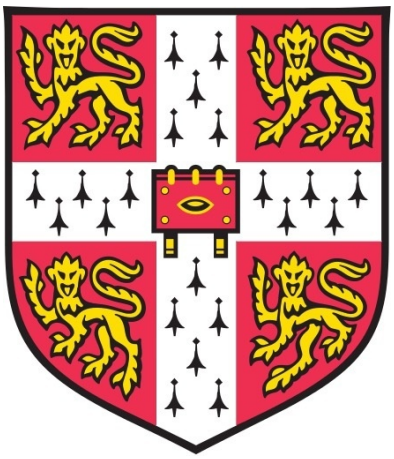
Understanding Disease and Disease Relationships Using Transcriptomic Data

Erin Oerton

St Catharine's College

September 2018



This dissertation is submitted for the degree of Doctor of Philosophy

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.
It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text
It does not exceed the prescribed word limit for the relevant Degree Committee.

Work done in collaboration:
Chapter 5: The text-mining for literature co-occurrence of diseases was carried out by Patrick Lewis, as specified in the text.

ACKNOWLEDGEMENTS

To my supervisor, Dr Andreas Bender, for his help and guidance over the course of my PhD.

To Dr Dezső Módos for his tireless enthusiasm, and for being a constant source of ideas and inspiration, particularly for the work described in Chapter 4.

To Dr Ian Roberts for being so generous with his time and expert advice, to Dr Tim Guilliams for his leadership and support, and to all at Healx for being a fantastic and welcoming team.

To Dr Avid Afzal for many helpful discussions, and to Ben Alexander-Dann for proof-reading and witty comments.

To members of the Bender group, past and present. Special thanks are due to Dr Avid Afzal, Ben Alexander-Dann, Stephanie Ashenden, Fatima Baldo, Dr Krishna Bulusu, Kathryn Giblin, and Dr Christoph Schlaffner; also to Susan Begg for keeping the CMI running.

To the BBSRC, for funding my PhD through the Doctoral Training Programme. To St Catharine's College and the Chemistry Department, for providing me with financial support to take up the fantastic opportunities that have arisen during my three years here.

To the crew at St Catharine's, most of all Isobel Everall for always finding the best places to write. To Fynn Krause, for never failing to brighten my day. To Dr Sarah Oerton, for her guidance and encouragement. Finally, to my mother, Janet Oerton, for supporting me always.

SUMMARY

As the volume of transcriptomic data continues to increase, so too does its potential to deepen our understanding of disease; for example, by revealing gene expression patterns shared between diseases. However, key questions remain around the strength of the transcriptomic signal of disease and the identification of meaningful commonalities between datasets, which are addressed in this thesis as follows.

The first chapter, *Concordance of Microarray Studies of Parkinson's Disease*, examines the agreement between differential expression signatures across 33 studies of Parkinson's disease. Comparison of these studies, which cover a range of microarray platforms, tissues, and disease models, reveals a characteristic pattern of differential expression in the most highly-affected tissues in human patients. Using correlation and clustering analyses to measure the representativeness of different study designs to human disease, the work described acts as a guideline for the comparison of microarray studies in the following chapters.

In the next chapter, *Using Dysregulated Signalling Paths to Understand Disease*, gene expression changes are linked on the human signalling network, enabling identification of network regions dysregulated in disease. Applying this method across a large dataset of 141 common and rare diseases identifies dysregulated processes shared between diverse conditions, which relate to known disease- and drug-sharing-relationships.

The final chapter, *Understanding and Predicting Disease Relationships Through Similarity Fusion*, explores the integration of gene expression with other data types – in this case, ontological, phenotypic, literature co-occurrence, genetic, and drug data – to understand relationships between diseases. A similarity fusion approach is proposed to overcome the differences in data type properties between each space, resulting in the identification of novel disease relationships spanning multiple bioinformatic levels. The similarity of disease relationships between each data type is considered, revealing that relationships in differential expression space are distinct from those in other molecular and clinical spaces.

In summary, the work described in this thesis sets out a framework for the comparative analysis of transcriptomic data in disease, including the integration of biological networks and other bioinformatic data types, in order to further our knowledge of diseases and the relationships between them.

LIST OF PUBLICATIONS

Oerton E, Bender A. (2017) Concordance analysis of microarray studies identifies representative gene expression changes in Parkinson's disease: a comparison of 33 human and animal studies. *BMC Neurology*; 17(1):58.

Cavalla D, Oerton E, Bender A. (2017) Drug Repurposing Review. In: *Comprehensive Medicinal Chemistry III*. Elsevier; 11-47.

Oerton E, Roberts I, Lewis PSH, Guilliams T, Bender A. Predicting disease relationships through similarity fusion. *Bioinformatics* Advance Access published Aug 30, 2018, doi:10.1093/bioinformatics/bty754.

Alexander-Dann B, Pruteanu L, Oerton E, Sharma N, Berindan-Neagoe I, Modos D, Bender A. (2018) Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. *Molecular Omics* 14(4):218–36.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 UNDERSTANDING DISEASE

1.1.1 Defining disease

Disease can be defined as the dysfunction of one or more of the systems in our body, resulting in the signs and symptoms which characterize a particular condition. Diseases may result from heritable genetic factors, lifestyle and environmental factors, external causes such as infection, or a mixture of these factors: the chance of developing disease may be determined by the combination of inherited genetic risk factors with risk modifiers such as age and lifestyle. Advances in molecular biology have deepened our understanding of the origins of disease, illuminating the flow of dysfunction from the molecular level through successive biological 'layers' (including cells, tissues, organs, systems, and the communications between them) that eventually give rise to the observable phenotype of a disease.

Recent advances in our understanding of disease have introduced several issues in the definition and classification of disease. One example is a blurring of the boundaries between health and disease. It has long been known that certain diseases, such as infectious and/or chronic diseases, have dormant or 'asymptomatic' states during which affected individuals do not display symptoms but still carry the disease, which may recur at any time. In recent years, however, advances in genetics and medical imaging have revealed the existence of what could be called a 'pre-symptomatic' state in certain diseases. In autosomal dominant hereditary disorders such as Huntington's disease, genetic testing can confirm the eventual development of the disease in at-risk individuals[1]; in Alzheimer's disease, accumulation of amyloid proteins in the brain begins years before the associated memory impairment[2]. In the pre-symptomatic state, an individual does not experience any of the physical effects of the disease, yet may not be fully classified as healthy[3], calling into question how disease can be defined in the absence of its symptoms.

A second issue relates to the specificity with which disease is defined. One example of this is the definition of cancer: a broad term referring to disruption of cellular proliferation leading to uncontrolled cell growth, which has the potential to spread throughout the body[4]. However, cancer also refers to a collection of diseases affecting different anatomical locations (lung cancer, breast cancer); different locations within an organ (small cell lung cancer, non-small

cell lung cancer); different tissue types (carcinoma, sarcoma); and different cell types (adenocarcinoma, squamous cell carcinoma). A specific cancer is usually defined according to all of these classifiers (e.g. non-small cell lung adenocarcinoma). Within these subtypes, however, cancer can now be further defined according to its molecular features, such as the estrogen receptor status in breast cancer, which determines which type of treatment is most likely to be effective[5]. Genomic features of tumours may also be an important determinant of drug response[6], leading to the development of 'personalised medicine' approaches to cancer treatment based on exploiting particular mutations. Whilst these molecular features may not affect the symptoms or histopathological appearance of cancer, they are therefore of great importance in defining appropriate treatment strategies for each patient.

These two examples illustrate the difficulty of defining disease by its clinical features alone. One solution to this, which has been made possible by recent developments in molecular biology and bioinformatics, is to understand disease through its molecular features (such as genetic factors) rather than through its symptoms or clinical presentation. By altering our understanding of how a disease is defined, these developments have the potential to revolutionise disease biology, changing the way that disease is prevented, managed and treated.

1.1.2 Using disease relationships to understand disease

Disease classifications describe disease by establishing the relationships between them, which may be based on clinical presentation, aetiology[7], or a combination of these factors. Existing classification systems, which are used by medics, health researchers, and economists[7], include the International Classification of Diseases (ICD)[8], the Systematized Nomenclature of Medicine (SNOMED)[9], the Disease Ontology (DO)[10], and Medical Subject Headings (MeSH)[11]. There are also specialized disease classification systems for particular disease areas, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM)[12] which classifies mental disorders, or Online Mendelian Inheritance In Man (OMIM)[13] for genetic disorders. Although each system is developed for a specific purpose, in general disease classification systems provide a shared computer-readable vocabulary to describe disease, and have a hierarchical structure which groups related diseases together under 'is-a' relationships (e.g. *lung cancer* is-a *cancer*).

These traditional classification systems are based on established disease relationships, and as such do not incorporate new evidence arising from recently developed bioinformatic data

types. At one end of the scale, the advent of large-scale electronic health record data (such as that held by insurance companies) allows the identification of disease comorbidities[14], or links between disease and particular lifestyle or environmental factors. At the other end of the scale, the development of '-omics' techniques, such as genomics, transcriptomics, and proteomics, allows diseases to be related at a molecular level, e.g. through shared genes[15].

These 'molecular-level' relationships between diseases are an important new way of understanding disease relationships, both to identify the shared features underlying known disease relationships and to identify unexpected connections between diseases. Two diseases which appear to be highly related may have very different molecular mechanisms: to adapt an example from Dudley et al.[16], hereditary pattern baldness and alopecia both result in the symptom 'hair loss', but result from different underlying causes (hormonal vs. immune-related). Conversely, two diseases which appear unrelated symptomatically may in fact be very similar at the cellular or molecular level – for instance, a particular type of cancer may produce different symptoms depending on which organ is affected, but the underlying process of uncontrolled cell growth and division is the same.

The identification of molecular commonalities between apparently unrelated diseases could not only shed new light on the pathogenesis of disease, but could help to identify potential treatments. A direct consequence of this would be the ability to reposition or repurpose drugs between related conditions. Drug repurposing (or drug repositioning), defined as the use of a drug in a new indication for which the drug was not originally developed[17], is a promising strategy in the identification of new treatments for diseases, dramatically reducing the cost and time taken to get a drug to market compared to *de novo* drug development[18]. Historically, drug repurposing opportunities have arisen through chance discoveries[17], but our deepening understanding of the molecular basis of diseases and disease relationships could enable the identification of novel repurposing opportunities in a more systematic way.

The comparison of diseases at a molecular level is therefore an invaluable tool for improving our understanding of diseases and their treatment. Current research into the molecular basis of disease is focused primarily on genomics and gene expression, as the technology for both data types is now mature and relatively cost-effective, with well-established methods and software for analysis and large amounts of publicly available data (compared to other -omics technologies such as proteomics and metabolomics). However, whilst DNA generally remains static over the course of an individual's lifespan, gene expression provides a dynamic 'snapshot' of cellular state at a particular point in time, and as such is a key molecular read-out

of disease. In the following sections, I will discuss gene expression data, its analysis, and its use in the comparison of diseases.

1.2 Introduction to Gene Expression Data

1.2.1 Gene expression and its regulation

Figure 1.1 Formation of a protein from the 'recipe' encoded by the DNA

DNA is transcribed into RNA through complementary base pairing, catalysed by RNA polymerase under the control of an assembly of transcription factors (not shown). The RNA molecule initially contains non-coding sequences called introns, which are removed in a process called splicing. The addition of a 5' cap and poly-adenlyated tail (not shown) complete the mature RNA molecule. Messenger RNA (mRNA) is processed at the ribosomes, where the base sequence encoded by the mRNA molecule is translated into a sequence of amino acids, with each group of three bases (a codon) encoding a specific amino acid. The figure shows how two different codons encode the amino acid threonine (Thr), illustrating the redundancy in the amino acid code. Finally, the amino acid chain is folded into the three-dimensional structure of the completed protein.

Gene expression is the process by which a gene encoded as DNA is converted into a functional gene product, such as a protein[19]. An overview of this process is given in Figure 1.1. The process begins with *transcription*: the synthesis of an RNA molecule from the 'recipe' contained in the DNA. To begin the process of transcription, molecules called transcription factors must assemble at specific regulatory binding sites situated upstream of the gene. The

transcription factors help to recruit and bind an enzyme, RNA polymerase, which catalyses the process of transcription. RNA polymerase unwinds the double-stranded DNA helix and works along a single DNA strand, base pairing the deoxyribonucleic acids of the DNA code with complementary free ribonucleic acids (RNAs) in the cell. This process continues until a termination site is reached, and RNA polymerase releases a single-stranded RNA molecule. Various processing is carried out on the RNA molecule before it leaves the nucleus of the cell, including splicing (the removal of sequences of intervening RNA called introns) and the addition of an RNA cap at the 5' end and a poly-adenylated tail at the other, marking the RNA as complete and intact.

The next step in the expression of protein-coding genes is the *translation* of messenger RNA (mRNA) into protein, which takes place on cellular structures called ribosomes. Beginning from a specific 'start' codon (a sequence of three nucleotide bases), mRNA is threaded through the ribosomes, where it binds to complementary transfer RNA (tRNA), which carries an amino acid corresponding to each codon. The amino acids are therefore joined in the sequence specified by the mRNA, continuing until the 'stop' codon is reached, at which point the completed polypeptide formed by the chain of amino acids is released and can fold into its three-dimensional structure. Protein-coding mRNA comprises only 3-5% of total RNA[20]; other RNA types are not translated but remain as RNA molecules in the cell. These include ribosomal RNAs, which form the core of the ribosomes; microRNAs, which play a role in the regulation of gene expression; and long non-coding RNAs, whose function is not completely understood but which may regulate diverse cell processes[20].

Although each cell contains all the DNA code required to make every protein in the human body, only a certain fraction (estimated at around 30-60%) of genes are expressed at any one time[20]. Control of gene expression underlies the differentiation of cell types, and allows cells to respond to environmental conditions and extracellular signals. Gene expression regulation is therefore absolutely crucial to development and homeostasis. Although gene expression can be regulated post-transcriptionally (for example, through complementary binding by microRNAs), the dominant mechanism of gene expression regulation is through transcriptional control[20]. Transcriptional control is mediated through *transcription factors*, which activate or repress transcription through gene-specific mechanisms, such as enhancing the binding of RNA polymerase (activation) or blocking the promoter (repression). Transcription factors can be activated in response to changing conditions in the cell: for example, the binding of a molecule to a specific transcription factor (which takes place when the molecule is present in sufficient

concentrations) changes the three-dimensional conformation of the transcription factor in such a manner that it can then bind to a repressor site on the DNA, blocking transcription[20]. Gene expression levels can therefore be seen as a response to the cellular context, providing a measurement of the cell's response to different conditions.

1.2.2 Motivation for studying gene expression

Improving our understanding of the cellular state in diseased tissue could provide increased insight into the pathogenesis of, and cellular response to, disease. One approach to measuring cellular state is proteomics, or the quantification of proteins. Proteomics is an appealing approach for the study of disease, because it is a direct measure of the 'final' product of gene expression; proteins associated with a disease could therefore function as biomarkers, or even new targets for drugs. However, the proteome is extremely complex, containing some 100,000 proteins according to some estimates[21]; it is also challenging to measure due to the need for quantification of the complex three-dimensional structure of proteins, which may also be post-translationally modified by phosphorylation, ubiquitination, or other processes.

A simpler way to measure gene expression is via RNA molecules, rather than proteins. Each RNA species can be uniquely described by its nucleotide sequence, so quantifying RNA is far less complex than quantifying proteins. In the last two decades, high-throughput methods (such as microarray or RNA-sequencing technologies, which will be discussed in more detail in Methods) have been developed which can measure the expression levels of tens of thousands of genes relatively cheaply and quickly[22]. Studies of this type, which enable gene expression to be studied at a large scale, are referred to as transcriptomics. Transcriptomic experiments generally focus on protein-coding mRNA, which can be interpreted as a proxy or rough estimate of the levels of protein present in the cell (although the correlation between the two is far from perfect, due to numerous factors including post-transcriptional regulation and post-translational modifications[23]; these are reviewed in in Maier et al.[24]). Measurement of other RNA types, such as miRNA, provide additional information about what is taking place inside the cell; indeed, these molecules may play a key part in disease[25]. However, these technologies are still developing and are less widely adopted. In this thesis, therefore, the terms 'transcriptomics' and (less precisely) 'gene expression data' will be used to refer to the quantification of mRNA levels in a high-throughput manner.

The suffix '-*ome*' in molecular biology is generally understood to indicate a global viewpoint – the *transcriptome* therefore refers to the mRNA expression levels of (close to) all genes in a cell. In contrast to classical molecular biology, where individual features may be studied in great detail, -omics approaches are most suited to drive hypothesis generation – for instance, indicating potential biomarkers or altered pathway activity – which may then be confirmed in a more focused manner. A very brief overview of the wide variety of applications of transcriptomics includes uses in developmental biology to study early embryonic development[26]; in agricultural and environmental biology to profile response to environmental stressors[27,28]; in drug development to study drug mechanism of action[29] and toxicity[30,31], and in microbiology to study antibiotic resistance[32] and host-pathogen interactions[33,34] (a more detailed review of these various applications is available in Lowe et al.[22]). In the study of disease, transcriptomics is used to understand more about disease biology, such as to identify perturbed biological pathways[35,36], discover disease subtypes[37] and predict prognosis[38]; transcriptomics is also used in drug discovery to predict drug sensitivity in cancer[39] and to identify drug repurposing opportunities[40,41]. Some of these applications will be discussed further in the following sections.

1.2.3 Measurement of gene expression

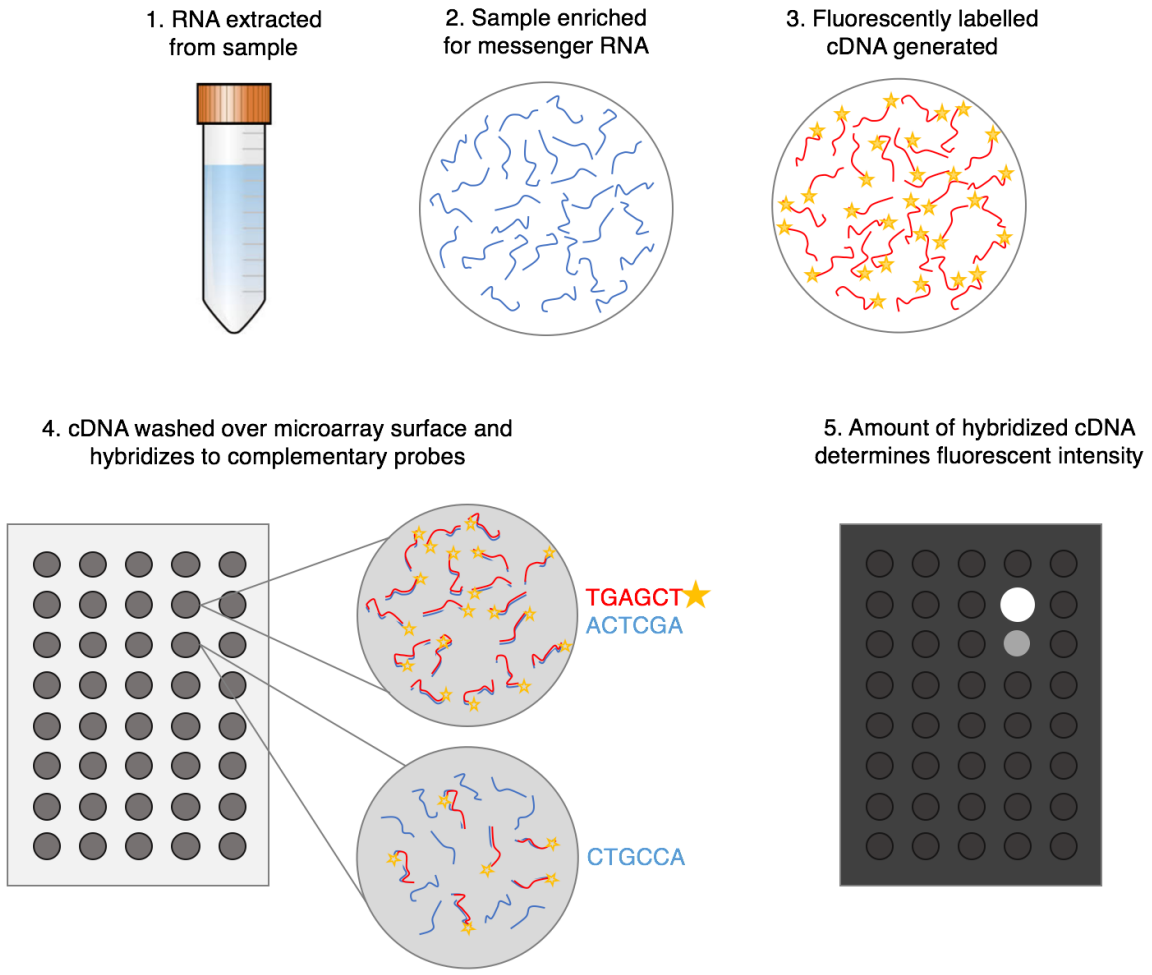


Figure 1.2 Quantification of RNA levels using a single-channel microarray

After extracting RNA from the sample and removing unwanted forms of RNA (e.g. ribosomal RNA, steps 1 and 2), stable complementary DNA (cDNA) is generated and fluorescently tagged (step 3), and is then washed over the array, which contains a number of probes complementary to the cDNA (matching the original RNA sequences). The amount of fluorescently tagged cDNA bound to each probe indicates the abundance of the corresponding RNA molecule in the original sample. As an example, a large amount of the RNA containing the base sequence 'ACTCGA' (top cut-out, step 4) is present in the original RNA sample, so a large amount of corresponding cDNA hybridizes to its matching probe on the array. By contrast, there is little RNA containing the base sequence 'CTGCCA' (bottom cut-out, step 4), resulting in less hybridization and a lower fluorescent intensity on the scanned array (step 5).

Older techniques to measure gene expression, such as Northern blotting and reverse-transcriptase quantitative polymerase chain reaction (rt-qPCR), are low-throughput, measuring only a few individual transcripts at a time. True transcriptomic methods able to measure the expression of thousands of transcripts at a time were introduced with the advent of microarrays

in the mid-1990s, followed by RNA-Seq a decade later. Both methods aim to quantify the amount of mRNA present in a sample and share similar initial steps: extraction of the RNA from the sample; purification to remove unwanted molecular fragments; enrichment for mRNA (this process involves removing the ribosomal RNA which can account for up to 98% of total RNA content in a cell[42]); and finally generation of stable cDNA via reverse transcriptase, which may be further amplified by PCR. The difference between microarrays and RNA-Seq lies in the method of quantification of the generated cDNA.

Microarrays consist of a specific arrangement of oligonucleotide sequences (or 'probes') on a solid surface, each probe being designed to match a particular transcript. In a microarray experiment, the generated cDNA is fluorescently labelled and, when washed over the microarray, will hybridize to complementary probes matching specific RNA sequences (Figure 1.2). Once the unbound cDNA is washed off, the amount of fluorescence at each probe gives an indication of the abundance of each mRNA species in the original sample (relative to other samples). In older 'dual-channel' microarray designs, test and control samples are labelled with different fluorophores (e.g. red and green dyes) and hybridized on the same array, and it is the ratio between the two colours which determines the relative amount of mRNA present in each condition.

RNA-Seq, by contrast, works by sequencing fragments of the cDNA. Each sequenced fragment is aligned to a reference transcriptome (which may be assembled from the genome or generated *de novo*), and the count of each fragment at each transcript location indicates the abundance of each mRNA species in the original sample. RNA-Seq has a number of advantages over microarrays: firstly, its dynamic range is substantially higher, allowing detection of very highly or lowly expressed transcripts (over five orders of magnitude[22]), whereas microarrays suffer from the limits of detection of fluorescence. A further advantage of RNA-Seq is that it doesn't use pre-designed probes, allowing identification of novel transcripts. For these reasons, RNA-Seq is now the dominant technique in transcriptomics, overtaking microarrays (in terms of number of publications referring to the technique) in 2015[22]. However, due to the popularity of microarrays in the preceding two decades, microarray datasets still far outnumber RNA-Seq in public repositories (discussed below); a further advantage is that during this time, techniques for pre-processing and analysis of microarray data have become highly developed and standardized. These techniques will be discussed further in Methods.

1.2.4 Obtaining gene expression data from public repositories

Since the advent of high-throughput technologies in the mid-1990s, the volume of gene expression data stored in public repositories has been increasing rapidly[43]. Developments in technology have enabled transcriptomic studies to be carried out more cheaply and easily than ever before; at the same time, requirements for data sharing put in place by funders and journals mean that this data is increasingly being made publicly available. The largest public gene expression repositories, Gene Expression Omnibus[44] (GEO) and ArrayExpress, contain tens of thousands of studies: at the time of writing, GEO contained 73,388 expression profiling studies, (of which 53,691 were microarray studies and 18,126 were high-throughput RNA sequencing studies[45]), and ArrayExpress contained 70,894 experiments[46] (ArrayExpress imports data from GEO, so there is substantial overlap between the two). These cover many study types, from early microarray studies to cutting-edge single-cell RNA-Seq experiments, and span diverse research areas including toxicology, pharmacology, ageing, and development. In particular, available studies cover hundreds of different diseases, from common, well-studied diseases to extremely rare conditions.

In addition to these general-purpose repositories, several application-specific gene expression databases are available. These include DrugMatrix[47] and Open TG-Gates[48] for toxicity, and the Connectivity Map (CMap)[49] and the Library of Interconnected Cellular Signatures L1000 dataset (LINCS)[50], which record gene expression in human cell lines in response to perturbation by drugs. There are also disease- and organ-specific databases such as the Oncomine database of gene expression in cancer[51] and the Allen Human Brain Atlas[52]. These databases generally have better annotation, curation, and integration than their general-purpose counterparts; the trade-off, however, is that they may not be as comprehensive, i.e. they are unlikely to include all studies related to a particular condition.

1.3 Using Gene Expression Data to Understand Disease

1.3.1 Differential expression analysis

1.3.1.1 Calculating differential expression

After the appropriate pre-processing steps (discussed in Methods, as these are platform-specific), a transcriptomic experiment results in a matrix containing the measured abundance of *i* genes in each of *j* samples. Whilst this 'baseline' gene expression can be informative (for instance, to assess the expression of a particular gene in a specific tissue or cell type, or to

examine the change in gene expression over a time series experiment), the measured abundance can be strongly affected by technical and laboratory-specific factors (discussed below). In many applications, it is more helpful to give gene expression as a ratio of expression between conditions, known as 'differential expression'. In differential expression analysis, genes are interpreted in terms of the expression difference between one group (the condition under study, e.g. patient or tumour samples) and another (the control group, e.g. healthy individuals or non-cancerous tissue). Differential expression has become a popular means of working with gene expression data, as it accounts for the fact that (unless special 'spike-in' controls are used) gene expression measurements are relative, rather than absolute. The magnitude of differential expression is expressed as a fold change (e.g. if a gene is twice as highly expressed in one group than the other, this would be a two-fold change), usually in $\log_2$-transformed space in order to provide a symmetric scale around zero.

Due to the noise inherent to gene expression data, the magnitude of fold change values should be considered in conjunction with a measure of significance – i.e., given the variance observed across samples, how likely is it that the observed difference in means could occur under the null hypothesis (that values in the two groups are drawn from the same distribution)? This is essentially a t-test; however, because of the particular properties of gene expression data (noisiness, gene- and sample-specific variance, low ratio of observations to features), several methods have been developed to estimate a 'moderated' t-statistic more suited to the analysis of transcriptomic data. Numerous packages exist to calculate these quantities including limma[53] (LInear Models for MicroArrays) and SAM[54] (Significance Analysis of Microarrays) for microarrays; and EdgeR[55] and DESeq[56]/DESeq2[57] for RNA-Seq data.

The basic outline of a limma analysis (limma being the most popular analysis package for microarray data) is as follows:

1. Specify the design matrix indicating which samples are to be compared against each other (in a simple case-vs-control differential expression analysis this simply requires assigning samples to case and control groups; additional steps are required for more complicated analyses).
2. Fit a linear model $Y = X\beta + \epsilon$ to each probe, where Y are the observed expression values in each sample, X is the design matrix, β are the co-efficients, and ϵ is an error term (limma function *lmFit*).
3. Use an empirical Bayes procedure (limma function *eBayes*) to smooth the probe-wise variances by borrowing information from other probes. The reasoning behind this is

that given the small sample sizes usually used in microarray experiments, the true variance will be difficult to estimate, so the *eBayes* call adjusts the observed variance towards the expected variance computed from the average of all variances. The smoothed variances are used to calculate a moderated t-statistic for each probe.

4. Summarize the fold changes, significance values, and other statistics (e.g. average expression) for each probe (limma function *topTable*).

An important consideration here is the correction of the obtained significance values for multiple testing. The traditional concept of hypothesis testing was designed to be used in the context of single experiments, but with an individual p-value for each gene, tens of thousands of hypotheses are effectively being tested at once. On a 10,000-gene microarray, an average of 500 genes would meet a significance threshold of $p<0.05$, whether or not any genes were truly differentially expressed. One way to address this problem is to use multiple testing correction to select a more stringent significance threshold, such as the Bonferroni method (which simply divides the chosen significance threshold by the number of hypotheses tested) or less stringent False Discovery Rate methods which specify an acceptable proportion of false discoveries, such as the widely-used Benjamini-Hochberg method.

Probes may then be mapped to their corresponding genes (discussed in Section 2.1.3). Once fold change and (multiple-testing corrected) significance values have been obtained for each gene, 'interesting' genes can be selected based on some combination of fold change magnitude and statistical significance. It is important to note that the fold change and significance threshold used to determine what counts as 'significant differential expression' are essentially arbitrary, and so hundreds or even thousands of genes may be classed as differentially expressed in a particular experiment depending on the selected threshold. The selection of an appropriate threshold is dependent on the context (what question is being answered with the analysis?) and the type of analysis carried out.

1.3.1.2 Using differential expression to understand disease

Classical differential expression analysis aims to understand DEG lists through the functions of individual dysregulated genes, identifying those which could form a compelling hypothesis for the changes underlying the development of or response to disease. An example of this type of analysis is the work of Trigueros-Motos et al.[58], who analysed differential expression in vascular regions prone to atherosclerosis compared to those resistant to atherosclerosis. They

found overexpression of a set of four homeobox genes (involved in anatomical specification in embryonic development), *Hox6-Hox10,* in murine athero-resistant aorta. The athero-prone regions in murine aorta and human smooth muscle cells showed higher activity of inflammatory mediators normally inhibited by homeobox genes, suggesting that the interplay between homeobox and inflammatory gene expression patterns could create an environment which allows the development of atherosclerosis.

The advantage of this type of analysis is that it can provide a mechanistic understanding of the observed gene expression patterns, which can then be followed up in detail by e.g. confirmation of high-throughput measurements using techniques such as qPCR. However, this relies on human interpretation of potentially tens of thousands of data points, necessitating the use of strict thresholds to reduce DEG lists to a reasonably interpretable size. This leads to a focus on only on the most strongly dysregulated genes, risking the exclusion of less strongly dysregulated genes which could nevertheless be relevant to the studied condition. Further, human interpretation is subject to bias, as investigators will naturally focus on the most recognised and well-studied genes, potentially overlooking important but less well-characterized genes. Other methods which have been developed for the analysis of gene expression data therefore aim to provide an interpretable summary of large gene lists by grouping genes into sets.

1.3.2 Analysis against known gene sets

1.3.2.1 Performing gene set analysis

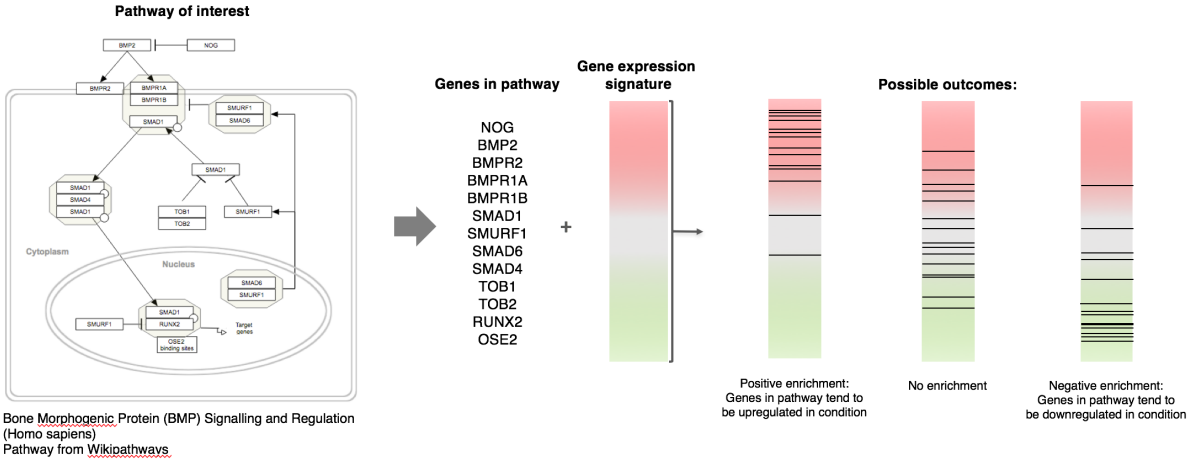


Figure 1.3 Association of a gene expression signature to the set of genes in a pathway

A gene set of interest (which may be e.g. a biological pathway or a set of genes known to be associated with a particular condition) is compared against a ranked gene expression profile (e.g. the most up-regulated to the most down-regulated genes in a particular condition). If the genes in the set are found towards the top or bottom of the ranked gene list (generally quantified using a Kolmogorov-Smirnov statistic), this gene set is considered to be 'enriched' in the gene expression profile. An unsigned version of this test can be carried out (for instance, if genes are ranked in order of p-value), in which case only enrichment towards the top of the list is of interest.

Given a selection of genes of interest, gene set analysis methods aim to associate (subsets of) these genes with known biological processes. This can be thought of as a 'translation' of the observed gene expression patterns to a higher biological level, reducing complexity and aiding interpretation. Typically, gene set methods calculate enrichment against biological pathways, including:

- Metabolic pathways, which describe a sequence of reactions taking place in the cell that transform (metabolise) substrates into new products[59]. Examples include glycolysis and the citric acid cycle.
- Gene regulatory pathways, which describe the interaction of genes, RNA, proteins, and other cofactors to regulate gene expression and protein production in the cell.
- Signalling pathways, which are involved in cellular communication and the transmission of cellular signals to and from the cell. Examples include pathways involved in homeostasis, such as the insulin signalling pathway.

- Some collections also define disease pathways, which detail perturbations to the healthy biological system which are associated with disease. Disease pathways are available for several well-studied diseases such as cancer, diabetes, and cystic fibrosis, but for many diseases pathways are not known; and where pathways are known these may be incomplete.

In the last two decades, numerous collections have been developed which relate genes to biological pathways. Popular resources include primary pathway databases such as KEGG Pathways[60], Reactome[61], and Panther[62] as well as commercial pathway databases. Aside from biological pathways, other ways to group genes include functional or local grouping – a well-known gene set collection of this type of is Gene Ontology[63], which annotates genes with lower-level biological processes (which can be considered the 'building blocks' of biological pathways), molecular functions, and cellular components. Other gene set collections include collated resources such as WikiPathways[64] and MSigDB[65], which bring together multiple collections into a single database.

The identification of gene sets relevant to the condition under study can be achieved through various methods which can be divided into two broad classes: overrepresentation and enrichment. *Overrepresentation* methods take a gene list as input, and test for a significant difference in the representation of the input list compared to the background or 'reference' list in each gene set. By contrast, *enrichment* analysis uses the whole measured expression profile, removing the need to define a DEG list using an (arbitrary) cut-off. The most widely-used enrichment method is gene set enrichment analysis (GSEA)[66], which uses a Kolgomorov-Smirnov-like statistic to test for the distribution of each gene set within the measured gene expression profile, which is rank-ordered by e.g. fold change or significance (Figure 1.3). Numerous tools exist to perform either type of gene set analysis, including online tools such as DAVID[67], ToppGene Suite[68], WebGestalt[69], and GOrilla[70], as well as R packages such as ReactomePA[71] and TopGO[72], and standalone software for GSEA. Whilst in theory the choice of method is independent of the gene set used, in practice, the tools are generally integrated with particular datasets – for instance, the aforementioned Panther includes a web tool to perform both overrepresentation and enrichment analysis on Panther pathways, although recent updates have added GO and Reactome annotations as well.

1.3.2.2 Using gene set analysis to understand disease

One example of the use of gene set analysis to understand diseases is the meta-analysis of 17 Parkinson's disease (PD) studies carried out by Zheng et al.[73], covering studies ranging from early subclinical to severe disease stages. The authors used GSEA to identify 10 MSigDB gene sets associated with PD across multiple studies, including 'electron transport chain', 'oxidative phosphorylation', and 'pyruvate metabolism'. The ten gene sets covered four distinct biological areas related to neuronal energy metabolism: electron transport, mitochondrial biogenesis, and glucose utilization and sensing. These findings suggested the intriguing hypothesis of Parkinson's disease as an alteration of normal cellular energetics to which dopaminergic neurons 'may be intrinsically more susceptible than other cells'[73].

Whilst gene set enrichment methods provide a high-level description of altered gene expression in terms of the biological functions affected, they hide information about the roles of individual genes. Analysis against known gene sets is also limited by the fact that our knowledge of biological pathways and processes is incomplete and incompletely accurate[74]. Recently, *de novo* gene set identification methods have been proposed as an intermediate, providing insight as to how genes work together (e.g. along signalling pathways) whilst retaining information on the activity of individual genes.

1.3.3 Gene network analysis

Rather than analysing gene expression against predefined gene sets, more recently developed methods of interpreting gene expression data are based on the identification of *de novo* gene groupings by linking genes whose expression is altered in the condition under study, forming condition-specific gene networks. Current network-based methods fall into one of two categories according to what the gene links (*edges* of the network) represent: the first is *co-expression* methods, which link genes whose expression varies similarly across a series of samples. Co-expression methods are based on the hypothesis that groups of genes whose expression varies together are associated with similar biological or regulatory processes[75]; they do not try to describe the causality or directionality of interactions between genes[76].

Conversely, methods in the second category (which can be termed *interaction*-based methods) link genes via known interactions between them (or between their products). These may include physical binding interactions and/or (indirect) functional (e.g. regulatory, metabolic, or signalling) interactions; all of these different interaction types can collectively be termed the

'interactome'[77]. Many databases are available which cover interactions of different types; some of the best-known are String[78], IntAct[79], ConsensusPathDB[80], and SignalLink[81], as well as more recent integrative efforts such as OmniPath[82] and HIPPIE[83]. Just as with gene set data, our knowledge of the interactome is incomplete[84]; however, the interactome can be treated as a 'scaffold' of prior knowledge to aid gene expression analysis. For example, Rakshit et al.[85] constructed a protein-protein interaction (PPI) network from differentially expressed genes in Parkinson's disease (PD). The authors suggested that genes which were topologically significant in the constructed PPI network (hub or bottleneck nodes) may represent possible therapeutic targets in PD. However, by focusing only on the most topologically significant nodes, key genes involved in disease – which tend not to be 'hub' genes[15] – may be missed.

An alternative method to analyse gene networks is to group the network into 'modules', representing sets of interlinked genes that may be related to a particular biological function. Module identification methods include clustering methods like MCODE[86] or GLay[87], which are based on network topology and are therefore independent of network type. Other approaches have been developed which are specific to either co-expression- or interaction-based networks, such as Weighted Gene Co-expression Network Analysis (WGCNA)[88] for co-expression networks, which performs co-expression analysis, network construction, and module identification.

An example of module-based analysis of gene networks is the work of Ray et al.[89], who investigated co-expression patterns across six brain regions in Alzheimer's disease. For each pair of brain regions, genes showing dysregulation in both regions were used to construct co-expression modules. Each module was analysed for preservation or perturbation between the two regions: the varying co-expression patterns of commonly differentially expressed genes in perturbed modules suggest regulatory variation between the two regions. A limitation of module-based network analysis is that the resulting modules may be large, necessitating further analysis in order to be interpretable – in this study, the authors used pathway analysis to explore the functions of the genes in each co-expression module. A key challenge for network-based analysis is therefore to highlight small 'active' network regions, forming a middle ground between analysis of individual topologically significant nodes and analysis of large modules. Developments in this area will be described in more detail in Chapter 4.

1.3.4 Integration of other data types

Gene expression represents only one dimension of the molecular changes associated with disease, and as the volume of bioinformatic data of all types continues to grow, methods which combine gene expression with other data types are becoming increasingly popular. A common choice is to integrate gene expression data with genetic variant data (heritable variants, e.g. SNPs, or somatic mutations) in order to explore potential relationships between genetic alterations and gene expression[90]. Integration approaches can be extended to different -omics data types including proteomics and metabolomics. In the last few years, large-scale datasets such as The Cancer Genome Atlas[91] which contain genomic, epigenomic, transcriptomic, and proteomic data relating to each sample[92], have enabled 'multi-omics' approaches which have the potential to provide a comprehensive molecular-level characterization of disease.

Molecular-level data can then be combined with data at the clinical level – such as survival, comorbidity, phenotype, and drug prescription – to identify links between molecular-level changes and clinical outcomes in disease. For example, Orozco et al.[93] used a 'systems genetics' approach to study the association between methylation data and clinical (including blood, fat, and insulin-related) and molecular (including metabolites and proteins as well as gene expression) traits in mice, finding that many associations could be identified by the methylation data ('epigenome-wide association') which were not identified using traditional genome-wide association methods. Methods for data integration are currently an active area of research, particularly for methods which integrate data types other than -omics data (such as electronic health record or co-morbidity data), as it is not yet clear how associations can best be modelled across diverse data types. This issue will be discussed further in Chapter 5.

1.4 Using Gene Expression and Other Bioinformatic Data for the Comparative Analysis of Diseases

1.4.1 Using comparative analysis to deepen our understanding of disease

The reuse and reanalysis of existing studies reduces duplication of effort and enables multiple researchers to analyse the same data, ensuring reproducibility. Further, previous studies can be combined to increase statistical power, enabling us to address existing biological questions with greater insight, and even to pose new questions motivated by greater data availability. A 2012 review of reuse of gene expression data found that a quarter of studies citing

ArrayExpress used it to address a biological question without generating any new experimental data[94]; a further quarter used public data to complement their own newly generated data (e.g. as a validation dataset). However, the high dimensionality of gene expression data, and its susceptibility to noise resulting from multiple biological[95] and technical[96] factors, means that comparison across gene expression datasets is challenging. In this section I will discuss how the comparative analysis of multiple studies can inform us about disease, including some of the potential pitfalls that must be considered.

1.4.1.1 Meta-analysis reveals commonalities across studies of the same condition

A common type of comparative analysis is meta-analysis: the integration of multiple studies of the same condition to effectively form a much larger study which is more generalizable (drawing results across different experimental designs and sample populations)[97] and has increased statistical power compared to any one individual study[98]. Meta-analysis approaches are particularly useful in the context of gene expression studies, where the high dimensionality and non-standardized (pre-)processing and analysis methods mean that there may be little agreement between individual differentially expressed gene lists[99] of the type discussed in Section 1.3.1.2. In a meta-analysis of 21 thyroid cancer studies[100], for example, only 107 of 755 genes in published gene lists showed consistent differential expression in more than one study. Those genes which are consistently differentially expressed across multiple studies, however, are then more likely to be associated with the condition under study rather than technical or experimental factors (discussed below).

Given the low agreement between gene expression studies of the same condition, a key issue in conducting a meta-analysis is how to define which studies are comparable, i.e. how to determine the best balance between study inclusion vs exclusion: the greater the number of studies that can be included in a meta-analysis, the greater the statistical power that can be achieved, but including too wide a range of studies risks diluting the detected signal due to 'noise' from studies which do not accurately reflect the condition under study.

1.4.1.2 Comparing gene expression reveals links between different conditions

The comparison of clinically related diseases can reveal common molecular mechanisms underpinning shared pathogenesis or phenotypes. Studies of gene expression patterns have illustrated the similarities between inflammatory bowel diseases[101], systemic autoimmune

diseases[102], neurodegenerative diseases[103], and mental health diseases[104] to name just a few examples. As well as identifying similarities between related diseases, this approach can also be applied to phenotypically diverse diseases to uncover similar mechanisms of dysregulation that may lead to distinct phenotypes due to e.g. differences in cellular context[105]. A 2009 study[40] compared 74 diseases using both 1) the correlation between their gene expression profiles and 2) the enrichment of genes significantly differentially expressed in one disease in the profile of another (analogous to gene set enrichment analysis). This approach revealed unexpected connections between diseases, such as between Crohn's disease and malaria, as well as known or emerging connections between diseases in the same ontological category, such as between actinic keratosis ('sun spots', which may be pre-cancerous) and multiple cancers. This illustrates how large-scale exploratory comparisons of gene expression profiles from clinically unrelated diseases can highlight unexpected similarities between diseases.

The genes and pathways shared between diseases can be used to generate new hypotheses about the molecular mechanisms of disease: for instance, Yang et al.[106] compared gene co-expression across 108 diseases, finding overlap of co-expressed genes between allergic asthma, type 2 diabetes, and chronic kidney disease. Many of the shared genes were involved in Wnt signalling, suggesting the possible involvement of a common pathway in these phenotypically distinct diseases. In the network as a whole, more than half (57%) of the 1326 disease-disease links are novel links according to traditional disease classifications, with 82% of these sharing disease-related genes or drugs, illustrating a relevant biological basis for the connection. This study provides an example of the use of co-expression network analysis to compare diseases; surprisingly, however, few approaches have compared diseases based on gene expression in an interaction network (the other type of network-based analysis discussed above), despite the potential utility of identifying shared network regions active in disease. This will be discussed further in Chapter 4.

Shifting focus from diseases to genes, other approaches have used large-scale disease comparisons to identify genes which are frequently dysregulated in different conditions. This analysis helps us to better understand observed gene expression patterns by identifying which patterns may reflect a general 'disease response' such as inflammation or immune system activation, and which appear specific to the disease. Suthram et al.[107] identified 'disease modules' by mapping differential gene expression in 54 diseases to pre-computed 'functional modules' (describing e.g. protein complexes in the human protein-protein interaction network). They found that 59 of the 4,620 functional modules were enriched in at least half of diseases

in the network, which the authors suggested form a 'common disease-state signature'. Further, these frequently dysregulated modules were found to be enriched for known drug targets, illustrating that these modules could form a useful basis for therapeutic options targeting common disease symptoms. Unfortunately, this study (which was carried out in 2009) is somewhat limited due to the restricted number of diseases studied, as well as the use of generic pre-computed modules, rather than disease-specific modules. A similar analysis over a wider range of diseases would therefore be of great interest for the analysis of gene expression profiles and how they reflect common disease responses.

1.4.1.3 Cross-condition comparison of gene expression signatures can be used for drug repurposing

Comparative analysis of gene expression signatures is a popular approach in transcriptomic drug repurposing, based on the idea that drugs inducing gene expression profiles which are 'opposite' to a disease-induced differential expression profile may be able to reverse the dysregulation associated with the disease[40,41]. A study comparing gene expression profiles in 100 diseases and 164 drugs (using drug-induced gene expression profiles from the CMap resource introduced in Section 1.2.4)[41] used the 'connectivity' between diseases and drugs (a concept similar to gene set enrichment, in which the top differentially expressed genes in disease form the 'gene set' against which the drug expression profile is tested) to identify potential candidates for drug repurposing. If a disease is strongly negatively enriched for a particular drug, this suggests that the drug may potentially be therapeutic against the disease. As well as known drug-disease connections, such as the corticosteroid prednisolone for inflammatory bowel diseases, the authors identified novel connections such as the anti-ulcer drug cimetidine for lung cancer, which experimental validation showed to be effective in a mouse model. Following the success of this study, many other approaches have been developed based on similar concepts[108–110].

