3):361–366.

- Yu H., Braun P., Yildirim M. A., Lemmens I., Venkatesan K., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.
- Zemojtel T., Köhler S., Mackenroth L., Jäger M., Hecht J., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine*, 6(252):1–10.
- Zhang B., Shi Z., Duncan D. T., Prodduturi N., Marnett L. J., et al. (2011). Relating protein adduction to gene expression changes: a systems approach. *Molecular Biosystems*, 7(7):2118–2127.
- Zhou X., Menche J., Barabási A.-L., & Sharma A. (2014). Human symptoms-disease network. *Nature Communications*, 5:4212.
- Zhou Y., Damsky C. H., & Fisher S. J. (1997). Preeclampsia is associated with failure of human cytotrophoblasts to mimic a vascular adhesion phenotype: One cause of defective endovascular invasion in this syndrome? *Journal of Clinical Investigation*, 99(9):2152–2164.

Appendix

Cell type	MeSH terms(s)
Acinar cell	Acinar cells
Alveolar epithelial cell	Pneumocytes
Amniotic epithelial cell	Epithelial cells/Amnion
Astrocyte	Astrocytes
Blood vessel endothelial cell	Endothelial cells/Blood vessels
Bronchial epithelial cell	Epithelial cells/Bronchi
Bronchial smooth muscle cell	Myocytes, smooth muscle/Bronchi
Cardiac fibroblast	Fibroblasts/Heart
Cardiac myocyte	Myocytes, cardiac
CD14+ monocyte derived progenitor cell	Monocytes
Chondrocyte	Chondrocytes
Ciliated epithelial cell	Epithelial cells
Corneal epithelial cell	Epithelial cells/Cornea
Dendritic cell - monocyte immature derived	Dendritic cells
Endothelial cell of hepatic sinusoid	Hepatocytes
Endothelial cell of lymphatic vessel	Endothelial cells/Lymphatic vessels
Enteric smooth muscle cell	Myocytes, smooth muscle/Colon
Epithelial cell of malassez	Epithelial cells/Jaw
Epithelial cell of prostate	Epithelial cells/Prostate
Fat cell	Adipocytes
Fibroblast of choroid plexus	Fibroblasts/Choroid plexus

Table continued on the next page.

Cell type	MeSH terms(s)
Fibroblast of gingiva	Fibroblasts/Gingiva
Fibroblast of lymphatic vessel	Fibroblasts/Lymphatic vessels
Fibroblast of periodontium	Fibroblasts/Periodontium
Fibroblast of tunica adventitia of artery	Fibroblasts/Adventitia
Gingival epithelial cell	Epithelial cells/Gingiva
Hair follicle cell	Mesenchymal stromal cells/Hair follicle
Hepatic stellate cell	Hepatic stellate cells
Immature Langerhans cell	Langerhans cells
Keratinocyte - epidermal	Keratinocytes
Keratocyte	Corneal keratocytes
Kidney epithelial cell	Epithelial cells/Kidney
Lens epithelial cell	Epithelial cells/Lenses
Lymphocyte of B lineage	B-lymphocytes
Macrophage	Macrophages
Mammary epithelial cell	Epithelial cells/Mammary glands, animal
Mast cell	Mast cells
Melanocyte - light	Melanocytes
Mesenchymal precursor cell	Mesenchymal stromal cells
Mesenchymal somatic cell	Mesenchymal stromal cells
Mesenchymal stem cell - adipose	Mesenchymal stromal cells/Adipose tissue
Mesenchymal stem cell - amniotic membrane	Mesenchymal stromal cells/Amnion
Mesenchymal stem cell - bone marrow	Bone marrow cells
Mesenchymal stem cell - hepatic	Mesenchymal stromal cells/Liver
Mesenchymal stem cell - umbilical	Mesenchymal stromal cells/Umbilical cord
Mesothelial cell	Epithelial cells
Migratory langerhans cell	Langerhans cells
Monocyte	Monocytes
Multipotent cord blood somatic stem cell	Mesenchymal stromal cells/Umbilical cord
Myoblast	Myoblasts

Table continued on the next page.

Cell type	MeSH terms(s)
Natural killer cell	Killer cells, natural
Neuron	Neurons
Neuronal stem cell	Neural stem cells
Neutrophil	Neutrophils
Osteoblast	Osteoblasts
Pericyte cell	Pericytes
Placental epithelial cell	Epithelial cells/Placenta
Preadipocyte	Fibroblasts/Adipose tissue
Reticulocyte	Reticulocytes
Retinal pigment epithelial cell	Epithelial cells/Retinal pigment epithelium
Sensory epithelial cell	Epithelial cells
Skeletal muscle cell	Muscle fibers, skeletal
Skin fibroblast	Fibroblasts/Skin
Small airway epithelial cell	Epithelial cells/Respiratory system
Smooth muscle cell of prostate	Myocytes, smooth muscle/Prostate
Smooth muscle cell of trachea	Myocytes, smooth muscle/Trachea
Stromal cell	Stromal cells
T cell	T-Lymphocytes
Tendon cell	Stromal cells/Tendons
Trabecular meshwork cell	Endothelial cells/Trabecular meshwork
Tracheal epithelial cell	Epithelial cells/Trachea
Urothelial cell	Epithelial cells/Urothelium
Vascular associated smooth muscle cell	Myocytes, smooth muscle/Muscle, smooth, vascular

Table A.1: The 73 cell types for which cell-type-specific PPI networks are generated in this thesis and each of the MeSH terms mapped to these cell types.