Despite the potential of these approaches for the identification of drug repurposing hypotheses, comparisons between drug and disease signatures suffer from the same limitations as comparisons across disease signatures, including incomplete understanding of issues such as how cell line models of disease (which are used to record drug response) can be compared to disease gene expression profiles from patients. Our understanding of potential repurposed

treatments for disease would greatly benefit from improved insight into the comparative analysis of gene expression profiles.

1.4.2 General issues in comparative analysis of gene expression

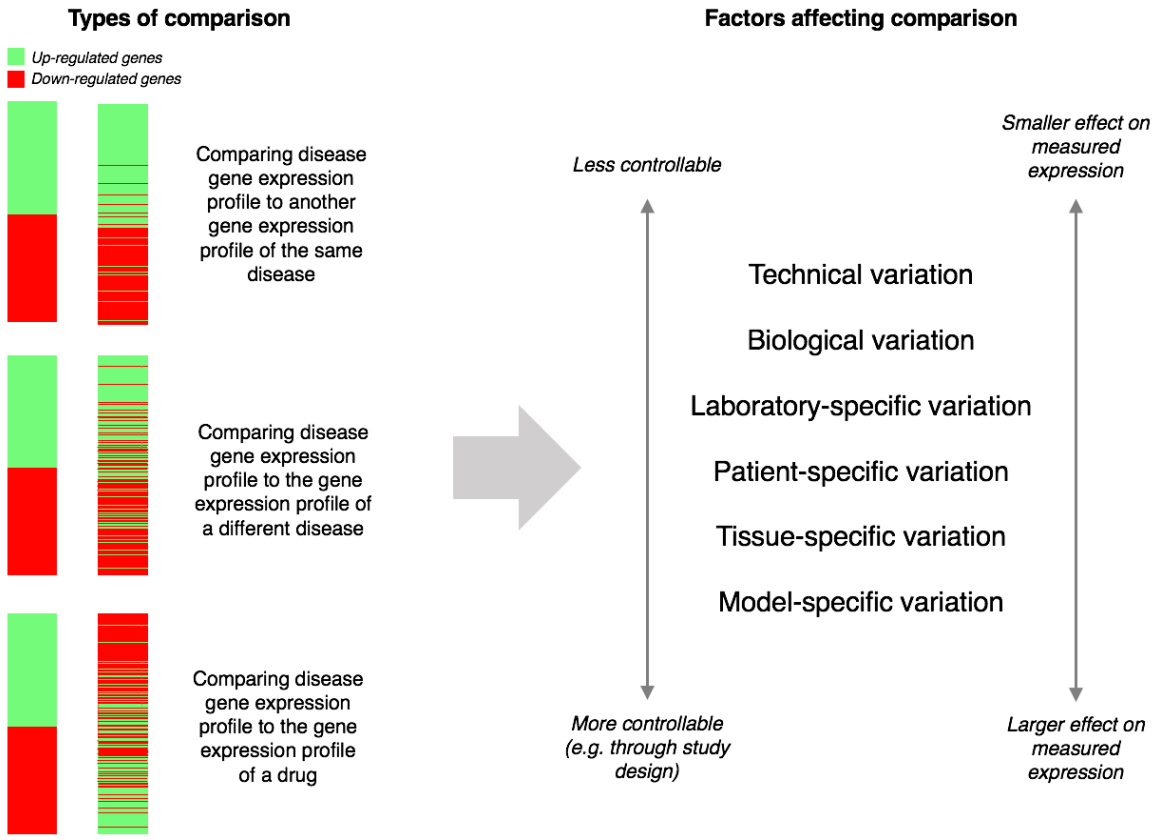


Figure 1.4 Types of comparative analysis of gene expression profiles of disease and factors affecting comparison

A number of factors must be taken into account when comparing measured gene expression profiles. These range from technical and biological variance, which are difficult to control but which have only a small influence on measured expression; to factors which can cause much larger differences in measured expression, such as patient selection (e.g. the drug treatment history of patients) or model type (e.g. comparing a human study to an animal model). A real example of the schematic diagram on the left is presented in the work of Dudley et al.[111].

Sources of variation between gene expression experiments can be roughly divided into two categories: variation arising from the noise inherent to gene expression measurement, which is pervasive and difficult to control; and variation arising from choices made in the design of the experiment, which is controlled according to the aims of the study.

Figure 1.4 summarises these types of variation, which will be discussed further in this section. Although this figure illustrates the comparison of differential gene expression profiles, which is the focus of this thesis, these factors also apply to the comparison of base gene expression. However, comparison of base expression profiles is further complicated by differences in e.g. average measured intensity, which are cancelled out (due to being reported as a ratio) in differential expression analysis.

1.4.2.1 Experimental factors affecting comparison of gene expression studies

The first source of variation can be described as *noise* – unintentional variation resulting from factors other than the condition under study – and can be further divided into biological and technical noise. *Biological noise* is an inherent property of gene expression, and includes noise resulting from the dynamic nature of transcription, including transcriptional bursting[112] and differences in transcription at different stages of the cell cycle[95]; these are of more concern in recently developed single-cell transcriptomics methods, as they should even out over RNA sampled from whole tissue. Another source of biological noise is variation in gene expression patterns between individuals[113], although these differences should also even out over an adequate number of samples.

Technical noise is noise resulting from experimental and measurement factors, such as sample preparation (RNA extraction, labelling and amplification[114]), probe hybridization[115] and array scanning[116]. This type of noise appears even between repeat measurements of the same sample. Numerous studies, most notably the large-scale studies co-ordinated by the Microarray Quality Control Consortium[117] and the Sequencing Quality Control Consortium[118], have found good agreement for relative (differential) gene expression measurements of the same sample, suggesting that repeats of an experiment should be highly concordant. Outside of these controlled large-scale studies, however, a further source of technical variation between gene expression studies arises from laboratory-specific differences in sample handling and the protocols and platform type used.

1.4.2.2 Study design factors affecting comparison of gene expression studies

The second group of factors that must be taken into account when comparing experiments relates to the design choices made for a given study. A key consideration in the study of disease

is the choice of tissue from which to sample: investigators may choose to sample gene expression in the tissue most relevant to the disease, or they may choose to use tissues that are more easily accessible, such as blood samples. This is particularly important in cases where sampling the actual tissue is invasive (e.g. colonic tissue) or where tissue cannot be sampled until post-mortem (e.g. brain tissue). In the case of progressive diseases such as neurodegenerative disease, different sub-tissues may be sampled to follow the progression of the disease. Patterns of tissue-specific baseline gene expression exist across healthy tissue[119], and it is reasonable to assume that the gene expression response to disease may also differ between tissues. A study by Dudley et al.[111] addressed the subject of whether the disease-specific signal across tissues is stronger than the tissue-specific response to disease. Across microarray studies representing 238 diseases and 122 tissues, under 84 combinations of workflow parameters (normalization, merging, and quantification methods), the authors concluded that although comparison across different tissues reduced the concordance between studies, 'the molecular signature of disease across tissues is overall more prominent than the signature of tissue expression across diseases'[111]. However, the question remains to what extent an experiment carried out in a 'surrogate' tissue is reflective of the gene expression changes that would be observed in the most directly affected tissue.

A further issue is the selection of patients (and equivalent healthy controls). Distinct from biological noise as discussed above, which would be found even in a controlled population (e.g. of genetically identical laboratory animals raised in identical conditions), human studies involve genetically diverse populations of patients of different ages and genders. These factors may affect both baseline gene expression[120] and the gene expression response to disease. A further consideration particular to the study of disease is the drug treatment history of patients: patients may follow many different drug treatment regimes, which will affect the observed expression patterns by diluting the 'disease signal' with the 'drug signal'. Complicating these issues further is the fact that this demographic information is often not supplied in the sample meta-data stored in public repositories.

A final issue in sample selection concerns the choice of disease model. The 'gold standard' for gene expression experiments in disease is samples directly from patients; however, some studies of disease use animal models or cell lines in order to perform experiments which would not be possible in human patients. The question of how well gene expression in e.g. cell lines taken from patients reflects gene expression sampled directly from the patient has not yet been answered definitively. The issues raised here will be addressed further in Chapter 3.

1.4.2.3 Factors specific to cross-condition comparison of gene expression studies

As well as the general issues in comparing gene expression data discussed above, there are further issues specific to comparing gene expression between different diseases. An initial consideration is that diseases affect different anatomical locations – their effects may be systemic, such as autoimmune or metabolic disorders, or they may be localized to a particular tissue or organ, such as skin diseases or cancers. As discussed above, gene expression responses to disease differ depending on the sampled tissue; one solution is simply to assume that the strongest and most 'representative' signal of disease is found in the most highly affected tissue, and to compare disease signatures in the most affected tissue for each. Given the availability of enough studies (which is not currently the case), diseases could be compared across studies in a single tissue such as blood. However, the strength of the transcriptomic signal of disease in e.g. blood compared to the affected tissue is not yet established.

A further consideration is that different diseases have different drug treatment programs, which will affect observed gene expression to different extents. The alternative is to limit the dataset to the few studies which state that they have been carried out in drug-naïve patients, but currently there are not enough transcriptomic studies available in explicitly drug-naïve patients for this to be practical. Despite these limitations, previous studies, such as those discussed in the preceding sections, have illustrated that comparing gene expression data across diseases can yield valuable insights into disease and its potential treatment.

1.5 PROPOSED RESEARCH

In the Introduction so far, I have discussed issues relating to the analysis of gene expression data, and explored the advantages and limitations relating specifically to the comparison of multiple gene expression data sets. In particular, I have focused on the use of gene expression data to compare different diseases. Despite the potential of this type of analysis to deepen our knowledge of diseases (and their possible treatments), there are key questions remaining to be answered in order to fully exploit the potential of this approach. I will now provide a more detailed summary of how these will be addressed in this thesis.

1.5.1 How do experimental factors affect measured gene expression in disease?

This question concerns the comparability of microarray data across different tissues and organisms. As discussed above, factors such as the choice of microarray platform and disease model in a transcriptomic experiment all affect the measured signal of a disease. Several studies have addressed these factors individually, but it remains unclear how these factors affect the 'representativeness' of the measured gene expression to the condition under study. This in turn affects the ability to compare gene expression from different studies – if the choice of e.g. microarray platform strongly influences the measured gene expression, is it meaningful to compare studies from different microarray platforms?

As an initial study in the comparison of transcriptomic data, I examine the concordance of microarray studies across different tissues, microarray platforms, and disease models, as well as the effect of sample sizes, using the neurodegenerative condition Parkinson's disease as a case study. After establishing the similarity of different study types to the 'gold-standard' studies of human brain tissue, I further examine the specificity of the measured signals through comparison with other brain diseases: Alzheimer's disease and cancer. As well as being the largest meta-analysis of gene expression in Parkinson's disease to date, this work provides guidelines for study selection e.g. in a meta-analysis context. This chapter forms a basis for the following work, helping to define the criteria for study inclusion in the larger disease dataset used in the following chapter.

1.5.2 How can shared gene expression patterns across different diseases be identified?

Previous sections of the Introduction discussed methods to aid the interpretation of gene expression data, including enrichment of differentially expressed genes against functional gene sets (e.g. biological pathways) and grouping into functional or co-expression networks. Whilst these methods can be invaluable for the analysis of individual gene expression datasets, they have several limitations for comparison of diseases. Biological pathways are a high-level description of a biological process, and comparison of diseases at this level discards the gene-level information that could help to identify more specific processes. Further, methods reliant on known biological pathways are restricted only to the identification of known processes, and

are thus incapable of identifying novel disease-associated processes. By contrast, methods based on interaction networks not only retain the gene-level information, but (depending on the interaction type) can also illustrate the potential flow of disease-related perturbation through the network. However, network-based methods tend to produce large modules that may include many genes solely on the basis of topology, which are not ideal for making comparisons between diseases.

In this chapter, I introduce a weighted shortest-paths method (based on the work of Sambarey et al.[121]) which identifies the most highly perturbed signalling paths in a disease. Each edge in the path links two differentially expressed genes, allowing comparison of diseases based on individual shared edges. Connecting the shared edges forms shared dysregulated networks, enabling the identification of processes showing altered expression in both diseases. I first confirm the biological relevance of the identified paths by showing their enrichment for disease-associated genes and drug-interacting genes compared to the use of simple (non-network based) differentially expressed gene lists. I then examine the properties of genes which are frequently in dysregulated paths across multiple diseases, identifying commonly dysregulated genes which may represent 'pressure points' in the human signalling network. Finally, I apply the method to the pairwise comparison of 141 studies of common and rare diseases, identifying disease pairs with significant similarity (i.e., a significant number of shared interacting gene pairs) and illustrating how shared dysregulated paths might be used to identify potential opportunities for drug repurposing.

1.5.3 How can gene expression data be integrated with other bioinformatic data types to make connections between diseases?

Previous sections of the Introduction outlined how the integration of other bioinformatic data types can aid the interpretation of gene expression in disease. However, gene expression represents just one possible information 'layer' at which diseases may be compared: diseases may share drug treatments, for instance, or a genetic variation conferring risk for the disease. Defining relationships between diseases across multiple biological layers could form a new bioinformatic classification of disease, incorporating new molecular data types as well as traditional disease relationships. Such multi-layer links between diseases could improve our understanding of disease biology and potentially identify opportunities for drug-sharing.

However, relatively few methods have been developed for the comparison of diseases across diverse data types. One reason for this may be that differences in data type properties (e.g. information content) make it difficult to directly compare disease relationships across spaces[122]. In the final section, I explore the use of additional data types in the comparison of diseases, developing a method based on the integration of pairwise disease similarities in an unbiased manner which accounts for differences in data type properties. The proposed method is used to integrate gene expression data with five other data types – ontology, phenotype, literature co-occurrence, genetic variation, and drug prescription – and is designed to easily incorporate additional data types. The integrated similarities reveal how links between diseases at the gene expression level relate to links at other levels, and are used to explore disease relationships that exist across multiple levels, particularly in relation to drug-sharing.

1.6 SUMMARY

In this thesis, I will use comparative analysis of gene expression data to understand diseases and the relationships between them, addressing issues including the comparability and interpretation of transcriptomic data, as well as the integration of other data types to aid identification of disease relationships. The remainder of the thesis is structured as follows:

In *Methods,* I discuss methodological aspects related to the selection, pre-processing, and analysis of gene expression studies.

In *Concordance of Microarray Studies of Parkinson's Disease*, I describe work investigating the comparability of microarray studies across different species, tissues, and microarray platforms.

In *Using Dysregulated Signalling Paths to Understand Disease*, I describe the integration of gene expression data with a signalling network to aid the interpretation and comparison of gene expression datasets across diseases.

In *Understanding and Predicting Disease Relationships Through Similarity Fusion*, I introduce a method for the integration of multiple bioinformatic data types, and apply it to reveal disease relationships spanning gene expression and other spaces.

Finally, in *Conclusions* I describe the significance and possible future directions of the research described in this thesis.

2 METHODS

2.1 ANALYSIS OF MICROARRAY DATA

2.1.1 Motivation for using microarray data in this project

In the Introduction, I discussed the advantages and disadvantages of microarray and RNA-Seq technology. Despite the many advantages of RNA-Seq, in this work I have chosen to focus exclusively on the analysis of gene expression data from microarrays. This is the most suitable data type for this project, which involves the analysis of hundreds of transcriptomic datasets, for two reasons. The first reason is the greater ease of analysis of microarray data: the large volumes of data required by this project necessitate a semi-automated analysis workflow that can be applied in a standardized manner across multiple datasets, which would be much more complex without the greater homogeneity and standardization amongst microarrays (at least within a platform type). The second reason is the greater number of microarray datasets: as detailed in the Introduction, microarray experiments make up much of the data stored in public repositories due to their longer history of use compared to RNA-Seq. As this project relies on the use of public datasets, focusing on microarray data then allows a greater number and variety of datasets to be included in the comparative analysis setting.

The other possibility is to analyse both data types together, with the bulk of the datasets coming from microarray experiments, supplemented with RNA-Seq experiments where necessary for greater coverage. However, given that the reported concordance between the two technologies is moderate[123] and dependent on transcript abundance[124], the risk is that the use of different technologies might introduce an extra confounding factor in the comparison of measured gene expression. Therefore, only microarray studies were used in this project.

2.1.2 Retrieval and pre-processing of microarray data

A broad range of microarray platforms are in use for the measurement of gene expression data. By far the most common microarray type in the datasets used by this project is the Affymetrix range of microarrays (see Appendices A, B, and C), in particular the Human Genome U133A and Human Genome U133A Plus 2.0 arrays; other platform types represented in the dataset

include Illumina BeadChip and Agilent SurePrint arrays. While there should be general consistency amongst these different platform types (as evidenced by e.g. the work of the Microarray Quality Control Consortium[117]), platform effects may still be detectable between the different technologies[125,126]. This is investigated further in Chapter 3. No matter the platform, differential expression analysis of microarray data follows the same basic steps: data retrieval; pre-processing and normalization; and determination of significantly differentially expressed genes.

As discussed in the Introduction, microarray data can be retrieved from public repositories such as Gene Expression Omnibus[44] or ArrayExpress[127]. The datasets used in this project were obtained from GEO, which has a number of associated R tools and also contains much of the data in ArrayExpress. GEO records often contain both raw and pre-processed data. Several different algorithms are available for pre-processing (which is required to convert the measured fluorescent intensities from each chip into comparable expression values), and in a comparative analysis setting it is therefore advisable to download the raw rather than submitter-supplied pre-processed data in order to exclude the possibility of bias resulting from different pre-processing methods.

Pre-processing methods are specific to each platform type: for Affymetrix data, the different pre-processing algorithms available include MAS-5.0, RMA, and gc-RMA. RMA has several advantages over the older MAS-5.0 algorithm, namely less noise and variance at lower expression levels, and is the currently the most widely-used approach[128], so I have used RMA for processing Affymetrix data. The three steps of RMA are: background correction, which models the observed expression as a function of signal and noise; quantile normalization, which fits the expression values on the chip to the same distribution; and finally summarization of the log-transformed values using a median polish algorithm, which iteratively subtracts chip- and probe-level medians to estimate chip- and probe-specific errors. Following this step, the measurements from each chip are comparable to each other, and further analysis can be undertaken.

2.1.3 Generating a differential expression profile

In the Introduction, I discussed differential expression analysis: the comparison of gene expression values between case and control samples, and the calculation of the magnitude and significance of the fold change value for each probe on the microarray. The next step in

generating a differential expression profile is mapping of probes to their corresponding gene(s) using the platform-specific annotation files supplied by the manufacturer, which can be obtained from repositories such as GEO. The relationship between probes and genes is not one-to-one. There are several methods to resolve this relationship (which are detailed in a review by Ramasamy et al.[129]); a straightforward method adopted for this project is to retain the probes with the highest p-value in the case of multiple probes mapping to a gene, and to duplicate the probe information in the case of probes mapping to multiple genes.

The final stage of differential expression analysis is to determine which genes are considered meaningfully differentially expressed. This requires the choice of an appropriate threshold of fold change and/or significance; as mentioned in the Introduction, the use of these thresholds will vary depending on the particular experiment and the goals of the analysis. Multiple testing correction is generally advisable in differential expression analysis of a single experiment, where a list of high-confidence differentially expressed genes with few false positives is the desired outcome.

By contrast, this project involves the comparative analysis of many datasets, with highly variable profiles of significance. For some of the studies used in this work, no genes remain significantly differentially expressed after Benjamini-Hochberg multiple testing correction (the default in limma) is applied: this is the case for 24 of 42 studies used in Chapter 3, 35 of 141 studies used in Chapter 4, and 17 of 84 studies used in Chapter 5. Rather than remove these studies from consideration entirely, I have applied a non-conservative significance cut-off of raw $p < 0.05$ to call differentially expressed genes throughout the dataset. In the comparative analysis setting, it is the genes that are shared between datasets that are of interest, rather than each individual gene in a gene list. The proportion of false positives is therefore less of a concern than in the analysis of a single experiment: a study which does not record any truly differentially expressed genes will likely not appear similar to any other experiment, and can be excluded from further consideration at this point.

Following the approach recommended by the Microarray Quality Control consortium (which found that combining a non-stringent significance cut-off with log-fold change ranking generates gene lists of higher reproducibility compared to methods such as p-value based ranking[130–132]), this significance cut-off is combined with further gene selection based on log-fold change magnitude. In Chapters 3 and 5, the top e.g. 100 most significantly differentially expressed genes are considered; in Chapter 4, a combination of log-fold change magnitude and signalling network interactions is used to select ‘interesting’ genes.

2.2 IDENTIFICATION OF SUITABLE MICROARRAY EXPERIMENTS

This work relies entirely on the re-use of publicly available microarray data. I used three main criteria for selecting suitable experiments from the Gene Expression Omnibus (GEO) repository:

- The study must include (healthy) controls, in order to calculate differential expression. The definition of a control varies from study to study: in many studies, the control samples are from different individuals, but there are also studies where the control samples are non-affected tissues from the same individual (e.g. acne-affected skin vs non-acne-affected skin); in some studies, both contrasts are provided (e.g. lobular breast tumour samples vs non-affected lobular tissue from lobular breast carcinoma patients or vs non-affected lobular tissue from ductal breast carcinoma patients). In the latter case samples from different individuals were used rather than 'non-affected' samples from the same patient, in order to avoid bias resulting from non-phenotypic gene expression changes that might be present even in apparently unaffected tissue.
- There must be at least two samples per condition. Whilst larger sample sizes are desirable for increased statistical power and reduction of biological noise, in the setting of large-scale comparative analysis the inclusion of a greater number of possibly noisy studies is preferable to fewer, more reliable studies, in order to increase the coverage of diseases in the dataset.
- There must not be another study in the dataset submitted by the same investigator, in order to minimize the chance of overlap between datasets resulting from technical factors. For Chapter 3 this criterion was relaxed to include only those studies submitted within a year of each other, in order to include as many Parkinson's disease studies as possible. This criterion was not applied to the drug response datasets used in Chapter 4, as they are not compared against other studies.
- Following the work described in Chapter 3, an additional condition in building the datasets used in the subsequent chapters was that the diseases must be recorded in human patients, rather than in animal models, and where possible they must be from whole tissues, rather than from cell lines. A few cell line studies were included where no patient tissue studies were available, i.e. in rare disease studies.

The Parkinson's disease dataset used for Chapter 3 is detailed in Appendix A.

The large dataset of common and rare diseases used for Chapter 4 is detailed in Appendix B.

A smaller dataset containing common diseases was used in Chapter 5, this is detailed in Appendix C.

Metadata recorded from each experiment included the disease, the submitter and institution name, the microarray platform, the tissue sampled, the number of cases and controls, and the samples which were included/excluded. Sample selection is necessary for studies which cover a number of different conditions, such as comparing two different types of arthritis to healthy controls; in this case only the relevant conditions were retained, and the other samples excluded from the analysis.

An early version of the disease dataset was based on the work of Yasaman Kalantar Motamedi on text-mining of GEO records to identify suitable experiments. However, as many of these experiments did not pass the above inclusion criteria, I chose instead to base my dataset on manual searching, which enabled me to find a higher number of datasets whilst applying strict quality control. Those datasets discovered by her text mining approach that did pass the quality criteria after manual checks were retained.

2.3 Development of an Automated Workflow for Processing of Raw Microarray Data

The processing of large numbers of datasets requires a standardized system to convert raw CEL files from GEO into differential expression profiles with minimal manual input. Given the large variety of microarray types (some of which occur only once in the dataset) and the need to specify e.g. the correct case-control designation of samples, a fully automated system for microarray data processing is not possible. With these limitations in mind, I constructed a workflow that requires minimal input for most cases.

The first part of the workflow (steps 1-4) takes the raw CEL files and calculates log-fold change and significance metrics for every probe on the microarray. The second part of the workflow (steps 5-6) maps the probe-level data to their corresponding genes, in order to compare data across different microarray platforms. All analyses were carried out in R version 3.3.2 running under OS X 10.11.6 (El Capitan)[133].

The protocol is as follows:

1. Download the raw CEL files from GEO. Delete any files that correspond to samples to be excluded.

2. Specify the design matrix corresponding to that particular experiment. This indicates (in order of filename) which files are cases, i.e. disease samples, and which are controls, i.e. healthy samples. The controls are designated as the reference (represented by 0), so that a positive log fold change indicates a gene that is more highly expressed in cases (represented by 1) than in controls.
3. Read in the CEL files and rma-normalize them using functions *ReadAffy* and *rma* from the package affy[134] (version 1.52.0).
4. Determine probe-level statistics (including log-fold change, p-value, adjusted p-value) using the functions *lmFit*, *eBayes*, and *topTable* from the package limma[53] (versions 3.26.7-3.30.13 depending on when the analysis was carried out). Statistics for all probes are retained by setting the parameter *number* = "LNF" (equivalent to setting to Inf).
5. Retrieve the appropriate platform annotation file from GEO using the *getGEO* function from the GEOQuery package[135] (version 2.40.0). Match the probe IDs to their corresponding gene ID and symbol (using code from the online differential expression service integrated with GEO, GEO2R[136]).
6. Where more than one probe maps to a gene, retain the probe with the smallest p-value[129]. Where a probe maps to more than one gene, duplicate the probe record, matching its information to both genes. Remove any probes that do not map to a gene, as these are non-informative for cross-platform comparison.

This describes the basic workflow constructed for Affymetrix arrays, which were the most common array type encountered in my dataset. Variations of the workflow for other microarray types include the following:

- Certain Affymetrix ST arrays cannot be processed by the *affy* package, in which cases the *oligo* package[137] (version 1.38.0) was used in step 3.
- For experiments that used Illumina platforms, the submitter-supplied non-normalized data was obtained from GEO, and step 3 was replaced by log-transformation and quantile-normalization (for consistency with the steps used in the Affymetrix-specific RMA normalization method).
- For the few experiments which used other platforms such as Agilent arrays, or where raw data was not provided, submitter-supplied normalized data was processed using GEO's web service GEO2R[136] in place of steps 1-5.

Additional processing steps that can be applied to further reduce noise in microarray data include array quality-checking to remove aberrant or outlier arrays; variance filtering to exclude those genes which show low variance across all experiments; and batch correction to account for the effects of sample handling and processing across different experimental batches. Batch correction was not considered for this analysis, as GEO records do not generally supply experimental batch information. It is possible to use the microarray scan date as a potential source of batch effects, but given the potential for sample clustering across these batches (e.g. all disease samples scanned on one date, all controls scanned on another), batch correction risks normalizing out biological signal and is therefore not appropriate in this setting.

Variance filtering and array quality checking methods were tested on the dataset described in Chapter 3, as this was the first piece of work undertaken and provided a simple metric with which to test the utility of various methods in the comparative analysis setting: a 'useful' method should increase the concordance between two studies of the same disease. Variance filtering was carried out using the package genefilter[138] (version 1.56.0) and array quality checking was carried out using the package ArrayQualityMetrics[139], version 3.30.0. These were found to make almost no difference to the observed concordance between studies of Parkinson's disease. Variance filtering was therefore not applied, in order to retain the maximum number of genes for comparison between experiments; the results of array quality checking were retained for the Parkinson's disease dataset (see description in Section 3.2.2) but were not applied to the larger dataset, as with such a large volume of datasets to process this would involve a significant amount of manual work for potentially very little benefit. Whilst these steps may be valuable for the analysis of individual datasets, for these reasons I decided not to apply these steps in the comparative analysis setting.

3 CONCORDANCE OF MICROARRAY STUDIES OF PARKINSON'S DISEASE

This work was previously published as Oerton E, Bender A. Concordance analysis of microarray studies identifies representative gene expression changes in Parkinson's disease: a comparison of 33 human and animal studies. BMC Neurol. 2017;17(1):58. doi:10.1186/s12883-017-0838-x.

All analyses, text, and figures were produced by the author, incorporating comments from co-authors.

SUMMARY

The reported lack of concordance between transcriptomic studies of the same condition raises questions about the representativeness of different study types, such as studies of surrogate tissues or animal models, to gene expression in the human disease. In a comparison of 33 microarray studies of Parkinson's disease, correlation and clustering analyses were used to investigate concordance between studies, including agreement between different tissue types, different microarray platforms, and between disease models and human Parkinson's disease.

Concordance over all studies is low, with correlation of only 0.05 between differential gene expression signatures on average, but increases within human patients and studies of the same tissue type, rising to 0.38 for studies of the substantia nigra region of the human brain. Studies of the substantia nigra in Parkinson's disease patients form a distinct group, showing patterns of differential gene expression noticeably different from that in non-brain tissues and animal models of Parkinson's disease. A meta-analysis of these 33 microarray studies demonstrates the greater ability of studies in humans and highly-affected tissues to identify expression changes in genes previously known to be associated with Parkinson's disease.

The observed clustering and concordance results suggest the existence of a 'characteristic' signal of Parkinson's disease found in significantly affected tissues in humans. These results help to account for the consistency (or lack thereof) so far observed in microarray studies of Parkinson's disease, and act as a guide to the selection of transcriptomic studies most representative of the underlying gene expression changes in the human disease.

3.1 INTRODUCTION

Parkinson's disease (PD) – a neurodegenerative disorder which causes the death of dopaminergic neurons in the substantia nigra, causing tremors and postural instability – has been well-studied at the level of gene expression, with numerous microarray studies available in public repositories. However, the concordance of differential gene expression between these studies has been reported to be low, even when standardized analysis is applied[140–143]. The observed discordance may result from multiple factors, including differences in the progression of the disease at time of post-mortem[142] and differing amounts of neuronal loss between the substantia nigra (SN) and other regions of the brain. Several meta-analyses of PD gene expression in human patients have been carried out[73,140,143] on datasets of up to 14 unique studies. Although meta-analyses generally focus on the commonalities between studies (in order to identify the genes most relevant to the condition under study), meta-analysis approaches can also be used to shed light on inconsistencies between studies. For instance, one such analysis of 11 human PD microarray studies highlighted tissue-specific differences between studies, demonstrating increased convergence within studies using samples from the SN[140].

Also demonstrated by an early microarray study of PD[144] is the difference between animal models of PD (reviewed in Blesa et al.[145]) and the human condition, which is of much practical relevance for therapeutic research. These models were developed to mimic the clinical symptoms of Parkinson's disease, and it is unclear to what extent the underlying patterns of gene expression will reflect those that take place in human PD. Studies comparing disease models to human patients have reported conflicting results: one study examined the consistency of gene expression between a mouse model of colorectal liver metastasis and human specimens, and found an overlap of 35% of differentially expressed genes, as opposed to 44% in normal liver tissue[146]. Another study of mouse models of inflammation found little transcriptomic agreement between human inflammatory conditions and their murine counterparts[147], although a re-analysis of this data using different statistical methods questioned this conclusion[148]. As the use of transcriptomics becomes more prevalent in medicine and drug development, it is important to establish whether gene expression in a model system can be treated as a proxy for gene expression in the human condition.

Choice of microarray platform is another factor that can affect concordance between studies. Notably, although some studies have reported good cross-platform reproducibility[117,149], an

early study of a mouse model of PD found very little concordance between Affymetrix and CodeLink platforms[150]. More recent studies in psoriasis[151] and in healthy tissues[119] still found detectable platform biases, indicating that this issue will not necessarily be resolved by the use of newer microarray technologies. The effect of sample size on study concordance should also be considered. Multiple studies have found that larger sample sizes in microarray experiments allow greater confidence in calling differentially expressed genes and produce more robust differentially expressed gene lists[130,152,153], but the effect of sample size in the context of average concordance across different datasets – i.e., the likelihood of being an unrepresentative 'outlier' study – has not been examined directly. This question is particularly important in the study of neurodegenerative diseases such as PD, given that large numbers of high-quality brain tissue samples are not always easy to obtain[154,155].

The concordance between different studies of the same condition will act as a measure of 'representativeness' of the recorded gene expression to true human PD, helping to establish whether animal models of disease are representative of the human condition at the transcriptomic level, and whether gene expression in more easily accessible surrogate tissues could be useful in PD research or diagnostics[156]. In this chapter, the effects of four factors – species, tissue, platform, and sample size – are analysed in relation to the observed inconsistency between microarray studies of PD. As well as the specific findings related to PD, the general findings from this work will serve as a basis for study selection in the datasets used for later chapters.

3.2 METHODS

3.2.1 Obtaining Parkinson's disease microarray studies

GEO was searched for suitable case-control studies of Parkinson's disease using combinations of PD keywords, i.e. "Parkinson's"/"Rotenone"/"MPTP" [rotenone and MPTP are neurotoxins used to model PD in animal studies] AND "homo sapiens"/"mammals"/"primate", using studies submitted up to February 2017.

Inclusion/exclusion criteria were as follows:

- Studies must be designed specifically for the investigation of PD or PD drug treatment.
- At least two Parkinson's disease (or equivalent model) samples and two healthy (wild-type/vehicle injected) control samples must be available for each condition.
- Gene expression must be measured using microarray technology, as too few studies are currently available on GEO using other methods of expression profiling (e.g. RNA-Seq) to be able to draw any conclusions about their use in PD.
- Human stem cell studies must be derived from PD patients and not just modelled by PD-associated mutations, in order to be comparable with human PD; equivalently, stem cells derived from PD patients compared to mutation corrected controls were excluded.

In order to minimise the impact of possible laboratory effects on concordance results, where multiple datasets were contributed by the same investigator and less than a year apart, only one of the two was retained (with the exception of two studies submitted as part of a meta-analysis that did not state whether the studies originated from the same experimental group, see Appendix A). Similarly, if a single study analysed multiple tissues, only one tissue was retained for analysis. In both cases, the retained study was chosen in order to provide the most balanced dataset; i.e. the most even split between tissues.

This gave a total of 33 PD studies. Four studies of Alzheimer's disease and five studies of brain tumours (glia- and astrocyte-derived) were included as disease controls, giving a total of 42 studies (see Appendix A). These studies were analysed in Section 3.3.6.

3.2.2 Processing of datasets

Following pre-processing and generation of a differential gene expression profile as described in Section 2.3, array quality was assessed using the ArrayQualityMetrics package[139], version

3.30.0; any samples which failed more than one of the three outlier tests (distances between arrays; boxplots; MA plots) were removed.

In order to make comparisons between gene expression in different species, all non-human studies were mapped to orthologous human genes using annotationTools 1.44.0[157]. As stated in Section 2.3, where a probe was associated with multiple genes, the probe information was retained for both genes in order to maximise the number of genes available for comparisons between different platforms, and it should be noted that this could artificially inflate concordance between studies, especially for those using the same platform.

3.2.3 Biological pathway enrichment

Biological pathway enrichment profiles were calculated from the differential gene expression profiles (generated above) against the Reactome pathway database with the *GSEA* function of the Bioconductor package ReactomePA 1.14.4[71], using the default settings of 1000 permutations to calculate significance and a minimum geneset size of 10. For mouse and rat studies, the original (animal) genes were used to calculate enrichment profiles using species-specific pathways provided by Reactome.

3.2.4 Calculation of pairwise concordance of differential gene expression

The 'agreement' between two microarray studies can be measured in many different ways, including comparison of lists of genes which are differentially expressed according to some cut-off (which can be published lists, or lists created by standardized analysis of published data)[140,141,144], comparison of ranked expression values (e.g. Spearman correlation)[111,158], and agreement of sign and/or magnitude of measured gene expression (e.g. Pearson correlation)[151,159], either over all measured genes, or over those defined as significant by some cut-off. These are reviewed in a 2009 paper by Lu *et al*.[160].

In this analysis, concordance between studies is defined as the Pearson correlation (as calculated by R's *cor* function[161]) of their differential gene expression *signatures*: an expression signature is here defined as the 50 genes showing the highest absolute log-fold change at a significance of $p<0.05$ in each study, from the set of 2,372 genes recorded by all 33 PD studies, or 2,310 over all 42 studies of brain disease. Similar concordance results were obtained when the expression signature was defined over 20, 100, or 250 genes for each study;

a value of 50 was chosen in order to capture the most relevant information while keeping the dimensionality relatively low (important in the following analyses). If correlation was calculated over the sign of the log-fold changes (i.e. considering only the direction and not the magnitude of fold changes), similar results were obtained; concordance in the SN was somewhat reduced from 0.3 to 0.22, but was still the highest-concordance tissue type, and so the measured log-fold change magnitudes were used in order to retain information.

3.2.5 Calculation of pairwise concordance of biological pathway enrichment

At the biological pathway enrichment level, pairwise concordance c_{ij} between two studies was defined as the Pearson correlation of the normalized enrichment scores (NES) of pathways that are significantly up- or down-regulated (at an FDR <0.25, as recommended by the Broad Institute's GSEA page[162]) in either experiment. In the case where a pathway is significant in one experiment but there is no score reported in the other, a NES of 0 was assigned for the missing pathway. If no significantly enriched pathways were reported for either experiment, the correlation was set to 0. The Pearson correlation is the most appropriate correlation measure to use given the high proportion of zeros amongst the normalized enrichment scores[163].

3.2.6 Calculation of average concordances within subsets of studies

The mean of the pairwise concordances c_{ij} of a study i with every other study j in a set of studies S gives a measure A_i of how well this study agrees with other studies in S. From the average agreement of each individual study, the average agreement A_S in a set can be measured (i.e. A_S is the average of each A_i). In this case, S is a subset of studies chosen to represent a particular factor of experimental design, such as the subset of studies using human specimens or the subset of studies run on a particular microarray platform, and the basis of this analysis is the comparison of A_S between these different subsets, specifically for subsets in which three or more studies share one of the experimental factors tissue, species, platform, or sample size. These four factors were chosen to analyse as they are nearly always specified in study meta-data, and such can be quickly determined in a meta-analysis context.

Note that smaller subsets may have larger numbers of shared genes, (e.g. due to sharing a platform which measures the same genes). Concordance over smaller subsets was calculated on the same expression signatures as for the set of all studies, i.e. expression signatures selected from the shared 2,372 genes, in order to ensure that A_S was not biased by the size of shared gene-sets in different subsets. If concordance was calculated over the full set of genes shared by each subset, results were not substantially different (see Appendix F).

Significance of average subgroup concordances was tested against the average concordances over randomly sampled subgroups of the 33 PD studies (to a maximum of 100,000) of each size. An observed average correlation is defined as significant if it is greater than the 95th percentile value. The smaller the subgroup size, the more likely that randomly chosen subgroups show high concordance by chance alone (the distribution of observed correlations is wider), and so the confidence threshold is higher for smaller subgroups (see Appendix E).

3.2.7 Principal component analysis and hierarchical clustering

Hierarchical clustering was performed using R's *hclust* function[164] using correlation distance. Correlation distance was chosen over the default Euclidean distance because it is not affected by scale (e.g., differences in average log fold change magnitude across experiments)[165]. Significance of the observed clusters was calculated using the R package pvclust[166], which uses multiscale bootstrap resampling to approximate a probability value for each observed cluster (probability values quoted are the Approximately Unbiased values). Principal component analysis was performed using R's *prcomp* function with centering and scaling[167]. At the differential gene expression level, the feature vector for each study was defined as its log-fold change values over the gene-set defined by the union of the 50 highest-ranking genes (the union of expression signatures; i.e. the 50 genes in each study at a significance of $p<0.05$ with the highest absolute log-fold change in every study in the set) – for the PD studies, this is 1,008 genes. For hierarchical clustering, where high dimensionality affects the stability of clusters, this was reduced to the union of the top 10 highest-ranking genes, which for the PD studies is 258 genes.

3.2.8 Meta-analysis of Parkinson's disease microarray studies

A meta-analysis over the 33 PD studies was carried out using a 'vote-counting' approach in which a gene was deemed to be of importance in a study if it was in the top 50 genes by absolute log-fold change, at a significance of $p<0.05$. A gene was deemed to be significant by the meta-analysis if it was considered to be of importance by more than three studies. This threshold was chosen due to the low agreement between studies (see Results). The results of the meta-analysis were compared against a list of 694 PD-associated genes downloaded from the Centre for Therapeutic Target Validation (CTTV)[168] on 8th March 2016. Note that this resource was an early version of the OpenTargets platform[169] used in Chapter 4. The list includes genes identified by genetic associations, by PD drugs, and by text-mining; targets identified through reprocessing of previous RNA expression studies were excluded, as the studies used in the CTTV analysis could potentially overlap with those used here. Similar results (in terms of the proportions of genes identified by each subgroup) were obtained when the meta-analysis was carried out over the top 10 or top 100 genes instead of the top 50.

3.3 RESULTS

3.3.1 Higher concordance within human studies and within tissue groups

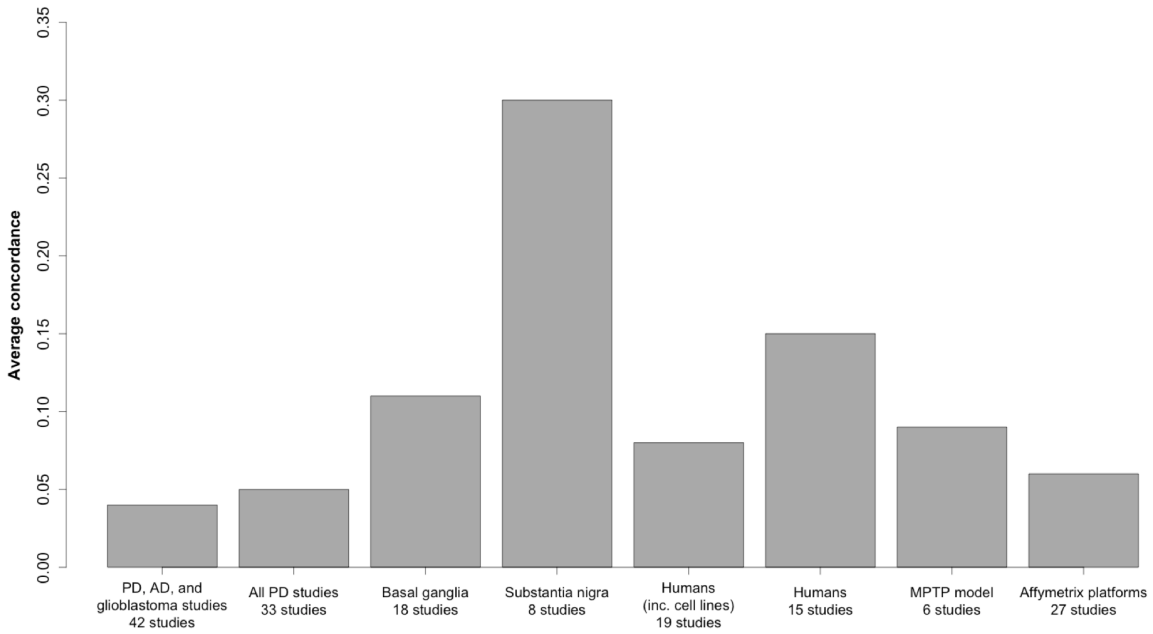


Figure 3.1: Average concordance of differential gene expression within subsets of shared factors

Average concordance over all studies is low, but increases within human patients and studies of the substantia nigra.

The mean average pairwise concordance of differential gene expression signatures (i.e. the Pearson correlation over the top 50 genes by absolute log fold change, see Methods) over all 33 Parkinson's disease studies was 0.05 (Fig 1), indicating little agreement between different studies. To identify how much of the observed inconsistency is due to differences in species, tissue, or microarray platform, concordance was examined within subgroups of studies that shared these characteristics (Table 3.1, Figure 3.1).

Table 3.1 Average concordance of differential gene expression signatures in microarray studies.

Asterisks indicate subgroups where concordance significant, i.e. is within the top 5% of concordance values over randomly sampled subgroups of PD studies. The threshold for significance varies with the number of studies in the subset (see Methods, Appendix E).

Subset	**Number of studies**	**Average concordance of expression signatures**
Whole dataset		
PD studies plus Alzheimer's disease and glioblastoma studies	42	0.04
All PD studies	33	0.05
Species		
Human (inc. human cell lines)	19	0.08
Human patients	15	0.15*
Mouse models	9	0.03
Rat models	4	-0.04
Disease model		
All neurotoxic models	12	0.03
MPTP	6	0.09
MPTP, mice only	5	0.10
6-OHDA	4	-0.03
Genetic models	3	0.12

Tissue		
Basal ganglia (SN (excluding isolated dopaminergic neurons) and striatum)	18	0.11*
SN: tissue	8	0.30*
SN: isolated dopaminergic neurons	4	0.03
Striatum	9	0.07
Platform		
Affymetrix	27	0.06
U133 and U133 Plus arrays (human studies only)	12	0.10

The first factor to be examined was species. The average concordance of differential gene expression signatures increased from 0.05 over all PD studies to 0.15 in human *in vivo* studies. In the subset of mouse studies, however, average concordance of differential gene expression decreased compared to the full dataset, at 0.03, and average concordance within the three rat studies was actually negative. This could be explained by the use of different disease models with distinct effects on gene expression: concordance within the MPTP and genetic models of PD was 0.09 and 0.12 respectively; although there was negative concordance between studies in the 6-OHDA group (Table 3.1).

The next factor considered (independently of species) was the tissue type sampled. Limiting the studies under consideration to those of an area highly affected in PD, the basal ganglia (here including studies of the striatum and the substantia nigra, which is functionally part of the basal ganglia), increased average gene-level concordance from 0.05 to 0.11, while further limiting the studies to just those of the substantia nigra yielded a substantial increase to 0.30 (Figure 3.1). This result is in agreement with a previous meta-analysis[140], which also reported an increase in concordance when the analysis was confined to studies of the substantia nigra. Concordance within striatal studies was lower than that over all tissues of the basal ganglia at 0.07; however, tissue selection was strongly associated with species, with substantia nigra

studies tending to be from humans (6 of 8 studies) and striatal studies tending to be from animal models (8 of 9 studies), and so the lower concordance within the striatal group perhaps reflects the general lower concordance between animal models. To deal with issues of species dependence in tissue choice and other experimental parameters, the following subgroup analyses focus on human studies.

3.3.2 High concordance of biological pathway enrichment in human PD

Given the low average concordance of differential gene expression, correlation was also calculated at the level of biological pathway enrichment (see Methods). As pathways are a higher-level biological concept, capturing concerted changes in gene expression, pathway enrichment might be expected to reveal higher concordance between studies, as previously shown in Sutherland *et al*.[140] Whilst the concordance values at the pathway level are not directly comparable to those at the gene expression level (due to the differing feature vectors used), biological pathway enrichment demonstrated relatively good agreement across human studies, from 0.22 over all human patient studies to 0.3 over studies of human brain tissue, indicating that measured differential expression reflects the activation of similar biological processes (Figure 3.2; see Appendix G for a list of significant pathways). In animals, in contrast to human studies, concordance at the pathway level was in most cases actually lower than that at differential expression level (Table 3.2). One reason for this may be incomplete annotation of non-human biological pathways in the database used in this study.