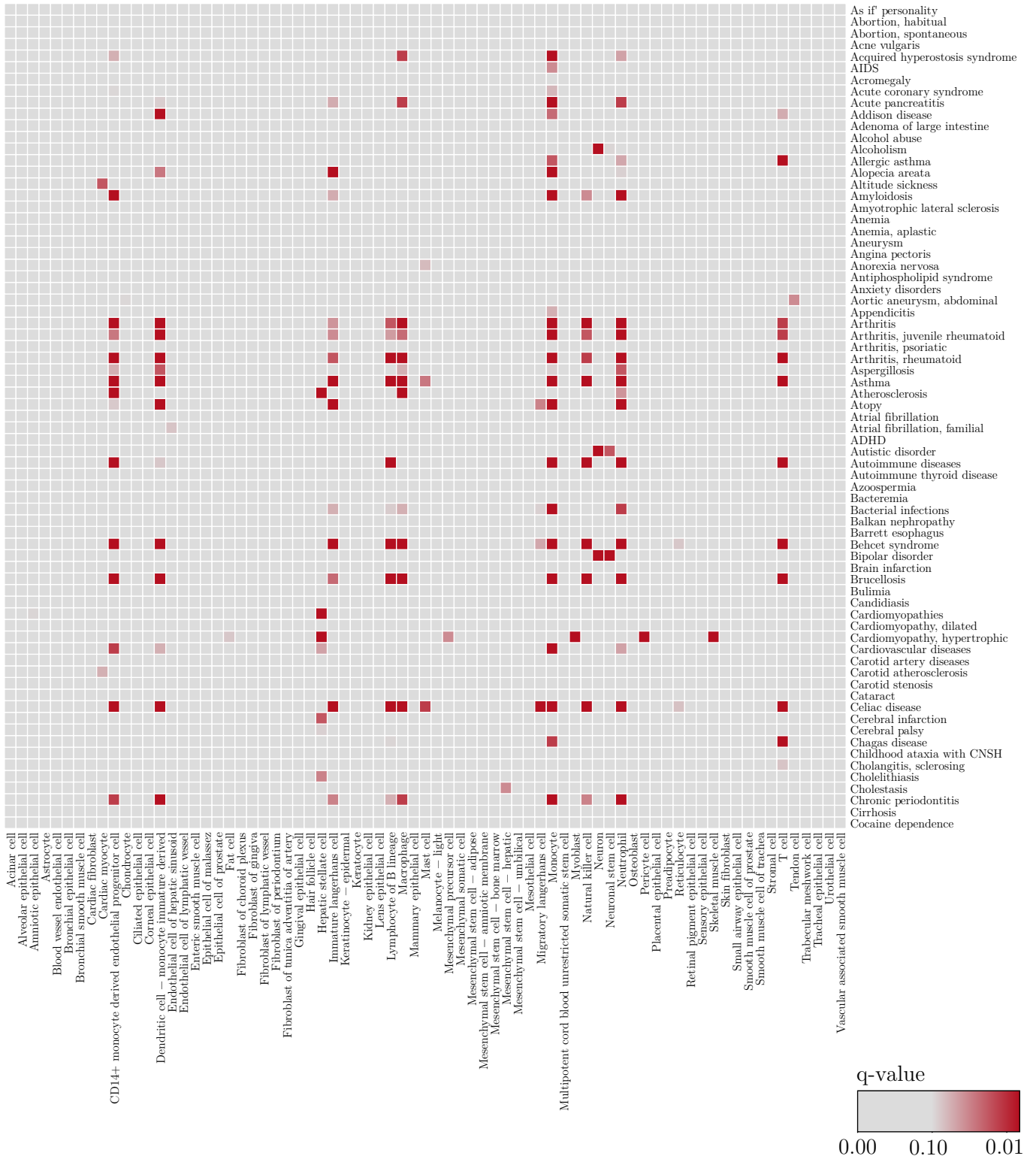


Figure A.1: Heat map of the disease-cell-type associations identified by the GSC method (Part 1). CNSH: central nervous system hypomyelination.

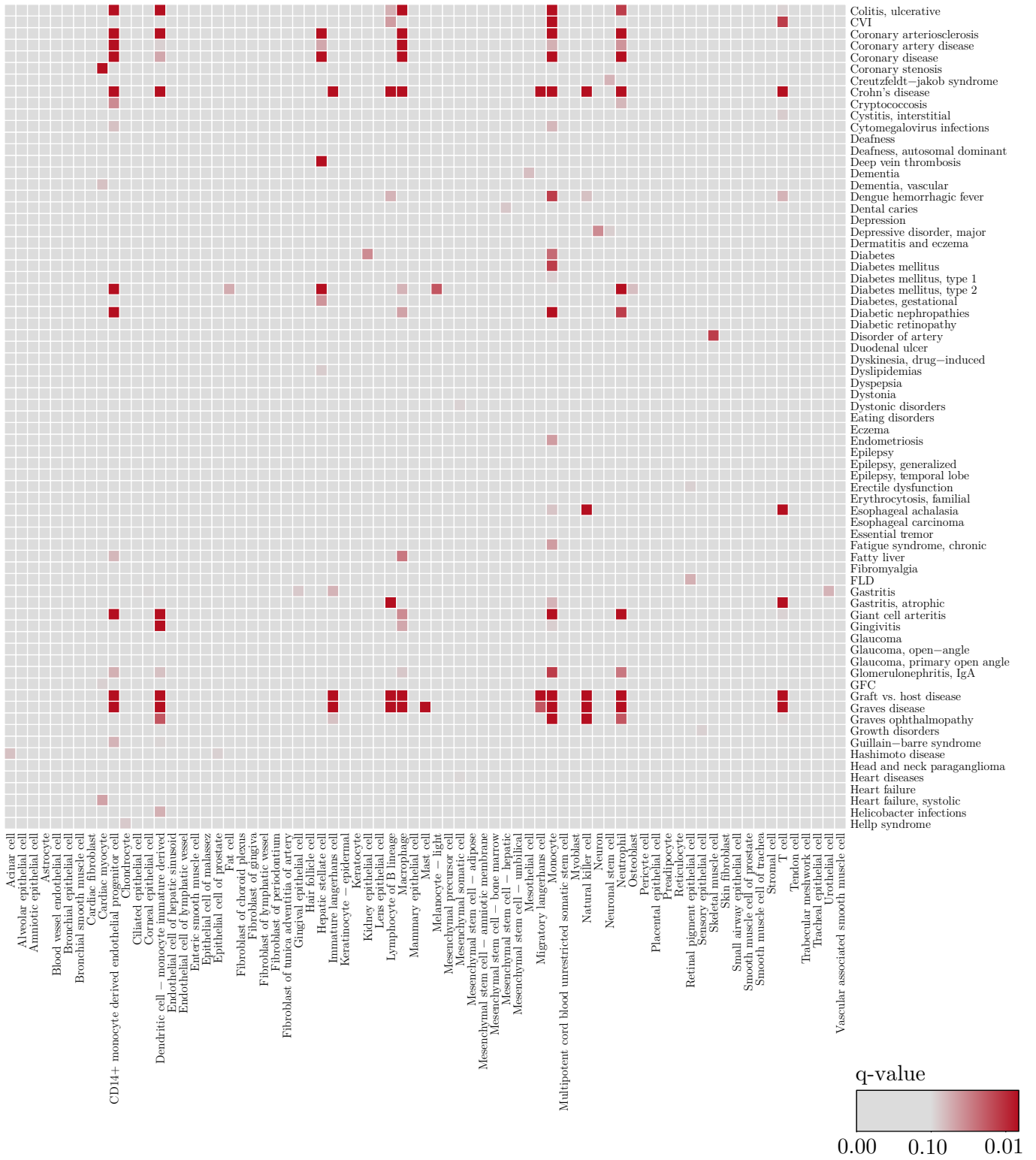


Figure A.1: Heat map of the disease-cell-type associations identified by the GSC method (Part 2). CVI: common variable immunodeficiency, FLD: frontotemporal lobar degeneration, GFC: glomerulosclerosis, focal segmental.