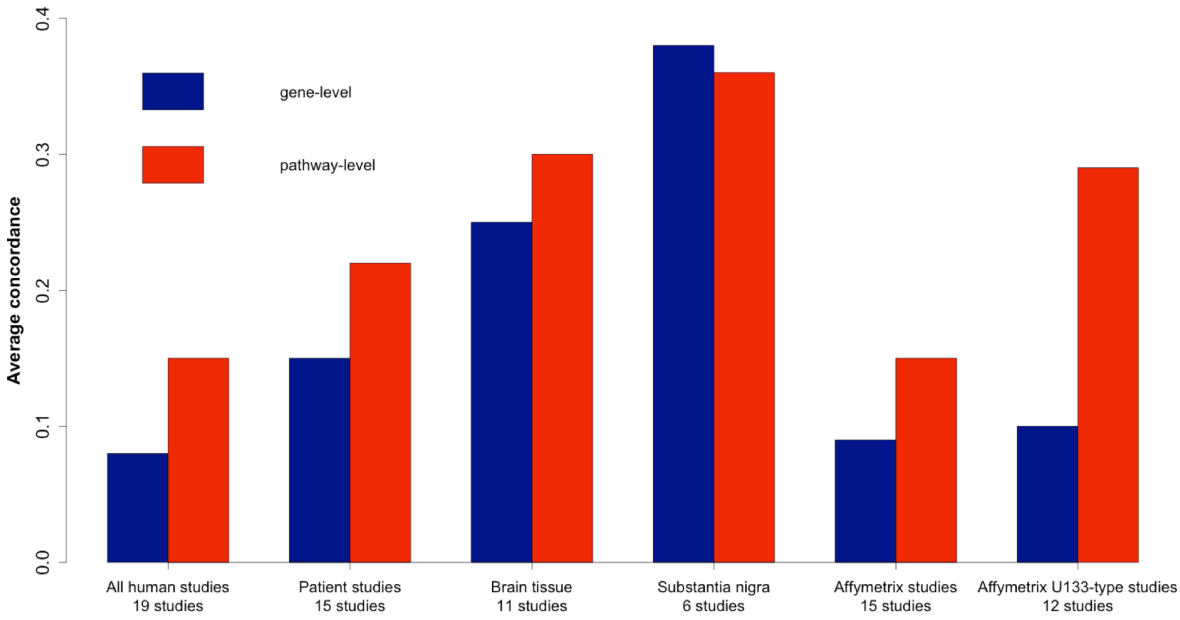


Figure 3.2 Average concordance within subgroups of human studies of PD

Gene- and pathway-level concordance increases in studies of human patients (i.e. excluding human cell line studies) and within tissue subgroups. The trend seen at the level of differential gene expression is replicated at the level of biological pathway enrichment, suggesting that subgroup-specific increases in concordance reflect enrichment of shared biological processes.

Table 3.2 Concordance results for different subgroups using biological pathway enrichment analysis

Pathway-level concordance in human studies reflects the patterns seen at the level of differential gene expression concordance; in animals, by contrast, pathway-level concordance is lower and does not increase by model type.

Subset	Number of studies	Average concordance of biological pathway signatures
Whole dataset		
PD studies plus Alzheimer's disease and glioblastoma studies	42	0.05
All PD studies	33	0.08
Species		
Human PD	19	0.15
Human PD, in vivo studies only	15	0.22
Mouse models	9	0.01
Rat models	4	-0.10
Disease model		
All neurotoxic models	12	0.02
MPTP	6	0.03
MPTP, mice only	5	-0.02
6-OHDA	4	-0.01
Genetic models	3	-0.02

Tissue		
Basal ganglia (SN (excluding isolated dopaminergic neurons), striatum, globus pallidus)	18	0.10
SN: tissue	8	0.24
SN: isolated dopaminergic neurons	4	0.01
Striatum	9	0.00
Platform		
Affymetrix	27	0.09
U133 and U133 Plus arrays (human studies only)	12	0.21

3.3.3 Microarray platform type has little effect on average concordance of human PD studies

The next factor examined was the effect of microarray platforms, which are intended to be species-specific (one macaque study run on the U133A platform was excluded from this analysis). There was a very slight concordance increase when selecting for platform types, from 0.08 over all 19 human studies to 0.09 over all Affymetrix platforms (15 studies) and 0.10 for those studies run on the most common platform types in the dataset, the Affymetrix U133A and U133 Plus 2.0 series (12 studies). It should be noted that although the U133 microarrays are distinct platforms, they are technically very similar, as the probe set of the U133A arrays represents a non-random subset of the U133 Plus 2.0 arrays[151], and so are considered as a single platform type for the purpose of this analysis. At the pathway level, the concordance increase within the U133 subgroup was much larger (Figure 3.2), and this may reflect the effect of a shared probeset in calculating pathway enrichment profiles, as the gene-set enrichment used here takes into account the expression of every measured gene.

3.3.4 Smaller PD studies do not show lower concordance of differential gene expression

The next factor examined was the study sample size. When the smallest 25% of human studies (five studies with sample sizes of less than 10) were excluded, concordance within the remaining larger studies increased slightly from 0.08 to 0.11 at the differential gene expression level and from 0.15 to 0.17 at the pathway level. Linear regression was then used to test whether there was an overall association between sample size and average concordance across all (human) datasets. There was no significant relationship between sample size (case plus control) and average concordance of differential gene expression signatures or of biological pathway enrichment (Figure 3.3), suggesting that smaller studies are no more likely to be discordant 'outlier' studies than larger studies.

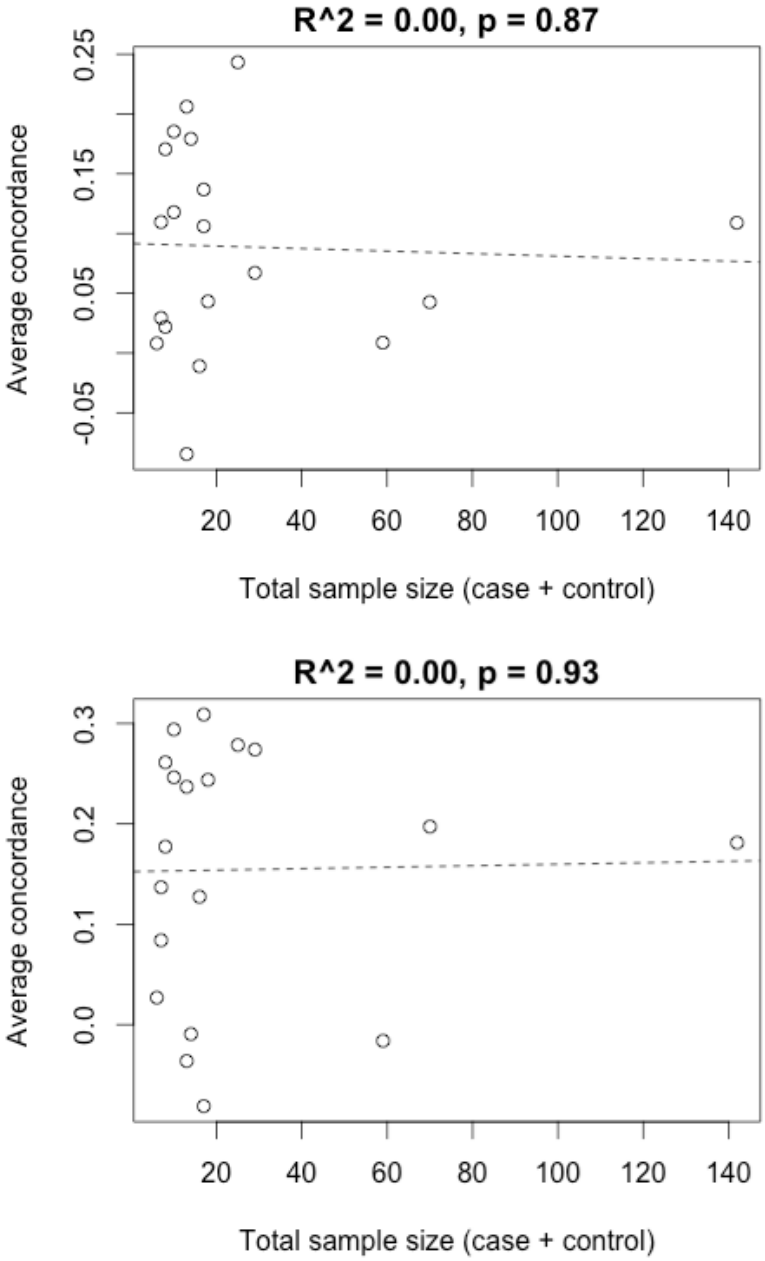


Figure 3.3 Average concordance against sample size for gene expression and biological pathway enrichment

There is no significant effect of sample size on average concordance of gene expression (top) or biological pathway enrichment (bottom).

3.3.5 Visualizing the gene expression landscape of PD studies reveals a distinct subset of human studies

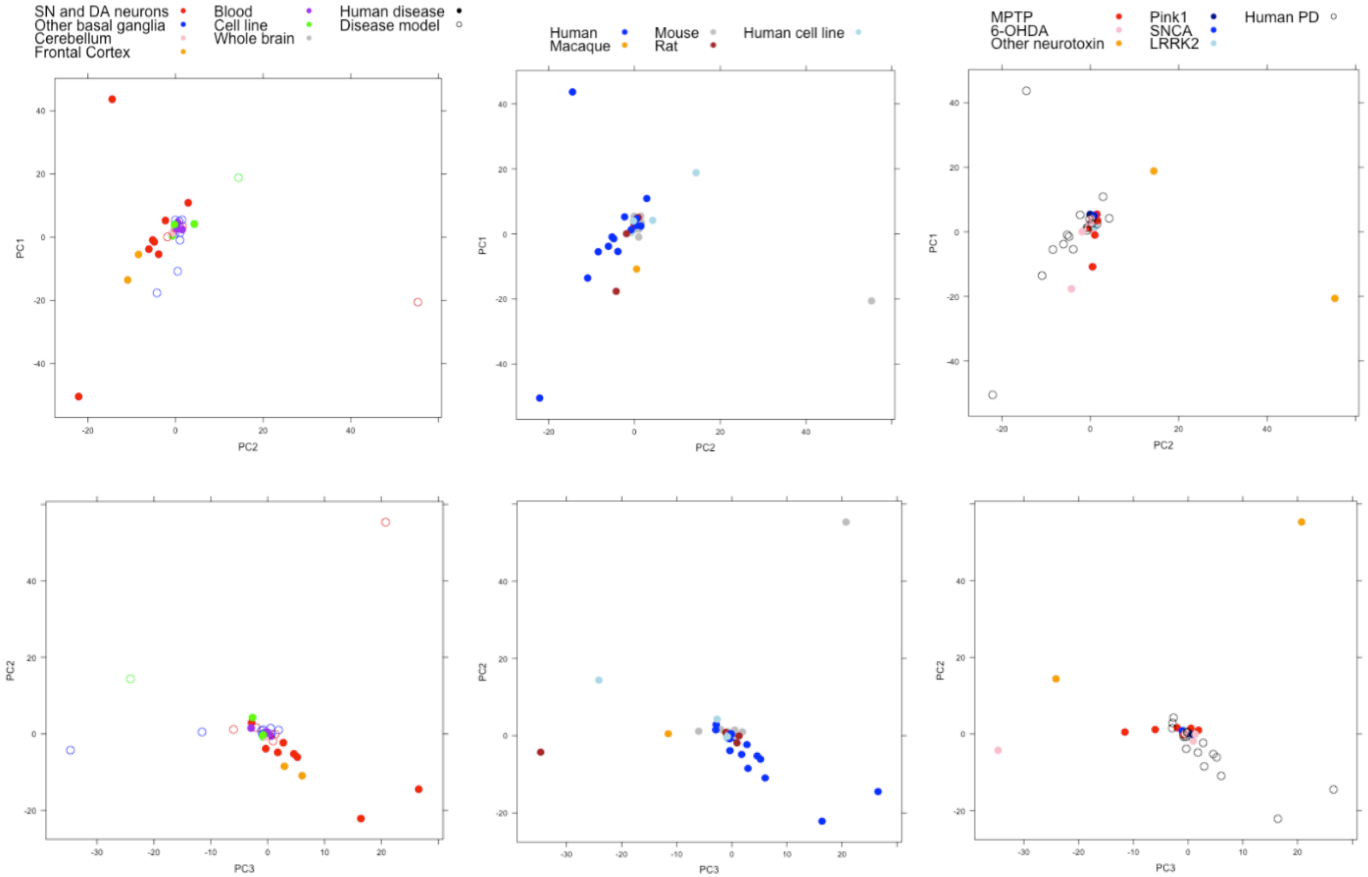


Figure 3.4 Principal component analysis of PD studies based on differential expression across the union of expression signatures

PCA of the 1,008 genes in the union of expression signatures across all 33 studies reveals a distinct group composed mainly of human studies (centre plots) of the substantia nigra and frontal cortex (left plots). This is most clearly seen in the second and third principal components (bottom row, top row displays the first and second components). There appears to be little separation between different disease model types (right); although the two studies using neurotoxins other than MPTP (rotenone and co-exposed maneb-paraquat) appear very distinct

The relationships between studies in differential gene expression space (here defined as the 1,008 genes in the union of expression signatures across all studies; see Section 3.2.7) were visualised using principal component analysis (PCA, Figure 3.4). PCA enables representation of the 1,008-dimensional expression signature space in a lower-dimensional space which captures the greatest amount of variance amongst studies[170]. The visualization of samples in this space shows an outlying group of human studies which appear distinct from other human and animal studies (Figure 3.4). This is most clearly seen in the second and third principal components, which together with the first component represent 44% of the variance.

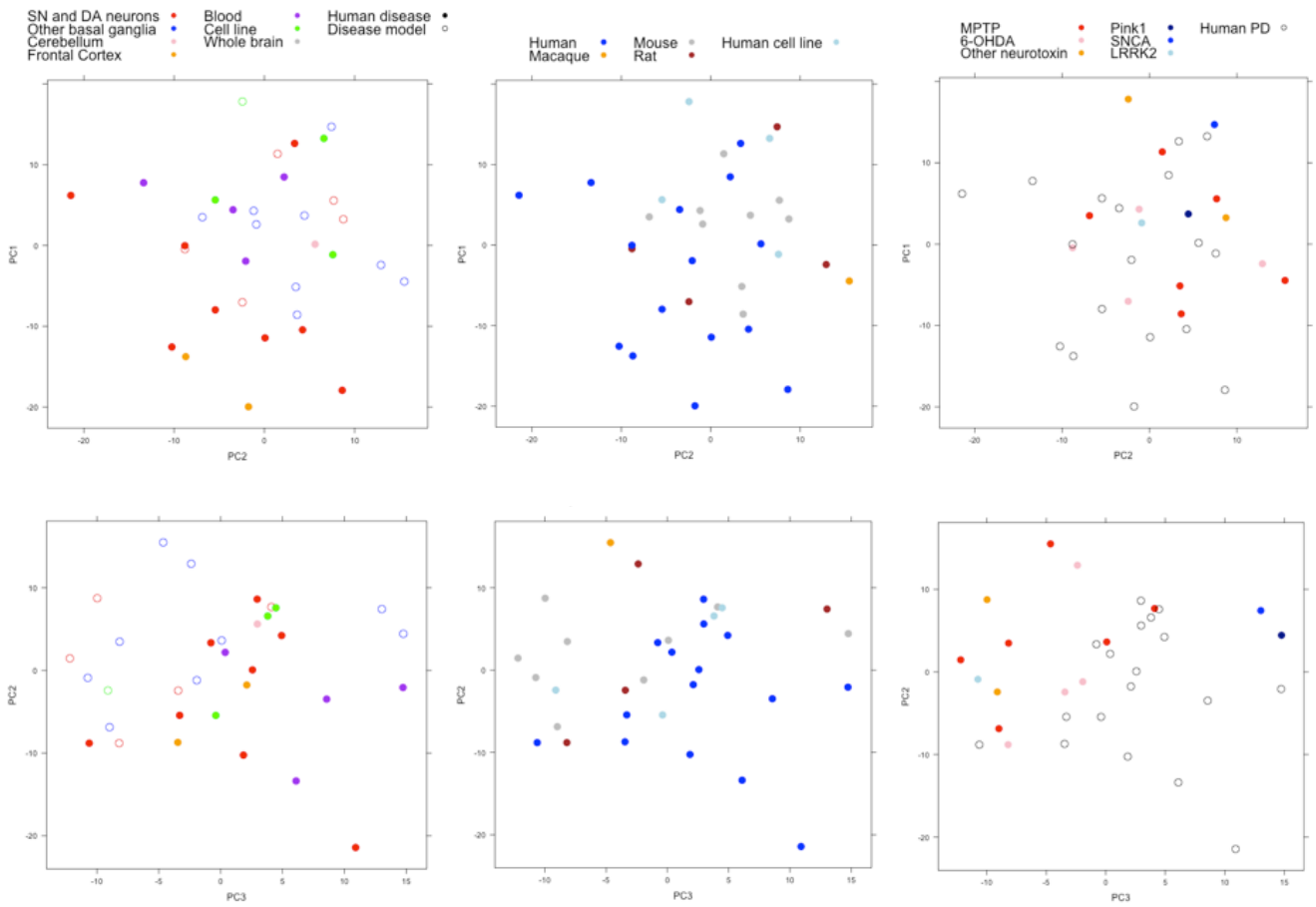


Figure 3.5 Principal component analysis of PD studies based on the sign of differential expression across the union of expression signatures

If PCA is carried out only on the sign of differential expression, ignoring the magnitude, separation between human and animal studies can still be seen, particularly in the second and third principal components (bottom right plot), although there is still some overlap between the two groups. Plots of the first and second principal components have been included for comparison, and here the separation of studies seems to also reflect tissue type in human studies (top left plot).

The principal component plots in Figure 3.4 show several studies which are outliers in principal component space, which may result from high average log-fold change magnitudes. One way to address this is to perform PCA on the sign of the differential expression signatures, discarding the magnitude information. These plots should be interpreted with caution, as they force the assignment of directionality to even very small gene expression changes, but the advantage is that the outlier effect is removed, allowing clearer visualisation of the separation of studies by tissue type and species, which is most clearly visible in the third principal component (Figure 3.5).

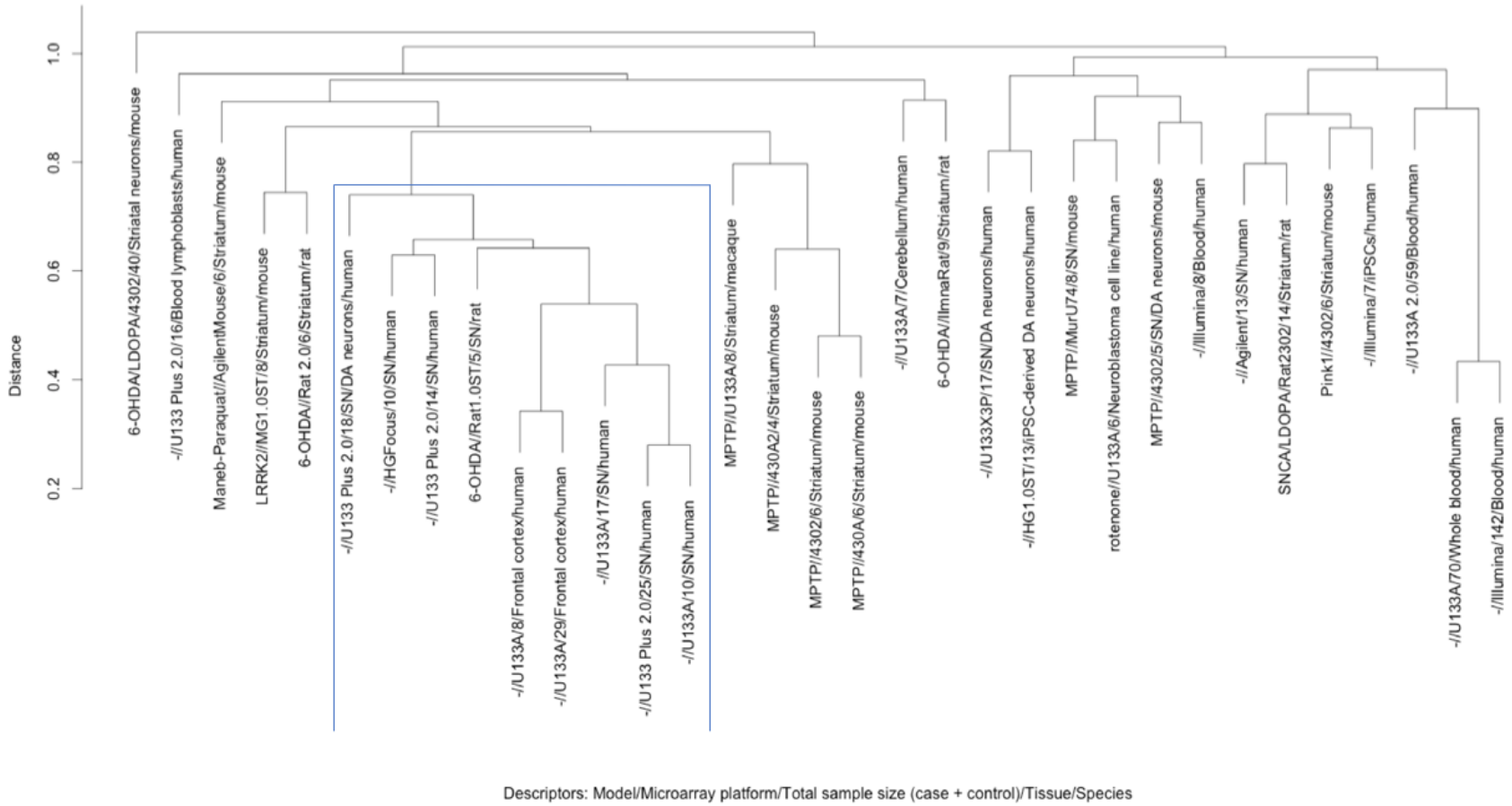


Figure 3.6 Hierarchical clustering of studies based on the most highly differentially expressed genes in each PD study
Clustering was performed based on differential expression of the 258 genes in the union of the top 10 genes (by absolute log-fold change) across the 33 studies. The highlighted cluster contains all but one of the human studies of the substantia nigra, as well as both human frontal cortex studies. This indicates that a distinct differential expression pattern is shared by these study types. However, this cluster also contains one rat study, indicating that it is possible for animal models to capture the expression patterns observed here. Aside from this group, there is no apparent clustering of other factors such as platform, disease model, or treatment (e.g. with L-DOPA), reflecting the low concordance seen in these groups.

To further examine the distinct group of human studies seen in Figure 3.4, hierarchical clustering was performed over the 258 genes in the union of the top 10 most differentially expressed genes over all 33 studies (Figure 3.6; see Appendix H for list of 258 genes). This shows a distinct cluster composed mainly of human studies of the substantia nigra (the most highly-affected tissue in PD) and studies of the cerebral cortex (SFG and PFC-Brodmann area 9)[171,172] (which are also affected in PD, although the cortex is affected at a later stage of disease[173]). The bootstrap probability value of the highlighted cluster (see Section 3.2.7) is 0.99, indicating that this cluster remains highly stable under resampling of the dataset. A heatmap of the differential expression signatures (Figure 3.7) reveals that studies in this cluster share downregulation in a set of genes which are enriched for the Panther pathway 'synaptic vesicle trafficking' (the pathway enrichment method used here is described in Section 4.2.6). It should be noted that a sixth study of the substantia nigra, which was run on an Agilent platform (all other studies were run on Affymetrix platforms), does not cluster with the others, showing a distinct differential expression pattern in which the majority of genes are up-regulated (Figure 3.7).

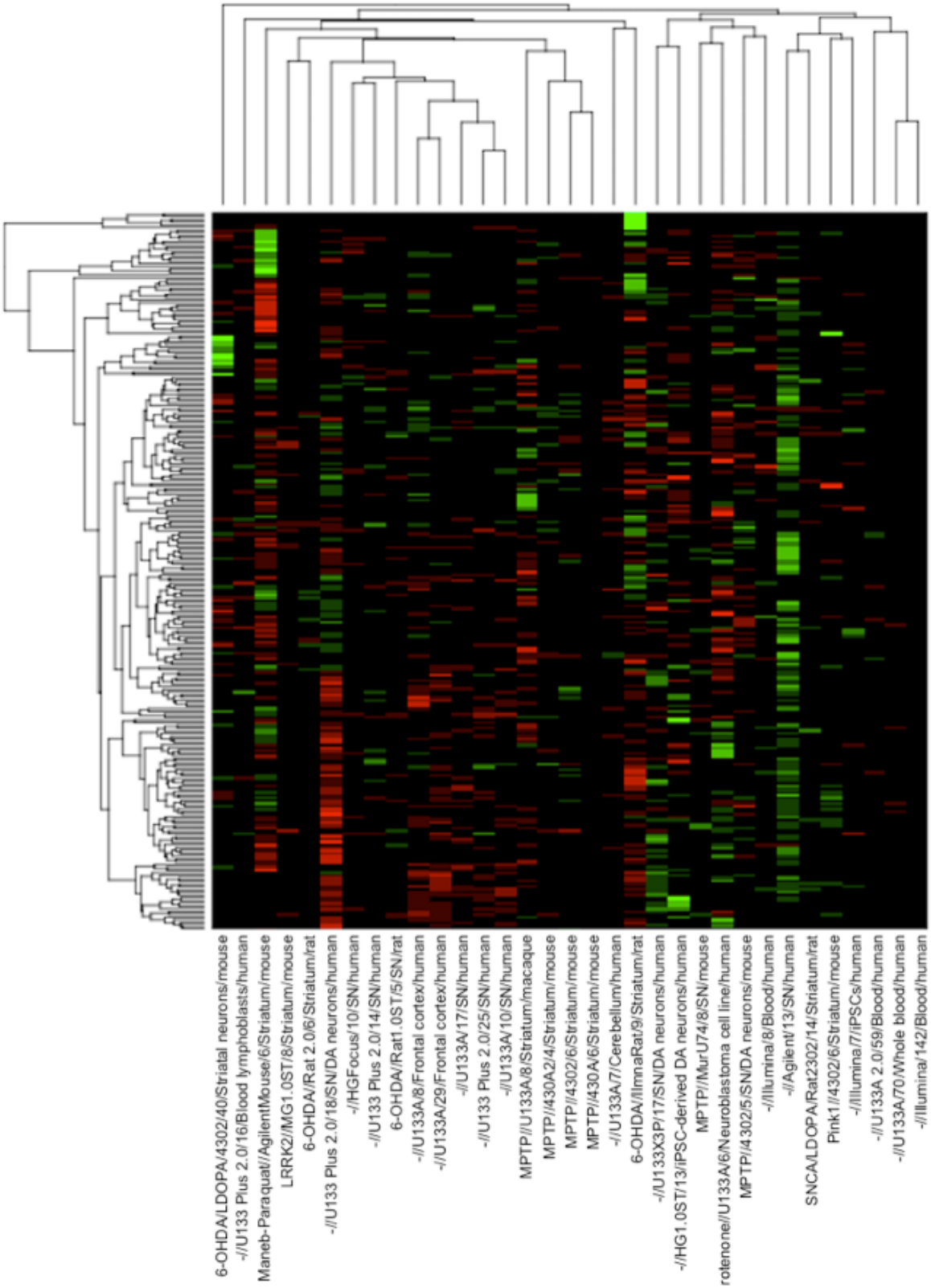


Figure 3.7 Heatmap of differential expression in Parkinson's disease studies

Green represents upregulated genes, red represents downregulated genes. The heatmap illustrates the differential expression patterns underlying the clustering shown in Figure 3.6, showing that the substantia nigra cluster of studies share downregulation in a group of genes towards the bottom of the plot (see Appendix H for row names).

The clustering in Figure 3.6 uses average linkage; when complete linkage is used, the six SN studies form a cluster on their own (bootstrap probability value 0.96), indicating that there are also expression patterns which are specific to the SN and not shared by the frontal cortex or dopaminergic neuron samples.

Other clusters that can be seen include 4 of the 6 MPTP models of PD, 3 of the 4 studies in blood, and clustering of iPSC studies with the appropriate tissue (dopaminergic neurons) or model (genetic animal models), although bootstrap probability values of these clusters are less than 95%, indicating a less stable clustering. Otherwise, there is no clear effect of any factor (such as microarray platform or treatment with L-DOPA) on study distribution within the clustering, reflecting the low concordance seen in these groups. Concordance in microarray studies of PD may therefore be partly explained by the different gene expression signals present in studies of human brains and in studies of peripheral areas or animal models.

3.3.6 Differential gene expression in human tissues highly-affected in PD is distinct from other brain diseases

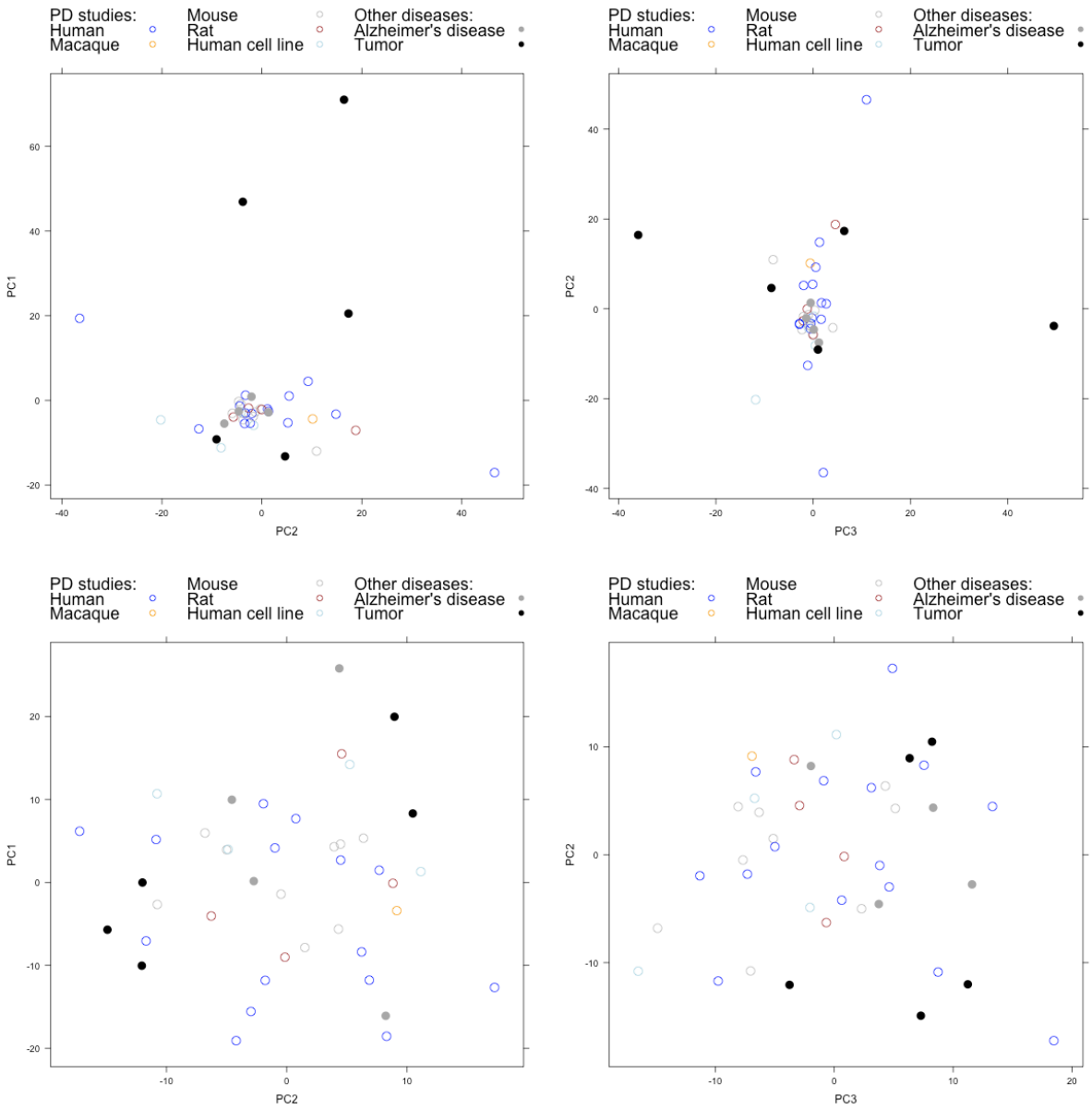


Figure 3.8 Principal component analysis of differential gene expression in Parkinson's disease, Alzheimer's disease and brain tumour studies

Top row: PCA based on differential expression over the 1,152 genes in the union of 42 expression signatures (including AD and tumour studies). Bottom row: PCA based on sign of differential expression (discarding the magnitude) over these 1,152 genes. The tumour studies appear highly distinct from AD and PD studies in the principal component representation of gene expression space. Even if only the sign of differential expression is taken into account (bottom row), the tumour studies appear at the outer edges of the second principal component, suggesting highly distinct patterns of gene expression in the two disease groups. By contrast, the four AD studies are not separated from the PD studies, illustrating that the variation between tumours and neurodegenerative diseases is much higher than that between the two neurodegenerative diseases.

In order to examine the disease specificity of gene expression in PD, PD studies were compared with studies of other diseases – namely Alzheimer's disease (AD), a neurodegenerative disorder which can present similar pathology to PD[172], and brain tumours (glioma), which are clinically unrelated to PD. As before, PCA was used to provide a low-dimensional visualisation of the separation of samples in differential expression space (Figure 3.8); the first three principal components here represent 42% of the variance. When magnitude of differential expression is taken into account, the cancer studies vary mostly across the first and third principal components, whilst the AD and PD studies vary across the second principal component, suggesting highly distinct patterns of gene expression in cancer compared to in the neurodegenerative diseases.

When only the sign of differential expression is taken into account (removing variation due to different magnitudes of expression changes, in order to focus on the general patterns of regulation), the cancer studies still appear distinct from the AD and PD studies, here showing the greatest variance across the second principal component. Interestingly, the cancer studies here split into two groups – one containing studies of human cerebellum, human whole brain, and mouse whole brain; the other containing studies of human blood and murine dorsal brain run on an Illumina platform – which appear at opposite edges of the second principal component. This means that when the magnitude information is discarded, there is greater variation within the cancer studies than between the cancer studies and the neurodegenerative diseases.

In all PCA plots, the AD studies appear less distinct from those of PD. The separation between human and animal studies seen in PCA of the PD studies only (Figure 3.4, Figure 3.5) appears reduced in this plot, illustrating that the variation between tumours and neurodegenerative diseases is much greater than the variation between human and non-human neurodegenerative diseases.

The relation between these studies can be further examined by clustering based on differential expression across the union of the top ten genes in all studies (as above). Interestingly, this plot shows one of the groups of tumour studies clustering near to the group of human substantia nigra studies (Figure 3.9). Examination of the heatmap reveals that their overall gene expression patterns are different, but that they share down-regulation in a cluster of genes at the bottom of the heatmap; pathway analysis reveals the Panther pathway 'synaptic vesicle trafficking' and GO biological process 'dopamine metabolic process' to be significantly enriched in these genes. A fourth tumour study taken from a blood sample clusters together

with all but one of the other blood-based studies on the outer edge of the heatmap, suggesting that blood displays overall different gene expression patterns from brain tissue, although the tumour blood gene expression study appears distinct within this group.

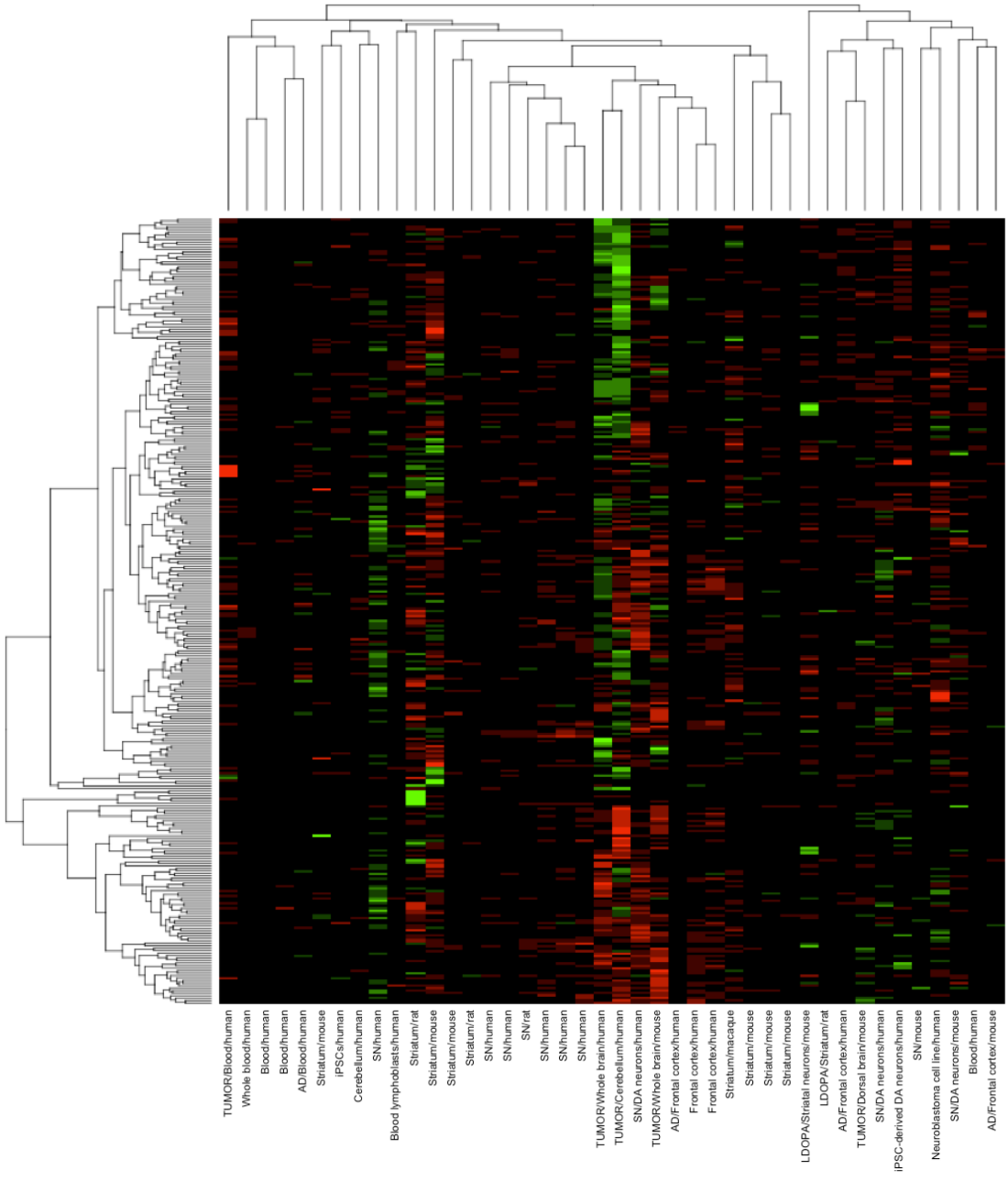


Figure 3.9 Heatmap of differential gene expression in Parkinson's disease, Alzheimer's disease, and brain tumour studies

The heatmap illustrates differential expression across the 315 genes in the union of the top 10 most differentially expressed genes over all studies (see Appendix I for rownames). The clustering reveals shared down-regulation between three of the tumour studies and the human substantia nigra studies in a group of genes related to dopamine transport (towards the bottom of the heatmap). Overall, however, the tumour studies appear distinct, with large differences in the magnitude of expression changes, particularly displaying strong upregulation of genes at the top of the heatmap. Distinct expression changes can also be seen in the cancer study taken from blood samples, which is clustered at the left of the heatmap together with other blood samples of PD and AD.

3.3.7 Inclusion of non-human and non-nigral tissue studies reduces the percentage of Parkinson's disease-associated genes identified in a meta-analysis

A key aim of this study is to determine whether gene expression in surrogate tissue (i.e. non-brain tissue) or in animal models of disease is reflective of gene expression in the brain of a human patient. In order to establish this, a meta-analysis was carried out across different subgroups of studies, where a gene was deemed to be significant if it was included in the top fifty most highly differentially expressed genes in more than three studies (this vote-counting methodology was chosen due to the low agreement between studies). The results of the meta-analysis were compared with a list of 694 potential PD-associated genes downloaded from the Centre for Therapeutic Target Validation[168]. These genes were selected on the basis of previous association with PD through genetic, drug target, or text-mining association (see Section 3.2.8) and represent numerous pathways including those involved in signal transduction (such as RAF/MAP kinase cascade and G alpha and AKT signalling events) and the immune system (such as interleukin-1 signalling and proteasome degradation).

Table 3.3 Genes highly differentially expressed in multiple Parkinson's disease studies

Table shows the number of times a gene is in the top 50 genes by absolute log-fold change in each study.

Gene	All studies	Human studies	Studies of the SN
Up-regulated			
HSPA1A	4	3	3
RELN	4	4	3
PTPRC	3	2	0
LCN2	3	0	0
PLIN4	3	0	0
MAFF	3	2	2
SLCO4A1	3	3	2

HSPA1B	3	3	3
IGF2BP2	3	0	0
CDKN1A	3	0	0
ENC1	3	2	1
Down-regulated			
EGR2	6	0	0
FOS	5	2	1
RGS4	5	5	3
TAC1	5	4	3
SLC6A3	4	3	3
AGTR1	4	4	3
FGF13	4	3	4
PCSK1	4	3	2
NPTX2	4	1	1
GABBR2	4	3	2
NR4A2	4	3	4
EIF1AY	3	2	2
SATB2	3	0	0
RET	3	1	2
SNCA	3	3	0

TTR	3	0	0
CCK	3	0	0
DDC	3	3	3
SLC18A2	3	3	3
ALDH1A1	3	3	3
KCNJ6	3	2	2
TMEM255A	3	3	3
SCG2	3	3	3
GPR26	3	2	3
DCLK1	3	2	0
DUSP1	3	2	1
HPCAL4	3	2	1
SYNGR3	3	3	2
PREPL	3	3	0
STMN2	3	3	2
VSNL1	3	3	2
NTS	3	2	3

The overall agreement in differentially expressed gene lists over all 33 studies was low, with no gene consistently differentially expressed in more than 6 studies (Table 3.3). Even if larger expression signatures including the top 100, 250 or even 500 most highly dysregulated genes (of a total 2,372 shared genes) were used, the findings were not much different, with no gene consistently regulated in more than 6, 8, and 10 studies respectively. The most common

findings shown in Table 3.3 include significant downregulations in genes including ALDH1A1, TTR, TAC1, and solute carrier genes SLC18A2 and SLC6A3, and upregulation of the heat shock protein genes HSPS1A and HSPS1B in multiple studies. This is consistent with the findings of a previous meta-analysis[140] of human datasets, who reported concordance as low as '20 genes… consistently differently regulated across 6 of 13 datasets', whilst cautioning that the downregulation seen in DDC and other genes could be the result of 'a disproportionate number of SN dopaminergic neurons between cases and controls'. Other findings include downregulation of FOS, which is more commonly associated with overexpression following L-DOPA treatment, in two animal (non-L-DOPA treated) and one human experiments. SNCA is also downregulated in multiple human studies, which previous studies have suggested may be related to long post-mortem intervals in PD cases[174].

Over all data sets, 26% of the 43 genes called significant by the meta-analysis (Table 3.3) were included in the list of previously PD-associated genes. If the meta-analysis was limited to human studies, however, 36% of the 22 significant genes had previous evidence of association with PD (Figure 3.10). The inclusion of non-human studies therefore reduced the enrichment of PD-associated genes in the list, i.e. the likelihood of each identified gene having a previously evidenced association with PD is lower. If the meta-analysis is limited to just animal models of PD, this was reduced to 10% of the 10 significant genes. There was a similarly noticeable difference between studies of different tissues. 32% of the 28 genes considered significant in a meta-analysis of the 18 basal ganglia studies (here including studies of the substantia nigra and striatum, excluding those which considered isolated dopaminergic neurons from the SN) had been previously associated with PD, and increasing to 40% when only substantia nigra studies were considered (Figure 3.10), suggesting that gene expression changes in these tissue types capture genes and gene products highly relevant to PD.

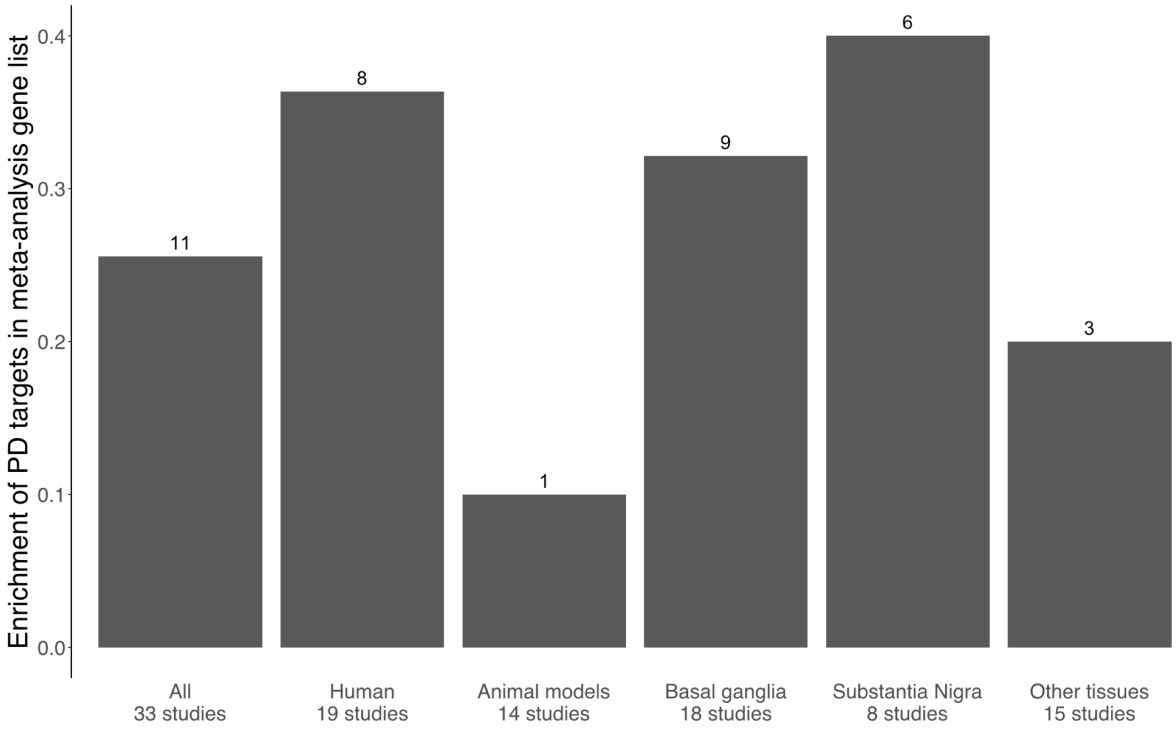


Figure 3.10 Percentage (bar) and number (number above bar) of potential PD targets amongst genes considered significant by a meta-analysis in each grouping

Gene lists from human studies and studies using tissue from the basal ganglia (here including studies of the striatum and substantia nigra) are more enriched for genes and proteins that have been associated with PD through genetic mutations, drugs, or literature-mining than those from animal models or studies using other tissues.

3.4 DISCUSSION

The low concordance between microarray studies of Parkinson's disease echoes recent concerns about the reproducibility of microarray studies between different labs[111,119,140,151] and between humans and animal models[144,146–148,175]. This study aimed to determine the major factors of study design influencing the observed lack of concordance.

The results presented here confirm that the differences between human studies and model systems, and between tissues, are larger than those caused by other experimental factors such as microarray platform or sample size (Figure 3.1, Figure 3.2, Figure 3.4). This analysis seems to indicate a split between human brain tissues and other study types (animal models and human studies of other tissues, including isolated dopaminergic neurons). It is possible that these human brain studies, particularly studies of the human substantia nigra, reflect a distinct 'characteristic' transcriptional signature specific to human PD; whereas the non-human studies and human studies of non-brain tissue reflect other, more general disease-associated molecular changes that take place in multiple tissues and systems (Figure 3.7). The inclusion in the 'characteristic' group of tissues affected later in the disease e.g. frontal cortex[173] (Figure 3.6) is noteworthy – given the progressive nature of PD, the late-affected tissues may display a signal of the early stages of neurodegeneration, which may be masked in the substantia nigra by the extent of cell death in this region at the time of post-mortem, as suggested by Sutherland *et al*.[140]

Although there are large differences between the results from animal models and human studies, it is encouraging to note that animal models (both genetic and neurotoxic) are not completely separated from human neurodegenerative disease in differential gene expression space (Figure 3.4, Figure 3.5). In particular, one of the two animal models sampling tissue from the SN appears very similar to human studies in hierarchical clustering (Figure 3.6). It is possible that these simply reflect the 'terminal cytoarchitectural differences'[140] related to neuronal loss in the SN. However, the observed similarity of studies of the frontal cortex – which does not show such severe neuronal loss[140] compared to the SN, where next to no dopaminergic neurons remain post-mortem[172] – to studies of the SN (Figure 3.4, Figure 3.5, Figure 3.6) points towards at least partly shared gene expression patterns which are reflective of other biological processes.