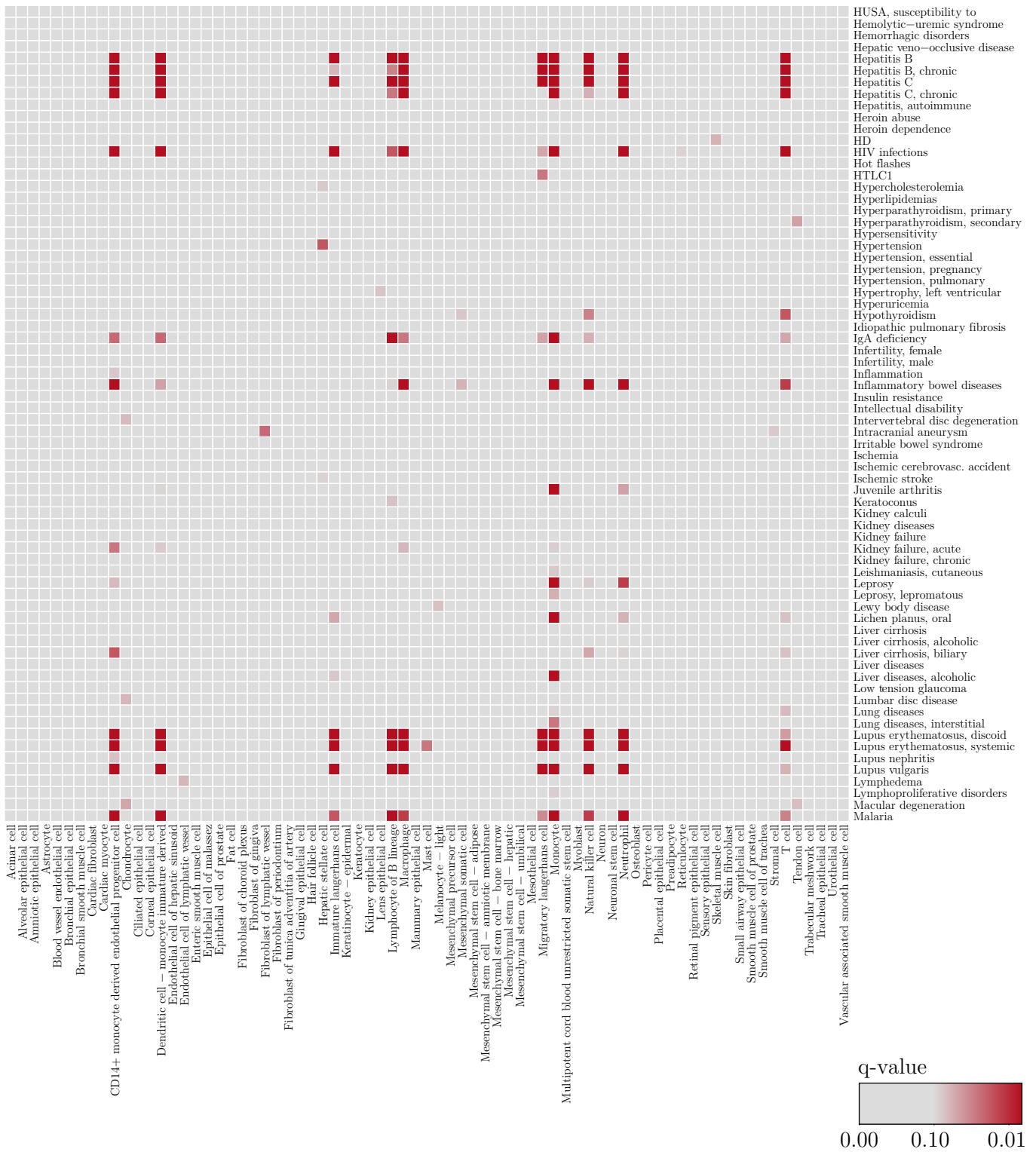


Figure A.1: Heat map of the disease-cell-type associations identified by the GSC method (Part 3). HUSA: hemolytic uremic syndrome, atypical, HD: Hirschsprung's disease, HTLC1: human T cell lymphotropic virus 1 infection.