There is much interest in the use of non-brain tissues for gene expression studies, as these can be relatively easily obtained pre-mortem and could reflect processes associated with early-

stage PD, as well as potentially offering direct patient benefit. Studies which use human cell lines, such as iPSCs derived from PD patients, do not appear to replicate the differential expression patterns found in tissue samples from PD patients. Using hierarchical clustering, however, iPSC-derived dopaminergic neurons appear similar to dopaminergic neurons isolated post-mortem, while iPSCs harbouring SNCA mutations cluster with Pink1- and SNCA-based genetic animal models of PD (Figure 3.6), suggesting the potential for these study types to replicate relevant gene expression patterns in PD. Similarly, studies in blood samples cluster together, appearing distinct from gene expression in brain tissue (Figure 3.6) but also appearing distinct from gene expression in blood studies of brain tumours (Figure 3.9), suggesting a common transcriptional pattern that could function as a marker of neurodegenerative disease. These are encouraging results for the development of surrogate tissue approaches for studying gene expression in PD.

In practice, the concordance between microarray studies from different experimental groups will never reach 100%. Experimental factors including sample amplification, labelling, array scanning, wash protocols, etc.[116,176,177] exert a significant effect on the results and reproducibility of studies; in the context of PD, there are a number of experimental factors which influence measured RNA expression in the brain including the impact of age, gender, and post-mortem interval[120,154,178] and other confounding factors including long-term anti-Parkinsonian drug treatment and the co-occurrence of other diseases such as Alzheimer's disease[141]. More detailed meta-data associated with studies uploaded to public repositories would be immensely helpful in aiding meta-analysis and identification of differences between studies. This could be disease-specific, such as distinguishing between idiopathic and genetic PD cases; or more general, such as distinguishing between drug-treated and drug-naïve patients, or providing a measure of RNA integrity such as RIN[179] (especially key in post-mortem studies where RNA quality is affected by the agonal state[180]).

Nevertheless, this study aims to illustrate the amount of agreement that can be expected between different microarray studies in the context of PD; further, its general conclusions may be equally applicable in studies of other conditions. This study acts as a guide to the 'representativeness' of different tissues and disease models to the human condition (which is of special significance in PD due to the inaccessibility of PD-affected tissues in living patients), and as a guide to the use of animal models, at a time of increasing importance of transcriptomics and other molecular-level analyses in drug discovery and development[181]. The identification of a specific 'characteristic' signal of PD in human brain tissues could explain the apparent

discordance between microarray studies of PD, and is hence of more general interest for the study of PD at the transcriptomic level.

4 USING DYSREGULATED SIGNALLING PATHS TO UNDERSTAND DISEASE

SUMMARY

Network-based methods of gene expression analysis have recently become popular, allowing gene expression changes to be interpreted by relating them on to each other on a known framework. However, gene networks produced by these methods are often large and difficult to interpret. In this chapter, a 'bottom-up' method of subnetwork identification based on weighted shortest-paths (termed 'path-set analysis') is described, which highlights smaller network regions which are dysregulated in disease. This method, in contrast to similar approaches, is based solely on differentially expressed genes. Each edge is therefore disease-specific, rather than including 'bystander' nodes resulting purely from network topology. This makes path-set analysis particularly suited to the comparison of expression changes in disease, allowing comparison at a granular (edge-wise) level instead of comparing large subnetworks.

The ability of path-set analysis to identify relevant dysregulated processes in disease is confirmed by the enrichment of known disease-associated genes in the returned paths. Comparing path-sets across the 141 transcriptomic studies in the dataset reveals commonly dysregulated genes which are included in the path-sets of multiple disease studies. There is a moderate relationship between the number of studies in which a gene is included and its network importance (quantified by degree), suggesting that path-set analysis identifies 'pressure points' in the network which can influence the progression of diverse disease types. Disease studies that share dysregulated paths are 2.5 times more likely to be in the same Disease Ontology subcategory than those that don't, and more than twice as likely to share drugs, confirming the relevance of path-sharing to known disease relationships. Over half of shared paths between disease pairs contain a drug-interacting gene, suggesting the utility of this approach in forming early-stage drug repurposing hypotheses.

This work represents the first application of path-based network analysis to the comparison of diseases, including understudied rare diseases. This analysis reveals the underlying processes dysregulated in disease, helping to develop our understanding of disease and disease relationships, which could ultimately lead to novel treatments.

4.1 INTRODUCTION

The analysis of gene expression data can give valuable insight into the underlying processes taking place in disease. However, as illustrated in Chapter 3, transcriptomic data can be noisy, displaying high variation even between studies of the same condition. Gene expression data is therefore often interpreted by translation to a higher biological level, such as biological pathways, under the assumption that different observed gene expression changes can reflect similar underlying processes. A limitation of this gene set analysis, however, is that canonical biological pathways represent inflexible, high-level descriptions of a complicated process which obscure the individual gene expression changes.

A middle ground between analysis of individual gene dysregulation and gene set analysis is network-based analysis, which relates genes on the interactome. The interactome describes the interactions between gene products (chiefly proteins) without reference to known pathways or processes, thereby enabling the discovery of novel regions of interest[182]. This type of analysis often involves the detection of dysregulated subnetworks of interacting genes and gene products that are active in disease. Distinct from 'functional module' approaches which partition the interactome into clusters based on topology[183,184], 'active' subnetwork approaches incorporate experimental information such as genomic or transcriptomic data (covered in a comprehensive review by Mitra et al.[185]).

An alternative to 'top-down' subnetwork identification is the 'bottom-up' identification of individual dysregulated *paths*: subunits of networks which have defined start and end points. The advantage of using paths is that unlike subnetworks, paths can represent isolated patterns of dysregulation which may involve only two or three gene products. This approach is commonly used to study drug response by identifying paths connecting drug targets to differentially expressed genes[186] or toxicity-related proteins[187]; or to connect known genetic variations in disease with differentially expressed genes[188]. However, it is not always possible (or desirable) to define these start and end points in advance, in which case *a priori* path identification approaches are required.

Of most relevance is the method developed by Sambarey et al.[121], who used a weighted shortest-paths approach to identify a common response network shared by multiple tuberculosis gene expression datasets. Briefly, dysregulated paths are identified by computing weighted shortest-paths on the signalling network, where the weights are inversely proportional to the differential expression magnitude of each gene. Those paths whose lengths are most

highly changed from the unweighted base network (representing signalling in a healthy person) therefore contain the most highly dysregulated genes. Using this method, the authors could identify commonalities between the datasets that were not found by traditional differentially expressed gene list analysis. They confirmed the specificity of the identified response network by comparing it to response networks in four other diseases (sarcoidosis, Still's disease, pneumonia, and systemic lupus erythematosus).

Whilst this approach looked for differences (specificity) between the tuberculosis response network and other diseases, such path-based approaches could also be valuable for the identification of commonalities between diseases. Previous studies have integrated multiple datasets from a shared phenotype to discover common differential-expression based subnetworks across related diseases[182,189,190]. For example, one study used topological network clustering of differentially expressed genes in oesophagitis and oesophageal cancer to discover functional modules common to both diseases[190]. Alroobi et al.[189] integrated datasets within phenotypic classes including 'gastroenteritis', 'carcinoma', 'neoplastic process' and 'cell or molecular dysfunction' to find subnetworks shared within these classes. However, few previous approaches have compared differential expression-based networks across different types of disease. One exception is the 2009 study by Suthram et al.[107] described in Section 1.4.1.2, who compared 54 diseases by mapping their gene expression to precomputed functional modules, finding significant disease correlations between e.g. Crohn's disease and malaria.

Here, a method inspired by the weighted shortest-paths approach of Sambarey et al.[121] (see Section 4.2.4 for discussion of the key differences) is applied to the comparison of 141 gene expression datasets representing 119 diseases, as well as 19 drug-induced gene expression profiles. In this method, which is termed *path-set analysis*, the initial network for each disease (or drug) is constructed from all genes with non-zero differential expression. The use of this non-conservative differential expression threshold enables path-set analysis to discover groups of genes which may not be highly differentially expressed individually, but which represent a flow of dysregulation along the network. Unlike traditional analysis methods, where the importance of a gene in a particular disease is determined solely by its individual log-fold change or significance value, in path-set analysis the importance of a gene is determined in a more holistic manner by taking into account the activity of its interacting genes.

An advantage of this approach (in contrast to the work of Sambarey et al., where paths combine differentially expressed genes with non-differentially expressed 'bystander' nodes) is that each

individual edge is disease-specific, i.e. each edge connects two genes which are differentially expressed in that disease, as opposed to connecting a differentially expressed gene with each of its neighbours. This property of edge specificity enables the comparison of diseases at a granular (edge-wise) level, rather than across whole (sub-) network modules as in the work of Suthram et al., allowing the identification of disease pairs which share dysregulated processes.

In this chapter, I use path-set analysis to interpret gene expression data across diverse conditions, including 141 microarray studies of disease representing 119 different diseases, and 19 drug response studies in human patients. The relevance of the paths for each disease is evaluated using the presence of known disease-associated genes and drug-interacting genes in each path. For the first time, I use path-based analysis to compare diseases, identifying paths which are shared between diseases and highlighting dysregulated processes common to diverse disease types. Finally, I explore the relevance of these common processes to the identification of potential drug repurposing opportunities.

4.2 METHODS

4.2.1 Gene expression dataset construction

Suitable gene expression datasets were identified by manually searching Gene Expression Omnibus[44] for specific diseases and by searching for the keywords '*disease*', '*syndrome*' and '*cancer*' and selecting those for which high-quality patient transcriptomic data were available, according to the criteria described in Section 2.2. This resulted in 141 datasets covering 119 distinct diseases, including 35 rare genetic diseases (defined as inheritable diseases with a prevalence of less than 1 in 10,000 where known). Due to limitations of the available data, some of the gene expression profiles are based on *in vitro* samples from cell lines, rather than *in vivo* samples directly from patient tissue (this is mostly for rare diseases where fewer studies are available). Where technical replicates e.g. multiple repeats of the same cell line are used, only the first is taken in order to have a consistent 'one-patient, one-sample' structure throughout each dataset. Datasets and sample selection are listed in Appendix B. In order to classify diseases, disease names were manually mapped to Disease Ontology[191] terms and their top- and second-level classes were recorded.

All datasets were downloaded and processed as described in Section 2.3, resulting in differential expression profiles of diseased vs healthy patients. A non-conservative threshold of $p<0.05$ was used to call significant differential expression; log-fold changes of non-significant genes were set to 0. This non-conservative threshold represents a departure from traditional gene expression analysis, which relies on the significance of individual gene expression changes to select a list of individually 'important' genes. Instead, this method considers the dysregulation of a gene in combination with that of its neighbours, representing a network-based view of 'importance'. At this threshold, a median of 27% of the genes in a disease experiment have non-zero differential expression, although only 1.6% of genes on average have an absolute log-fold change value greater than 1.

Drug gene expression datasets were identified by manually searching Gene Expression Omnibus using keywords including 'drug', 'treatment', 'compound', 'placebo'. This resulted in 19 datasets covering 16 different drugs. These were downloaded and processed as described in Section 2.3, resulting in differential profiles from e.g. the patient after drug treatment vs after taking a placebo, or the patient after drug treatment vs before drug treatment.

4.2.2 Identifying disease-associated and drug-interacting genes

Data on disease-associated genes was obtained from the OpenTargets platform in December 2016 using the provided REST API[169]. Diseases were mapped to their closest disease concepts in OpenTargets, although in some cases the match is to a less specific concept (e.g. 'breast lobular carcinoma' maps to 'breast carcinoma'; 'non-small cell lung carcinoma' maps to 'lung cancer', 'teratozoospermia' maps to 'male infertility'). Genes with an evidence score >0.2 in 'genetic association' or 'somatic mutation' were defined as disease-associated genes. For drug response datasets, disease genes corresponding to the disease in which the drug was tested were used.

To identify drug-interacting genes related to diseases (for use with the disease datasets), drug indication data was obtained from ChEMBL version 22.1[192] (https://www.ebi.ac.uk/chembl/downloads, files chembl_drug_indication and chembl_mol_dict), and approved drugs or drugs in Phase III clinical trials were retained. Genes related to these drugs were downloaded from the Drug Gene Interaction Database (http://www.dgidb.org/) in November 2016. Genes corresponding to primary targets of drugs (for use with the drug response datasets) were identified using the 'Mechanisms of action' information from the OpenTargets platform, which translates ChEMBL mechanism of action information into target space.

4.2.3 Signalling pathway network construction

OmniPath[82] was used as the basis of the signalling network. OmniPath is a recently published resource containing 'literature-curated human signalling interactions'[82] from 27 different resources including SignaLink, Reactome, IntAct, WikiPathways, Signor and others, resulting in coverage of '~39% of the human proteome'[82]. Only those interactions of known direction (causal interactions, which are the basis of signalling pathways) are retained, resulting in a network of 6,942 nodes.

In this work, as in common in canonical pathway analysis, changes in gene expression (i.e. mRNA abundance) are treated as a proxy measure of changes in pathway activity, recognising that gene expression is not directly correlated to the abundance of the corresponding proteins[193], but that they may function as a broad indicator of dysregulation in a particular pathway. The proteins in the network are therefore represented by their corresponding genes,

with the UniProt IDs used in OmniPath converted to gene symbols supplied by the Hugo Gene Nomenclature Committee (conversion tool downloaded from https://www.genenames.org/cgi-bin/download). Where a UniProt ID maps to more than one gene (or vice versa), both mappings were kept in order to retain the maximum number of interactions. Duplicate entries and self-loops were removed.

In order to test the dependence of the results on the underlying network, the analysis was repeated with an independent network resource, HIPPIE (Human Integrated Protein-Protein Interaction rEference)[83]. Unlike OmniPath, HIPPIE is not based on signalling interactions, but on experimentally determined protein-protein interactions. HIPPIE provides a confidence score based on the available evidence supporting each interaction, allowing the filtering of the network to retain only high-confidence interactions (score >=0.73). Following this filtering, HIPPIE contains 62,615 interactions between 12,162 proteins compared to the 43,693 interactions between 6,972 proteins in OmniPath. Although the resulting path-sets contained different nodes (due to the small overlap between interactions in HIPPIE and OmniPath) and contained slightly fewer disease-associated genes, overall properties of HIPPIE path-set analysis in terms of shared edges and enrichment of disease-associated genes compared to other methods were not substantially different to those obtained with OmniPath (discussed in Appendix J), indicating that the results presented in this chapter are not dependent on the specific topology of the OmniPath network.

4.2.4 Identification of dysregulated signal paths in each disease

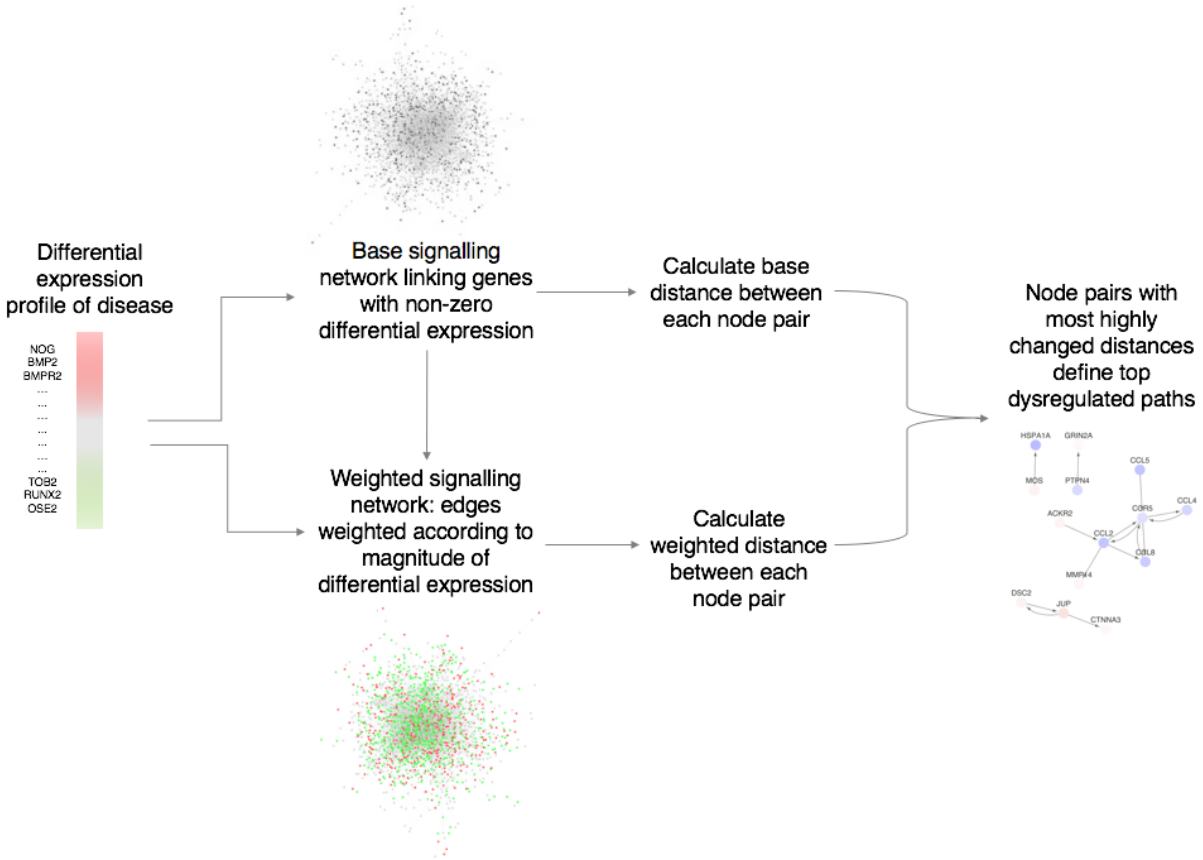


Figure 4.1 Identification of dysregulated signal paths

Path-set analysis uses weighted shortest-paths on a directed signalling network to find paths which are strongly dysregulated in disease. For each disease, all genes with non-zero differential expression changes are identified, and an initial base network is created from these genes. A modified version of the base network is then created in which the outgoing edges of a gene are weighted according to the magnitude of differential expression for that gene, so that high differential expression 'reduces' the distances to the gene's neighbours. The length of the path between each node pair is then calculated for both networks. Similar to the wok of Sambarey et al.[121], paths containing many highly differentially expressed genes will be much shorter in the weighted network than in the base network. These paths with highly changed distances form the path-set for a disease.

Path-set analysis calculates, for each disease, the paths which link the most highly differentially expressed nodes. This process is summarised in Figure 4.1 and detailed in full below. Steps 1-3 describe the construction of the base signalling network, steps 4-6 describe the construction of the weighted signalling network, and steps 7-11 describe the selection of dysregulated paths.

1. First, an initial base network is created from OmniPath by excluding nodes which are not differentially expressed in the disease (at a threshold of $p<0.05$). This prevents

the algorithm from simply returning all possible paths between two non-neighbouring genes, ensuring that each edge is disease-specific rather than based on network topology.

2. All nodes in the base network must have at least one outgoing edge, because it is the outgoing edge that is considered in the weighted shortest path calculation (step 6). Any 'leaf' nodes which do not have children in this network are therefore connected to a 'dummy' child node.
3. The resulting interactions list is converted to a graph representation using the *graph_from_edgelist* function from the R package igraph[194] (version 1.1.2). Unweighted shortest path distances are computed along this base network, using igraph's *distances* function with mode=out, resulting in a distance matrix where first neighbours have a distance of 1, neighbours of first neighbours have a distance of 2, and so on, which is the base distance matrix, *baseDist*.
4. An edge weight vector *initialWeights* is initialised where for each edge p->q, the weight is set to the absolute log-fold change value of the source node p.
5. The weights are inverted:

$$weights = \max(initialWeights) - initialWeights$$

so that the most highly differentially expressed gene has a weight of 0 (a small positive value δ is added to this gene to avoid having a 0-length path) and therefore the smallest path cost, and the least highly differentially expressed gene has the highest weight (equal to the highest absolute log-fold change) and therefore the highest path cost.

6. The distance computation via *distances* is repeated with the weights from the previous step, yielding the weighted shortest path distance matrix, *weightedDist*.
7. In order to be comparable with the weighted distances, *baseDist* is multiplied by a constant equal to the maximum log-fold change observed in the disease. This creates a new distance matrix, *scaledBaseDist*, which is equivalent to performing weighted shortest-paths on a network where all differential expression values are zero (and therefore has longer path lengths than any in *weightedDist*).
8. Path dysregulation scores are calculated as the difference between the scaled base distance matrix and the weighted distance matrix, normalized by the length of the path:

$$difference = (scaledBaseDist - weightedDist)/baseDist$$

Note that the normalization here could equally be division by *scaledBaseDist*.

9. A threshold is set which specifies the number of paths to consider for each disease, here set to the 100th or 500th highest value of the *difference* matrix depending on the application. All differences less than this threshold are set to 0, so that in further analysis the (roughly) 100/500 most highly dysregulated paths are considered.
10. For each node pair whose path dysregulation score is above this threshold, the vertices of the shortest path(s) between them are returned by *igraph*'s *all_shortest_paths* function, again with mode=out and weights as calculated in step 5.
11. The returned paths are then pruned according to the following criteria:
    a. Paths must link at least two nodes after removal of the dummy node.
    b. Any paths which are shorter subsets of other paths are removed.

Diseases are then represented by the set of edges resulting from the union of these paths, the *path-set* for each disease.

A threshold of the 100th highest distance was used for the first part of the analysis, which focuses on paths in individual diseases. A more relaxed threshold of the 500th highest distance was used for the second part of the analysis (Sections 4.3.5 and 4.3.6), in order to increase the possible number of paths shared between diseases.

One potential limitation of the shortest-paths method is that by recording only the *shortest* path between two nodes, the algorithm may potentially miss other paths between the two nodes which may also meet the threshold. An adjustment of the algorithm is possible which takes this into account by incorporating every path between two nodes which meets the threshold, but this increases the run-time of the algorithm exponentially without adding many new nodes to the path-set. The shortest-path formulation was therefore retained for this version of the algorithm.

In order to compare path-set analysis to traditional differential expression analysis, an *LFC-set* is also constructed for each disease. The LFC-set is simply the top *n* most highly differentially expressed genes (where *n* is the number of nodes in the path-set of that disease) by absolute log-fold change at $p<0.05$, restricted to those genes which are contained in OmniPath. Genes in OmniPath may have different properties than those genes which are not in OmniPath (possibly associated with being more well-studied), so this allows a fairer comparison between the two methods. In order to examine the contribution of network information in the absence of log-fold change information, 100 *random-path-sets* are also constructed. These are constructed as the real path-sets, but with permuted log-fold change values within each disease.

Path-set analysis is inspired by the method proposed by Sambarey et al.[121] who used weighted shortest-paths to identify response networks in tuberculosis, but differs in several key details including the calculation of the node and edge weights and the use of directed interactions on a signalling network to represent signal flow; and further in the use of only differentially expressed genes to define the network, so that the resulting paths are composed solely of differentially expressed genes rather than including 'bystander' neighbour nodes.

4.2.5 Identification of shared dysregulated signalling paths between diseases

In order to compare dysregulated signalling paths across diseases, a *common-path-set* was additionally constructed for each disease. The common-path-set is constructed as above, but takes into account the diversity of platform types (each measuring different sets of genes) in the dataset by restricting the analysis to the 3,724 genes measured on all platforms, 2,306 of which are in OmniPath. The common-path-set was used for analyses involving comparison across disease datasets (Sections 4.3.5, 4.3.6).

Using the common-path-sets, the number of edges shared between two diseases is calculated. An edge is shared between two diseases if:

1. The edge is in the common-path-set of both diseases
2. The direction of the log-fold change associated with the nodes linked by the edge is the same in both diseases.

A random permutation test is used to calculate if the number of shared edges between two diseases is significant. For each disease, a random edge-set is created which contains the same number of edges as in the original path-set by sampling edges from the shared base network according to the frequency of these edges over all common-path-sets. The number of overlapping random edges between each disease pair is then calculated as above (randomly assigning a direction of log-fold change to each node). This procedure is repeated 1000 times for each disease pair, and the highest random overlap is taken as the significance threshold for each disease pair.

Finally, the disease similarity score is calculated as the number of shared edges divided by the total number of edges in the path-set of each disease. Where the number of shared edges is less than the significance threshold, the similarity score is set to 0.

4.2.6 Pathway enrichment analysis

Pathway enrichment analysis was carried out using the Panther Classification System[62] version 13.1 (http://www.pantherdb.org). Gene lists were uploaded to Panther and an overrepresentation test was performed specifying an appropriate reference list as the background (i.e. all genes in the network). Gene lists were analysed against the Panther GO-Slim biological process termset. The Fisher's exact test was used to determine significance; pathways with a Benjamini-Hochberg FDR < 0.05 (the default reported by Panther) were reported as significant.

4.3 RESULTS

4.3.1 Path-set analysis of gene expression changes in disease reveals shared dysregulation amongst interacting gene products

To quantify the extent to which path-set analysis differs from traditional log-fold change analysis (which considers the most strongly differentially expressed genes in each disease), the *n* genes comprising a disease's path-set were compared to the disease's *n* most highly differentially expressed genes in OmniPath (the LFC-set, see Methods). Across all diseases, a median average of 22% of genes are common to the path-set and the LFC-set of a disease, i.e. the path-sets capture 22% of the disease's most highly differentially expressed genes.

Simply mapping the highly differentially expressed genes in the LFC-sets to the OmniPath network results in very small network sizes: most of these genes do not interact with each other, meaning that the resulting networks cover a median of only 18% of the genes in the LFC-sets. On the other hand, trying to connect more of the LFC-set genes e.g. by including all first-neighbour genes results in a very large network: whilst the median LFC-set size is 82 genes, the median resulting network size is 659 genes, which is infeasible to analyse visually.

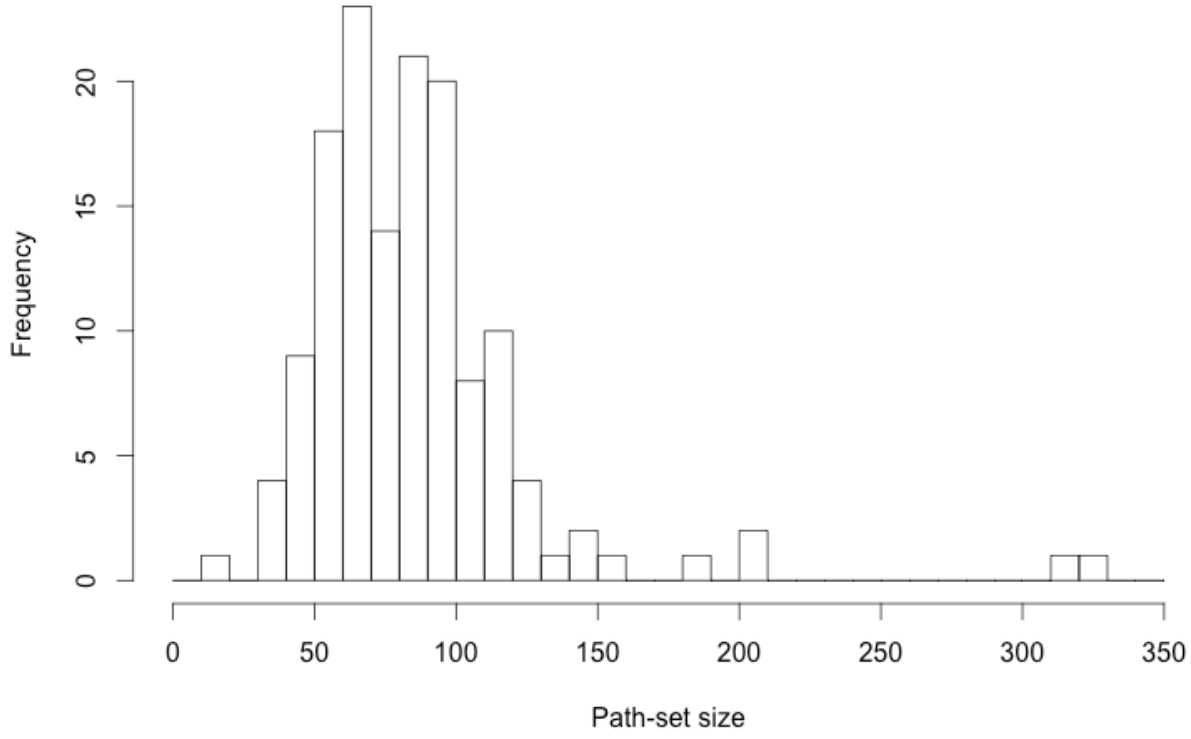


Figure 4.2 Distribution of path-set size

At a threshold of the 100 top paths, most diseases have a path-set comprising around 50-100 dysregulated nodes, although a few path-sets (for hepatocellular carcinoma and Turner syndrome) are much larger.

Path-set analysis represents an alternative to network-based analysis of DEG lists, which rather than selecting the top most differentially expressed genes, highlights genes which interact in a dysregulated biological process (in this case, signalling). The initial network for each disease is built from all genes with non-zero differential expression; across all diseases, there is a median of 4,714 genes with non-zero differential expression per disease, resulting in a median initial network size of 1,408 nodes. The top most dysregulated paths are then selected from this network (see Section 4.2.4), resulting in a median path-set size of 82 nodes (Figure 4.2).

This type of analysis can be particularly useful for experiments where few genes show high log-fold changes. One example of this is the experiment for asthma (see Appendix B for details), where the highest absolute log-fold change (at $p<0.05$) is only 0.56. By comparison, the median highest absolute log-fold change across all diseases in the dataset is 3.91.

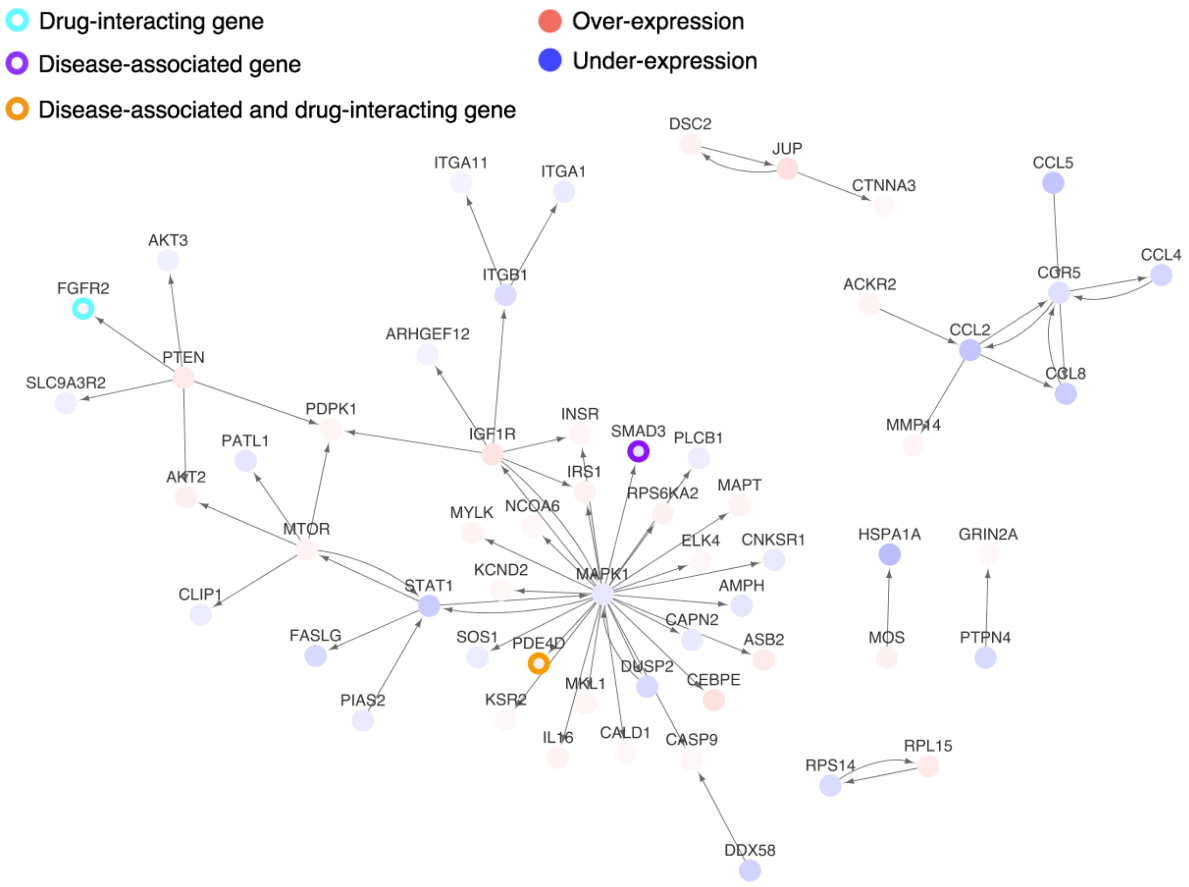


Figure 4.3 Dysregulated signalling paths in asthma

The path-set of asthma (based on the top 100 paths) includes two connected components, one centred around the network 'hub' gene MAPK1 and its downstream genes, and the other centred around the inflammatory chemokine receptor CCR5 and its ligands. 18 of the 58 genes in the path-set are also in the set of most highly differentially expressed genes (the LFC-set), but the LFC-set does not include the key nodes MAPK1 and CCR5 which form the basis of these components. The pale node colours indicate low magnitude of differential expression in comparison to later figures, which are plotted on the same colour scale.

Figure 4.3 shows the genes comprising the dysregulated paths (the path-set) for the asthma dataset; overlaying the individual paths onto the OmniPath network in this way shows how the expression changes are related to each other. 18 of the 58 genes in the asthma path-set are also captured by its LFC-set; however, the LFC-set does not capture the key network-specific 'hub genes' mitogen-activated protein kinase 1 (MAPK1) and C-C chemokine receptor type 5 (CCR5). MAPK1 in particular has a central position in the human signalling network (it has a degree of 476, one of the highest-degree nodes in OmniPath) and so it is unsurprising to see that this hub structure is retained in the asthma path-set.

As OmniPath is a signalling network, links between dysregulated genes are not a casual explanation (e.g. under-expression of gene A leads to under-expression of gene B). Rather, path-set analysis highlights groups of interacting genes with shared dysregulation, suggesting broader underlying dysregulation e.g. of a particular signalling pathway. It should also be noted that as mRNA expression levels are not necessarily correlated with protein abundance[193], expression dysregulation provides at best a rough suggestion of what might be taking place at the protein level. Bearing these points in mind, the observed expression changes can be interpreted as indicators of signalling events that may be dysregulated in asthma. Throughout this chapter, protein symbols will be given in italics to distinguish them from gene symbols.

For instance, the hub gene MAPK1 is shown to interact with PDE4D, a subtype of phosphodiesterase 4. Its product, mitogen-activated protein kinase 1, phosphorylates the product of PDE4D, but whether this has an activating or inhibitory effect depends on the exact isoform of PDE4D that is translated (PDE4D has an intron variant in asthma which could potentially affect this)[195]. The upregulation of PDE4D, if this translates to increased levels of its protein product, chimes with the fact that inhibitors of *PDE4* are currently being developed for the treatment of asthma (although the *PDE4D* subtype specifically is associated with the side effect of nausea and vomiting)[196].

Another component of interest is formed by the inflammatory chemokines CCL2, CCL4, and CCL5, and the chemokine receptor CCR5. Previous studies have found levels of their corresponding proteins to be increased in asthma[197], however, here all four chemokines and their receptors are downregulated. The fact that these chemokines interact with each other suggests that their downregulation is not just a chance observance but is part of a co-ordinated process, which is particularly important to establish as the observed magnitude of the fold changes is so low. Interestingly, whilst the LFC-set captures the four ligands CCL2, CCL4, CCL5, and CCL8, indicating that they are some of the most highly dysregulated genes in asthma, the receptor CCR5 is not strongly differentially expressed enough to be included in this set, illustrating how analysis based on fold change alone excludes relevant gene expression changes.

4.3.2 Dysregulated paths are enriched for disease-associated genes and drug-interacting genes

Table 4.1 Known disease-associated genes and drug-interacting genes (KDGs) in path-sets

Path-sets are enriched for KDGs compared to selecting the same number of genes by log-fold change (LFC-set) or random sampling (random-gene-set), capturing KDGs for 58% of diseases (values reported are mean averages across diseases which have at least one KDG in OmniPath). The high percentage of KDGs found by the random-path-set (which is created by permuting the LFC values for each disease) suggests that this is due to the ability of path-sets to capture genes of higher degree, which are more likely to be KDGs.

	Path-set	**LFC-set**	**Random-path-set**	**Random-gene-set**
What proportion of sets contained at least one KDG?	0.58	0.56	0.54	0.43
How many KDGs were found per set on average?	2.95	1.89	2.40	1.18
What percentage of genes in the set were KDGs on average?	3.2%	2.4%	2.5%	1.3%

The presence of genes associated with a disease, and genes which interact with drugs for this disease (together 'known disease-associated genes'; KDGs) in a path-set can indicate the interaction of the dysregulated paths with causative or therapeutic processes taking place in the disease. The presence of KDGs in each path-set was therefore used as a proxy measure to evaluate the biological relevance of the discovered paths (Table 4.1). In this study, disease-associated genes are defined as genes which have a variation or somatic mutation which has been previously associated with the disease. Drug-interacting genes are defined as genes which interact in some way with drugs prescribed for the disease (e.g. a drug inhibits a product of this gene) (see Section 4.2.2).

Across all diseases with at least one KDG (130 diseases), 58% contained at least one KDG in their path-set, with a mean average of 2.95 KDGs per path-set. If the less strict threshold of 500 is used, these values increase to 72% of path-sets, containing a mean average of 7 KDGs. Path-sets capture more KDGs than LFC-sets, which find only 1.89 KDGs per disease on average. Surprisingly, the *random-path-sets* (in which the log-fold change values for each gene are permuted, repeated 100 times) also capture many KDGs in disease. Although the difference between the real and random path-sets is statistically significant (t-test p-value $<2.2e^{-16}$ for percentage of datasets containing at least one KDG and for number of KDGs found), the magnitude of the difference is small. This is not seen in the *random-gene-sets* (gene-sets of the same length as the path-sets randomly selected from the genes in OmniPath, repeated 1000 times), suggesting that the path-set method inherently selects for KDGs, regardless of gene expression information.

The enrichment of KDGs in random path-sets can be explained by the tendency of path-sets to return genes of high network importance, quantified here by the *degree* of the node, i.e. the number of nodes with which it interacts (calculated using igraph's *degree* function). The median degree of genes in the real path-sets is 15.5, and 19 in the random-path-sets. By comparison, this figure is only 5 for both the LFC-sets and the random-gene-sets. This illustrates how the path-set analysis method selects for genes which have a greater number of interactions, which have greater chance to be included in a dysregulated path.

KDGs also tend to have higher degree on average than non-KDGs (median degree of 7 for drug-interacting genes vs 4 for non-drug-interacting genes, Wilcox p-value $<4.46e^{-12}$; median degree of 7 for disease-associated genes vs 4 for non-disease-associated genes, Wilcox p-value $<2.20e^{-16}$). The performance of the random-path-sets at identifying KDGs suggests that the ability of path-sets to identify KDGs may be partly based on the incorporation of the network structure information. However, the LFC-sets do much better than the random-gene-sets at finding KDGs, suggesting that fold change information is also important for identifying paths which contain KDGs; it is therefore unsurprising that path-sets, which combine fold-change and network information, are most enriched for KDGs.

4.3.3 Dysregulated paths interact with known disease-associated genes and drug-interacting genes

It should be noted that not all of a disease's KDGs can be captured by path-set analysis. Only KDGs which are:

1. Measured by the gene expression platform used for that disease
2. In OmniPath (median of 30% of measured genes in each experiment)
3. Differentially expressed in the associated disease (median of 36% of KDGs, compared to the 27% of all measured genes with non-zero differential expression on average across diseases)

can be returned by this method. Whilst path-set analysis focuses only on differentially expressed genes (in order to improve understanding of differentially expressed gene lists in disease), there will also be many genes involved in disease which are not dysregulated, such as the 64% of KDGs which are not differentially expressed in their associated disease. By relating genes in the path-sets to first-neighbour KDGs, other 'key players' can be captured which despite not showing changes in their expression levels may influence or be influenced by the dysregulated processes captured in the path-sets. A median of 12.5 KDGs per path-set are captured in the first-neighbour genes.

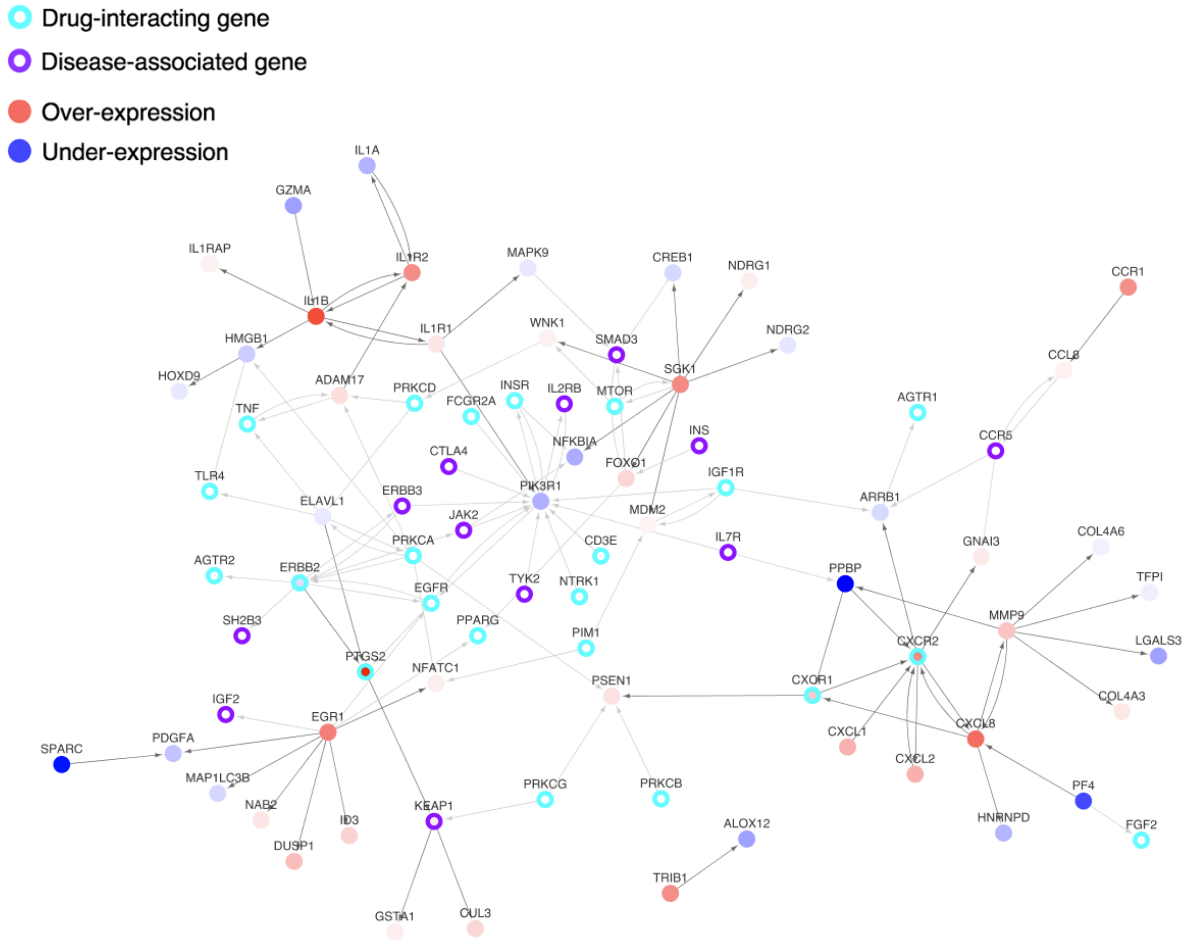


Figure 4.4 Dysregulated paths in type 1 diabetes interact with non-dysregulated disease-associated genes and drug-interacting genes

Visualising the first-neighbour KDGs of the 53 genes in the type 1 diabetes dysregulated path-set shows how genes in the path-set interact with disease-relevant proteins in the signalling network. The figure also shows 5 KDGs directly in the path-set, including disease-associated genes KEAP1 (purple border) and four drug-interacting genes (cyan border)

Figure 4.4 shows an example visualisation of a path-set and (non-differentially expressed) first-neighbour KDGs in type 1 diabetes mellitus (T1D), an autoimmune condition in which insulin-producing beta cells in the pancreas are destroyed by the body's own immune system, resulting in an inability to control glucose levels in the bloodstream. Again, the Figure reveals interactions between groups of dysregulated genes, such as the coordinate overexpression between the chemokine receptors CXCR1 and CXCR2 (referred to as 'master regulators of diabetes pathogenesis'[198] due to their role in the autoimmune destruction of insulin-secreting beta cells[199]) and their ligands, CXCL1, CXCL2, and CXCL8 (shown on the right-hand side of the Figure).

The Figure also illustrates how the nodes in the path-set interact with (non-differentially expressed) first-neighbour KDGs. The path-set together with first-neighbours now captures 23.6% of the 93 drug-interacting genes associated with T1D, and 9.8% of the 122 known disease-associated genes. Again, examining the relations between genes in the path-set and their first-neighbours can reveal more about the patterns of dysregulation seen in disease. One example is PIK3R1, which has only one neighbour in the path-set, but which takes on a more central 'hub' position when the first-neighbour KDGs are added, suggesting its potential importance in T1D. *PIK3R1* is the p85α regulatory subunit of *PI3K*, which mediates insulin signalling[200,201]; lowered PIK3R1 expression has been found to prevent insulin resistance in obese mice[202], and so the under-expression seen in this dataset might be associated with lower blood insulin levels by increasing insulin sensitivity.

4.3.4 Path-sets reveal genes frequently dysregulated in disease

Comparing the path-sets across the 141 experiments in the dataset (here using the shared-genes path-sets, which are based on only the 3724 genes measured in all experiments) allows identification of genes which are frequently dysregulated in many different diseases. These genes may represent a 'stress response' which is not specific to the disease, but which form a more general response to disease e.g. involvement of the immune system. Note that 15 diseases are represented twice in the dataset, 2 diseases represented thrice, and one disease (multiple sclerosis) being represented four times, leaving a total of 119 unique diseases. Given the difference that can exist between two measurements of the same disease (as has been shown in Chapter 3) these replicate experiments were retained for the following analysis, which therefore more properly refers to genes dysregulated in multiple *experiments*, rather than diseases.

Ignoring those genes which do not appear in any path-sets, each gene appears in a median of 4 path-sets; however, some genes are in many more path-sets, with 54 genes appearing in the path-sets of 25 or more experiments (Appendix K). One gene, the epidermal growth factor receptor (EGFR), is in the path-sets of 66 experiments, which include cancers (EGFR is known to play a role in many cancers[203]) as well as other disease types including skin diseases and a number of rare syndromes.

Gene Ontology biological process (GO BP) enrichment analysis of these 54 genes returned 14 terms, mostly related to signalling, including *biological regulation*, *signal transduction* and

cell communication. This is unsurprising given that the dysregulated paths are based on the OmniPath network, which is designed to reflect signalling interactions; although other significant terms were not directly related to signalling, including *negative regulation of apoptotic process* and *metabolic process* (see Appendix K). These terms suggest mechanisms through which these commonly dysregulated genes may influence the general response to disease.

This analysis was repeated for genes in multiple LFC-sets and random-path-sets. There is less overlap between the LFC-sets than between the path-sets, with genes in LFC-sets being dysregulated in a mean of 5.5 LFC-sets each, compared to the mean of 6.1 path-sets (although the median value for both is 4), and no gene appearing in the LFC-sets of more than 34 experiments (compared to 14 genes appearing in more than 34 path-sets). One explanation for this is that whereas LFC-sets focus on the individual genes with the highest differential expression, which will vary between diseases and experiments, path-set analysis returns sets of interacting genes which show co-ordinated expression dysregulation, which may be more likely to be replicated in multiple experiments due to e.g. the involvement of these gene sets in particular biological processes.

Interestingly, despite the selection of genes in random path-sets having a strong relationship to their degree (Figure 4.5), there is comparatively little overlap between the random path-sets: taking the total number of occurrences of a gene in the 100 random path sets and dividing by 100, genes are included in a mean of 4.5 random-path-sets each. This suggests that the overlap seen in the path-sets is not simply due to the repeated selection of high-degree nodes across multiple experiments.

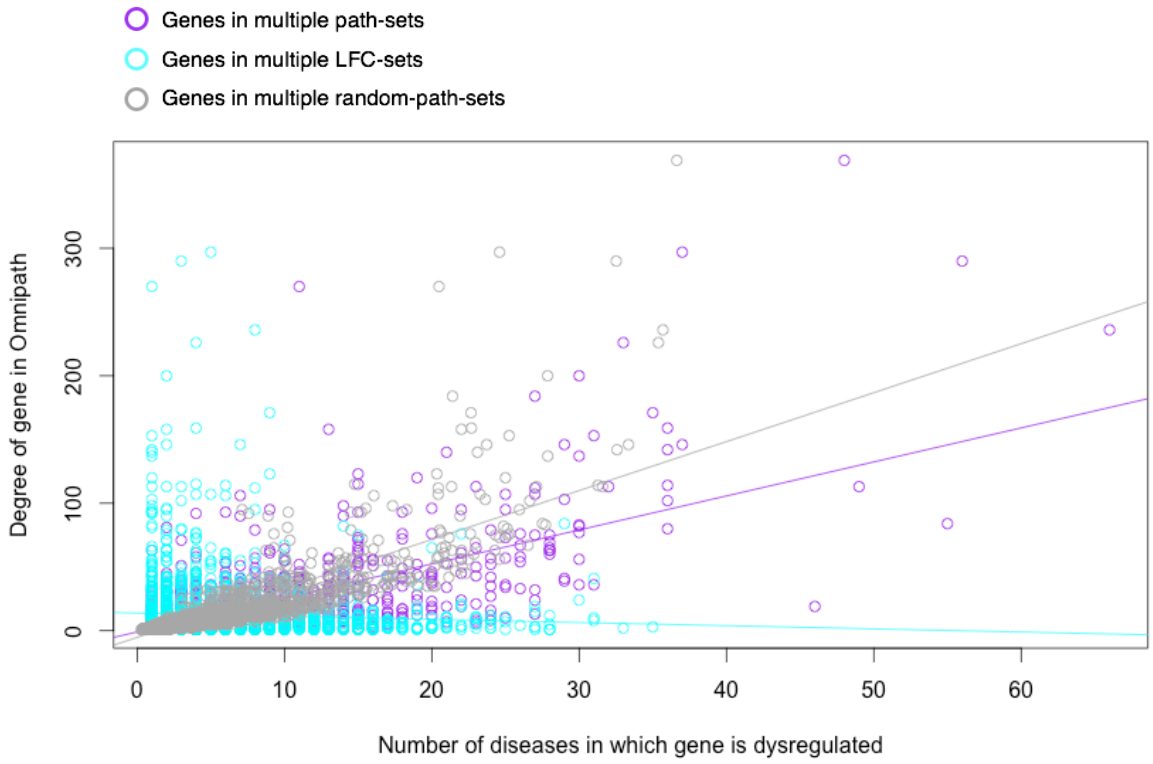


Figure 4.5 Moderate relationship between the number of dysregulated path-sets in which a gene is included and the degree of the gene in OmniPath

The figure shows the number of genes which are dysregulated in multiple experiments according to membership of the dysregulated path-set (purple points), membership of the same number of most highly dysregulated genes (LFC-set; cyan points), or membership of path-sets based on permuted log-fold change profiles (random-path-set; grey points). Genes which are included in more path-sets tend to be of higher degree in the OmniPath signalling network, although this relationship is weaker than in the random-path-sets (which represent the absence of log-fold change information). The plot also shows that there are some genes which are in the dysregulated path-sets of over 40 experiments, suggesting an influential role of these genes in the disease response of diverse disease types. Note that the x-axis is multiplied by 100 for the random-path-sets, as the randomization was repeated 100 times across the 141 diseases.

GO BP enrichment analysis of the 54 genes most frequently in LFC-sets did not return any significant terms. Enrichment analysis of the 54 genes most frequently in random-path-sets returned 8 terms, all of which were signalling pathways also returned by the analysis for the real path-sets (e.g. *biological regulation*, *signal transduction* and *cell communication*). This could be explained by the relatively strong relationship between the number of random-path-sets containing a gene and the degree of that gene (R^2 of 0.679; Figure 4.5): high-degree nodes in a signalling network will tend to be linked to signalling-related functions. This relationship confirms that in the absence of coherent log-fold change information, path-set analysis tends to select genes of high degree.

By contrast, there is almost no relationship between the number of LFC-sets containing a gene and the degree of that gene (R^2 of 0.003). The line of best fit shown in Figure 4.5 even suggests a slight negative trend, i.e. genes with high log-fold change values in multiple experiments do not tend to be central in the network. This in contrast to the genes frequently in path-sets, which show a moderate trend towards higher degree (R^2 of 0.470) but not as strongly as the random-path-sets. These genes may represent 'pressure points' in the network whose dysregulation is often associated with further dysregulation in their network neighbours. These genes would be expected to have higher-than-average degree, as they must interact with many different genes in order to influence diverse biological processes in different diseases; at the same time, the most important (highest degree) nodes in the network are unlikely to be frequently dysregulated, as these will be the critical nodes, dysfunction in which could be lethal to the cell.

4.3.5 Shared edges between diseases reveal unexpected disease relationships which are enriched for shared drugs and drug-interacting genes

Path-set analysis enables the comparison of diseases through shared dysregulated edges. A shared edge is an interaction between two genes that is contained in the path-set of both diseases, where each gene is regulated in the same direction in both diseases. Whilst shared dysregulated genes might represent isolated points on the biological network, shared edges are a stricter method of comparison which helps to make a stronger case for common mechanisms between two diseases.

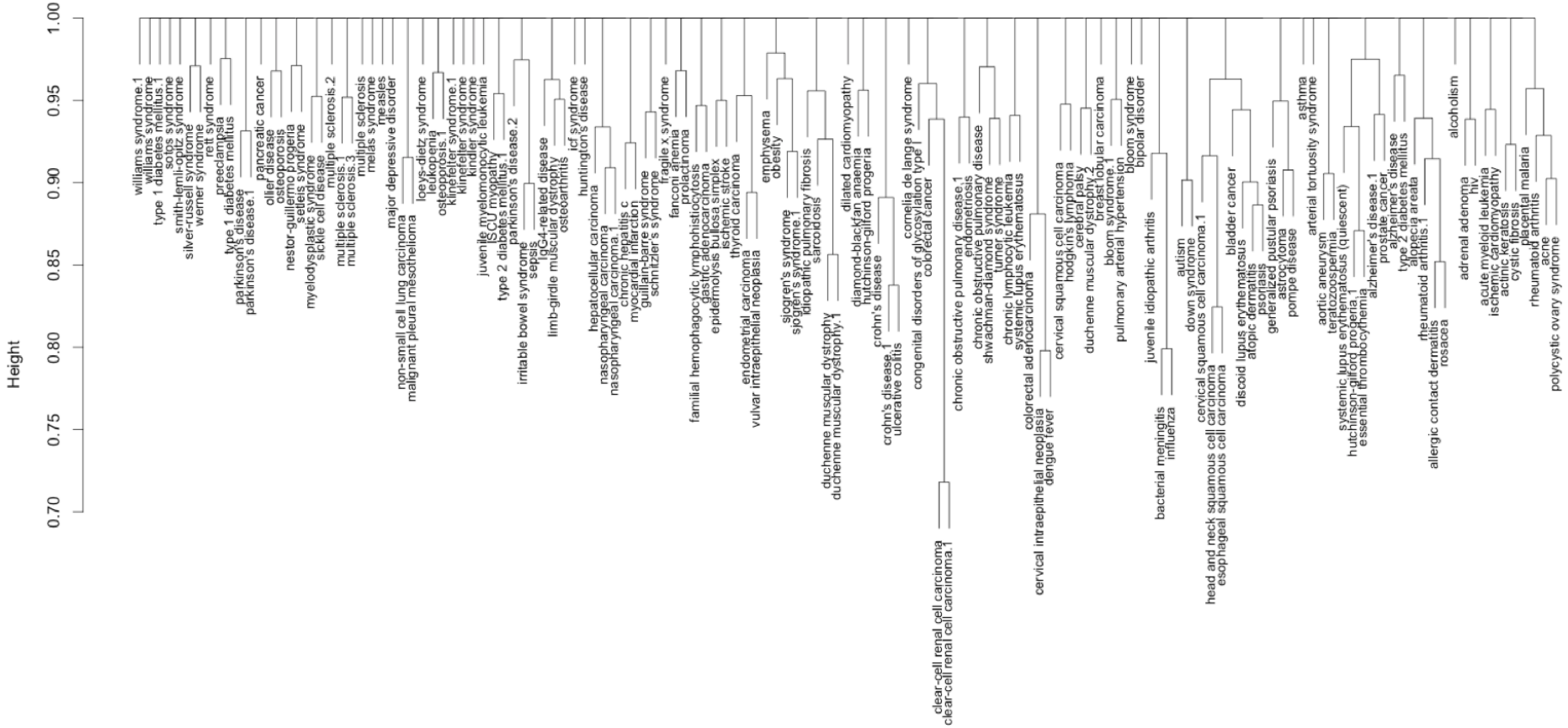


Figure 4.6 Disease similarity based on number of shared dysregulated edges shows known and novel similarities between diseases

Many of the most significant similarities are between different experiments measuring the same disease (such as clear-cell renal carcinoma, Duchenne muscular dystrophy, and Parkinson's disease) or between diseases of similar pathogenesis or anatomical location (such as head and neck squamous cell carcinoma and esophageal squamous cell carcinoma). Other similarities, such as between cervical intraepithelial neoplasia and dengue fever, are less expected.

Figure 4.6 shows a clustering of diseases according the size of their shared-edge-set, based on 676 significant disease pairs of a possible 9,870 pairs (see Methods for details of the similarity score calculation, including the permutation test to discover if the number of edges shared between two diseases is greater than would be expected by random chance). The shared-edge-sets were based on path-sets calculated from the 3,724 genes measured in all diseases, and using a more lenient threshold of 500 in order to capture as many shared edges between diseases as possible. If a threshold of 100 was used for this analysis, proportionately fewer shared edges were found (Appendix L).

Many of the clusters in Figure 4.6 reflect known disease relationships: there are strong edge similarities between multiple experiments measuring the same or similar diseases (renal clear cell carcinoma; nasopharyngeal carcinoma; Sjogren's syndrome; Parkinson's disease; ulcerative colitis and Crohn's disease; Duchenne muscular dystrophy) as well as between diseases which are symptomatically (polycystic ovary syndrome and acne) or pathologically (head and neck and esophageal squamous cell carcinomas) related. Overall, 45% of significant pairs are in the same Disease Ontology (DO) top-level class (e.g. '*disease of anatomical entity*'); 18% in the same DO sub-class (e.g. '*musculoskeletal system disease*'), making them 2.5 times more likely to be in the same sub-class than diseases that do not share edges (Table 4.2). Interestingly, 17 of the 27 same-disease pairs (two experiments measuring the same disease) do not share significant numbers of edges, again illustrating the discordance between different measurements of the same condition.

Table 4.2 Biological relevance of shared edges

Disease pairs that share a significant number of dysregulated edges are more likely to be in the same ontological class, more likely to share disease-associated genes, and are more likely to be treated with the same drugs, than diseases that do not share dysregulated edges. Note that this analysis includes disease pairs which are different experiments in the same disease (e.g. Parkinson's disease, Parkinsons' disease.1).

	Disease pairs with shared paths	**Disease pairs without shared paths**	**Fisher test *p*-value**
In the same Disease Ontology top-level class	44.7%	26.3%	$<2.20e^{-16}$
In the same Disease Ontology sub-class	18.1%	7.2%	$<2.20e^{-16}$
Share drugs (in Phase III clinical trials or approved)	17.2%	8.0%	$1.33e^{-13}$
Share disease-associated genes	19.2%	10.8%	$5.27e^{-10}$

However, many disease pairs which share edges are not obviously related, such as the connections between diseases in different DO classes shown in Table 4.3. Some of these relationships are not entirely unexpected – for instance, acne is a symptom of polycystic ovary syndrome, so it is unsurprising to find that these two share some dysregulated processes. Other relationships are less obvious, such as between cervical intraepithelial neoplasia (CIN) and Dengue fever. Table 4.3 shows that their shared-edge-set contains five drug-interacting genes for CIN, suggesting that the shared edges are strongly linked to a known (druggable) biological process dysregulated in CIN, which may also be relevant in Dengue fever.

Table 4.3 Selected disease pairs with significant numbers of shared edges

65% of significant shared-edge-sets contain disease-associated genes or drug-interacting genes (KDGs) for one or both of the diseases. Under the guilt-by-association hypothesis, a KDG for one disease in the shared-edge-set with another disease may indicate possible relevance of that KDG in the other disease. This is especially of interest for novel disease connections (those in different Disease Ontology (DO) classes or subclasses) and links between common and rare diseases (italicized).

Disease pair	**Number of genes in shared dysregulated edges**	**Disease-associated genes in shared dysregulated genes**	**Drug-interacting genes in shared dysregulated genes**
In the same DO subclass			
Crohn's disease.1, ulcerative colitis	98	STAT3 PLAU IL1R1 IL7R	ENG ANXA1 MMP9
Atopic dermatitis, psoriasis	51	STAT3	CCND1 STAT3
Cervical squamous cell carcinoma.1, esophageal squamous cell carcinoma	96	RB1	TOP2A ITGB3
In different DO (sub)classes			
Hutchinson-gilford progeria.1, thrombocythemia	54	PTPN11	-
Epidermolysis bullosa simplex, prostate cancer	15	SMA4 GNAQ	ADRA1A PRKCA
Cervical intraepithelial neoplasia, dengue fever	77	-	BIRC5 STMN1 BRCA1 RRM1 RRM2
Acne, polycystic ovary syndrome	48	IRF1	ITGB2
Actinic keratosis, sepsis	33	-	PRKCA

Overall, 42% of significant shared-edge-sets contain a disease gene for at least one of the diseases, and 52% of significant shared-edge-sets contain a drug-interacting gene for at least one of the diseases. This figure increases the greater the number of edges that are shared above random, so for the top 100 most significant disease pairings (of the 676 significant pairs, the 100 pairs with the greatest difference in the number of edges shared from the median number of shared edges observed in the random permutation test – this is actually 106 pairs here, due to ties), 62% of shared-edge-sets contain a disease-associated gene and 75% a drug-interacting gene. Excluding same-disease pairs (e.g. Parkinson's disease, Parkinson's disease.1), 7.4% and 15.9% of these genes respectively are relevant to both diseases.

Table 4.4 Disease-associated and drug-interacting genes in shared edges
Where the shared-edge-set of two diseases includes a disease-associated or drug-interacting gene, this suggests that the shared edges are capturing a disease-relevant process, particularly where this gene is associated with both diseases. For the top 100 most significant disease pairs (those with the highest number of shared edges compared to random expectation), 16% of drug-interacting genes in shared edges are relevant to both diseases. This suggests that some of the remaining genes (currently associated with only one disease) could also be relevant drug-interacting genes in the other disease, leading to potential drug repurposing suggestions.

	All 676 significant disease pairs	**Top 100 most significant disease pairs**
Percentage of significant disease pairs which include a disease-associated gene for either disease in their shared-edge-set	41.7%	62.2%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	6.4%	7.4%
Percentage of significant disease pairs which include a drug-interacting gene for either disease in their shared-edge-set	51.9%	74.5%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	9.8%	15.9%

The percentage of drug-interacting genes in shared edges which are applicable to both diseases suggests that within the shared-edge-sets, there may be other drug-interacting genes currently associated with only one of the diseases which could be applicable in the other disease. This could imply the possibility for drugs to be repurposed from one disease to the other. In fact, diseases that share edges are 1.78x as likely to share drugs than diseases that don't share edges (Table 4.4), rising to 3.38x as likely for the top 100 most significant disease pairings.

4.3.6 Shared paths highlight shared mechanisms between rare and common diseases which may be used for drug repurposing

The shared-edges approach is particularly useful for investigating connections between common and rare diseases, enabling the discovery of potential disease-associated genes through an association transfer approach: common diseases are usually better studied than rare diseases, so are more likely to have known disease-associated genes. If these KDGs are associated with processes shared by the two diseases, then they could feasibly also be important in the rare disease.

One example of a connection between a common and a rare disease is the link between polycystic ovary syndrome and Pompe disease. Polycystic ovary syndrome (PCOS) is a common condition in which elevated levels of androgens are produced in the ovaries, resulting in anovulation, irregular periods, and difficulty conceiving. The exact cause is not known, but it is thought to be associated with high blood insulin levels caused by insulin resistance. Pompe disease (otherwise known as Type II glycogen storage disease) is an inherited metabolic disease caused by deficiency in the acid alpha-glucosidase enzyme, which results in the accumulation of glycogen inside cellular lysosomes, causing progressive muscle wasting, liver enlargement, and respiratory difficulties. Interestingly, polycystic ovaries are known to appear in females with glycogen storage disease at a much higher prevalence than in healthy females[204], although without necessarily displaying related symptoms such as amenorrhea[205]; the link is thought to be due to impaired glucose tolerance in Pompe disease patients.

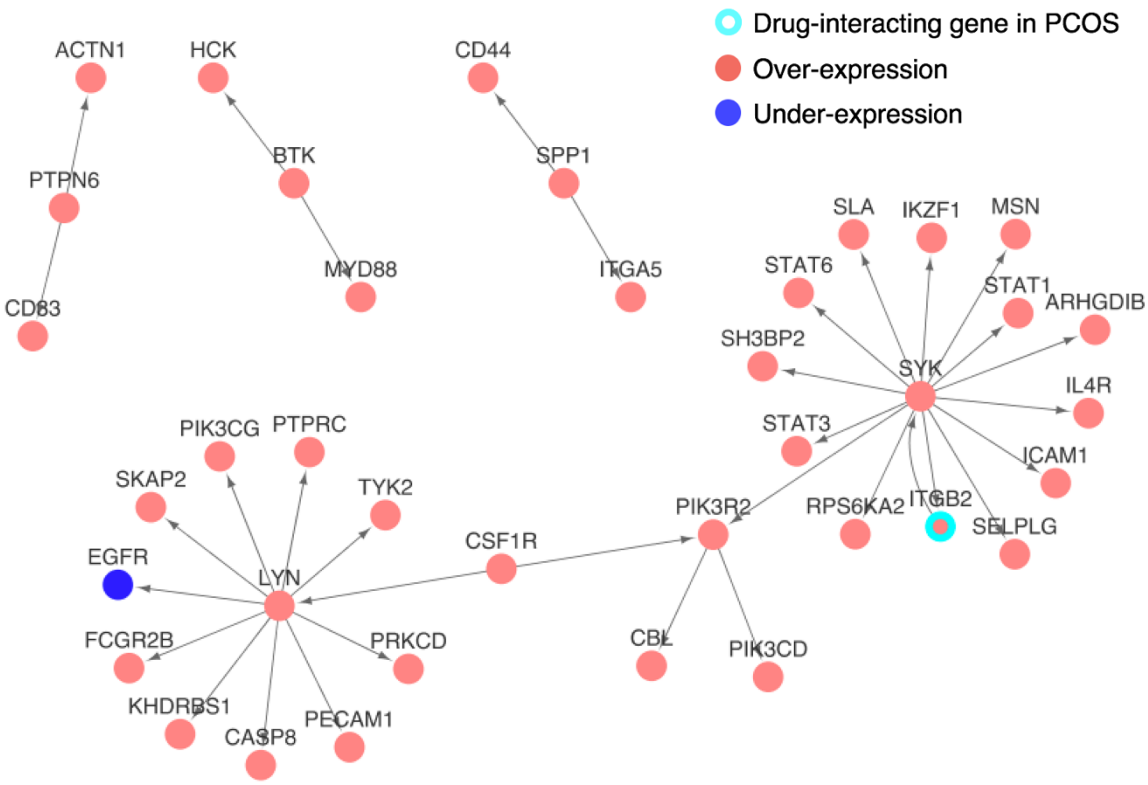


Figure 4.7 Paths dysregulated in both polycystic ovary syndrome and Pompe disease

The shared paths are centred around the tyrosine kinases LYN and SYK, and show the interaction of SYK with integrin-beta 2 (ITGB2), a target of the PCOS drug simvastatin. The entire network component is upregulated in both diseases, aside from a single down-regulated gene, EGFR.

Figure 4.7 shows the shared edges dysregulated in Pompe disease and PCOS. The shared dysregulated components are centred around two hub genes, LYN and SYK. Activation of spleen tyrosine kinase (*Syk*)[206], and possibly tyrosine-protein kinase Lyn (*Lyn*)[207], together with insulin-mediated activation of PI3-kinases[208] (in Figure 4.7, represented by PIK3R2, PIK3CG, PIK3CD), is involved in activation of *Akt* (protein kinase B). *Akt* inactivates glycogen synthase kinase 3 (*GSK-3*)[207], inducing glycogen synthesis. The upregulation in the genes corresponding to these kinases might therefore be indicative of increased glycogen synthesis, which would seem counter-intuitive: in PCOS, decreased insulin-stimulated glycogen synthesis has previously been reported in granulosa cells[209]. It should be noted, however, that 3 of the 7 PCOS samples in the gene expression study used here were from non-insulin-resistant patients, which might be one explanation for the observed upregulation in this part of the glycogen synthesis pathway. In Pompe disease, increased glycogen synthesis also seems counter-intuitive, but could be related to the inability to break down stored glycogen – glycogen storage has previously been found to correlate with an increase in glycogen synthesis-promoting factors with in a murine model of Pompe disease[210].

Also highlighted in Figure 4.7 is the upregulated integrin-beta 2 (ITGB2), a drug-interacting gene in PCOS which participates in a two-way interaction with SYK. *ITGB2* is the beta subunit of the integrin *LFA-1*, which is inhibited by simvastatin[211]. Statins, such as simvastatin, act by inhibiting HMG-CoA reductase, which has two possible therapeutic mechanisms in PCOS. The first is the reduction of cholesterol synthesis, which may in turn result in decreased androgen production; the second is through the reduced production of another product of the same pathway, dolichol[212]. Dolichol is required for maturation of insulin receptors, so decreasing its levels may therefore reduce the effects of excess insulin in PCOS[212]; in Pompe disease, this could potentially support the reduction of glycogen synthesis through reduced insulin receptor levels. Unfortunately, a literature search reveals that this finding may not be particularly promising due to the potential of statins to cause myopathy in patients with Pompe disease[213]. It is also worth noting that the activity of simvastatin on the shared target *ITGB2* specifically is thought to produce an anti-inflammatory effect[214] rather than being involved in the HMG-CoA pathway, therefore the shared path-set shown in Figure 4.7 might not be directly relevant to the desired mechanism of action.

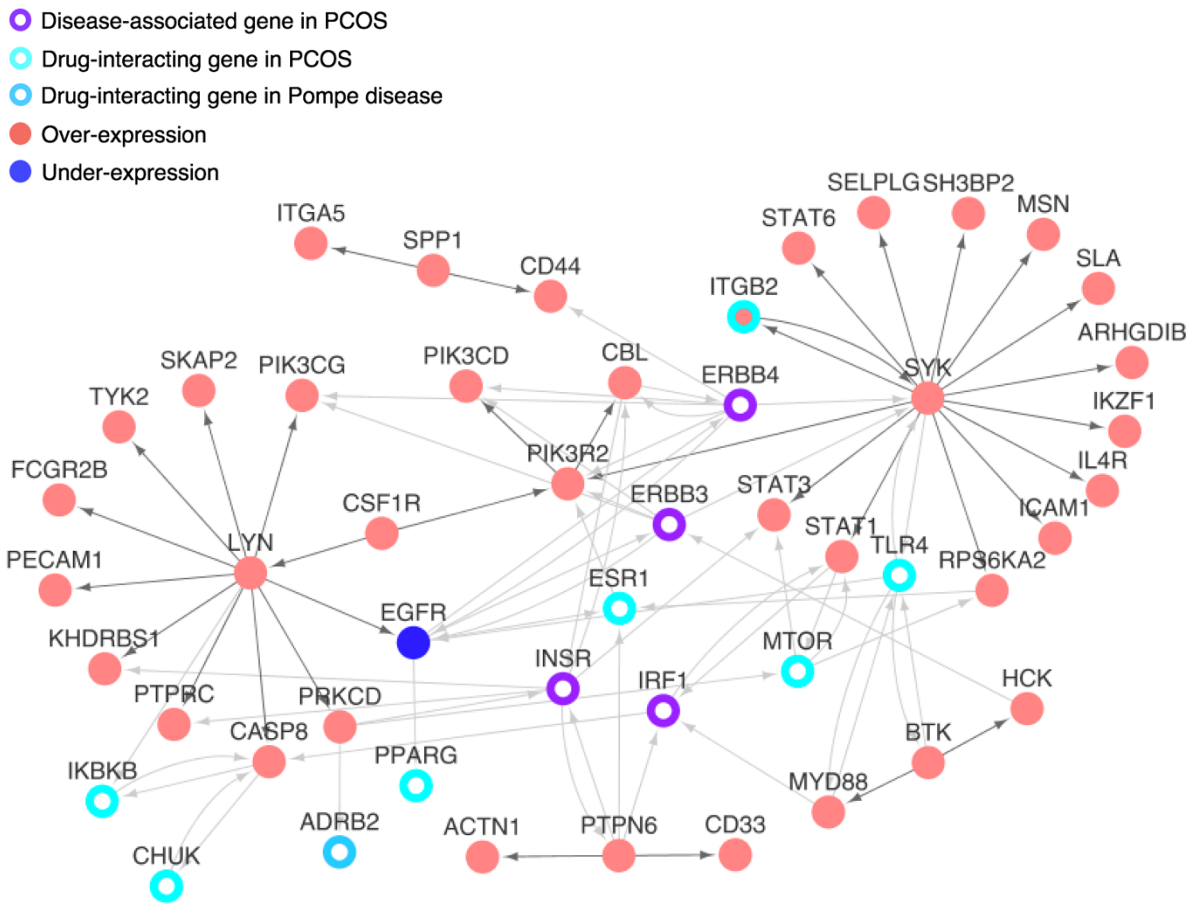


Figure 4.8 Paths dysregulated in polycystic ovary syndrome and Pompe disease, including first-neighbour known disease-associated genes

The first-neighbours of genes in the shared-edge-set include multiple genes which are known genetic variants or drug-interacting genes in PCOS. Only one of these first-neighbour genes is related to Pompe disease, which is somewhat expected due to the rarer disease being less well studied – Pompe disease has only 3 known disease-associated genes, compared to 78 for PCOS.

As discussed in Section 4.3.3, KDGs that are not differentially expressed in disease can be associated with dysregulated paths by searching the first-neighbour genes. Amongst the first-neighbour KDGs of the shared-edge set of PCOS and Pompe disease (Figure 4.8) is MTOR. The mechanistic target of rapamycin complex 1 (*mTORc1*), of which *mTOR* is a core component, is inhibited by metformin, a drug commonly prescribed for type II diabetes due to its ability to decrease high blood glucose levels, but which may also be used in PCOS[215]. Chronic activation of *mTORc1* is known to play a role in insulin resistance due to feedback inhibition of insulin signalling, and so its inhibition by metformin may improve the metabolic profile in insulin-resistant individuals[216].

In Pompe disease the problem is not insulin resistance but glycogen accumulation. Here, *mTOR* inhibition is also of therapeutic interest: inhibition of *mTOR* via rapamycin has been shown to block amino-acid induced inactivation of glycogen synthase kinase 3, decreasing glycogen synthesis[217], making it a potential therapeutic avenue for glycogen storage diseases. Rapamycin therapy has demonstrated some benefit in a canine model of glycogen storage disease type III[218] and may also be potentially useful in Pompe disease (glycogen storage disease type II)[219], although this is not without controversy[220]. Given the similarity between PCOS and Pompe disease discussed here, perhaps metformin could also be considered as a useful *mTOR* inhibitor in Pompe disease.

4.3.7 Path-set analysis captures the mechanism of action of cediranib

Table 4.5 Known disease-associated genes and drug targets (KDGs) in path-sets compared to in the same number of OmniPath genes by absolute log-fold change or random selection.

Path-sets capture more KDGs compared to selecting the same number of genes by log-fold change (LFC-set) or random selection.

		Path-set	**LFC-set**	**Random-path-set**	**Random-gene-set**
What proportion of sets contained at least one KDG?	*Drug targets only*	0.18	0.24	0.05	0.02
	Disease- and drug targets	0.47	0.32	0.34	0.23
How many KDGs were found per set on average?	*Drug targets only*	0.47	0.35	0.06	0.02
	Disease- and drug targets	1.05	0.74	0.60	0.31

Path-set analysis was also applied to human drug response datasets, to investigate whether a drug's path-set may be able to reveal details of its mechanism of action. The analysis in Section 4.3.2 was repeated for the drug response datasets, although here, the definition of a relevant gene is slightly different. For drug response, a relevant gene is defined as a the gene corresponding to a target related to the drug's primary mechanism of action (as defined in Section 4.2.2). Given that most drugs in this dataset have only one primary target, and that only 33% of target genes show non-zero differential expression in the corresponding dataset, it is not surprising to find that the primary target could only be discovered in three of the 17 drug path-sets for which a primary target was available – cediranib, sunitinib, and tamoxifen.

However, the definition of a relevant target can also be extended to include disease-associated genes for the indication in which the drug experiment took place, under the assumption that an effective drug might target proteins located near to disease genes in the interactome (a loosening of the disease modules concept described in the work of Guney et al.[221]). This could provide a way to evaluate whether the drug path-set contains genes related to its mode of action in cases where the target itself is not differentially expressed. Under this definition, drug-relevant targets are discovered in 9 of the 19 datasets, finding a mean of 1.05 targets per path-set (compared to a mean of 0.74 targets per LFC-set).

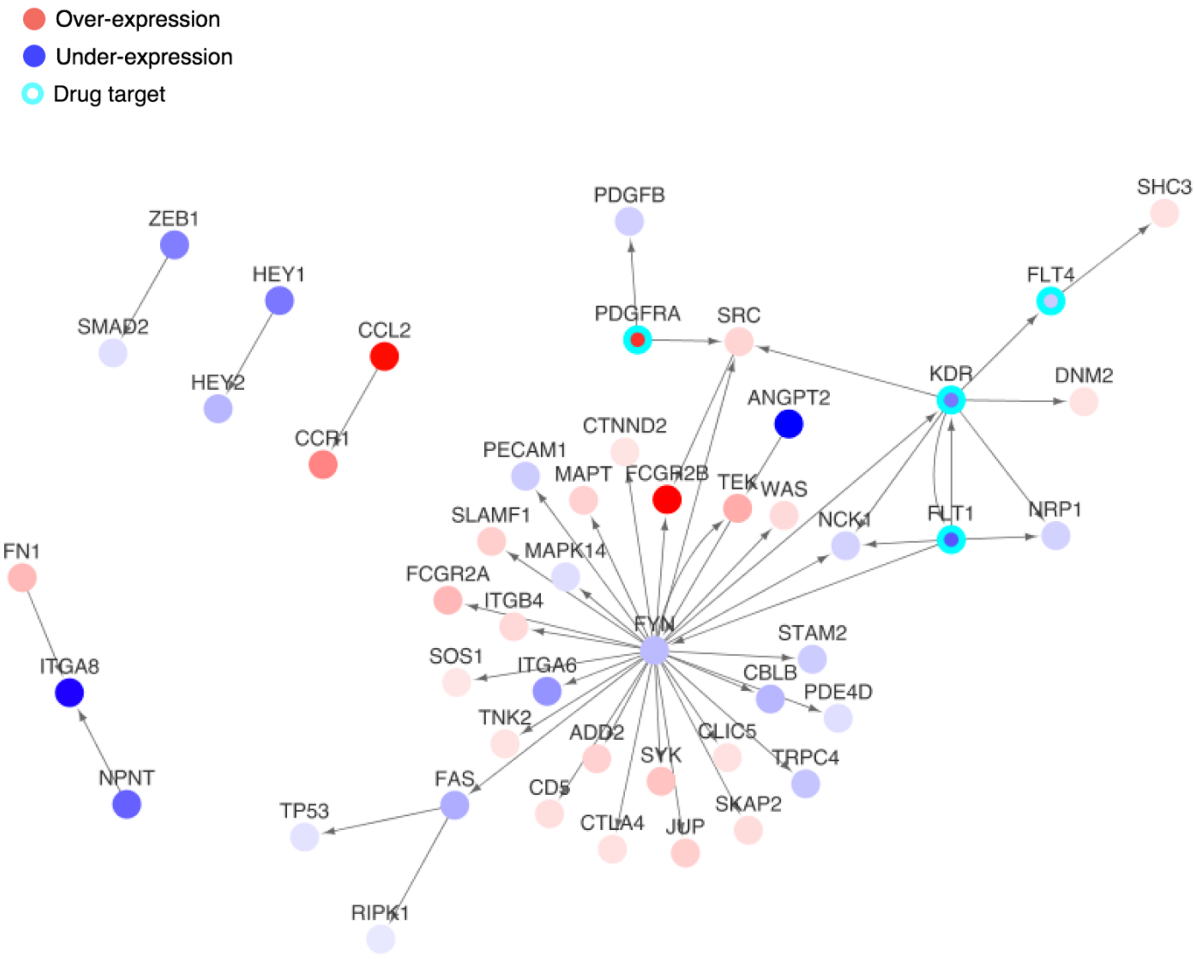


Figure 4.9 Signalling paths dysregulated following cediranib administration
The dysregulated path-set for the vascular endothelial growth factor receptor (VEGFR) inhibitor cediranib shows downregulation of its VEGFR targets (FLT1, FLT4, and KDR) and dysregulation of downstream nodes, which may be mediated through FYN kinase.

The path-set of cediranib, a vascular endothelial growth factor receptor (*VEGFR*) inhibitor for the treatment of various cancers, was examined as a case study. No known disease-associated genes of alveolar soft part sarcoma (the rare cancer in which cediranib response was studied) are seen in the cediranib path-set, but genes corresponding to its target *VEGFR*s (FLT1, FLT4, and KDR) are shown to be down-regulated (Figure 4.9). The kinase *Fyn*, whose corresponding gene forms a hub in this path-set, is activated by *VEGFR1* (FLT1) signalling[222]; FYN expression is decreased in this study, suggesting possible transcriptomic disturbance following *VEGFR1* inhibition. By contrast, the LFC-set of cediranib does not include its target FLT4, and does not capture the hub gene FYN, thereby missing a possible network structure through which transcriptomic signals of *VEGFR* inhibition may be transmitted in the cell. Another cediranib target whose corresponding gene is shown in Figure 4.9 is platelet-derived growth factor receptor alpha (*PDGFRA*), which is also inhibited by cediranib due to its structural similarity to *VEGFR*[223]. In contrast to the *VEGFR* genes, however, PDGFRA is upregulated following cediranib administration, suggesting a role for different transcriptomic feedback mechanisms in cediranib response.

4.4 DISCUSSION

Given the noise inherent to gene expression data, new methods of analysis are needed which bridge the gap between individual fold change analysis and high-level biological pathway methods. This chapter describes the use of path-set analysis to relate gene expression changes on the human signalling network, enabling the discovery of dysregulated signalling paths in disease and drug response. The adaptation of the original framework proposed by Sambarey et al.[121] constrains the paths to differentially expressed genes, forming disease-specific subpaths which can be used to compare dysregulated paths across diseases.

The dysregulated paths are enriched for known disease-associated genes and drug-interacting genes (KDGs) compared to a gene-set of the same size based on magnitude of differential expression (Table 4.1). Whilst the aim of path-set analysis is not to find putative disease-associated genes (in which case limiting the path-sets to differentially expressed genes would be counter-productive), their presence confirms the relevance of the returned paths, and aids the interpretation of the observed gene expression by showing how dysregulated genes are interacting with genes known to influence the pathogenesis or treatment of disease. In drug response datasets, while only a few path-sets captured the primary targets of the drug (possibly due to the low likelihood of the drug's target to be differentially expressed in response to drug administration[224,225]), path-set analysis captured known disease variant genes for the disease in which the drug was given for 9 of the 19 drug path-sets. This could be used as a measure of efficacy of a drug in a given disease, i.e., an 'effective' drug (one that treats the underlying cause of a disease rather than just the symptoms) is one that affects the expression of disease-associated proteins. This concept was introduced by Guney et al.[221], who used the proximity of a drug's targets to disease-associated proteins as a measure of efficacy, successfully applying this measure to make suggestions for drug repurposing.

Applying path-set analysis over multiple diseases allows the identification of genes which are frequently dysregulated in disease. These genes may be acting as points of influence in the network whose modulation is associated with the development of a disease, either causal (resulting from a disruption in function through a variant or mutation) or opposing (alleviating symptoms of the disease through pharmacological interference). The frequency with which these genes are involved in dysregulated paths shows a moderate correlation with their degree (one measure of network importance), which is not evident for genes which show high log-fold changes in many diseases (Figure 4.5). Previous research on the network topology of

disease-associated processes has been focused on genetic variants, and this has produced somewhat differing results[226]. One key study found that 'essential' genes required for survival tended to be network hubs, whereas disease-associated genes tended to be non-essential, non-hubs[15]. However, genes which show somatic mutations in disease (i.e., cancer-associated genes) tend to have a more central network position[15,226]; and research looking specifically at rare diseases has found that genes associated with rare diseases do tend to be hubs[227]. Similarly, drug targets should be 'highly influential in, but not toxic to, the functioning of the entire network'[228]; drugs targeting hub genes tend to have more side effects[229]. The moderate correlation between a gene's frequency in path-sets and its degree may be explained by the same principle: genes that are moderately important in signalling are frequently in path-sets, but the most important (highest-degree) genes which are essential to the functioning of the cell tend not to show expression dysregulation, and are therefore not so likely to be included in path-sets.

As every gene in a disease's path-set is differentially expressed in the disease, path-set analysis can also be used to identify dysregulated paths which are shared between diseases. Disease pairs that share edges are 2.5 times more likely to be in the same Disease Ontology subcategory than those that don't (Table 4.2), confirming that shared edges reflect known disease relationships. Potentially more interesting are the cases where dysregulated paths are shared between two diseases which are not known to be related. The case study of PCOS and Pompe disease illustrated how path-set analysis can highlight dysregulated processes shared between phenotypically distinct diseases, and how drug-interacting genes in shared paths might be used to suggest potential drug-sharing options between the two diseases. With 52% of disease pairs containing a drug-interacting gene for at least one of the diseases, this analysis could be extended to many other disease pairs. This analysis can also be applied to non-differentially expressed first-neighbour genes, as shown with the example of metformin in PCOS and Pompe disease. Targeting first-neighbours of differentially expressed genes has been proposed as a drug repurposing strategy in cancer[230], and network proximity to differentially expressed genes has been shown to be a good predictor of potential drug targets for repurposing[231], suggesting that this could be a viable repurposing strategy where diseases share dysregulated edges.

A potential limitation of path-set analysis is the applicability of the general human signalling network on which this method is based. The interactome is known to vary across different tissues and may be altered in disease[232]; future developments of path-set analysis could incorporate tissue- or disease-specific interactomes which would provide a more accurate

picture of interactions taking place in individual diseases. Unfortunately, our knowledge of the general human interactome (much less condition-specific alterations) is highly incomplete[84], and although this work is based on the most comprehensive signalling network published to date[82], it will necessarily contain inaccuracies and omissions. Related to this, genes with no known interactions are necessarily excluded from the analysis, meaning that important genes may potentially be omitted. This situation should improve in future, as our knowledge of the interactome develops further. Despite its current limitations, network data represents a valuable 'extra dimension' whose integration with gene expression data can result in improved interpretability and ease of analysis.

Path-set analysis represents a valuable addition to the transcriptomic analysis toolkit, which can be used to identify interactions between dysregulated genes, genetic variants, and drug-interacting genes in disease. Here, path-set analysis was used to discover dysregulated processes shared between diseases, highlighting common molecular mechanisms underlying disease and revealing new connections between conditions. As the vast amount of transcriptomic data continues to grow, this type of analysis will be key to improving our understanding of gene expression in disease.

5 UNDERSTANDING AND PREDICTING DISEASE RELATIONSHIPS THROUGH SIMILARITY FUSION

This work was previously published as Erin Oerton, Ian Roberts*, Patrick S. H. Lewis*, Tim Guilliams*, Andreas Bender. Predicting disease relationships through similarity fusion. *Bioinformatics* Advance Access published Aug 30, 2018, doi:10.1093/bioinformatics/bty754.

* Healx Ltd, Park House, Castle Park, Cambridge CB3 0DU, United Kingdom

The work represents the result of a collaboration with the listed co-authors. All work described here was carried out by the author except the text-mining for literature co-occurrence of diseases, as noted in the text. All analyses, text, and figures were produced by the author, incorporating comments from co-authors.

SUMMARY

Relationships between diseases can be defined on multiple levels, from the observable phenotype down to molecular-level events. Combining information across these levels could yield a systems-level view of disease relationships, aiding our understanding of common biological processes taking place in disease. However, each of these levels differs in features and information content, and it is unclear how they could be most effectively combined. In this chapter, a similarity fusion approach is proposed which enables comparison of diverse data types. This method is applied to 6 different data types (ontological, phenotypic, literature co-occurrence, genetic association, gene expression, and drug indication data) for 84 diseases to create a 'disease map': a network of diseases connected at one or more biological levels.

The fused similarities are used to classify diseases into known categories from the Disease Ontology. With a mean Random Forest AUROC of 0.95 for these two tasks, the disease map scores over 10% higher than the mean of its component spaces, confirming that the fused values are good predictors of known disease relationships. As well as known relationships, 15% of links in the disease map are novel links that span traditional ontological classes, such as between psoriasis and inflammatory bowel disease. 62% of diseases linked in the disease map share drugs (approved or in Phase III clinical trials), illustrating the relevance of the disease map to the identification of potential therapeutic relationships. The analysis presented here illustrates how similarity fusion can give greater insight into shared disease biology than individual data types alone.

5.1 INTRODUCTION

Establishing relationships between diseases increases our understanding of disease biology, aiding the identification of shared mechanisms or development of new treatments, for example through drug repurposing. As discussed in Chapter 1, existing disease classification systems such as the International Classification of Diseases[8] and Medical Subject Headings[11] are based on established clinical relationships between diseases. There is therefore great biological and pharmacological interest in the identification of novel disease relationships using new types of evidence arising from the development of bioinformatics technologies.

As well as the gene-expression based approaches to relating diseases covered in Section 1.4.1.2[40,107], other -omics data types that have been used to explore disease relationships include disease-associated genes[15,233,234], protein interaction networks[84], pathways[235] and biological processes[236]. Rather than examining each of these different data types in isolation, however, recent studies have related diseases by considering multiple data types simultaneously. These data integration approaches can provide a more comprehensive understanding of disease, potentially reflecting interactions between the different layers of the biological system[237] where links at one layer (e.g. genetic variance) are associated with changes at another layer (e.g. gene expression or phenotype). Recent examples have demonstrated how this can be achieved through the use of heterogeneous networks, such as the DiseaseConnect web server developed by Liu et al.[238], or through matrix factorization approaches, such as that presented by Zitnik et al.[239].

However, these approaches do not quantify the overall strength of the relationship across multiple levels. Defining a measure of disease similarity that takes into account multiple data types is not straightforward, as such a measure must consider differences between properties such as information content[122]. Sun et al.[240] evaluated disease similarity by defining a feature vector for each disease in which every element (genes, chemicals, pathways, and GO terms) was weighted according to its information content. The downside of this approach is that it requires an entry for each entity in the feature universe, needing a feature vector of tens of thousands of dimensions to represent just four spaces. Computing similarity across multiple spaces by this approach therefore does not scale readily to large numbers of feature spaces.

In this work, this issue is addressed by translating the feature vectors in each space into pairwise disease similarities, thus capturing disease relationships in a lower-dimensional space before performing the integration step to define an overall measure of similarity. This

'similarity fusion' approach has been successfully applied to integrate data in drug repurposing[241,242], gene prioritization[243] and patient subtyping and survival analysis[244,245]. Yet there have been few applications of this approach to quantify disease similarity. In one study, disease similarities were computed by integrating literature-based similarity of diseases with protein interaction network topology-based similarity of their associated genes[246]; more recent work related diseases through 'meta-correlation', combining similarity amongst gene expression and electronic health record profiles of diseases[247]. Another study integrated similarity in nine different spaces according to a pre-defined 'importance', with the resulting relationships weighted towards genetic similarities[248]. Although the relative 'importance' of each relationship type naturally depends on the context in which the map is used, no study has yet defined a general method for the combination of multiple disease similarities in an unbiased manner. In particular, unbiased combination of spaces requires consideration of the underlying distributions of similarity in each space. Here, quantile normalization (usually associated with microarray statistics) is used to adjust the distributions of similarity in each space, enabling balanced comparison and combination of disease similarities across multiple spaces.

In summary, in this chapter the proposed similarity fusion approach is applied to six different data types – ontological, phenotypic, literature co-occurrence, genetic association, gene expression, and drug data – to create a disease map: a network of diseases connected at one or more biological levels. The disease links revealed by the map are explored, with a focus on disease pairs not previously known to be related, and evaluated against their relation to existing disease classifications and drug-sharing relationships.