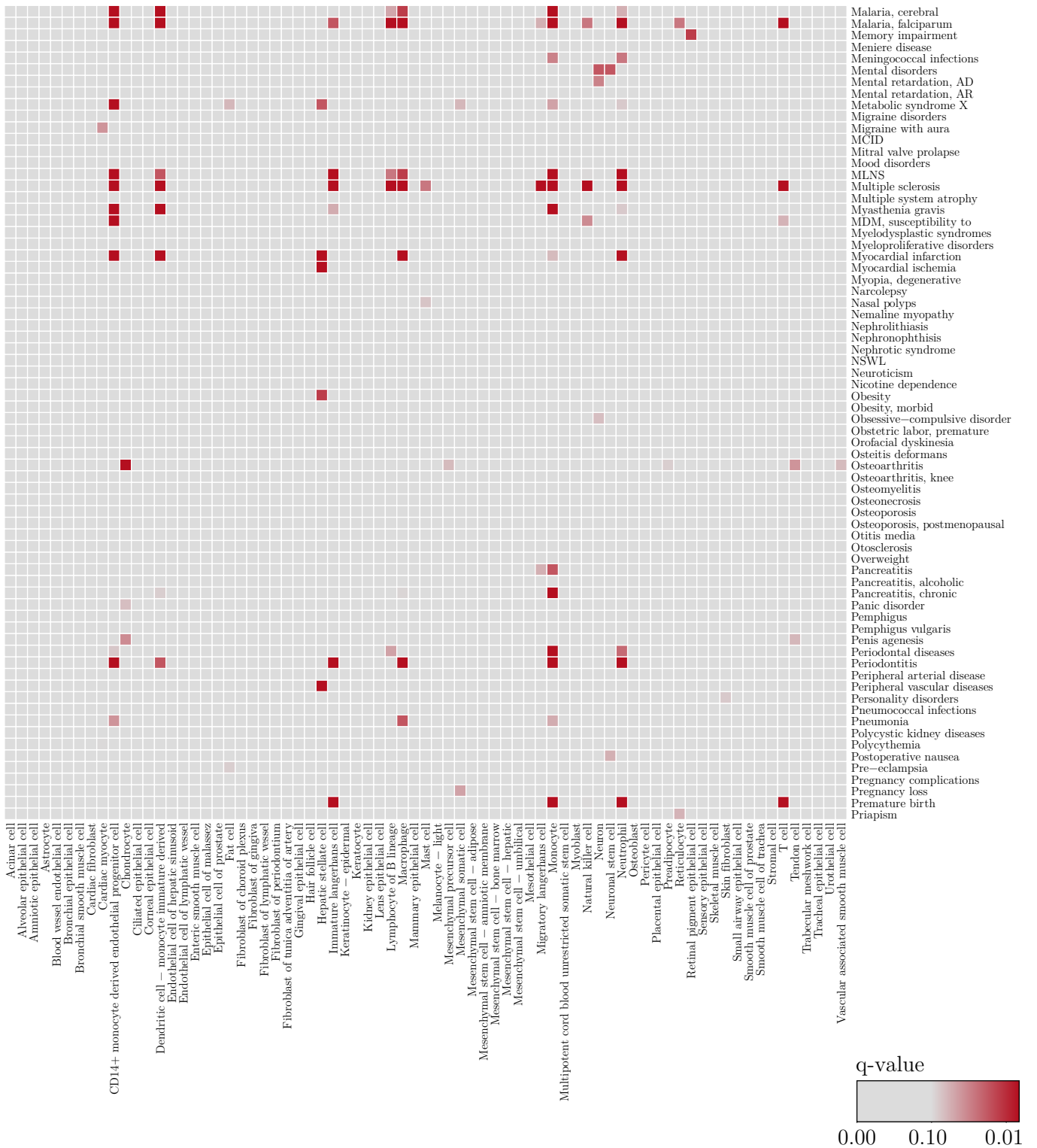


Figure A.1: Heat map of the disease-cell-type associations identified by the GSC method (Part 4). AD: autosomal dominant, AR: autosomal recessive, MCID: mitochondrial complex I deficiency, MLNS: mucocutaneous lymph node syndrome, MDM: mycobacterial disease, Mendelian, NSWL: nephrotic syndrome with lesion of minimal change glomerulonephritis.