5.2 METHODS

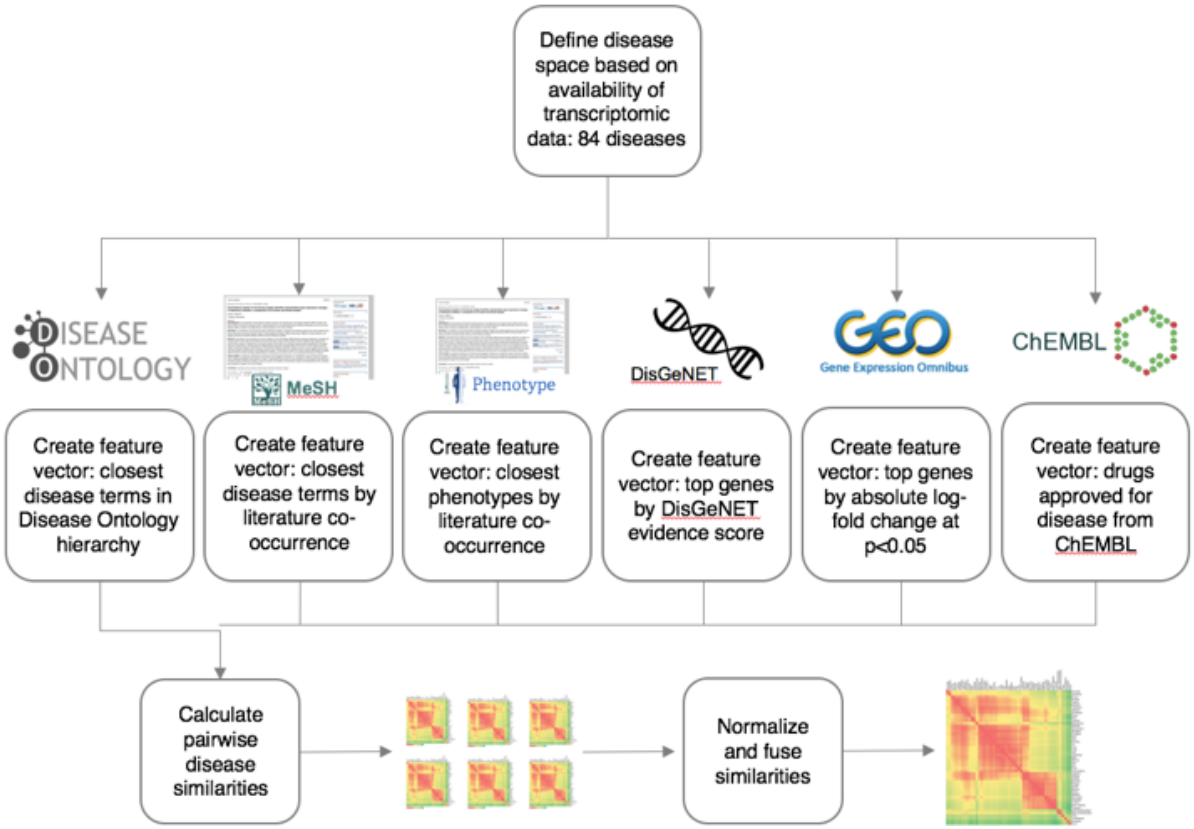


Figure 5.1 Disease similarity fusion workflow

Disease data from six different 'feature spaces' are transformed into symmetric similarity matrix representations. Feature sets for each disease are formed of the approximately 100 top features in each space, although the exact number varies slightly dependent on the available data. Similarity matrices representing each individual feature space are then normalized and combined into a single fused similarity matrix. The disease relationships represented by this matrix can be analysed to find novel links between diseases, or links which may represent drug-sharing opportunities.

5.2.1 Disease dataset construction

The disease dataset which formed the basis of this work was compiled based on available transcriptomic experiments as described in Section 2.2. In order to ensure data coverage across all six spaces, this was limited to common diseases. Where multiple transcriptomic experiments were present for a disease, one experiment was randomly selected to represent the disease. This resulted in a dataset of 84 diseases, some of which were closely related (e.g. asthma and allergic asthma; see Appendix C). 39 of these diseases are in the Disease Ontology (DO) class *disease of anatomical entity*, 25 are in the DO class *disease of cellular*

proliferation, and the remainder are distributed across the other top level DO classes, with the exception of class *physical disorder* (no diseases in the dataset belong to this class). These diseases were mapped to the most closely matching disease terms in each space (e.g. 'teratozoospermia' may map to 'azoospermia' or simply 'male infertility', depending on what representation is available in each space; see Appendix D for details).

Feature sets for each disease were then constructed as detailed below in each of ontological, phenotypic, literature co-occurrence, genetic, transcriptomic, and drug spaces (Figure 5.1) using R version 3.3.2[133]. The feature set size of phenotypic space was restricted by the dataset used; the feature set size of drug space was restricted by the limited number of drugs prescribed for each disease. For the remaining spaces, different feature set sizes of 20, 50, 100, and 200 were tested. A feature set size of 100 was chosen, as this captured sufficient information in each space whilst not being overly large compared to the fixed-size feature spaces (see Appendix M for exploration of different feature set sizes). The exact number may be slightly more or less than 100 for some diseases due to e.g. ties in the data, this is detailed for each feature space below.

Ontological feature space: The Disease Ontology[191] was downloaded from http://ontologies.berkeleybop.org/doid.obo in December 2016 and used to match disease names to their Disease Ontology ID. The DOSE package[249] was used to measure semantic similarity between Disease Ontology IDs using Lin's measure[250], which minimizes the number of ties between terms. The feature set for each disease was then calculated as the top 100 most similar diseases according to this metric, excluding self-similarity. When there are ties for the top 100th similarity value, all tied diseases are retained, so some diseases have more than 100 features (up to a maximum of 135). Two diseases (irritable bowel syndrome and polycystic ovary syndrome) have no similar diseases according to DOSE, so these are assigned a feature vector of size zero in this space.

Phenotypic feature space: Disease-phenotype associations were taken from the work of Hoehndorf et al.[251] who matched diseases with terms from the Human and Mammalian Phenotype Ontologies based on literature co-occurrence. This dataset, which comprises the 21 most highly-associated Human/Mammalian Phenotype Ontology terms for each disease, was obtained from http://aber-owl.net/aber-owl/diseasephenotypes/data/ in November 2016, and this formed the feature set for the phenotypic feature space. Duplicated phenotypes in the supplied data (e.g. the term 'myocarditis' is found in both the Human and Mammalian

Phenotype ontologies) were removed, resulting in 16 of the 84 diseases having only 19 or 20 features in this space.

Literature co-occurrence feature space: 13 million Medline abstracts (dating between 2000 and 2016) were annotated with MeSH term identifiers using a recently published named entity recognition system, TaggerOne[252]. The normalized pointwise mutual information (NPMI) score between MeSH terms in these abstracts was calculated as a measure of co-occurrence. This work was carried out by Patrick Lewis, as named in the author list at the beginning of the chapter. These terms mostly represent MeSH disease concepts, although they also include a few more general concepts such as 'body weight' or 'infection', or higher level disease terms such as 'nervous system diseases' or 'musculoskeletal abnormalities'. The feature set for this space was the top 100 most highly co-occurring MeSH terms by NPMI score, excluding self-similarity. Three diseases had less than 100 co-occurring terms (allergic contact dermatitis with 23, male infertility with 55, and juvenile rheumatoid arthritis with 80). Certain MeSH terms overlap with the Human/Mammalian Phenotype Ontology terms used for the phenotypic space (e.g. 'diabetes mellitus', 'neoplasms' and 'carcinoma' are examples of terms that are included in both sets), so there is a degree of overlap between these two spaces.

Genetic feature space: Disease-gene associations were downloaded from DisGeNET[253] (http://www.disgenet.org/web/DisGeNET/menu/downloads) in November 2015. These mostly represent associations of type 'genetic variation', which includes susceptibility mutations, causal mutations, and modifying mutations; there are also a small number of associations of type 'post-translational modification' and 'therapeutic'. Associations of type 'AlteredExpression' were removed to avoid overlap with the transcriptomic feature space, and entries for the non-gene 'NEWENTRY' were removed. The feature set was composed of the top 100 associated genes by evidence score. There was high variation in the number of genes associated to each disease: 19 diseases have less than 100 associated genes (10 diseases have less than 20), but there were also a large number of ties in the data due to the calculation of the evidence score for each gene, leading some diseases to have more than 100 associated genes. 10 diseases had more than 100 genes, up to a maximum of 207 genes for malignant pleural mesothelioma.

Transcriptomic feature space: Gene expression microarray experiments were selected as described in Section 2.2, and differential expression profiles for each probe were generated as described in Section 2.3. The feature 'universe' was defined as the set of 4,482 genes measured

in all experiments; the feature set for each disease was then calculated as the top 100 of these genes by absolute log-fold change at a p-value threshold of <0.05.

Drug feature space: Drug indication data was downloaded from ChEMBL version 22.1[192] (https://www.ebi.ac.uk/chembl/downloads), as in Section 4.2.2. The feature set comprised approved drugs for each condition. The number of approved drugs listed for each disease in ChEMBL ranges from 0 (for 11 conditions, including four diseases – dengue fever, leukopenia, limb-girdle muscular dystrophy, and measles – which could not be mapped to EFO or MeSH terms used by ChEMBL) to 72 (for type II diabetes).

The spaces have different sparsities: phenotype is the most sparse space, with only 7.5% of disease pairs having any overlap in their phenotypes; followed by ontological at 12.9%, drug at 13.6%, co-occurrence at 58.5%, genetic at 83.3%, and finally transcriptomic space, with 84.4% of disease pairs having some overlap. This is related to the size of the feature set in each space relative to the size of the feature universe.

5.2.2 Independent comorbidity dataset

Comorbidity associations based on Medicare records of 13 million patients[254] were downloaded from sbi.imim.es/data/hudine in July 2018. Diseases are recorded in this data as ICD9 3-digit codes; mapping the set of 84 diseases to these codes resulted in duplicated codes for 14 diseases (e.g. type 1 and type 2 diabetes mellitus both map to 250 *diabetes mellitus*; bipolar disorder and major depressive disorder both map to 296 *episodic mood disorders*; see Appendix D for mappings).

Disease pairs with less than 100 co-occurrences were filtered out (as the relative risk (RR) comorbidity measure tends to overestimate for pairs with small numbers of observed associations[255]), leaving 88,347 disease pairs for which comorbidity data was recorded. 800 of these observations related to disease pairs in the dataset of 84 diseases, which when the 14 duplicate mappings are included covers 938 (27%) of the 3,486 disease pairs in the dataset. Relative risk (RR) was used to quantify comorbidity, where a RR of 1 indicates that diseases occur together as often as expected by chance. The lower quantile, median, and upper quantile values of RR in the 800 recorded pairs were 0.76, 1.07, and 1.62.

RR thresholds of 1.5, 2, and 5 were used to define comorbid disease pairs, with 239, 125 and 32 of the 800 observed disease pairs respectively meeting these thresholds. Of the 938 disease pairs in the dataset for which comorbidity data exists, the percentage of the 63 disease pairs

linked in the map which were comorbid at these thresholds were compared to the percentage of the 875 disease pairs not linked in the map which were comorbid at these thresholds. In both cases, duplicate pairs were counted twice (e.g. "type 1 diabetes-obesity" and "type 2 diabetes-obesity" were counted as two separate pairs although they both map to "250-278" at the ICD code level).

5.2.3 Similarity fusion

Pairwise similarity scores between each of the 84 diseases were calculated based on the Jaccard index[256] of their feature sets. In the case of transcriptomic data, the up- and down-regulated genes are considered separately, and so the Jaccard score was calculated as a weighted average of Jaccard scores for the two sets.

As the distributions of similarity scores within each space are uneven, fusion of the raw similarity scores would cause those spaces with higher average scores to dominate the fused similarity. Even if the scores are normalized to the same sum, the fused similarities would still be affected by the differences in distribution of similarity values in each space (e.g. causing sparse spaces to dominate the fused scores at high similarities). Quantile normalization[257] was therefore applied to adjust the distributions of similarity scores towards each other, enabling comparison and combination of each space independently of their distributions.

Quantile normalization, which has been previously mentioned in the context of microarray normalization (Section 2.1.2), involves replacing each value with the mean value of the same rank across each space. In the example shown in Table 5.1, the maximum similarity scores are 0.556 in phenotypic space and 1 in drug space, and so the maximum similarity score in both spaces is replaced with the mean of these two values: 0.778. The second highest values are 0.481 in phenotypic space and 0.500 in drug space, so these scores are replaced in each space with the mean of 0.4905, and so on. Adjustment for ties is used, so that tied ranks are replaced with the mean of the quantile normalized value across those ranks (orange values in Table 5.1).

Table 5.1 Example of quantile normalization with adjustment for ties

Where there are ties for a particular rank in one space, the values are replaced by the mean of the quantile-normalized values for those ranks: here the tied values in phenotypic space are replaced by the mean of 0.475, 0.470, and 0.465 = 0.470.

	Raw similarity scores, drug space	**Raw similarity scores, phenotypic space**	**New value in drug space after quantile normalization**	**New value in phenotypic space after quantile normalization**
Highest value	1	0.556	0.778	0.778
2nd highest value	0.500	0.481	0.491	0.491
3rd highest value	0.490	0.460	0.475	0.470
…	0.480	0.460	0.470	0.470
…	0.470	0.460	0.465	0.470

Following quantile normalization of the similarity values using limma's *normalizeQuantiles* function, a single 'fused' similarity score was computed by taking the mean of the quantile-normalized similarity values for each disease across each space, resulting in a 3,486-dimensional similarity vector (or an 84*84 symmetric similarity matrix) forming the basis of the disease map. Figure 5.1 shows an overview of this process. The majority of the analysis presented here is based on an unweighted mean of spaces, although the method allows the specification of weights in order to adjust the influence of each space on the fused similarities, in which case a weighted mean of spaces is calculated.

5.2.4 Defining a significance threshold for disease similarity

To construct the disease map, a threshold of significant similarity t was defined above which diseases are linked, based on 1000 random similarity matrices. Randomized feature vectors were constructed for each disease by sampling from the feature universe, defined as the union of all features in that space across all diseases in the dataset, according to their distribution (frequency) in the dataset. Using these random feature vectors, 1000 random fused similarity matrices were created. The 99.99^{th} percentile of the random similarity scores (equivalently, the maximum similarity observed in 83% of the random matrices) was taken as the threshold

of similarity above which diseases were considered to be linked. 6.9% of similarity values in the network were above this threshold. Cytoscape[258] was used for network visualisation.

5.2.5 Evaluating the fused similarity scores

An initial evaluation of the fused similarity scores was carried out against the independent disease comorbidity dataset[254] described in Section 5.2.2, which covers 938 of the 3,486 disease pairs in the dataset. Any disease-related evaluation data covering all diseases could also be used as a feature space, and so for more detailed evaluation a 'hold-out' style of evaluation was used measuring how well one feature space is represented in the remaining five. Two feature spaces were chosen to capture different aspects of disease-relatedness.

Firstly, drug approval information (obtained from ChEMBL as described in Section 5.2.2, although here drugs in Phase III clinical trials were also included) measures whether similarity between two diseases might indicate drug-sharing potential. Secondly, membership of Disease Ontology top-level classes (e.g. *disease of anatomical entity*, *disease of cellular proliferation*) measures how closely disease associations match established notions of clinical similarity. This was evaluated by training a random forest classifier on the pairwise similarity values, using the R package randomForest[259] with default parameters. To ensure availability of sufficient training data, DO class prediction was split into two binary tasks – membership of *disease of anatomical entity*, and membership of *disease of cellular proliferation* (as these are the two largest classes within the dataset). Model performance was evaluated using stratified Monte Carlo cross-validation, with an 80-20 split into training and test sets. The true positive rate (TPR), false positive rate (FPR), and area under the ROC curve (AUROC) were calculated using the function *performance* from the package ROCR[260] averaged over 1,000 runs. In order to display ROC curves, TPR and FPR were averaged only over those runs where the mode average number of data points were recorded.

5.3 RESULTS

5.3.1 Exploratory disease map analysis identifies existing and novel disease relationships

Similarity fusion (Figure 5.1), which enables comparison and combination of heterogeneous data types, was used to create a 'disease map': a network of diseases that are linked at multiple biological levels. Links in the disease map represent similarities above a threshold of significance (calculated as described in Methods) between the 84 diseases analysed here, as shown in Figure 5.2. 81 of the 84 diseases are included in the map, with cystic fibrosis, teratozoospermia, and placental malaria not showing any significant links to other diseases. Many links in the map correspond to the traditional ontological classes represented by the Disease Ontology (DO) – particularly within the DO classes *disease of cellular proliferation, disease by infectious agent* and *respiratory system disease* – but many novel links were additionally observed that span traditional disease categories, here defined as disease pairs which are not in the same top-level DO class. These novel links, which are listed in Table 5.2, make up 15% of the links in the disease map.

The network consists of two densely connected areas, the first containing cancers (yellow nodes) and the second composed of inflammatory bowel diseases, skin diseases, and immune system diseases (blue and purple nodes). The strong interconnection between cancers has been noted in other disease similarity studies such as the Human Disease Network of Goh et al.[261], which found that cancers were highly interconnected due to common involvement of tumour repressor genes; the disease map confirms that this commonality is replicated across different spaces, with different cancers presenting e.g. similar phenotypes, similar gene expression responses, and potential to be treated with similar drugs. The second densely connected area links diseases which are less obviously related, which will be explored in Section 5.3.2.

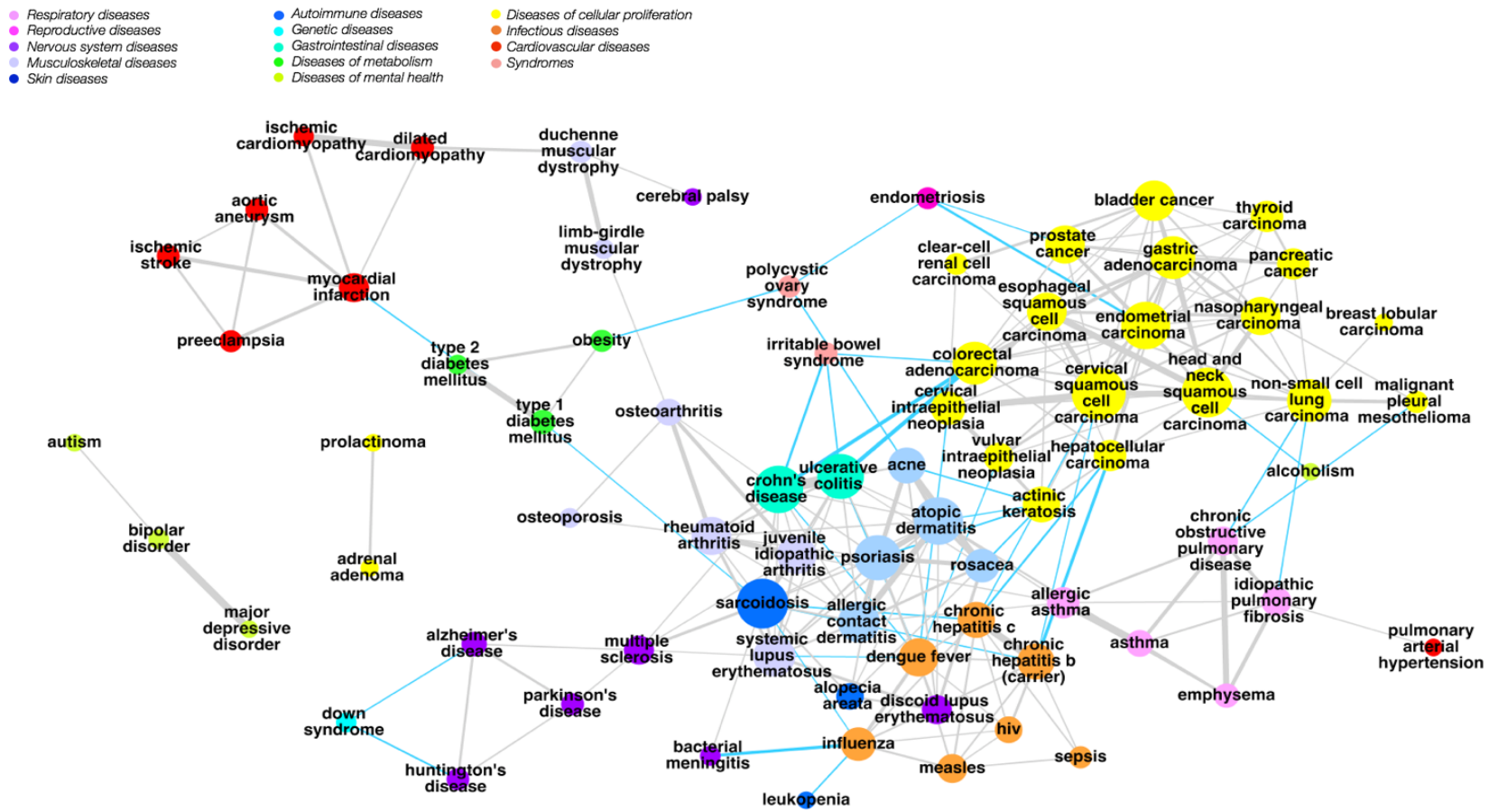


Figure 5.2 Disease network resulting from fused similarity scores.

Connecting the most similar diseases defines a disease network, where edges represent similarity in multiple feature spaces. The network shown here is constructed from all 6 feature spaces, and connects diseases not only within DO classes (the class 'Disease of anatomical entity' has been split into subclasses for clarity) but shows novel links (highlighted in blue) which are in different DO classes. More detail on these novel links is given in Table 5.2. The network shown is based on a force-directed layout, with minor adjustments to node position for greater readability.

Many of the novel links in the network represent diseases of distinct aetiology which share similar clinical presentations, such as actinic keratosis and psoriasis, or chronic obstructive pulmonary disease and malignant pleural mesothelioma. The shared features between each of these links can help to understand how these similarities arise: given any linked disease pair, the shared features which contribute to their similarity can be identified by simply taking the intersection of their feature vectors in each individual space, as demonstrated in Table 5.2.

Table 5.2 Novel links between diseases in different Disease Ontology classes

Novel links are disease pairs which have similarity higher than the significance threshold, but which are not related by the Disease Ontology top-level classes. Many novel links are related in multiple feature spaces, indicating similarities on different biological levels. Some relationships which fall under this definition are expected, such as the connection between inflammatory bowel disease (DO class 'disease of anatomical entity') and colorectal cancer (DO class 'disease of cellular proliferation'), as they affect the same organ system. Other connections, such as between hepatitis B and cervical cancer, seem surprising, and in such cases it is helpful to interpret the features shared between the two diseases.

Novel link		**Number of shared:**				
		Phenotypes	MeSH terms co-occurring in literature	Genetic associations	Dysregulated genes	Drugs
Acne	Actinic keratosis	1	15	1	18	0
Acne	Polycystic ovary syndrome	1	16	0	15	0
Actinic keratosis	Atopic dermatitis	0	10	0	23	0
Actinic keratosis	Psoriasis	0	15	6	25	0
Actinic keratosis	Rosacea	3	15	4	13	0
Alcoholism	Head and neck squamous cell carcinoma	0	33	2	2	0
Alzheimer's disease	Down syndrome	0	4	17	3	2
Bacterial meningitis	Influenza	1	12	4	35	1
Cervical squamous cell carcinoma	Chronic hepatitis b (carrier)	0	6	42	0	0
Cervical squamous cell carcinoma	Chronic hepatitis c	0	5	41	1	0

Cervical intraepithelial neoplasia	Dengue fever	0	0	5	29	0
Chronic obstructive pulmonary disease	Non-small cell lung carcinoma	2	4	1	20	0
Chronic obstructive pulmonary disease	Malignant pleural mesothelioma	2	2	2	19	0
Colorectal adenocarcinoma	Crohn's disease	3	34	12	21	0
Colorectal adenocarcinoma	Irritable bowel syndrome	0	29	9	7	0
Colorectal adenocarcinoma	Ulcerative colitis	3	35	16	26	0
Crohn's disease	Dengue fever	5	1	17	4	0
Crohn's disease	Irritable bowel syndrome	0	22	27	6	4
Dengue fever	Systemic lupus erythematosus	0	10	15	22	0
Dengue fever	Ulcerative colitis	5	2	15	3	0
Dengue fever	Vulvar intraepithelial neoplasia	0	0	0	24	0
Down syndrome	Huntington's disease	0	4	12	0	1
Endometrial carcinoma	Endometriosis	7	21	8	1	0
Endometriosis	Polycystic ovary syndrome	3	12	7	1	0
Endometriosis	Prostate cancer	3	3	6	1	1
Chronic hepatitis b (carrier)	Hepatocellular carcinoma	3	29	0	2	2
Chronic hepatitis b (carrier)	Sarcoidosis	0	2	35	9	1
Chronic hepatitis c	Hepatocellular carcinoma	3	28	0	0	0
Chronic hepatitis c	Sarcoidosis	0	2	31	15	0
Idiopathic pulmonary fibrosis	Non-small cell lung carcinoma	2	6	7	19	0
Influenza	Leukopenia	2	5	9	21	0
Influenza	Sarcoidosis	0	1	16	21	0

Irritable bowel syndrome	Ulcerative colitis	0	24	27	8	1
Myocardial infarction	Type 1 diabetes mellitus	0	0	17	26	2
Obesity	Polycystic ovary syndrome	0	27	4	1	3
Sarcoidosis	Type 1 diabetes mellitus	0	2	26	15	0
Sickle cell disease	Essential thrombocythemia	0	10	14	5	1

As an example of this analysis, the unexpected connection between cervical squamous cell carcinoma and hepatitis B can be examined. Table 5.2 shows that they share 42 genetic associations, including genes in the human leukocyte antigen system involved in antigen presentation (HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLA-DRB1); this suggests that the link between the two diseases is driven by shared aspects of immunological response. This may reflect the involvement of the human papillomavirus (HPV) in the majority of cervical cancer cases, with the immune response playing a key role in the development of cervical cancer from an initial HPV infection[262]. Likewise, the hepatitis B virus is a causal risk factor in the development of hepatocellular cancer[263], so shared processes between the two diseases could also reflect the interface between infection and carcinogenesis of these two DNA viruses.

If diseases linked in the map are pathologically related, they may be more likely to co-occur in the same patient. Links in the disease map were therefore compared to disease comorbidities based on the medical records of 13 million patients[254]. The 63 links for which comorbidity scores are available had a median relative risk (RR) of 2.35 (i.e. diseases are 2.35 times more likely to co-occur than expected by chance), compared to a median of 1.06 for the 875 disease pairs (for which comorbidity scores are available) that are not linked in the disease map. 71% of these links co-occur in patients at a RR threshold above 1.5, compared to 27% of the non-linked pairs, or 2.6 times more often. At higher RR thresholds of 2 and 5, this ratio increases to 4.6 and 10.6 respectively. This relationship suggests that links in the disease map represent clinically relevant associations.

5.3.2 Case study: psoriasis

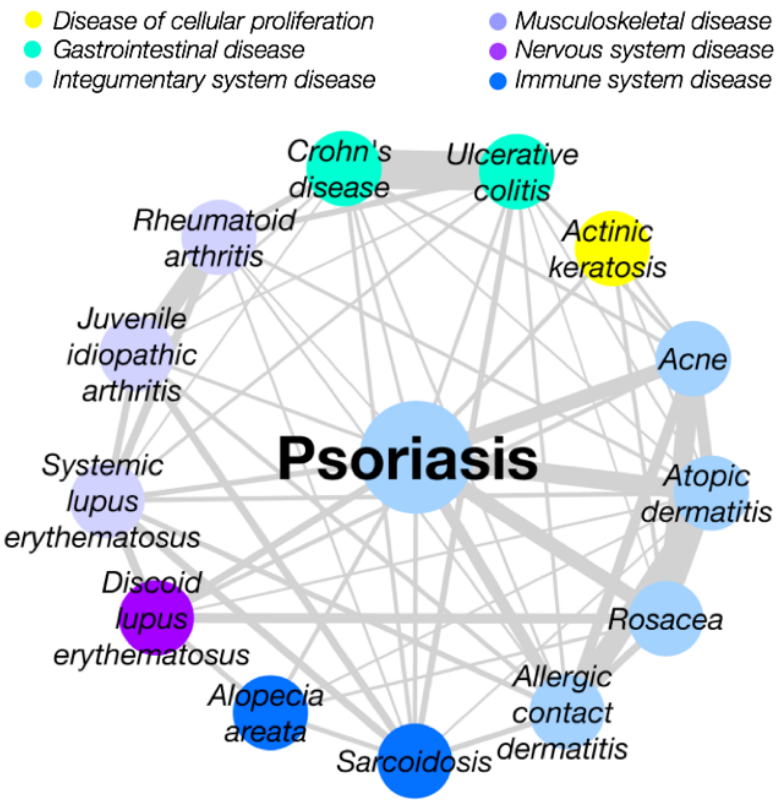


Figure 5.3 Diseases related to psoriasis

As well as known links to other skin diseases (light blue nodes), psoriasis has links to a number of phenotypically distinct diseases with an autoimmune component, such as alopecia, arthritis, and lupus, as well as inflammatory bowel diseases (turquoise nodes), with which it shares genetic features related to drugs that can be used to treat both conditions. There is a high degree of interconnection amongst this group of diseases, which form one of the most densely connected areas in the network.

The disease map also allows us to focus on connections of a disease of interest. As a case study, I examine psoriasis and its related diseases, which form a densely-connected region of the map (Figure 5.2). Psoriasis is classified as a skin condition in Disease Ontology, but is known to have immune and hereditary components[264]. This is reflected in the disease map, which links psoriasis to a number of autoimmune diseases as well as to other skin diseases (Figure 5.3). One example is the relationship between psoriasis and the inflammatory bowel diseases Crohn's disease (CD) and ulcerative colitis (UC). The inflammatory bowel diseases are phenotypically distinct from psoriasis, but both diseases involve an autoimmune component, and in fact show a degree of co-occurrence in patients[265].

Examining the feature sets of psoriasis, CD, and UC shows that they share a number of associations in genetic space, including Interleukin family genes IL12B and IL23R, involved in cytokine-mediated immune response; STAT3, which is activated by the interleukin IL6 (also shared) to produce inflammatory T-cells[266]; and (in psoriasis and UC) human leukocyte antigen HLA-B, which also plays an important role in the immune system. Psoriasis, CD, and UC also show shared dysregulation in the expression of several genes including upregulation in the pro-inflammatory S100 family (S100A8, S100A9) and CXC chemokines CXCL8, CXCL9, and CXCL10 (associated with immune system activation). Importantly, some of their shared features are relevant to the drugs prescribed for these diseases: the monoclonal antibodies adalimumab and infliximab are antagonists of tumor necrosis factor[267], a pro-inflammatory cytokine whose corresponding gene, TNF, shows genetic variation in a number of diseases including CD, UC, and psoriasis.

5.3.3 Similarity conversion allows comparison of information content between feature spaces

The use of quantile normalization allows the direct comparison of disease relationships present in the individual and fused feature spaces. This can be quantified by the Pearson correlation between the pairwise disease similarities in each space (Figure 5.4). The most similar spaces are phenotype and literature co-occurrence, with a Pearson correlation of 0.56. Both spaces are based on literature-mining, and there is also a degree of overlap between MeSH disease terms and phenotypes (e.g. 'diabetes mellitus' is both a MeSH disease term and a phenotype in the Human Phenotype Ontology) so the two spaces are not completely orthogonal. The ontological space also has high correlation with these two spaces, suggesting that these spaces capture 'traditional' knowledge of disease relationships. By contrast, the low correlation (<0.2) across the three 'non-traditional' representations (genetic association, gene expression, and drug approval) indicate that disease relationships are highly distinct in each of these spaces.

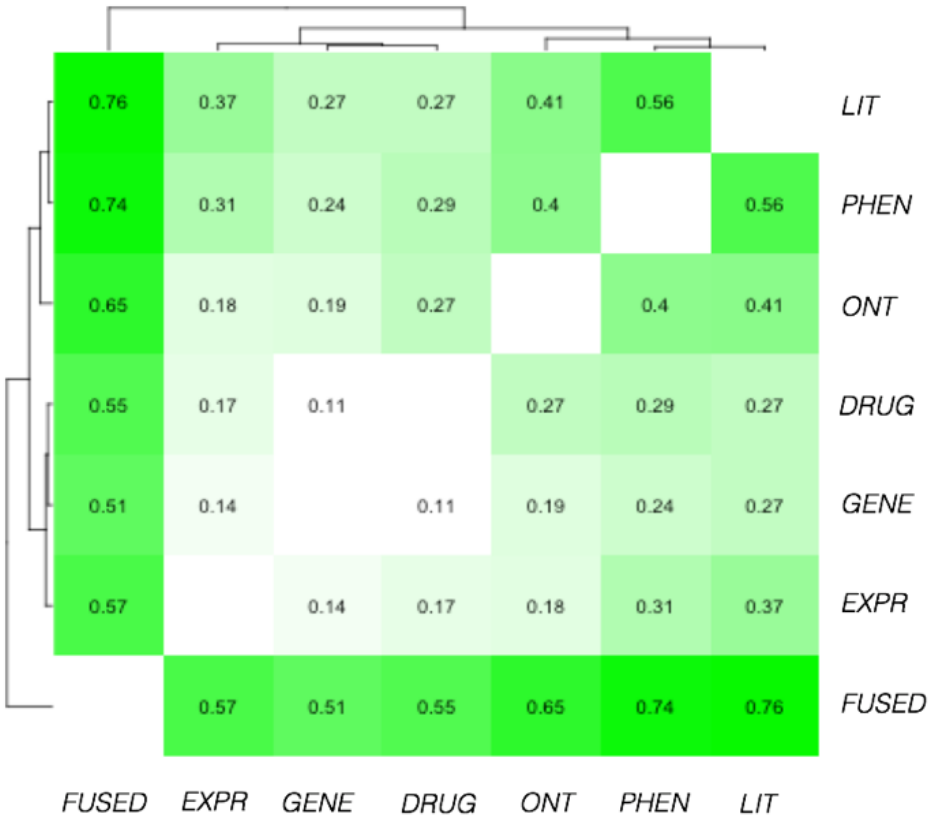


Figure 5.4 Correlation of pairwise similarity scores between feature spaces
The high correlation between phenotypic-, ontological-, and literature-based similarity indicates that relationships in these 'traditional' spaces are relatively similar to each other, whereas there is little resemblance between relationships in genetic association, gene expression, and drug spaces. The fused space resembles relationships in all spaces, but appears more similar to the 'traditional' spaces due to the multiple representation of relationships shared between these spaces. Lit = Literature co-occurrence; Ont = Disease Ontology; Drug = Approved drugs; Phen = Phenotype; Gene = Genetic association; Expr = Gene expression; Fused = fused similarity scores from six spaces.

Whilst the fused similarities have high correlation with each of the individual spaces, the fused space seems to resemble the three 'traditional' spaces more than the others, despite each space contributing equally to the fused similarities. As may be anticipated, shared similarities in the 'traditional' spaces cause the averaged similarities in the fused space to reflect these shared similarities more highly. This can be adjusted by down-weighting these spaces so that they have less influence on the fused similarities. Weighting the 'traditional' spaces so that they together make up one-third of the total similarity (instead of half), the similarity of the 'traditional' spaces to the fused becomes 0.56, 0.65, and 0.68 for ontological, phenotypic, and literature-based spaces respectively; and 0.58, 0.63, and 0.61 for genetic, expression, and drug spaces. Despite doubling the contribution of the 'non-traditional' spaces, the resulting disease

map does not appear substantially different (see Appendix N), suggesting that the disease map is not overly affected by the similarity of the 'traditional' spaces. The disease map therefore fundamentally resembles these traditional spaces, whilst inclusion of the diverse relationships from the genetic association, gene expression, and drug spaces adds novel similarities which distinguish the disease map from traditional classification systems.

5.3.4 Top disease links in the fused space show high overlap in shared drugs relative to the individual spaces

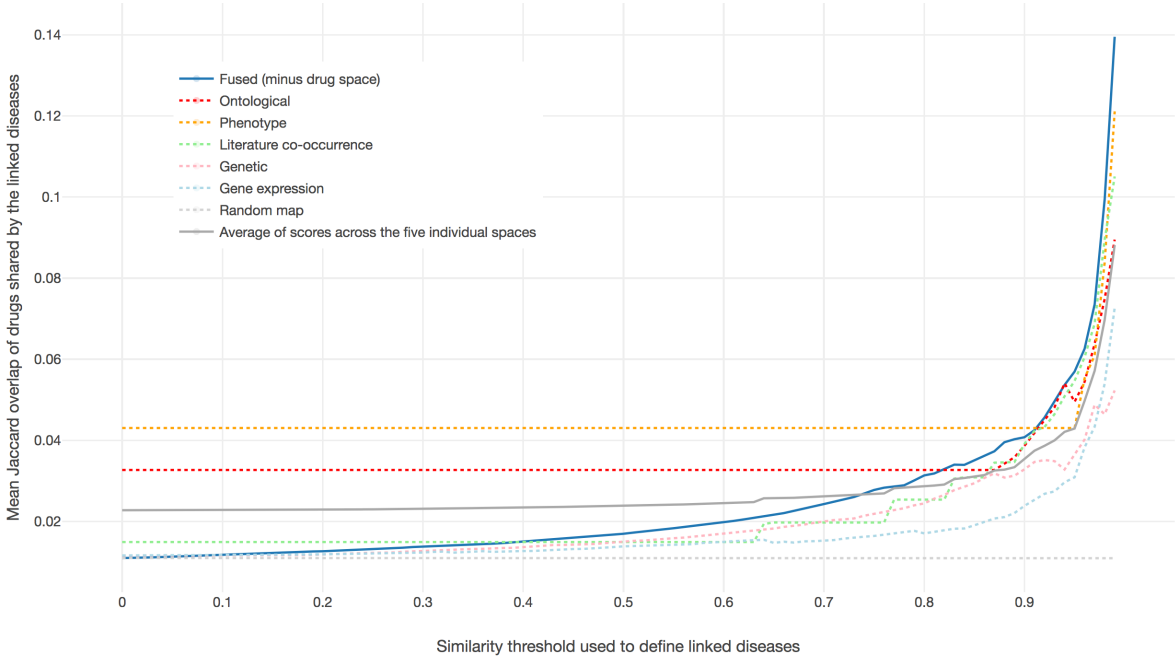


Figure 5.5 Mean Jaccard overlap of drugs (approved or in Phase III clinical trials) between disease pairs linked at different thresholds of similarity

Diseases that are highly similar in the fused space (constructed without drug information) are more likely to share approved or trialled drugs than diseases that are highly similar in the five individual spaces on average (grey line). Drug overlap in the sparse feature spaces, which have comparably few links between diseases, is static until higher thresholds of similarity are used (noticeable for the ontological and phenotypic spaces).

One aim of the disease map is the identification of similarities between diseases that could indicate where two diseases might be treated with the same drug. The extent to which links in the disease map correspond to drug-sharing relationships was therefore evaluated, including drugs that are in phase 3 clinical trials (as opposed to approved drugs only, which were used to construct the drug feature space). 61.6% of the links in the full disease map share drugs (with

44.2% of links sharing approved drugs only). Even if the information from drug space is excluded, still 50.8% of links in the top 6.9% of the non-drug fused values (the significance cut-off used to construct the full disease map) share drugs.

Rather than simply looking at the percentage of links which share at least one drug, the mean Jaccard drug overlap of diseases linked by the map can be evaluated. This accounts for differences in the number of drugs prescribed for each disease, as well as the number of drugs shared. However, this score is less intuitive and is best understood in comparison with the individual disease maps. Excluding any information from drug space, the remaining individual spaces were therefore compared to a disease map constructed from the fusion of these five spaces. At the cut-off of the top 6.9% of similarity values, links in the non-drug fused space have a higher Jaccard overlap of drugs approved and in Phase III trials (0.050) than any of the individual spaces (mean of 0.040).

This analysis was repeated across multiple similarity thresholds, from all values to the top 1% highest similarity scores (Figure 5.5). As expected, the higher the similarity threshold used, the greater the mean Jaccard drug score of diseases linked in the resulting map. Indeed, at the top thresholds of similarity (the top 5% or above), links in the fused map show greater mean drug overlap than links in any of the maps constructed from individual spaces, although the difference is relatively small. Importantly, drug overlap at the top thresholds is higher for the fused similarities than the mean over the five spaces (grey line on Figure 5.5), despite the fact that the fused map is constructed from the mean of similarities in each space. A similar result was also seen when considering only approved drugs (Figure 5.6) and for the weighted disease map (Appendix N), although for these cases ontological and/or literature spaces slightly outperform the fused space at the top similarity thresholds.

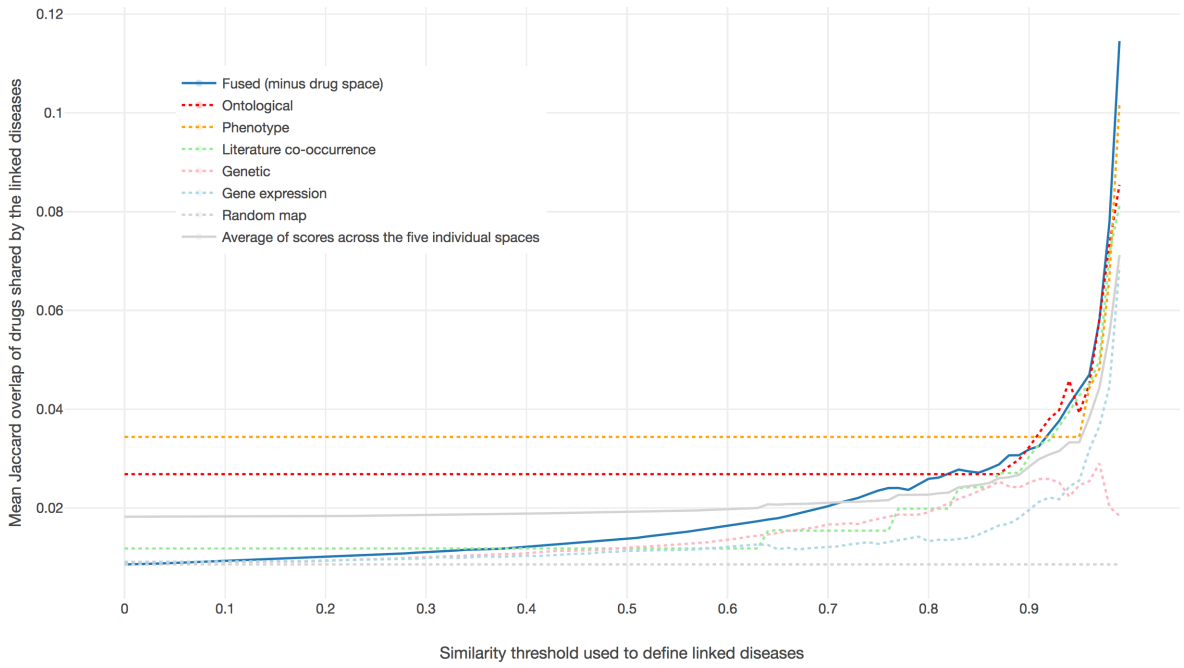


Figure 5.6 Mean drug overlap (by Jaccard score) of diseases linked at different thresholds of similarity, approved drugs only

The fused space has a high proportion of links which share approved drugs relative to other spaces. At the threshold of the top 6.9% most similar values (the threshold used for the disease map), the fused matrix (mean Jaccard score 0.038) is outperformed only by the ontological space (mean Jaccard score 0.040).

If only novel links (those that are in different top-level Disease Ontology classes) are considered, the ontological and literature co-occurrence spaces outperform the fused space. Ontological space does contain some disease pairs which are given high similarity (according to Lin's similarity measure) despite being in different top-level classes, but as expected this number is very small compared to the other spaces, with e.g. 24 novel links at a similarity threshold of 0.9 compared to 92 for the non-drug matrix and 117 for the literature co-occurrence matrix. Examination of the novel links at a threshold of 0.9 suggests that the reduced performance of the (non-drug) fused space is due to the failure to identify links between neurodegenerative and mental disorders which share drugs, such as major depressive disorder/bipolar and Parkinson's disease (identified in literature co-occurrence space) or Alzheimer's/Parkinson's/Huntington's diseases and Down syndrome (identified in ontological space). Literature co-occurrence space identifies a number of additional novel drug-sharing pairs not in the fused space, such as polycystic ovary syndrome and type II diabetes, or cystic fibrosis and chronic obstructive pulmonary disease.

5.3.5 Fused similarities outperform individual similarities in the prediction of Disease Ontology classes

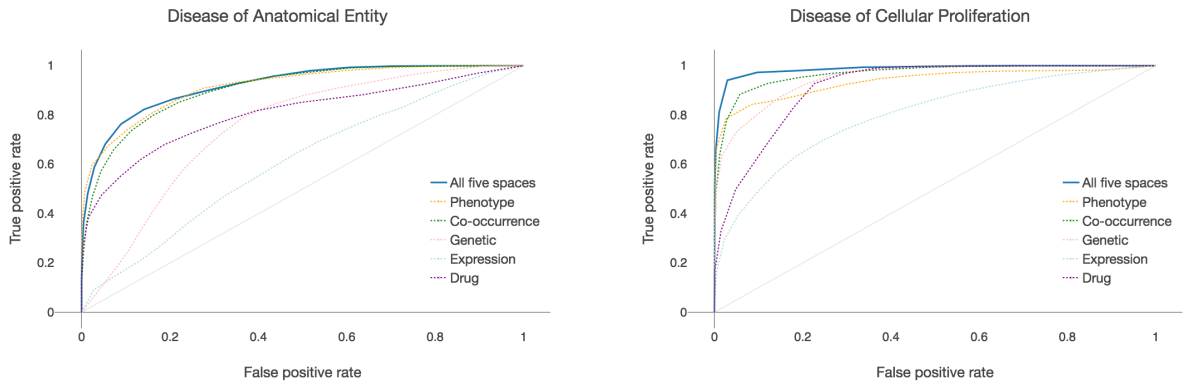


Figure 5.7 Ability of similarity scores from fused and individual spaces to predict Disease Ontology classes

Individual spaces differ widely in their predictive ability, with literature-based similarity and phenotypic similarity performing particularly well. The fused similarity scores outperform all individual spaces for the prediction of 'disease of cellular proliferation' (AUROC 0.986, right-hand plot). The fused similarity scores also outperform the individual spaces for predicting 'disease of anatomical entity' (AUROC 0.920, left-hand plot), although for this class (which contains more diverse disease types) phenotype and literature co-occurrence perform almost as well (AUROC 0.901 and 0.905 respectively).

A Random Forest classifier was used to examine how well the similarities in fused and individual spaces (excluding the ontological space) correspond to known disease categories (see Methods), reasoning that the ability of the fused similarities to reconstruct known categories would grant greater confidence that any novel relationships are likely to be biologically relevant. To ensure the existence of sufficient training data to build a robust classifier, the two largest Disease Ontology classes *disease of anatomical entity* and *disease of cellular proliferation* were predicted. Receiver Operating Characteristic curves for each space show that there is high variation between each space, although all spaces did better than random (Figure 5.7). Of the individual spaces, literature-based similarities were best able to classify diseases into known categories, with an AUROC of 0.905 for *disease of anatomical entity* and 0.968 for *disease of cellular proliferation*. Phenotypic similarities were also good predictors of disease classes, with an AUROC of 0.901 and 0.927 for *disease of anatomical entity* and *disease of cellular proliferation* respectively. Genetic and transcriptomic spaces do not closely correlate with the known categorizations (Figure 5.7), which is expected as traditional disease classifications (such as DO) do not take into account genetic or transcriptomic similarities.

The fused kernel outperformed any of the individual kernels, with AUROC scores of 0.920 for *disease of anatomical entity* and 0.986 for *disease of cellular proliferation*, despite the integration of spaces which are not such good predictors of disease classes. The mean performance over the five individual spaces was 0.795 for the prediction of *disease of anatomical entity* and 0.910 for the prediction of *disease of cellular proliferation*, meaning that the fused similarities in the disease map outperformed the individual similarities by 10% on average (mean AUROC over both tasks of 0.953 for the fused kernel vs 0.852 for the individual kernels). Largely similar classification results were seen for different feature vector sizes (Appendix M), although phenotypic and/or literature spaces slightly outperformed the fused space at some feature set sizes. Weighting the fused similarities so that the overlapping phenotype and literature co-occurrence spaces accounted for only 25% of the fused similarities (instead of 40%, as ontological space is excluded) did not significantly affect classification of *disease of cellular entity*, but slightly reduced the AUROC score to 0.891 for *disease of anatomical entity* (slightly less than the AUROC score of phenotypic and literature similarities, see Appendix N).

5.4 DISCUSSION

This chapter introduced a method, *similarity fusion*, to integrate biological data across multiple domains through conversion of features into normalized similarity scores, such that each space contributes evenly to the fused similarity. For the first time, the similarity fusion approach was applied across six feature spaces (ontological, phenotypic, literature co-occurrence, genetic association, gene expression, and drug indication data) in an unbiased manner. Following the normalization step, spaces may be weighted according to the desired application of the map (in terms of the importance placed on finding novel links vs reflecting known links, for instance). Here, a balanced fusion of disease relationships was used to create a 'disease map': a network linking diseases with significant similarities across multiple spaces.

The disease map reveals novel connections between diseases in different ontological categories (Figure 5.2), and highlights shared features between diseases – for example, shared gene expression patterns which may underlie an observed common phenotype. The case study of psoriasis illustrated how genetic variants shared with inflammatory bowel diseases were also targeted by drugs used for both conditions, illustrating how the identification of similarities between diseases at a 'molecular' level can indicate potential opportunities for sharing drugs, and to generate potential drug repurposing hypotheses in a 'guilt-by-association' approach[268]. Similar links have been identified in previous studies of -omics data integration, such as the DiseaseConnect web server[238] (association between psoriasis and inflammatory bowel disease); the interactome-based approach of Menche et al.[84] (association between psoriasis and other autoimmune diseases, their Supplementary Material); and the Integrated Disease Network[240] (connections between Crohn's and autoimmune conditions including parapsoriasis and psoriatic arthritis). This example illustrates how 'molecular' (e.g. genetic- and gene-expression based) approaches to disease similarity can identify disease relationships which are not captured by traditional disease classifications: the link between psoriasis and autoimmune disease, for example, is present in SNOMED but absent from other major classifications including MeSH, DO, and ICD.

Through the fusion of multiple data types, the disease map gives a new perspective on disease relationships, where aspects of disease (such as genetics and gene expression) not ordinarily considered by established classification systems reveal novel similarities between diseases. These spaces contain similarities not captured in our 'traditional' understanding of disease relationships (Figure 5.4), and therefore contribute greater depth of interest to the disease map. From this perspective, the more data types that can be included in the map, the more complete

the description of the biological system becomes. Using a 'hold-out' evaluation style, all six data types included in this study could be incorporated into the map, without designating any data types as reserved for evaluation purposes.

In agreement with previous studies showing how inclusion of more data types leads to greater accuracy in the prediction of disease relationships[239,240], the first evaluation task showed that the integrated disease map outperformed any individual space in predicting disease class membership, despite the inclusion of spaces that individually had little relation to known disease classes (Figure 5.7). In fact, the disease map, which is based on averaging similarity values, outperformed the average of individual similarity values by a mean of 10% across the two classes. One explanation for this is that the two spaces that are most similar to the ontological space (phenotypic and literature co-occurrence spaces) are also the most similar to each other (Figure 5.4), as they are based on literature mining of phenotype terms and MeSH terms respectively, and there is some overlap between these term sets. The similar disease relationships contained in these spaces therefore reinforce each other in the fused similarities. However, even if these two spaces are down-weighted (with a corresponding increase in influence of the 'non-traditional' spaces), the fused similarities still markedly outperform the average of other spaces in the prediction of disease classes (Appendix N). This suggests that the classification performance is not driven purely by these spaces; rather, the benefit in similarity fusion lies in the prioritization of disease relationships common to multiple spaces.

The second evaluation measure was the sharing of drugs (either approved or in Phase 3 clinical trials) between diseases linked by the disease map. Although the drug sharing space is highly distinct from any of the other spaces (Figure 5.4), drug sharing relationships were captured well by the fused space, which had a high mean Jaccard overlap of drugs shared amongst its most similar disease pairs relative to the individual spaces (Figure 5.5). This not only increases confidence in the biological relevance of the linked diseases, it further illustrates the value of incorporating multiple data types into the disease map. This pattern fits what has generally been seen in computational drug repurposing approaches: while approaches based on individual data types such as genome-wide association studies[269] or transcriptomics[270–272] are possible, successful drug repurposing methods often incorporate multiple data types[241,273]; data fusion may therefore become an increasingly important approach in drug discovery.

In summary, this chapter has demonstrated the utility of similarity fusion for integrating different types of biological data in the analysis of disease relationships, showing that the fused

data is not only able to reconstruct known disease and drug-sharing associations, but also offers the possibility of highlighting new relationships between diseases. The similarity-based approach proposed here will be particularly suited for the integration of high-throughput data sets where dimensionality would otherwise pose a problem, such as proteomics and metabolomics data, as the technology matures and the data becomes available for a large enough number of diseases. This approach could be extended to any number of spaces, leading to the possibility of a fully comprehensive disease map. Such a map could transform our current understanding of disease and disease relationships, revealing shared mechanisms behind diverse diseases which could eventually help to drive novel drug repurposing and treatment opportunities.

6 CONCLUSIONS

6.1 SUMMARY OF FINDINGS

The aim of this thesis was to explore the comparative analysis of gene expression data to understand disease and disease relationships, including the comparability of the transcriptomic disease signal across different study types; the analysis and comparison of gene expression changes through translation to the level of signaling interactions; and finally, the integration of gene expression data with other data types to explore how disease relationships extend across different biological levels.

This work began with the research described in Chapter 3, *Concordance of Microarray Studies of Parkinson's Disease*, which showed that gene expression studies of the substantia nigra in Parkinson's disease patients shared expression patterns which were distinct from those in studies of other tissues and disease models. Previous research has studied the effects of individual factors such as microarray platform type on the resulting gene expression profile[117,149–151], and examined the concordance between disease models and human patients[144,146–148]. However, in a comparative analysis setting, multiple factors come into play when selecting experiments that will be representative of the condition under study. The work described in this chapter therefore considered the effects of four key factors – tissue, platform, sample size, and disease model – using the agreement between studies of the same disease to understand how these affect the gene expression representation of disease. The improved concordance within the most highly-affected tissue in human patients suggested that these studies formed a characteristic representation of gene expression in PD; also notable was a lack of effect of factors such as the platform type (at least within Affymetrix-type microarrays) on concordance. The general study selection guidelines set out in this research were employed in selecting studies to be included in the larger disease dataset used for the next chapters.

The relatively low concordance observed between even the two most closely related PD studies illustrates the noisiness inherent to gene expression data. In the next chapter, *Using Dysregulated Signalling Paths to Understand Disease*, the weighted shortest-paths method of Sambarey et al.[121] was adapted in order to make more informative comparisons between diseases. Each disease was represented by a set of dysregulated paths on the human signaling network, providing a middle ground between raw gene expression data and canonical

biological pathways. The work described in this chapter is the first to use a path-based approach to compare multiple diseases, and reveals the existence of shared signaling processes between common and rare diseases. 52% of the paths shared between disease pairs contained a drug-interacting gene for at least one of the diseases, suggesting that this approach could be used to identify drug repurposing hypotheses, as detailed in a case study of the link between polycystic ovary syndrome and the rare condition Pompe disease. As well as comparing diseases, this chapter also examined more general properties of gene expression, finding that genes in dysregulated networks of multiple diseases have a moderate tendency to have higher degree, suggesting an influential (but not central) role in dysregulated signaling networks across diverse disease types.

In the final chapter, *Understanding and Predicting Disease Relationships Through Similarity Fusion*, the comparison of diseases was extended to multiple bioinformatics spaces. Following the direct network integration approach used in the previous chapter, a more generalized integration method was introduced to link diseases across different biological levels, from the molecular (genetic variation, gene expression, and drug indication) to the clinical (phenotype, literature co-occurrence, and ontological relationships). When quantifying the strength of relationships across such different data types, it is imperative to take into account the differing properties of each space[122], such as sparsity and the size of the feature universe. However, current methods to address this (such as the work of Sun et al.[240]) are limited in terms of the number of feature spaces that can be included. The similarity fusion method introduced in this chapter uses quantile normalization to adjust the distributions of pairwise similarity vectors towards each other, making them comparable and combinable. While the disease map contained mostly known relationships, 15% of links were novel links between diseases not related by traditional classification systems, such as between psoriasis and inflammatory bowel disease, or cervical cancer and hepatitis B. Importantly, links in the disease map were indicative of drug-sharing, with 62% of the links in the full disease map sharing drugs; this provides further indication of the potential of disease similarity approaches for drug repurposing.

Although the focus of this thesis has been comparative analysis of gene expression across diseases, the findings will be of interest in broader settings. The conclusions of this work translate most clearly to meta-analysis design, as many of the same considerations (such as the effect of tissue, platform, and disease model) apply to determining study selection criteria. By illustrating the agreement that can be expected between studies of the same disease and of

different diseases, this work also contributes to wider questions around the strength and specificity of gene expression representations of disease. Following on from this, the method developed in Chapter 4 is not only applicable to comparing disease, but is of broader interest for the interpretation of individual gene expression datasets (including drug response data, as shown here) in the context of signaling pathways. Finally, the data integration method developed in Chapter 5 can be applied to diverse data types as a general data integration method, particularly for the analysis of high-dimensional data.

6.2 LIMITATIONS

The research in this thesis confirms what countless previous studies have demonstrated: gene expression is noisy, representing a snapshot of the disease state under highly specific conditions. This is most clearly illustrated by the study of gene expression datasets in Parkinson's disease, where the low agreement between studies of the same condition may reflect the effect of multiple factors. These range from deliberate factors such as the choice of tissue to sample (as different tissues will show a different transcriptomic response to disease), to inherent factors such as biological and technical variation. Further complicating the interpretation of gene expression data is the well-established lack of predictivity of gene expression to the abundance of its corresponding protein product[193]. Gene expression data is therefore at best a partial representation of cellular state, which gives limited information compared to other measurements such as the proteome. The second and third chapters of this work therefore incorporate other data types in order to provide additional evidence for disease relationships.

Even assuming that a 'representative' gene expression profile is available, a further issue in gene expression data analysis (one that is common to other dynamic -omics measurements including the proteome and metabolome) is that measured gene expression may reflect not only disease pathogenesis, but also the body's compensatory responses (e.g. immune system activation). This further complicates the comparison of diseases: shared gene expression profiles may not necessarily represent shared mechanisms of disease, but may simply represent a shared response to different pathological processes, as shown with the case study of polycystic ovary syndrome and Pompe disease in Chapter 4. This can be useful – for example, the inflammatory response underlying the symptoms of many diseases can be treated with non-disease-specific anti-inflammatory drugs such as aspirin – but means that it is not possible to

establish a causal relationship between gene expression and disease, as is (sometimes) possible with e.g. genetic data.

In addition to the general limitations of comparative analysis of microarray data discussed here and in the Introduction, a broader limitation of this work stems from its reliance on public data. Despite the rapidly increasing number of gene expression profiles of disease, research questions are inevitably constrained by the availability of suitable studies: in Chapter 3, for instance, the scarcity of non-Affymetrix microarray studies of Parkinson's disease prevented drawing a stronger conclusion on the effect of microarray platform type on study concordance. As the popularity of transcriptomic analysis continues to grow in line with its decreasing cost, the amount of public transcriptomic data should further increase, providing greater statistical power and enabling a wider range of questions to be answered from the re-analysis of existing data.

6.3 Future Directions

Given the limitations of gene expression data, an obvious extension of this research would be to employ multi-omics data to study disease relationships in more detail. A basic version of this idea is already implemented in Chapter 4, with the inclusion of genetic variant data in dysregulated signaling pathways helping to give context to the observed gene expression changes; this could easily be extended to other -omic data types and/or incorporated into the path-finding algorithm directly. This would help to determine whether shared gene expression patterns between two diseases are associated with shared genetic, epigenetic, and/or proteomic features, leading to a greater understanding of shared disease biology. Multi-omics approaches are already being applied in disease subtyping[244,274] and precision medicine[275,276], although available multi-omics profiles are so far largely confined to cancer datasets such as The Cancer Genome Atlas[91]. It would be an interesting next step to apply this type of analysis to identify similarities between different cancers or, when the data becomes available, more diverse disease types.

Whilst it may be several years until this type of analysis is feasible on a large scale, one topic that would be of more immediate benefit is the identification of patterns of drug-induced gene expression in transcriptomic studies of disease. As mentioned previously, gene expression in disease may be affected by the drug treatment history of the patient; one way to identify this would be to compare patient differential expression profiles with drug-induced

differential expression profiles from large-scale resources such as CMAP[49] and LINCS[50], which contain *in vitro* drug response data from thousands of drugs tested in multiple cancer cell lines. The methods used here to study disease gene expression can equally be applied to understand gene expression in response to drugs, as shown by the drug-response case studies in Chapter 4; the next step would then be to directly compare these patterns between diseases and the drugs that are prescribed for them. Preliminary studies of the path-set analysis from Chapter 4 on CMAP data did not show substantial overlap between drug-induced differential expression profiles and those of their corresponding diseases; possible explanations for this include differences between drug response *in vivo* and *in vitro*, or the variable 'transcriptional activity'[30,277] of drugs. Given the prevalence of drug response as a confounding factor in transcriptomic studies of disease, a more detailed investigation would be of great benefit for this field.

The ultimate goal in studying disease is to find treatment to overcome it, and so a key consideration in this work is how the comparative analysis of gene expression data can be applied to discovering drug treatments for disease, specifically through drug repurposing. Examples in the preceding chapters have demonstrated how the methods introduced here can be used to suggest potential drug repurposing hypotheses, both gene-expression based (Chapter 4) and data integration-based (Chapter 5). The next step would be to ascertain which of the potential drug-sharing links identified here might be a promising candidate for experimental validation. This could be done at the small-scale, which would require disease experts to closely examine the feasibility of the proposed mechanism in the new indication. However, given the number of potential hypotheses, this could be a laborious task. Most successful bioinformatics drug repurposing approaches also use drug-side (cheminformatics) evidence such as structure and bioactivity data[17], giving a more comprehensive characterization of the potential match between a drug and a disease. The methods proposed in this thesis could easily be extended to incorporate drug-side information: for instance, the data integration method in Chapter 5 could be applied to combine different types of drug data, in an approach similar to the PREDICT method[241], which combined disease similarity with drug similarity in order to predict drug indications.

The comparison of gene expression across diseases not only enables the identification of shared disease biology, in turn revealing more about processes taking place in the individual diseases, but also helps to develop a greater understanding of the properties of gene expression data as a representation of disease. Given the increasing availability of gene

expression studies, the power of comparative analysis (including other approaches which re-use public data, such as meta-analysis) is steadily increasing, both in terms of statistical power and in terms of the breadth and depth of questions that can be answered. The findings in this thesis therefore aim to serve as a timely investigation into some of the issues surrounding this new approach to understanding disease. Whilst the work presented here represents only a small subset of the vast field of gene expression data analysis, I hope that the framework set out here for comparative analysis of diseases could eventually aid our understanding of disease and its treatment.

7 BIBLIOGRAPHY

1. Walker FO. Huntington's disease. *Lancet*. 2007;369(9557):218-228. doi:10.1016/S0140-6736(07)60111-1.
2. Beason-Held LL, Goh JO, An Y, et al. Changes in brain function occur years before the onset of cognitive impairment. *J Neurosci*. 2013;33(46):18008-18014. doi:10.1523/JNEUROSCI.1402-13.2013.
3. Scully JL. What is a disease? *EMBO Rep*. 2004;5(7):650-653. doi:10.1038/sj.embor.7400195.
4. National Cancer Institute. What Is Cancer? https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Accessed June 6, 2018.
5. Berry DA, Cirrincione C, Henderson IC, et al. Estrogen-Receptor Status and Outcomes of Modern Chemotherapy for Patients With Node-Positive Breast Cancer. *JAMA*. 2006;295(14):1658. doi:10.1001/jama.295.14.1658.
6. Garman KS, Nevins JR, Potti A. Genomic strategies for personalized cancer therapy. *Hum Mol Genet*. 2007;16(R2):R226-R232. doi:10.1093/hmg/ddm184.
7. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat Genet*. 2017;49:1319-1325. doi:10.1038/ng.3931.
8. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, Tenth Revision. 1990. http://www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf. Accessed August 17, 2017.
9. Côté RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA*. 243(8):756-762. http://www.ncbi.nlm.nih.gov/pubmed/6986000. Accessed August 17, 2017.
10. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 2015;43(D1):D1071-D1078. doi:10.1093/nar/gku1011.

11. NCBI. Medical Subject Headings. https://www.nlm.nih.gov/mesh/. Accessed March 4, 2016.

12. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association; 2013. doi:10.1176/appi.books.9780890425596.

13. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*. 2009;37. doi:10.1093/nar/gkn665.

14. Bagley SC, Altman RB. Computing disease incidence, prevalence and comorbidity from electronic medical records. *J Biomed Inform*. 2016;63:108-111. doi:10.1016/J.JBI.2016.08.005.

15. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104(21):8685-8690. doi:10.1073/pnas.0701361104.

16. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform*. 2011. doi:10.1093/bib/bbr013.

17. Cavalla D, Oerton E, Bender A. Drug Repurposing Review. In: *Comprehensive Medicinal Chemistry III*. Elsevier; 2017:11-47. doi:10.1016/B978-0-12-409547-2.12283-8.

18. Tobinick EL. The value of drug repositioning in the current pharmaceutical market. *Drug News Perspect*. 2009;22(2):119-125. doi:10.1358/dnp.2009.22.2.1303818.

19. What is gene expression? | Facts | yourgenome.org. https://www.yourgenome.org/facts/what-is-gene-expression. Accessed April 5, 2018.

20. Alberts B, Johnson A, Lewis J, et al. *Molecular Biology of the Cell, Sixth Edition*. Garland Science; 2014.

21. Horgan RP, Kenny LC. "Omic" technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet Gynaecol*. 2011;1313. doi:10.1576/toag.13.3.189.27672.

22. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. doi:10.1371/journal.pcbi.1005457.

23. Dhingra V, Gupta M, Andacht T, Fu ZF. New frontiers in proteomics research: A perspective. *Int J Pharm*. 2005;299:1-18. doi:10.1016/j.ijpharm.2005.04.010.

24. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. 2009. doi:10.1016/j.febslet.2009.10.036.

25. Jiang Q, Wang Y, Hao Y, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*. 2009;37(Database issue):D98-104. doi:10.1093/nar/gkn714.

26. Park S-J, Komata M, Inoue F, et al. Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. *Genes Dev*. 2013;27(24):2736-2748. doi:10.1101/gad.227926.113.

27. Gracey AY. Interpreting physiological responses to environmental change through gene expression profiling. *J Exp Biol*. 2007;210(9):1584-1592. doi:10.1242/jeb.004333.

28. Garg R, Shankar R, Thakkar B, et al. Transcriptome analyses reveal genotype- and developmental stage-specific molecular responses to drought and salinity stresses in chickpea. *Sci Rep*. 2016;6(1):19228. doi:10.1038/srep19228.

29. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A*. 2010;107(33):14621-14626. doi:10.1073/pnas.1000138107.

30. Babcock JJ, Du F, Xu K, Wheelan SJ, Li M. Integrated Analysis of Drug-Induced Gene Expression Profiles Predicts Novel hERG Inhibitors. *PLoS One*. 2013;8(7). doi:10.1371/journal.pone.0069513.

31. Kim J, Shin M. An integrative model of multi-organ drug-induced toxicity prediction using gene-expression data. *BMC Bioinformatics*. 2014;15(Suppl 16):S2. doi:10.1186/1471-2105-15-S16-S2.

32. Suzuki S, Horinouchi T, Furusawa C. Prediction of antibiotic resistance by gene expression profiles. *Nat Commun*. 2014;5:5792. doi:10.1038/ncomms6792.

33. Jenner RG, Young RA. Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol*. 2005;3(4):281-294. doi:10.1038/nrmicro1126.

34. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*. 2012;10(9):618-630. doi:10.1038/nrmicro2852.

35. Ahn T, Lee E, Huh N, Park T. Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics*. 2014;30(17):i422-i429. doi:10.1093/bioinformatics/btu449.

36. Hass HG, Vogel U, Scheurlen M, Jobst J. Gene-expression Analysis Identifies Specific Patterns of Dysregulated Molecular Pathways and Genetic Subgroups of Human Hepatocellular Carcinoma. *Anticancer Res*. 2016;36(10):5087-5095. doi:10.21873/anticanres.11078.

37. Chen P, Fan Y, Man T, Hung YS, Lau CC, Wong STC. A gene signature based method for identifying subtypes and subtype-specific drivers in cancer with an application to medulloblastoma. *BMC Bioinformatics*. 2013;14 Suppl 18(Suppl 18):S1. doi:10.1186/1471-2105-14-S18-S1.

38. Sandhu R, Parker JS, Jones WD, Livasy CA, Coleman WB. Microarray-Based Gene Expression Profiling for Molecular Classification of Breast Cancer and Identification of New Targets for Therapy. *Lab Med*. 2010;41(6):364-372. doi:10.1309/LMLIK0VIE3CJK0WD.

39. Lee BKB, Tiong KH, Chang JK, et al. DeSigN: connecting gene expression with therapeutics for drug repurposing and development. *BMC Genomics*. 2017;18(Suppl 1):934. doi:10.1186/s12864-016-3260-7.

40. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. *PLoS One*. 2009;4(8). doi:10.1371/journal.pone.0006536.

41. Sirota M, Dudley JT, Kim J, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med*. 2011;3(96).

42. Bryant S, Manning DL. Isolation of Messenger RNA. In: *RNA Isolation and Characterization Protocols*. Vol 86. Totowa, NJ: Humana Press; 1998:61-64. doi:10.1385/0-89603-494-1:61.

43. Baker M. Gene data to hit milestone. *Nature*. 2012;487(7407):282-283. doi:10.1038/487282a.

44. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-5. doi:10.1093/nar/gks1193.

45. GEO - NCBI. GEO Summary. https://www.ncbi.nlm.nih.gov/geo/summary/. Accessed April 7, 2018.

46. EMBL-EBI. ArrayExpress. https://www.ebi.ac.uk/arrayexpress/. Accessed April 7, 2018.

47. Ganter B, Tugendreich S, Pearson CI, et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J Biotechnol*. 2005;119(3):219-244. doi:10.1016/j.jbiotec.2005.03.022.

48. Igarashi Y, Nakatsu N, Yamashita T, et al. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res*. 2015;43(D1):D921-D927. doi:10.1093/nar/gku955.

49. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map : Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-)*. 2006;313(September):1929-1935. doi:10.1126/science.1132939.

50. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. 2017;171(6):1437-1452.e17. doi:10.1016/j.cell.2017.10.049.

51. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6(1):1-6. http://www.ncbi.nlm.nih.gov/pubmed/15068665. Accessed June 14, 2018.

52. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*. 2012;489(7416):391-399. doi:10.1038/nature11405.

53. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. January 2015. doi:10.1093/nar/gkv007.

54. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116-5121. doi:10.1073/pnas.091062498.

55. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616.

56. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

57. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

58. Trigueros-Motos L, González-Granado JM, Cheung C, et al. Embryological-origin-dependent differences in homeobox expression in adult aorta: role in regional phenotypic variability and regulation of NF-κB activity. *Arterioscler Thromb Vasc Biol*. 2013;33(6):1248-1256. doi:10.1161/ATVBAHA.112.300539.

59. Nelson DL, Lehninger AL, Cox MM. *Lehninger Principles of Biochemistry*. 5th ed. New York: W.H. Freeman; 2008.

60. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353-D361. doi:10.1093/nar/gkw1092.

61. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649-D655. doi:10.1093/nar/gkx1132.

62. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc*. 2013;8(8):1551-1566. doi:10.1038/nprot.2013.092.

63. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556.

64. Slenter DN, Kutmon M, Hanspers K, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*. 2018;46(D1):D661-D667. doi:10.1093/nar/gkx1064.

65. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260.

66. Subramanian A, Subramanian A, Tamayo P, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.

67. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.

68. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009;37(Web Server issue):W305-11. doi:10.1093/nar/gkp427.

69. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res*. 2017;45(W1):W130-W137. doi:10.1093/nar/gkx356.

70. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*. 2009;10(1):48. doi:10.1186/1471-2105-10-48.

71. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol BioSyst*. 2015;12(2):477-479. doi:10.1039/C5MB00663E.

72. Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22(13):1600-1607. doi:10.1093/bioinformatics/btl140.

73. Zheng B, Liao Z, Locascio J. PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. *Sci Transl*. 2010;2(52):ra73. doi:10.1126/scitranslmed.3001059.PGC-1.

74. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. doi:10.1371/journal.pcbi.1002375.

75. Weirauch MT. Gene Coexpression Networks for the Analysis of DNA Microarray Data. In: *Applied Statistics for Network Biology*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2011:215-250. doi:10.1002/9783527638079.ch11.

76. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression

analysis for functional classification and gene–disease predictions. *Brief Bioinform*. January 2017:bbw139. doi:10.1093/bib/bbw139.

77. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. *Cell*. 2011;144(6):986-998. doi:10.1016/j.cell.2011.02.016.

78. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;45(D1):D362-D368. doi:10.1093/nar/gkw937.

79. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(Database issue):D841-6. doi:10.1093/nar/gkr1088.

80. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*. 2013;41(D1):D793-D800. doi:10.1093/nar/gks1055.

81. Fazekas D, Koltai M, Türei D, et al. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol*. 2013;7:7. doi:10.1186/1752-0509-7-7.

82. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods*. 2016;13(12):966-967. doi:10.1038/nmeth.4077.

83. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*. 2012;7(2):e31826. doi:10.1371/journal.pone.0031826.

84. Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science (80-)*. 2015;347(6224):1257601. doi:10.1126/science.1116608.

85. Rakshit H, Rathi N, Roy D. Construction and Analysis of the Protein-Protein Interaction Networks Based on Gene Expression Profiles of Parkinson's Disease. Fleming S, ed. *PLoS One*. 2014;9(8):e103047. doi:10.1371/journal.pone.0103047.

86. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4(1):2. doi:10.1186/1471-2105-4-2.

87. Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: community structure analysis of biological networks. *Bioinformatics*. 2010;26(24):3135-3137. doi:10.1093/bioinformatics/btq596.

88. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559.

89. Ray S, Hossain SMM, Khatun L, Mukhopadhyay A. A comprehensive analysis on preservation patterns of gene co-expression networks during Alzheimer's disease progression. *BMC Bioinformatics*. 2017;18(1):579. doi:10.1186/s12859-017-1946-8.

90. Williams RBH, Chan EKF, Cowley MJ, Little PFR. The influence of genetic variation on gene expression. *Genome Res*. 2007;17(12):1707-1716. doi:10.1101/gr.6981507.

91. NIH. The Cancer Genome Atlas. https://cancergenome.nih.gov/. Accessed April 12, 2018.

92. Vasaikar S V, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res*. 2018;46(D1):D956-D963. doi:10.1093/nar/gkx1090.

93. Orozco LD, Morselli M, Rubbi L, et al. Epigenome-Wide Association of Liver Methylation Patterns and Complex Metabolic Traits in Mice. *Cell Metab*. 2015;21(6):905-917. doi:10.1016/j.cmet.2015.04.025.

94. Rung J, Brazma A. Reuse of public genome-wide gene expression data. *Nat Rev Genet*. 2013;14(2):89-99. doi:10.1038/nrg3394.

95. Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. *Science*. 2005;309(5743):2010-2013. doi:10.1126/science.1105891.

96. Zakharkin SO, Kim K, Mehta T, et al. Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*. 2005;6(1):214. doi:10.1186/1471-2105-6-214.

97. Sweeney TE, Haynes WA, Vallania F, Ioannidis JP, Khatri P. Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Res*. 2016;45(1):gkw797. doi:10.1093/nar/gkw797.

98. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*. 2012;40(9):3785-3799. doi:10.1093/nar/gkr1265.

99. Zhang M, Yao C, Guo Z, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*. 2008;24(18):2057-2063. doi:10.1093/bioinformatics/btn365.

100. Griffith OL, Melck A, Jones SJM, Wiseman SM. Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. *J Clin Oncol*. 2006;24(31):5043-5051. doi:10.1200/JCO.2006.06.7330.

101. Van Beelen Granlund A, Flatberg A, Østvik AE, et al. Whole Genome Gene Expression Meta-Analysis of Inflammatory Bowel Disease Colon Mucosa Demonstrates Lack of Major Differences between Crohn's Disease and Ulcerative Colitis. http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0056818&type=printable. Accessed March 30, 2018.

102. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. 2014:1-8. doi:10.1186/s13075-014-0489-x.

103. Li MD, Burns TC, Morgan AA, Khatri P. Integrated multi-cohort transcriptional meta-analysis of neurodegenerative diseases. doi:10.1186/s40478-014-0093-y.

104. Shao L, Vawter MP. Shared Gene Expression Alterations in Schizophrenia and Bipolar Disorder. *Biol Psychiatry*. 2008;64(2):89-97. doi:10.1016/J.BIOPSYCH.2007.11.010.

105. Hirsch HA, Iliopoulos D, Joshi A, et al. A Transcriptional Signature and Common Gene Networks Link Cancer with Lipid Metabolism and Diverse Human Diseases. *Cancer Cell*. 17:348-361. doi:10.1016/j.ccr.2010.01.022.

106. Yang J, Wu SJ, Dai WT, Li YX, Li YY. The human disease network in terms of dysfunctional regulatory mechanisms. *Biol Direct*. 2015;10. doi:10.1186/s13062-015-0088-z.

107. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*. 2010. doi:10.1371/journal.pcbi.1000662.

108. Shigemizu D, Hu Z, Hung JH, Huang CL, Wang Y, DeLisi C. Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer. *PLoS Comput Biol*. 2012;8(2):1-9. doi:10.1371/journal.pcbi.1002347.

109. Kunkel SD, Suneja M, Ebert SM, et al. mRNA Expression Signatures of Human Skeletal Muscle Atrophy Identify a Natural Compound that Increases Muscle Mass. 2012;13(6):627-638. doi:10.1016/j.cmet.2011.03.020.mRNA.

110. van Noort V, Schölch S, Iskar M, et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res*. 2014;74(20):5690-5699. doi:10.1158/0008-5472.CAN-13-3540.

111. Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. *Mol Syst Biol*. 2009;5(307):307. doi:10.1038/msb.2009.66.

112. Sanchez A, Golding I. Genetic determinants and cellular constraints in noisy gene expression. *Science*. 2013;342(6163):1188-1193. doi:10.1126/science.1242975.

113. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-Expression Variation Within and Among Human Populations. *Am J Hum Genet*. 2007;80(3):502-509. doi:10.1086/512017.

114. Kitchen RR, Sabine VS, Simen AA, Dixon JM, Bartlett JM, Sims AH. Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. doi:10.1186/1471-2164-12-589.

115. Tu Y, Stolovitzky G, Klein U. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A*. 2002;99(22):14031-14036. doi:10.1073/pnas.222164199.

116. Ach RA, Floore A, Curry B, et al. Factors influencing reproducibility of gene expression measurements using DNA microarrays. *Cancer Res*. 2005;65(9_Supplement):102-. http://cancerres.aacrjournals.org/content/65/9_Supplement/102.1. Accessed April 2, 2016.

117. Shi L, Reid LH, Jones WD, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24(9):1151-1161. doi:10.1038/nbt1239.

118. Su Z, Łabaj PP, Li S, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*. 2014;32(9):903-914. doi:10.1038/nbt.2957.

119. Russ J, Futschik ME. Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics*. 2010;11:305. doi:10.1186/1471-2164-11-305.

120. Mistry M, Pavlidis P. A cross-laboratory comparison of expression profiling data from normal human postmortem brain. *Neuroscience*. 2010;167(2):384-395. doi:10.1016/j.neuroscience.2010.01.016.

121. Sambarey A, Devaprasad A, Baloni P, et al. Meta-analysis of host response networks identifies a common core in tuberculosis. *npj Syst Biol Appl*. 2017;3(1):4. doi:10.1038/s41540-017-0005-4.

122. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface*. 2015;12(112):20150571-. doi:10.1098/rsif.2015.0571.

123. Trost B, Moir CA, Gillespie ZE, Kusalik A, Mitchell JA, Eskiw CH. Concordance between RNA-sequencing data and DNA microarray data in transcriptome analysis of proliferative and quiescent fibroblasts. *R Soc Open Sci*. 2015;2(9):150402. doi:10.1098/rsos.150402.

124. Wang C, Gong B, Bushel P, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol*. 2014;32:926–932. doi:10.1038/nbt.3001.

125. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res*. 2005;33(18):5914-5923. doi:10.1093/nar/gki890.

126. Wang H, He X, Band M, Wilson C, Liu L. A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics*. 2005;6(1):71. doi:10.1186/1471-2164-6-71.

127. Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update--simplifying data submissions. *Nucleic Acids Res*. 2015;43(Database issue):D1113-6. doi:10.1093/nar/gku1057.

128. Sutherland JJ, Jolly RA, Goldstein KM, Stevens JL. Assessing Concordance of Drug-Induced Transcriptional Response in Rodent Liver and Cultured Hepatocytes. *PLOS Comput Biol*. 2016;12(3):e1004847. doi:10.1371/journal.pcbi.1004847.

129. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*. 2008;5(9):1320-1332. doi:10.1371/journal.pmed.0050184.

130. Guo L, Lobenhofer EK, Wang C, et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol*. 2006;24(9):1162-1169. doi:10.1038/nbt1238.

131. Kadota K, Nakai Y, Shimizu K. Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. *Algorithms Mol Biol*. 2009;4(1):7. doi:10.1186/1748-7188-4-7.

132. Zhang L, Zhang J, Yang G, et al. Investigating the concordance of Gene Ontology terms reveals the intra- and inter-platform reproducibility of enrichment analysis. *BMC Bioinformatics*. 2013;14(1):143. doi:10.1186/1471-2105-14-143.

133. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/. Published 2015. Accessed February 17, 2016.

134. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3):307-315. doi:10.1093/bioinformatics/btg405.

135. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846-1847. doi:10.1093/bioinformatics/btm254.

136. NCBI. GEO2R. http://www.ncbi.nlm.nih.gov/geo/geo2r/. Accessed April 21, 2015.

137. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010;26(19):2363-2367. doi:10.1093/bioinformatics/btq431.

138. Gentleman R, Carey V, Huber W, Hahne F. genefilter: methods for filtering genes from high-throughput experiments. 2017. https://bioconductor.org/packages/release/bioc/html/genefilter.html. Accessed April

24, 2018.

139. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. *Bioinformatics*. 2009;25(3):415-416. doi:10.1093/bioinformatics/btn647.

140. Sutherland GT, Matigian N a., Chalk AM, et al. A cross-study transcriptional analysis of Parkinson's disease. *PLoS One*. 2009;4(3):1-8. doi:10.1371/journal.pone.0004955.

141. Dumitriu A, Latourelle JC, Hadzi TC, et al. Gene expression profiles in Parkinson disease prefrontal cortex implicate FOXO1 and genes under its transcriptional regulation. *PLoS Genet*. 2012;8(6). doi:10.1371/journal.pgen.1002794.

142. Edwards YJK, Beecham GW, Scott WK, et al. Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. *PLoS One*. 2011;6(2). doi:10.1371/journal.pone.0016917.

143. Cruz-Monteagudo M, Borges F, Paz-Y-Mino C, et al. Efficient and biologically relevant consensus strategy for Parkinson's disease gene prioritization. *BMC Med Genomics*. 2016;9:12. doi:10.1186/s12920-016-0173-x.

144. Miller RM, Federoff HJ. Altered gene expression profiles reveal similarities and differences between Parkinson disease and model systems. *Neuroscientist*. 2005;11(6):539-549. doi:10.1177/1073858405278330.

145. Blesa J, Phani S, Jackson-Lewis V, Przedborski S. Classic and new animal models of Parkinson's disease. *J Biomed Biotechnol*. 2012;2012. doi:10.1155/2012/845618.

146. Bandapalli OR, Kahlert C, Hellstern V, et al. Cross-species comparison of biological themes and underlying genes on a global gene expression scale in a mouse model of colorectal liver metastasis and in clinical specimens. *BMC Genomics*. 2008;9:448. doi:10.1186/1471-2164-9-448.

147. Seok J, Warren HS, Cuenca AG, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A*. 2013;110(9):3507-3512. doi:10.1073/pnas.1222878110.

148. Takao K, Miyakawa T. Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc Natl Acad Sci*. 2014;112(4):1401965111-. doi:10.1073/pnas.1401965111.

149. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. *Nat Methods*. 2005;2(5):337-344. doi:10.1038/nmeth757.

150. Miller R, Callhan L, Casaceli C, et al. Dysregulation of Gene Expression in the 1-Mehyl-4-Phenyl- 1,2,3,6-Tetrahydropyridine-Lesioned Mouse Substantia Nigra. *J Neurosci*. 2004;24(34):7445-7454. doi:10.1523/JNEUROSCI.4204-03.2004.

151. Bigler J, Rand HA, Kerkof K, Timour M, Russell CB. Cross-Study Homogeneity of Psoriasis Gene Expression in Skin across a Large Expression Range. *PLoS One*. 2013;8(1). doi:10.1371/journal.pone.0052242.

152. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*. 2006;103(15):5923-5928. doi:10.1073/pnas.0601231103.

153. Stretch C, Khan S, Asgarian N, et al. Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature. *PLoS One*. 2013;8(6):6-11. doi:10.1371/journal.pone.0065380.

154. Preece P, Cairns NJ. Quantifying mRNA in postmortem human brain: influence of gender, age at death, postmortem interval, brain pH, agonal state and inter-lobe mRNA variance. *Brain Res Mol Brain Res*. 2003;118(1-2):60-71. http://www.ncbi.nlm.nih.gov/pubmed/14559355. Accessed April 2, 2016.

155. Atz M, Walsh D, Cartagena P, et al. Methodological considerations for gene expression profiling of human brain. *J Neurosci Methods*. 2007;163(2):295-309. doi:10.1016/j.jneumeth.2007.03.022.

156. Kasim A, Shkedy Z, Lin D, et al. Translation of disease associated gene signatures across tissues. *Int J Data Min Bioinform*. 2015;11(3):301-313.

157. Kuhn A, Luthi-Carter R, Delorenzi M. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package “annotationTools”. *BMC Bioinformatics*. 2008;9(1):26. doi:10.1186/1471-2105-9-26.

158. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. *Nature*. 2013;504(7480):389-393. doi:10.1038/nature12831.

159. Kilpinen S, Autio R, Ojala K, et al. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol*. 2008;9(9):R139. doi:10.1186/gb-2008-9-9-r139.

160. Lu Y, Huggins P, Bar-Joseph Z. Cross species analysis of microarray expression data. *Bioinformatics*. 2009. doi:10.1093/bioinformatics/btp247.

161. R Foundation. cor {stats}. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html. Accessed June 12, 2016.

162. The Broad Institute. GSEA FAQ. http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ. Accessed February 17, 2016.

163. Huson LW. Performance of some correlation coefficients when applied to zero-clustered data. *J Mod Appl Stat Methods*. 2007;6(2):530-536.

164. R Foundation. hclust {stats}. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html. Accessed June 12, 2016.

165. D'Haeseleer P. How does gene expression clustering work? *Nat Biotech*. 2005;23(12):1499-1501. http://dx.doi.org/10.1038/nbt1205-1499.

166. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. 2006;22(12):1540-1542. doi:10.1093/bioinformatics/btl117.

167. prcomp {stats}. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html. Accessed June 12, 2016.

168. The CTTV Target Validation Platform. https://www.targetvalidation.org/data_sources. Accessed March 8, 2016.

169. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res*. 2017;45(D1):D985-D994. doi:10.1093/nar/gkw1055.

170. Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*. 2010;11(12):R124. doi:10.1186/gb-2010-11-12-r124.

171. Calabresi P, Picconi B, Tozzi A, Ghiglieri V, Di Filippo M. Direct and indirect

pathways of basal ganglia: a critical reappraisal. *Nat Neurosci*. 2014;17(8):1022-1030. doi:10.1038/nn.3743.

172. Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol Aging*. 24(2):197-211. http://www.ncbi.nlm.nih.gov/pubmed/12498954. Accessed January 27, 2015.

173. Davie CA. A review of Parkinson's disease. *Br Med Bull*. 2008;86(1):109-127. doi:10.1093/bmb/ldn013.

174. Dumitriu A, Moser C, Hadzi TC, et al. Postmortem Interval Influences α -Synuclein Expression in Parkinson Disease Brain. *Parkinsons Dis*. 2012;2012:1-8. doi:10.1155/2012/614212.

175. Lam SH, Wu YL, Vega VB, et al. Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat Biotechnol*. 2006;24(1):73-75. doi:10.1038/nbt1169.

176. van Hijum SAFT, de Jong A, Baerends RJS, et al. A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. *BMC Genomics*. 2005;6(1):77. doi:10.1186/1471-2164-6-77.

177. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M. Microarray experiments and factors which affect their reliability. *Biol Direct*. 2015;10:46. doi:10.1186/s13062-015-0077-2.

178. Cantuti-Castelvetri I, Keller-McGandy C, Bouzou B, et al. Effects of gender on nigral gene expression and parkinson disease. *Neurobiol Dis*. 2007;26(3):606-614. doi:10.1016/j.nbd.2007.02.009.

179. Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol*. 2006;7(1):3. doi:10.1186/1471-2199-7-3.

180. Lewis PA, Cookson MR. Gene expression in the Parkinson's disease brain. *Brain Res Bull*. 2012;88(4):302-312. doi:10.1016/j.brainresbull.2011.11.016.

181. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Brief Bioinform*. 2015;(March):1-11. doi:10.1093/bib/bbv020.

182. Ansari S, Donato M, Saberian N, Draghici S. An approach to infer putative disease-specific mechanisms using neighboring gene networks. doi:10.1093/bioinformatics/btx097.

183. Saha A, Choon Tan A, Kang J. Automatic Context-Specific Subnetwork Discovery from Large Interaction Networks. *PLoS One*. 2014;9(1). doi:10.1371/.

184. Altschuler GM, Hofmann O, Kalatskaya I, et al. Pathprinting: An integrative approach to understand the functional basis of disease. *Genome Med*. 2013;5:68. doi:10.1186/gm472.

185. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*. 2013;14(10):719-732. doi:10.1038/nrg3552.

186. Ruffalo M, Stojanov P, Pillutla VK, Varma R, Bar-Joseph Z. Reconstructing cancer drug response networks using multitask learning. *BMC Syst Biol*. 2017;11. doi:10.1186/s12918-017-0471-8.

187. Desai K, Brott D, Hu X, Christianson A. Mining Protein Interactions and Gene Expression Data to Gain Insights into Drug-induced Toxicity Mechanisms. In: *2011 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE; 2011:652-657. doi:10.1109/BIBM.2011.86.

188. Kim Y-A, Wuchty S, Przytycka TM. Identifying Causal Genes and Dysregulated Pathways in Complex Diseases. *PLoS Comput Biol*. 2011;7(3). doi:10.1371/journal.pcbi.1001095.

189. Alroobi R, Salem S. Discovering dysregulated phenotype-related gene patterns. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14*. New York, New York, USA: ACM Press; 2014:524-532. doi:10.1145/2649387.2649441.

190. Liu N, Li C, Huang Y, et al. A functional module-based exploration between inflammation and cancer in esophagus. *Sci Rep*. 2015;5(1):15340. doi:10.1038/srep15340.

191. Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40(Database issue):D940-6. doi:10.1093/nar/gkr972.

192. Bento AP, Gaulton A, Hersey A, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. 2014;42(D1):D1083-D1090. doi:10.1093/nar/gkt1031.

193. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*. 2016;165(3):535-550. doi:10.1016/j.cell.2016.03.014.

194. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Syst*. 2006;1695(5):1-9. http://igraph.org.

195. MacKenzie SJ, Baillie GS, McPhee I, Bolger GB, Houslay MD. ERK2 mitogen-activated protein kinase binding, phosphorylation, and regulation of the PDE4D cAMP-specific phosphodiesterases. The involvement of COOH-terminal docking sites and NH2-terminal UCR regions. *J Biol Chem*. 2000;275(22):16609-16617. doi:10.1074/jbc.275.22.16609.

196. Lipworth BJ. Phosphodiesterase-4 inhibitors for asthma and chronic obstructive pulmonary disease. *Lancet*. 2005;365(9454):167-175. doi:10.1016/S0140-6736(05)17708-3.

197. Lukacs NW. Role of chemokines in the pathogenesis of asthma. *Nat Rev Immunol*. 2001;1(2):108-116. doi:10.1038/35100503.

198. Citro A, Valle A, Cantarelli E, et al. CXCR1/2 inhibition blocks and reverses type 1 diabetes in mice. *Diabetes*. 2015;64(4):1329-1340. doi:10.2337/db14-0443.

199. Diana J, Lehuen A. Macrophages and beta-cells are responsible for CXCR2-mediated neutrophil infiltration of the pancreas during autoimmune diabetes. *EMBO Mol Med*. 2014;6(8):1090-1104. doi:10.15252/emmm.201404144.

200. Huang-Doran I, Tomlinson P, Payne F, et al. Insulin resistance uncoupled from dyslipidemia due to C-terminal PIK3R1 mutations. *JCI Insight*. 2016;1(17):e88766. doi:10.1172/jci.insight.88766.

201. Boucher J, Kleinridders A, Kahn CR. Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harb Perspect Biol*. 2014;6(1). doi:10.1101/cshperspect.a009191.

202. McCurdy CE, Schenk S, Holliday MJ, et al. Attenuated Pik3r1 expression prevents insulin resistance and adipose tissue macrophage accumulation in diet-induced obese mice. *Diabetes*. 2012;61(10):2495-2505. doi:10.2337/db11-1433.

203. Normanno N, De Luca A, Bianco C, et al. Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*. 2006;366(1):2-16. doi:10.1016/J.GENE.2005.10.018.

204. Lee PI, Patel A, Hindmarsh PC, Brook CGD, Leonard J V. Polycystic ovaries and glucose tolerance in hepatic glycogen storage disease. *Pediatr Res*. 1993;33(S5):S14-S14. doi:10.1203/00006450-199305001-00063.

205. Lee RJ, Leonard J V. The hepatic glycogen storage diseases - problems beyond childhood. *J Inher Metab Dis*. 1995;18:462-472. https://link.springer.com/content/pdf/10.1007%2FBF00710057.pdf. Accessed February 8, 2018.

206. Li HL, Davis WW, Whiteman EL, Birnbaum MJ, Puré E. The tyrosine kinases Syk and Lyn exert opposing effects on the activation of protein kinase Akt/PKB in B lymphocytes. *Proc Natl Acad Sci U S A*. 1999;96(12):6890-6895. doi:10.1073/PNAS.96.12.6890.

207. Gold MR, Scheid MP, Santos L, et al. The B Cell Antigen Receptor Activates the Akt (Protein Kinase B)/Glycogen Synthase Kinase-3 Signaling Pathway Via Phosphatidylinositol 3-Kinase. *J Immunol*. 1999;163:1894-1905. http://www.jimmunol.org/content/163/4/1894. Accessed March 2, 2018.

208. Guo S. Insulin signaling, resistance, and the metabolic syndrome: insights from mouse models into disease mechanisms. *J Endocrinol*. 2014;220(2):T1-T23. doi:10.1530/JOE-13-0327.

209. Wu XK, Zhou SY, Liu JX, et al. Selective ovary resistance to insulin signaling in women with polycystic ovary syndrome. *Fertil Steril*. 2003;80(4):954-965. http://www.ncbi.nlm.nih.gov/pubmed/14556818. Accessed August 20, 2018.

210. Taylor KM, Meyers E, Phipps M, et al. Dysregulation of Multiple Facets of Glycogen Metabolism in a Murine Model of Pompe Disease. Müller M, ed. *PLoS One*. 2013;8(2):e56181. doi:10.1371/journal.pone.0056181.

211. Katano H, Pesnicak L, Cohen JI. Simvastatin induces apoptosis of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines and delays development of EBV lymphomas. *Proc Natl Acad Sci U S A*. 2004;101(14):4960-4965. doi:10.1073/pnas.0305149101.

212. Sathyapalan T, Atkin SL. Evidence for statin therapy in polycystic ovary syndrome.

Ther Adv Endocrinol Metab. 2010;1(1):15-22. doi:10.1177/2042018810367984.