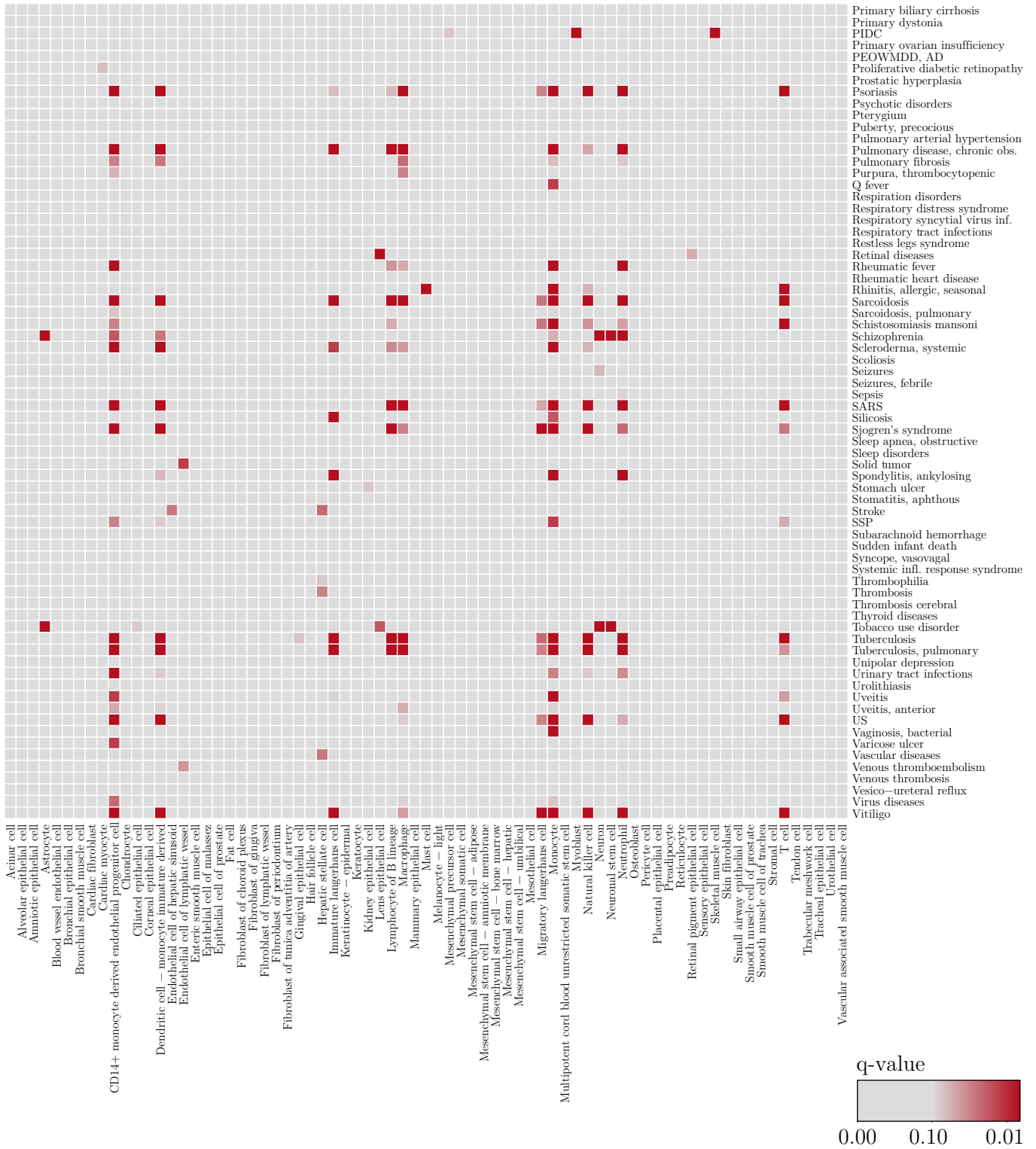


Figure A.1: Heat map of the disease-cell-type associations identified by the GSC method (Part 5). PIDC: primary idiopathic dilated cardiomyopathy, PEOWMDD: progressive external ophthalmoplegia with mitochondrial DNA deletions, SSP: subacute sclerosing panencephalitis, US: uveomeningoencephalitic syndrome.

Resource	Version	Date	URL
1KGP	Pilot	2015-03-30	ftp://ftp.1000genomes.ebi.ac.uk
BioGRID	3.4.131	2015-12-13	http://thebiogrid.org
CL	-	2014-07-08	http://obofoundry.org/ontology/cl
ClinVar	-	2015-03-03	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar
DisGeNET	2.1	2014-07-23	http://www.disgenet.org
Ensembl	75	2015-04-04	ftp://ftp.ensembl.org/pub/release-75
FANTOM5	-	2014-08-21	http://fantom.gsc.riken.jp
GWASdb	-	2015-10-25	http://jjwanglab.org/gwasdb
HapMap	II	2015-02-19	http://hapmap.ncbi.nlm.nih.gov
HI-II-14	-	2015-11-27	http://interactome.dfci.harvard.edu
Hoehndorf et al. scores	-	2015-11-11	http://aber-owl.net
HPO	-	2015-12-22	http://human-phenotype-ontology.github.io
IntAct	-	2016-01-04	ftp://ftp.ebi.ac.uk/pub/databases/intact
MGD	-	2015-11-14	ftp://ftp.informatics.jax.org/pub
OMIM	-	2015-05-19	http://www.omim.org
PrePPI	-	2016-01-04	https://bhapp.c2b2.columbia.edu/PrePPI
STRING	9.1	2015-03-09	http://string91.embl.de
STRING	10.0	2015-11-29	http://string10.embl.de
Uberon	-	2014-06-21	http://obofoundry.org/ontology/uberon
Uberpheno	-	2015-11-08	http://purl.obolibrary.org/obo/hp/uberpheno
UniProtKB/Swiss-Prot	-	2015-03-02	ftp://ftp.uniprot.org/pub/databases

Table A.2: Details of the resources downloaded and used in this thesis. Given is the name of the resource, the version downloaded (if available), the date on which I downloaded the resource and the URL of the resource.