213. Voermans NC, Lammens M, Wevers RA, Hermus AR, Engelen BG. Statin-disclosed acid maltase deficiency. *J Intern Med*. 2005;258(2):196-197. doi:10.1111/j.1365-2796.2005.01515.x.

214. Rezaie-Majd A, Prager GW, Bucek RA, et al. Simvastatin reduces the expression of adhesion molecules in circulating monocytes from hypercholesterolemic patients. *Arterioscler Thromb Vasc Biol*. 2003;23(3):397-403. doi:10.1161/01.ATV.0000059384.34874.F0.

215. Johnson NP. Metformin use in women with polycystic ovary syndrome. *Ann Transl Med*. 2014;2(6):56. doi:10.3978/j.issn.2305-5839.2014.04.15.

216. Laplante M, Sabatini DM. mTOR Signaling in Growth Control and Disease. *Cell*. 2012;149(2):274-293. doi:10.1016/J.CELL.2012.03.017.

217. Armstrong JL, Bonavaud SM, Toole BJ, Yeaman SJ. Regulation of glycogen synthesis by amino acids in cultured human muscle cells. *J Biol Chem*. 2001;276(2):952-956. doi:10.1074/jbc.M004812200.

218. Yi H, Brooks ED, Thurberg BL, Fyfe JC, Kishnani PS, Sun B. Correction of glycogen storage disease type III with rapamycin in a canine model. *J Mol Med*. 2014;92(6):641-650. doi:10.1007/s00109-014-1127-4.

219. Ashe KM, Taylor KM, Chu Q, et al. Inhibition of glycogen biosynthesis via mTORC1 suppression as an adjunct therapy for Pompe disease. *Mol Genet Metab*. 2010;100(4):309-315. doi:10.1016/j.ymgme.2010.05.001.

220. Lim J-A, Li L, Shirihai OS, Trudeau KM, Puertollano R, Raben N. Modulation of mTOR signaling as a strategy for the treatment of Pompe disease. *EMBO Mol Med*. 2017;9:353-370. doi:10.15252/emmm.

221. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. *Nat Commun*. 2016;7(May 2015):10331. doi:10.1038/ncomms10331.

222. Chou MT, Wang J, Fujita DJ. Src kinase becomes preferentially associated with the VEGFR, KDR/Flk-1, following VEGF stimulation of vascular endothelial cells. *BMC Biochem*. 2002;3:32. http://www.ncbi.nlm.nih.gov/pubmed/12509223. Accessed March 3, 2018.

223. Wedge SR, Kendrew J, Hennequin LF, et al. AZD2171: a highly potent, orally bioavailable, vascular endothelial growth factor receptor-2 tyrosine kinase inhibitor for the treatment of cancer. *Cancer Res*. 2005;65(10):4389-4400. doi:10.1158/0008-5472.CAN-04-4409.

224. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V, Bork P. Drug-induced regulation of target expression. *PLoS Comput Biol*. 2010;6(9). doi:10.1371/journal.pcbi.1000925.

225. Isik Z, Baldow C, Cannistraci CV, et al. Drug target prioritization by perturbed gene expression and network information. *Sci Rep*. 2015;5:17417. doi:10.1038/srep17417.

226. Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Sci Rep*. 2016;6. doi:10.1038/srep24570.

227. Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG. The Orphan Disease Networks. *Am J Hum Genet*. 2011;88:755-766. doi:10.1016/j.ajhg.2011.05.006.

228. Peng Q, Schork NJ. Utility of network integrity methods in therapeutic target identification. *Front Genet*. 2014;5(FEB). doi:10.3389/fgene.2014.00012.

229. Wang X, Thijssen B, Yu H. Target Essentiality and Centrality Characterize Drug Side Effects. Tucker-Kellogg G, ed. *PLoS Comput Biol*. 2013;9(7):e1003119. doi:10.1371/journal.pcbi.1003119.

230. Módos D, Bulusu KC, Fazekas D, et al. Neighbours of cancer-related proteins have key influence on pathogenesis and could increase the drug target space for anticancer therapies. *npj Syst Biol Appl*. 2017;3(2). doi:10.1038/s41540-017-0003-6.

231. Emig D, Ivliev A, Pustovalova O, et al. Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS One*. 2013;8(4). doi:10.1371/journal.pone.0060618.

232. Yeger-Lotem E, Sharan R. Human protein interaction networks across tissues and diseases. *Front Genet*. 2015;6:257. doi:10.3389/fgene.2015.00257.

233. Cheng L, Li J, Ju P, Peng J, Wang Y. SemFunSim: A New Method for Measuring Disease Similarity by Integrating Semantic and Gene Functional Association. *PLoS One*. 2014;9(6):e99415. doi:10.1371/journal.pone.0099415.

234. Sun K, Gonçalves JP, Larminie C, et al. Predicting disease associations via biological network analysis. *BMC Bioinformatics*. 2014;15(1):304. doi:10.1186/1471-2105-15-304.

235. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. *PLoS One*. 2009;4. doi:10.1371/journal.pone.0004346.

236. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. *J Biomed Inform*. 2012;45(2):363-371. doi:10.1016/j.jbi.2011.11.017.

237. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868.

238. Liu CC, Tseng YT, Li W, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res*. 2014;42. doi:10.1093/nar/gku412.

239. Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N. Discovering disease-disease associations by fusing systems-level molecular data. *Sci Rep*. 2013;3:3202. doi:10.1038/srep03202.

240. Sun K, Buchan N, Larminie C, Przulj N. The integrated disease network. *Integr Biol*. 2014;6(11):1069-1079. http://pubs.rsc.org/en/Content/ArticleLanding/2014/IB/C4IB00122B.

241. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*. 2011;7(496):496. doi:10.1038/msb.2011.26.

242. Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One*. 2013;8(11):e78518. doi:10.1371/journal.pone.0078518.

243. Chen Y, Wang W, Zhou Y, et al. In Silico Gene Prioritization by Integrating Multiple Data Sources. Gravenor MB, ed. *PLoS One*. 2011;6(6):e21137. doi:10.1371/journal.pone.0021137.

244. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333-337. doi:10.1038/nmeth.2810.

245. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*. 2015;31(12):i268-i275. doi:10.1093/bioinformatics/btv244.

246. Li P, Nie Y, Yu J. Fusing literature and full network data improves disease similarity computation. *BMC Bioinformatics*. 2016;17(1):326. doi:10.1186/s12859-016-1205-4.

247. Haynes W, Vashisht R, Vallania F, et al. Integrated molecular and clinical analysis for understanding human disease relationships. *bioRxiv*. November 2017:214833. doi:10.1101/214833.

248. Jalili M, Salehzadeh-Yazdi A, Yaghmaie M, Ghavamzadeh A, Alimoghaddam K. Cancerome: A hidden informative subnetwork of the diseasome. *Comput Biol Med*. 2016;76:173-177. doi:10.1016/J.COMPBIOMED.2016.07.010.

249. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. 2015;31(4):608-609. doi:10.1093/bioinformatics/btu684.

250. Lin D. An Information-Theoretic Definition of Similarity. *Proc Fifteenth Int Conf Mach Learn*. 1998:296-304. https://pdfs.semanticscholar.org/3216/3f8f5114beea5576c93b2ce21ec1e48988ce.pdf. Accessed May 18, 2018.

251. Hoehndorf R, Schofield PN, Gkoutos G V. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Sci Rep*. 2015;5(October 2014):10888. doi:10.1038/srep10888.

252. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*. 2016;32(18):2839-2846. doi:10.1093/bioinformatics/btw343.

253. Pinero J, Queralt-Rosinach N, Bravo A, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*. 2015;2015:bav028-bav028. doi:10.1093/database/bav028.

254. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*. 2009;5. doi:10.1371/journal.pcbi.1000353.

255. Rubio-Perez C, Guney E, Aguilar D, et al. Genetic and functional characterization of disease associations explains comorbidity. *Sci Rep*. 2017;7(1):6207. doi:10.1038/s41598-017-04939-4.

256. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull la Société Vaudoise des Sci Nat*. 1901;37:547-579. http://www.bibsonomy.org/bibtex/2224e882aa05e46ae13556fa145dacb06/asalber. Accessed March 4, 2016.

257. Bolstad BM, Irizarry R., Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185-193. doi:10.1093/bioinformatics/19.2.185.

258. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303.

259. Liaw A, Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22.

260. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-3941. doi:10.1093/bioinformatics/bti623.

261. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104. doi:10.1073/pnas.0701361104.

262. Mehta AM, Mooij M, Branković I, Ouburg S, Morré SA, Jordanova ES. Cervical Carcinogenesis and Immune Response Gene Polymorphisms: A Review. *J Immunol Res*. 2017;2017:1-12. doi:10.1155/2017/8913860.

263. Chang M-H. Cancer prevention by vaccination against hepatitis B. *Recent Results Cancer Res*. 2009;181:85-94. http://www.ncbi.nlm.nih.gov/pubmed/19213561. Accessed November 29, 2017.

264. Boehncke W-H, Schön MP. Psoriasis. *Lancet*. 2015;386(9997):983-994. doi:10.1016/S0140-6736(14)61909-7.

265. Egeberg A, Mallbris L, Warren RB, et al. Association between psoriasis and inflammatory bowel disease: a Danish nationwide cohort study. *Br J Dermatol*.

2016;175(3):487-492. doi:10.1111/bjd.14528.

266. Yang XO, Panopoulos AD, Nurieva R, et al. STAT3 regulates cytokine-mediated generation of inflammatory helper T cells. *J Biol Chem*. 2007;282(13):9358-9363. doi:10.1074/jbc.C600321200.

267. Park SC, Jeen YT. Current and emerging biologics for ulcerative colitis. *Gut Liver*. 2015;9(1):18-27. doi:10.5009/gnl14226.

268. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther*. 2009;86(5):507-510. doi:10.1038/clpt.2009.103.

269. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014;506(7488):376-381. http://dx.doi.org/10.1038/nature12873. Accessed April 15, 2015.

270. Jahchan NS, Dudley JT, Mazur PK, et al. A Drug Repositioning Approach Identifies Tricyclic Antidepressants as Inhibitors of Small Cell Lung Cancer and Other Neuroendocrine Tumors. *Cancer Discov*. September 2013. doi:10.1158/2159-8290.CD-13-0183.

271. Dudley JT, Sirota M, Shenoy M, et al. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Sci Transl Med*. 2011;3(96):96ra76. doi:10.1126/scitranslmed.3002648.

272. van Noort V, Schölch S, Iskar M, et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res*. 2014;74(20):5690-5699. http://cancerres.aacrjournals.org/content/early/2014/07/18/0008-5472.CAN-13-3540.abstract. Accessed September 16, 2015.

273. Iwata H, Sawada R, Mizutani S, Yamanishi Y. Systematic Drug Repositioning for a Wide Range of Diseases with Integrative Analyses of Phenotypic and Molecular Data. *J Chem Inf Model*. 2014. doi:10.1021/ci500670q.

274. Nguyen T, Tagett R, Diaz D, Draghici S. A novel approach for data integration and disease subtyping. *Genome Res*. 2017;27(12):2025-2039. doi:10.1101/gr.215129.116.

275. Tong M, Zheng W, Li H, et al. Multi-omics landscapes of colorectal cancer subtypes

discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. *Oncogenesis*. 2016;5(7):e242. doi:10.1038/oncsis.2016.51.

276. Lee S-I, Celik S, Logsdon BA, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun*. 2018;9(1):42. doi:10.1038/s41467-017-02465-5.

277. Sirci F, Napolitano F, Pisonero-Vaquero S, Carrella D, Medina DL, di Bernardo D. Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses. *npj Syst Biol Appl*. 2017;3(1):23. doi:10.1038/s41540-017-0022-3.

APPENDIX A: DATASET USED FOR CHAPTER 3

Note: in order to minimize the impact of laboratory effects on the concordance analysis, where multiple datasets were contributed by the same investigator and less than a year apart, only one of the two was retained. However, the two studies GSE20141 and GSE20153 contributed by Middleton were first reported in the associated meta-analysis; it is not stated whether they originate from the same group so both have been retained.

GEO ID	Contributor Lead author (where different) PMID (where applicable)	Submission date	Platform	Tissue	# case	# control	Species	Model	Sample selection
GSE6613	Scherzer CR 17215369	Dec 2006	U133A	Whole blood	50	22	Human	-	Parkinson's disease v healthy control; GSM153411 and GSM153454 removed as did not pass quality checks
GSE7621	Mullen JF Lesnick TG 17571925	Apr 2007	U133 Plus 2.0	Substantia nigra	16	9	Human	-	-
GSE20141	Middleton FA Zheng B 20926834	Feb 2010	U133 Plus 2.0	Laser-dissected substantia nigra pars compacta neurons	10	8	Human	-	-
GSE20153	Middleton FA Zheng B 20926834	Feb 2010	U133 Plus 2.0	B lymphocytes from peripheral blood	8	8	Human	-	-
GSE20163	Miller RM Zheng B 20926834	Feb 2010	U133A	Substantia nigra	8	9	Human	-	-

GSE20164	Hauser MA Zheng B 20926834	Feb 2010	U133A	Substantia nigra	6	5	Human	-	GSM506020 removed as did not pass quality checks
GSE20168	Middleton FA Zhang Y 15965975 Zheng B 20926834	Feb 2010; Originally published Aug 2005	U133A	Prefrontal cortex area 9 (Brodmann area 9)	14	15	Human	-	-
GSE20314	Wüllner U Zheng B 20926834	Feb 2010	U133A	Cerebellum	4	4	Human	-	GSM509109 removed as did not pass quality checks
GSE20333	Edna G	Feb 2010	HGFocus	Substantia nigra	6	6	Human	-	GSM509556 and GSM509557 removed as did not pass quality checks
GSE24378	Cantuti-Castelvetri I Zheng B 20926834	Sep 2010	X3P	Dopaminergic neurons isolated from substantia nigra	8	9	Human	-	-
GSE43490	Corradini BR 25525598	Jan 2013	AgilentPN	Substantia nigra	8	5	Human	-	SN parkinson's disease v SN control
GSE4788	Miller RM 15329391	May 2006	MurU74	Substantia nigra	4	4	Mouse	MPTP	MPTP MML v saline
GSE24233	Cadet JL	Sep 2010	IlluminaRat	Striatum	6	4	Rat	6-OHDA	Saline lesioned v saline control; GSM596030 removed as did not pass quality checks
GSE4550	Nahon J Storvik M 20206263	Mar 2006	U133A	Putamen	4	4	Macaque	MPTP	Putamen MPTP day 25 v putamen saline
GSE58710	Lipton JW Kanaan NM 25992874	Jun 2014	Rat1.0ST	Substantia nigra	3	3	Rat	6-OHDA	Wk4 6-OHDA v Wk4 vehicle; GSM1417209 removed as did not pass quality checks
GSE8030	Chin MH 18173235	Jun 2007	430A	Striatum	3	3	Mouse	MPTP	MPTP v control
GSE7707	Sforza DM	May 2007	4302	Striatum	3	3	Mouse	MPTP	-

GSE17542	Phani S 20462502	Aug 2009	4302	Dopaminergic neurons isolated from substantia nigra	3	3	Mouse	MPTP	10 day MPTP SN v control SN; GSM437382 removed as did not pass quality checks
GSE35642	Cabeza-Arvelaiz 22970289	Feb 2012	U133A	Neuroblastoma cell line	3	3	Human	rotenone	50nm rotenone 4 week v 0nm rotenone 4 week
GSE8397	Moran LB 16344956	Jul 2007	U133A	Frontal cerebral cortex - superior frontal gyrus	5	3	Human	-	-
GSE31458	Soreq L 22198569	Aug 2011	430A2	Striatal caudate-putamen	2	2	Mouse	MPTP	CPU MPTP FVB/N v CPU naive FVB/N Note: pooled design, each 'sample' contains RNA from 3-4 of 6 mice per condition.
GSE52584	Dorval V 24427314	Nov 2013	MG1.0ST	Striatum	4	4	Mouse	LRRK2	LRRK2 KO vs WT
GSE60413	Kurz A Gispert S 25296918	Aug 2014	4302	Striatum	3	3	Mouse	Pink1	KO 6 week v WT 6 week
GSE18309	Chen K	Sep 2009	U133 Plus 2.0	Peripheral blood mononuclear cells	3	3	Human	AD	Alzheimer's disease v normal
GSE48350	Berchtold NC 18832152 (see GEO series record for other associated studies)	Jun 2013	U133 Plus 2.0	Superior frontal gyrus	21	22	Human	AD	Superior frontal gyrus, age >= 70; GSM300250 removed as did not pass quality checks
GSE36980	Nakabeppu Y Hokama M 23595620	Apr 2012	HG1.0ST	Frontal cortex	15	18	Human	AD	Frontal cortex

GSE74995	Friedman B Srinivasan K 27097852	Nov 2015	AgilentMouse v2	Whole cortex	5	5	Mouse	AD	PS2APP 3 month vs WT 3 month
GSE15824	Morin PJ Grzmil M 21406405	Apr 2009	U133 Plus 2.0	Whole brain	12	2	Human	Tumour	Glioblastoma v normal brain
GSE44971	Lambert SR 23660940	Mar 2013	U133 Plus 2.0	Cerebellum	49	9	Human	Tumour	-
GSE31095	Nilsson RJ 21832279	Aug 2011	AgilentFN	Blood platelets	8	12	Human	Tumour	-
GSE64230	Giachino C 26669487	Dec 2014	MG1.0ST	Whole brain	4	4	Mouse	Tumour	PDGF+p53-/- tumour v control
GSE57036	Sheila AL	Apr 2014	IlluminaMouse	Dorsal brain	5	5	Mouse	Tumour	Tumour v dorsal control
GSE74382	Loiodice S	Oct 2015	Rat2302	Dorsal striatum	7	7	Rat	SNCA	Lesion L-dopa saline vs sham saline saline
GSE55096	Heiman M 24599591	Feb 2014	4302	Striatal neurons	20	20	Mouse	6-OHDA	low L-dopa 6-OHDA v saline ascorbate, Drd1a and Drd2 neurons
GSE72267	Roncaglia P Calligaris R 26510930	Aug 2015	U133A 2.0	Blood	40	19	Human	-	-
GSE54536	Alieva AK 24804238	Jan 2014	Illumina HT12v4	Blood	4	4	Human	-	Exclude pooled RNA
GSE93695	Chen G	Jan 2017	Rat 2.0	Striatum	3	3	Rat	6-OHDA	PD v normal
GSE89562	Kumar A 27884192	Nov 2016	AgilentMouse	Striatum	3	3	Mouse	Maneb-paraquat coexposure	WT-MP v WT
GSE57475	Scherzer CR Locascio JJ 26220939	May 2014	Illumina HT12v3	Blood	93	49	Human	-	-
GSE49036	Dijkstra AA 26087293	Jul 2013	U133 Plus 2.0	SN	6	8	Human	-	Braak stages III and IV v control; GSM1192710_BR34_7SN

									removed as did not pass quality checks
GSE51922	Ezquerra M Fernandez-Santiago R 26516212	Oct 2013	HG1.0ST	iPSC-derived DA neurons	9	4	Human	-	GSM1255326_SP02 removed as did not pass quality checks
GSE89883	Haenseler W	Nov 2016	Illumina HT12v4	iPSCs	4	3	Human	-	Use only first clone from each patient

Abbreviations:

U133A: Affymetrix Human Genome U133A Array
U133A 2.0: Affymetrix Human Genome U133A 2.0 Array
U133 Plus 2.0: Affymetrix Human Genome U133 Plus 2.0 Array
Illumina HT12v3: Illumina HumanHT-12 V3.0 expression beadchip
Illumina HT12v4: Illumina HumanHT-12 V4.0 expression beadchip
HGFocus: Affymetrix Human HG-Focus Target Array
X3P: Affymetrix Human X3P Array
AgilentPN: Agilent Whole Human Genome Microarray 4x44K (Probe name version)
AgilentFN: Agilent Whole Human Genome Microarray 4x44K (Feature number version)
MurU74: Affymetrix Murine Genome U74A Array
IlluminaRat: Illumina ratRef-12 v1.0 expression beadchip
Rat1.0ST: Affymetrix Rat Gene 1.0 ST Array
430A: Affymetrix Mouse Expression 430A Array
4302: Affymetrix Mouse Genome 430 2.0 Array
430A2: Affymetrix Mouse Genome 430A 2.0 Array
MG1.0ST: Affymetrix Mouse Gene 1.0 ST Array
HG1.0ST: Affymetrix Human Gene 1.0 ST Array
AgilentMouse (v2): Agilent Whole Mouse Genome Microarray 4x44K (v2)
IlluminaMouse: Illumina MouseWG-6 v2.0 expression beadchip
Rat2302: Affymetrix Rat Genome 230 2.0 Array

APPENDIX B DATASET USED FOR CHAPTER 4

– in sample selection indicates that all provided samples were used. Platform identifiers are GEO identifiers for individual microarray types, e.g. GPL96 is the Affymetrix Human Genome U133A Array.

Disease dataset

GSE Accession	Condition	Platform	Tissue	Sample selection
GSE6475	acne	GPL571	skin	Lesion v non-acne-patient normal skin (6 case, 6 control)
GSE63107	actinic keratosis	GPL570	skin	AK pre-treatment v uninvolved skin pre-treatment (6 case, 6 control)
GSE9476	acute myeloid leukemia	GPL96	peripheral blood	Peripheral blood only (19 case, 10 control)
GSE8514	adrenal adenoma	GPL570	adrenal gland	-
GSE44456	alcoholism	GPL6244	hippocampus	-
GSE6281	allergic contact dermatitis	GPL570	skin	96 hour timepoint only (6 case, 4 control)
GSE45512	alopecia areata	GPL570	skin	-
GSE36980	alzheimer's disease	GPL6244	hippocampus	Hippocampus only (7 case, 10 control)
GSE28146	alzheimer's disease	GPL570	hippocampus	Severe v control (7 case, 8 control)
GSE26969	aortic aneurysm	GPL570	cranial artery	-
GSE70683	arterial tortuosity syndrome	GPL6244	dermal fibroblasts	-
GSE35571	asthma	GPL570	peripheral blood	Exclude NA (60 case, 64 control)
GSE44971	astrocytoma	GPL570	cerebellum	-
GSE5667	atopic dermatitis	GPL96	skin	Lesional atopic dermatitis v normal healthy (6 case, 5 control)
GSE18123	autism	GPL570	peripheral blood	Autism v control (31 case, 33 control)

GSE40586	bacterial meningitis	GPL6244	peripheral blood	-
GSE12654	bipolar disorder	GPL8300	prefrontal cortex	Bipolar disorder v control (11 case, 15 control)
GSE3167	bladder cancer	GPL96	bladder	Exclude cystectomy samples (9 case, 41 control)
GSE19205	bloom syndrome	GPL571	fibroblasts	Bloom and control only (3 case, 3 control)
GSE54502	bloom syndrome	GPL5175	fibroblasts	Exclude BLM fibroblasts (14 case, 12 control)
GSE5764	breast lobular carcinoma	GPL570	lobular cells	Tumour lobular v ductal carcinoma normal lobular (5 case, 5 control)
GSE31243	cerebral palsy	GPL571	skeletal muscle	Gracilis only (10 case, 10 control)
GSE7803	cervical squamous cell carcinoma	GPL96	cervix	Squamous cell carcinoma v normal cervix (21 case, 10 control)
GSE9750	cervical squamous cell carcinoma	GPL96	cervix	Cervical cancer v normal cervix (33 case, 12 control)
GSE63514	cervical intraepithelial neoplasia	GPL570	cervix	CIN 3 v control (40 case, 24 control)
GSE26725	chronic lymphocytic leukemia	GPL570	peripheral blood	-
GSE42057	chronic obstructive pulmonary disease	GPL570	peripheral blood mononuclear cells	-
GSE8581	chronic obstructive pulmonary disease	GPL570	lung tissue	Exclude unclassified samples (16 case, 19 control)
GSE8671	colorectal adenocarcinoma	GPL570	colonic mucosa	-
GSE4107	colorectal cancer	GPL570	colon mucosa	-
GSE8440	congenital disorders of glycosylation type I	GPL96	dermal fibroblasts	Rep 1 only (9 case, 3 control)
GSE64034	cornelia de lange syndrome	GPL17889	dermal fibroblasts	First replicate only; exclude CHOPS syndrome (2 case, 4 control)
GSE6731	crohn's disease	GPL8300	colon	Crohn's affected v normal (7 case, 4 control)
GSE59071	crohn's disease	GPL6244	colon	CD v control (8 case, 11 control)
GSE15568	cystic fibrosis	GPL96	rectum	-

GSE51808	dengue fever	GPL13158	whole blood	Dengue fever v control (18 case, 9 control)
GSE14335	diamond-blackfan anaemia	GPL571	fibroblasts	-
GSE3585	dilated cardiomyopathy	GPL96	left ventricular tissue	-
GSE52471	discoid lupus erythematosus	GPL571	skin	Discoid lupus v Normal Affy (7 case, 10 control)
GSE5390	down syndrome	GPL96	dorsolateral prefrontal cortex	-
GSE6011	duchenne muscular dystrophy	GPL96	skeletal muscle	Exclude technical rep (22 case, 14 control)
GSE38417	duchenne muscular dystrophy	GPL570	skeletal muscle	-
GSE1004	duchenne muscular dystrophy	GPL8300	skeletal muscle	Samples run on GPL8300 only (12 case, 11 control)
GSE1122	emphysema	GPL80	lung	Exclude AAD-related emphysema (5 case, 5 control)
GSE63678	endometrial carcinoma	GPL571	endometrium	Endometrial cancer v normal endometrium (7 case, 5 control)
GSE6364	endometriosis	GPL570	endometrium	Mid-secretory phase only (9 case, 8 control)
GSE28315	epidermolysis bullosa simplex	GPL6244	epidermis	-
GSE26050	familial hemophagocytic lymphohistiocytosis	GPL570	peripheral blood mononuclear cells	-
GSE16334	fanconi anemia	GPL96	bone marrow	-
GSE62721	fragile x syndrome	GPL6244	fibroblasts	Exclude iPSCs and neurons (3 case, 2 control)
GSE79973	gastric adenocarcinoma	GPL570	gastric mucosa	-
GSE79704	generalized pustular psoriasis	GPL19983	skin	Exclude plaque psoriasis (32 case, 20 control)
GSE31014	guillain-barre syndrome	GPL96	peripheral blood leukocytes	-
GSE6631	head and neck squamous cell carcinoma	GPL8300	head & neck mucosa	-
GSE49954	chronic hepatitis c	GPL570	t lymphocytes	High viral load, CD4+ cells v control (5 case, 5 control)

GSE62232	hepatocellular carcinoma	GPL570	liver	-
GSE2171	hiv	GPL201	peripheral blood mononuclear cells	-
GSE47044	hodgkin's lymphoma	GPL6244	b cells	NLPHL v control (10 case, 5 control)
GSE8762	huntington's disease	GPL570	blood lymphocyte	-
GSE3860	hutchinson-gilford progeria	GPL96	fibroblasts	Sample 1 only (3 case, 3 control)
GSE69391	hutchinson-gilford progeria	GPL570	dermal fibroblasts	Exclude old healthy (6 case, 3 control)
GSE9499	icf syndrome	GPL96	lymphoblastoid cell line	Biological rep 1 only (3 case, 5 control)
GSE24206	idiopathic pulmonary fibrosis	GPL570	lung	Advanced IPF, upper lobe v control (5 case, 6 control)
GSE40568	IgG4-related disease	GPL570	labial salivary glands	Exclude Sjogren's (5 case, 3 control)
GSE27131	influenza	GPL6244	peripheral blood	Day 0 v control (7 case, 7 control
GSE36701	irritable bowel syndrome	GPL570	rectal colon	Part 1 only: IBS-C and IBS-D v healthy volunteers (87 case, 40 control)
GSE42955	ischemic cardiomyopathy	GPL6244	left ventricular tissue	Ischemic cardiomyopathy v normal (12 case, 5 control)
GSE22255	ischemic stroke	GPL570	peripheral blood mononuclear cells	-
GSE48574	ISCU myopathy	GPL570	muscle	-
GSE80060	juvenile idiopathic arthritis	GPL570	whole blood	Day 1 placebo v control (22 case, 22 control)
GSE71935	juvenile myelomonocytic leukemia	GPL570	bone marrow	Exclude peripheral blood (33 case, 9 control)
GSE47642	kindler syndrome	Illumina	skin	-
GSE42331	klinefelter syndrome	GPL6244	whole blood	Exclude female controls (35 case, 15 control)
GSE47584	klinefelter syndrome	Agilent	peripheral blood	-

GSE16020	leukopenia	GPL570	blood leukocytes	-
GSE11681	limb-girdle muscular dystrophy	GPL96	muscle	-
GSE38961	loeys-dietz syndrome	GPL570	blood endothelial cells	-
GSE19804	non-small cell lung carcinoma	GPL570	lung	-
GSE44593	major depressive disorder	GPL570	amygdala	-
GSE6872	teratozoospermia	GPL570	semen	-
GSE51024	malignant pleural mesothelioma	GPL570	lung	-
GSE5808	measles	GPL96	peripheral blood	Entry v control (5 case, 3 control)
GSE14882	melas syndrome	GPL96	peripheral blood	Note: pooled controls
GSE23832	multiple sclerosis	GPL6244	peripheral blood mononuclear cells	-
GSE21942	multiple sclerosis	GPL570	peripheral blood mononuclear cells	Exclude technical rep (12 case, 15 control)
GSE16461	multiple sclerosis	GPL570	blood CD8+ t cells	-
GSE43591	multiple sclerosis	GPL570	blood t-cells	-
GSE58831	myelodysplastic syndrome	GPL570	bone marrow CD34+ cells	-
GSE48060	myocardial infarction	GPL570	peripheral blood	-
GSE13597	nasopharyngeal carcinoma	GPL96	nasopharynx	-
GSE12452	nasopharyngeal carcinoma	GPL570	nasopharynx	-
GSE65170	nestor-guillermo progeria	GPL5175	iPSCs	Exclude fibroblasts (4 case, 2 control)
GSE9624	obesity	GPL570	adipose tissue	-
GSE20347	esophageal squamous cell carcinoma	GPL571	oesophagus	-

GSE22855	ollier disease	Illumina	cartilage	-
GSE55235	osteoarthritis	GPL96	synovium	Exclude rheumatoid arthritis samples (10 case, 10 control). Note same publication as GSE55457 but different submitter and institution.
GSE7429	osteoporosis	GPL96	blood lymphocytes	-
GSE56815	osteoporosis	GPL96	blood monocytes	-
GSE14245	pancreatic cancer	GPL570	saliva	-
GSE7621	parkinson's disease	GPL570	substantia nigra	-
GSE8397	parkinson's disease	GPL96	substantia nigra	Medial substantia nigra (15 case, 8 control)
GSE20291	parkinson's disease	GPL96	striatum	-
GSE7586	placental malaria	GPL570	placenta	Exclude past malaria (10 case, 5 control)
GSE34526	polycystic ovary syndrome	GPL570	ovarian follicle (granulosa cells)	-
GSE38680	pompe disease	GPL570	biceps	Exclude quadriceps and control 10 (later diagnosed with MELAS) (9 control, 9 disease)
GSE12767	preeclampsia	GPL570	placenta	-
GSE36314	prolactinoma	GPL8300	pituitary gland	-
GSE55945	prostate cancer	GPL570	prostate	Exclude corrupted files GSM1348937 and GSM1348948 (12 case, 7 control)
GSE14905	psoriasis	GPL570	skin	Exclude 'Uninvolved skin' (33 case, 21 control)
GSE53408	pulmonary arterial hypertension	GPL6244	lung	-
GSE36895	clear-cell renal cell carcinoma	GPL570	kidney	Exclude mouse tumourgraft (29 case, 23 control)
GSE6344	clear-cell renal cell carcinoma	GPL96	kidney	-
GSE75303	rett syndrome	Illumina	frontal cortex	Exclude temporal cortex (3 case, 3 control)
GSE77298	rheumatoid arthritis	GPL570	synovium	-

GSE55457	rheumatoid arthritis	GPL96	synovium	Exclude osteoarthritis samples (13 case, 10 control). Note same publication as GSE55235 but different submitter and institution.
GSE65914	rosacea	GPL570	skin	-
GSE19314	sarcoidosis	GPL570	whole blood	Sarcoidosis v control (38 case, 20 control)
GSE70019	schnitzler's syndrome	Illumina	peripheral blood mononuclear cells	Exclude drug-treated (3 case, 3 control)
GSE13205	sepsis	GPL570	skeletal muscle	-
GSE16524	setleis syndrome	GPL570	dermal fibroblasts	-
GSE32057	shwachman-diamond syndrome	GPL570	bone marrow mononuclear cells	-
GSE11524	sickle cell disease	GPL570	blood platelets	-
GSE61120	silver-russell syndrome	GPL13667	dermal fibroblasts	Exclude hypomethylated clones (4 case, 4 control)
GSE66795	sjogren's syndrome	Illumina	whole blood	Mid fatigue level only (74 case, 29 control)
GSE48378	sjogren's syndrome	GPL5175	Peripheral blood mononuclear cells	-
GSE61203	smith-lemli-opitz syndrome	GPL5175	iPSCs	Cholesterol-deficient 7d only (4 case, 4 control)
GSE27200	sotos syndrome	GPL570	dermal fibroblasts	Exclude retinoic acid (9 case, 9 control)
GSE10325	systemic lupus erythematosus	GPL96	perhipheral blood	-
GSE30153	systemic lupus erythematosus (quiescent)	GPL570	blood lymphocytes	-
GSE26049	essential thrombocythemia	GPL570	whole blood	Essential thrombocythemia v control, RMA only (19 case, 21 control)
GSE3678	thyroid carcinoma	GPL570	thyroid	-
GSE46687	turner syndrome	GPL570	peripheral blood mononuclear cells	Exclude paternal inherited

GSE55098	type 1 diabetes mellitus	GPL570	peripheral blood mononuclear cells	-
GSE9006	type 1 diabetes mellitus	GPL96	peripheral blood mononuclear cells	Newly diagnosed T1D only (43 case, 20 control)
GSE38642	type 2 diabetes mellitus	GPL6244	pancreatic islets	-
GSE25724	type 2 diabetes mellitus	GPL96	pancreatic islets	-
GSE38713	ulcerative colitis	GPL570	colon	Active UC, involved mucosa v control (15 case, 13 control)
GSE5563	vulvar intraepithelial neoplasia	GPL570	vulva	-
GSE48761	werner syndrome	GPL6244	dermal fibroblasts	Exclude iPSC/ESC; rep1 only (5 case, 5 control)
GSE16715	williams syndrome	GPL570	skin fibroblasts	-
GSE71664	williams syndrome	Illumina	iPSCs	iPSCs only; rep 1 only (2 case, 2 control)

Drug response dataset

GSE Accession	Drug	Condition	Platform	Tissue	Sample selection	Dose and time
GSE32569	cediranib	alveolar soft part sarcoma	GPL570	tumor	-	Baseline vs 3-5 days after treatment
GSE60540	everolimus	acute lymphoblastic leukemia	Illumina	peripheral blood	-	Baseline vs 24 hours after treatment
GSE10433	isotretinoin	acne	GPL571	skin	-	Baseline vs 1 week after treatment
GSE5462	letrozole	breast carcinoma	GPL96	breast	-	*NA*
GSE45867	methotrexate	rheumatoid arthritis	GPL570	synovium	Exclude methotrexate	*NA*
GSE19136	paclitaxel	-	GPL570	artery	Exclude control (non-stented)	Paclitaxel stent vs bare metal stent after 48h
GSE32357	resveratrol	obesity	GPL11532	vastus lateralis muscle	-	30 days placebo vs 30 days reseveratrol
GSE68421	resveratrol	non-alcoholic fatty liver disease	GPL16686	liver	Exclude placebo	Baseline vs 6 months treatment
GSE38663	ribavirin	hepatitis c	GPL570	liver	Exclude IFN, IFN+RBV	*NA*
GSE58837	sunitinib	breast carcinoma	GPL6244	breast	Exclude T3	*NA*
GSE12665	tamoxifen	breast carcinoma	Agilent	arm	Exclude ER-negative	Baseline vs treatment
GSE80060	canakinumab	juvenile idiopathic arthritis	GPL570	whole blood	Day 3 vs placebo	Day 3 vs placebo

GSE54629	rituximab	rheumatoid arthritis	GPL6244	whole blood	Week 24 vs baseline	Week 24 vs baseline
GSE58558	cyclosporine	atopic dermatitis	GPL570	skin	Week 12 vs baseline; lesional skin only	Week 12 vs baseline; lesional skin only
GSE45468	infliximab	major depressive disorder	Illumina	whole blood	Week 12 vs baseline	Week 12 vs baseline
GSE24742	rituximab	rheumatoid arthritis	GPL570	synovium	Week 12 vs baseline	Week 12 vs baseline
GSE16879	infliximab	ulcerative colitis	GPL570	colonic mucosa	Week 5 vs baseline; exclude CD	Week 5 vs baseline
GSE45867	tocilizumab	rheumatoid arthritis	GPL570	synovium	Week 12 vs baseline	Week 12 vs baseline
GSE83530	valproic acid	breast carcinoma	GPL571	breast tumor	Day 10 vs baseline	Day 10 vs baseline

APPENDIX C: DATASET USED FOR CHAPTER 5

– in sample selection indicates that all provided samples were used. Platform identifiers are GEO identifiers for individual microarray types, e.g. GPL96 is the Affymetrix Human Genome U133A Array.

Accession	Condition	Platform	Tissue	Sample selection
GSE6475	acne	GPL571	skin	Lesion v non-acne-patient normal skin (6 case, 6 control)
GSE63107	actinic keratosis	GPL570	skin	AK pre-treatment v uninvolved skin pre-treatment (6 case, 6 control)
GSE9476	acute myeloid leukemia	GPL96	peripheral blood	Peripheral blood only (19 case, 10 control)
GSE8514	adrenal adenoma	GPL570	adrenal gland	-
GSE44456	alcoholism	GPL6244	hippocampus	-
GSE41649	allergic asthma	GPL96	bronchus	-
GSE6281	allergic contact dermatitis	GPL570	skin	96 hour timepoint only (6 case, 4 control)
GSE45512	alopecia areata	GPL570	skin	-
GSE28146	alzheimer's disease	GPL570	hippocampus	Severe v control (7 case, 8 control)
GSE26969	aortic aneurysm	GPL570	cranial artery	-
GSE35571	asthma	GPL570	peripheral blood	Exclude NA (60 case, 64 control)
GSE44971	astrocytoma	GPL570	cerebellum	-

GSE5667	atopic dermatitis	GPL96	skin	Lesional atopic dermatitis v normal healthy (6 case, 5 control)
GSE18123	autism	GPL570	peripheral blood	Autism v control (31 case, 33 control)
GSE40586	bacterial meningitis	GPL6244	peripheral blood	-
GSE12654	bipolar disorder	GPL8300	prefrontal cortex	Bipolar disorder v control (11 case, 15 control)
GSE3167	bladder cancer	GPL96	bladder	Exclude cystectomy samples (9 case, 41 control)
GSE5764	breast lobular carcinoma	GPL570	lobular cells	Tumour lobular v ductal carcinoma normal lobular (5 case, 5 control)
GSE31243	cerebral palsy	GPL571	skeletal muscle	Gracilis only (10 case, 10 control)
GSE9750	cervical squamous cell carcinoma	GPL96	cervix	Cervical cancer v normal cervix (33 case, 12 control)
GSE63514	cervical intraepithelial neoplasia	GPL570	cervix	CIN 3 v control (40 case, 24 control)
GSE26725	chronic lymphocytic leukemia	GPL570	peripheral blood	-
GSE8581	chronic obstructive pulmonary disease	GPL570	lung tissue	Exclude unclassified samples (16 case, 19 control)
GSE8671	colorectal adenocarcinoma	GPL570	colonic mucosa	-
GSE59071	crohn's disease	GPL6244	colon	CD v control (8 case, 11 control)
GSE15568	cystic fibrosis	GPL96	rectum	-
GSE51808	dengue fever	GPL13158	whole blood	Dengue fever v control (18 case, 9 control)
GSE3585	dilated cardiomyopathy	GPL96	left ventricular tissue	-

GSE52471	discoid lupus erythematosus	GPL571	skin	Discoid lupus v Normal Affy (7 case, 10 control)
GSE5390	down syndrome	GPL96	dorsolateral prefrontal cortex	-
GSE38417	duchenne muscular dystrophy	GPL570	skeletal muscle	-
GSE1122	emphysema	GPL80	lung	Exclude AAD-related emphysema (5 case, 5 control)
GSE63678	endometrial carcinoma	GPL571	endometrium	Endometrial cancer v normal endometrium (7 case, 5 control)
GSE6364	endometriosis	GPL570	endometrium	Mid-secretory phase only (9 case, 8 control)
GSE79973	gastric adenocarcinoma	GPL570	gastric mucosa	-
GSE6631	head and neck squamous cell carcinoma	GPL8300	head & neck mucosa	-
GSE58208	chronic hepatitis b (carrier)	GPL570	peripheral blood mononuclear cells	Exclude hepatocellular carcinoma (12 case, 5 control)
GSE49954	chronic hepatitis c	GPL570	t lymphocytes	High viral load, CD4+ cells v control (5 case, 5 control)
GSE62232	hepatocellular carcinoma	GPL570	liver	-
GSE2171	hiv	GPL201	peripheral blood mononuclear cells	-
GSE47044	hodgkin's lymphoma	GPL6244	b cells	NLPHL v control (10 case, 5 control)
GSE8762	huntington's disease	GPL570	blood lymphocyte	-
GSE24206	idiopathic pulmonary fibrosis	GPL570	lung	Advanced IPF, upper lobe v control (5 case, 6 control)

GSE27131	influenza	GPL6244	peripheral blood	Day 0 v control (7 case, 7 control
GSE36701	irritable bowel syndrome	GPL570	rectal colon	Part 1 only: IBS-C and IBS-D v healthy volunteers (87 case, 40 control)
GSE42955	ischemic cardiomyopathy	GPL6244	left ventricular tissue	Ischemic cardiomyopathy v normal (12 case, 5 control)
GSE22255	ischemic stroke	GPL570	peripheral blood mononuclear cells	-
GSE80060	juvenile idiopathic arthritis	GPL570	whole blood	Day 1 placebo v control (22 case, 22 control)
GSE16020	leukopenia	GPL570	blood leukocytes	-
GSE11681	limb-girdle muscular dystrophy	GPL96	muscle	-
GSE19804	non-small cell lung carcinoma	GPL570	lung	-
GSE44593	major depressive disorder	GPL570	amygdala	-
GSE6872	teratozoospermia	GPL570	semen	-
GSE51024	malignant pleural mesothelioma	GPL570	lung	-
GSE5808	measles	GPL96	peripheral blood	Entry v control (5 case, 3 control)
GSE21942	multiple sclerosis	GPL570	peripheral blood mononuclear cells	Exclude technical rep (12 case, 15 control)
GSE48060	myocardial infarction	GPL570	peripheral blood	-
GSE12452	nasopharyngeal carcinoma	GPL570	nasopharynx	-
GSE9624	obesity	GPL570	adipose tissue	-
GSE20347	esophageal squamous cell carcinoma	GPL571	oesophagus	-

GSE55235	osteoarthritis	GPL96	synovium	Exclude rheumatoid arthritis samples (10 case, 10 control).
GSE56815	osteoporosis	GPL96	blood monocytes	-
GSE14245	pancreatic cancer	GPL570	saliva	-
GSE7621	parkinson's disease	GPL570	substantia nigra	-
GSE7586	placental malaria	GPL570	placenta	Exclude past malaria (10 case, 5 control)
GSE34526	polycystic ovary syndrome	GPL570	ovarian follicle	-
GSE12767	preeclampsia	GPL570	placenta	-
GSE36314	prolactinoma	GPL8300	pituitary gland	-
GSE55945	prostate cancer	GPL570	prostate	Exclude corrupted files GSM1348937 and GSM1348948 (12 case, 7 control)
GSE14905	psoriasis	GPL570	skin	Exclude 'Uninvolved skin' (33 case, 21 control)
GSE53408	pulmonary arterial hypertension	GPL6244	lung	-
GSE36895	clear-cell renal cell carcinoma	GPL570	kidney	Exclude mouse tumourgraft (29 case, 23 control)
GSE77298	rheumatoid arthritis	GPL570	synovium	-
GSE65914	rosacea	GPL570	skin	-
GSE19314	sarcoidosis	GPL570	whole blood	Sarcoidosis v control (38 case, 20 control)
GSE13205	sepsis	GPL570	skeletal muscle	-
GSE11524	sickle cell disease	GPL570	blood platelets	-

GSE10325	systemic lupus erythematosus	GPL96	perhipheral blood	-
GSE26049	essential thrombocythemia	GPL570	whole blood	Essential thrombocythemia v control, RMA only (19 case, 21 control)
GSE3678	thyroid carcinoma	GPL570	thyroid	-
GSE55098	type 1 diabetes mellitus	GPL570	peripheral blood mononuclear cells	-
GSE25724	type 2 diabetes mellitus	GPL96	pancreatic islets	-
GSE38713	ulcerative colitis	GPL570	colon	Active UC, involved mucosa v control (15 case, 13 control)
GSE5563	vulvar intraepithelial neoplasia	GPL570	vulva	-

APPENDIX D: DISEASE NAME MAPPING USED FOR CHAPTER 5

Transcriptomic space	Ontological and phenotypic spaces	Literature co-occurrence space	Genetic space	Drug space	ICD 3-digit code	ICD name
acne	acne	acne vulgaris	acne	acne	706	diseases of sebaceous glands
actinic keratosis	actinic keratosis	actinic keratosis	keratosis, actinic	actinic keratosis	702	other dermatoses
acute myeloid leukemia	acute myeloid leukemia	acute myeloid leukemia	leukemia, myeloid, acute	acute myeloid leukemia	205	myeloid leukemia
adrenal adenoma	adrenal adenoma	adrenocortical adenoma	adrenocortical adenoma	adrenocortical carcinoma	227	benign neoplasm of other endocrine glands and related structures
alcoholism	alcohol dependence	alcohol dependence	alcohol abuse or dependence	alcohol dependence	303	alcohol dependence syndrome
allergic asthma	allergic asthma	asthma, aspirin-induced	allergic asthma	asthma	493	asthma
allergic contact dermatitis	allergic contact dermatitis	allergic contact dermatitis	dermatitis, allergic contact	contact dermatitis	692	contact dermatitis and other eczema
alopecia areata	alopecia areata	alopecia areata	alopecia areata	alopecia areata	704	diseases of hair and hair follicles
alzheimer's disease	alzheimer's disease	alzheimer's disease	alzheimer disease	alzheimers disease	331	other cerebral degenerations

aortic aneurysm	aortic aneurysm	aortic aneurysm	aortic aneurysm	aortic aneurysm	441	aortic aneurysm and dissection
asthma	asthma	asthma	asthma	asthma	493	asthma
astrocytoma	astrocytoma	astrocytoma	astrocytoma	astrocytoma	191	malignant neoplasm of brain
atopic dermatitis	atopic dermatitis	atopic dermatitis	adult atopic dermatitis	atopic eczema	691	atopic dermatitis and related conditions
autism	autistic disorder	autistic disorder	autistic disorder	autism	299	pervasive developmental disorders
bacterial meningitis	bacterial meningitis	bacterial meningitis	meningitis, bacterial	bacterial meningitis	320	bacterial meningitis
bipolar disorder	bipolar disorder	bipolar disorder	bipolar disorder	bipolar disorder	296	episodic mood disorders
bladder cancer	urinary bladder cancer	urinary bladder cancer	carcinoma of bladder	bladder carcinoma	188	malignant neoplasm of bladder
breast lobular carcinoma	invasive lobular carcinoma	breast cancer	breast cancer, lobular	breast carcinoma	174	malignant neoplasm of female breast
cerebral palsy	cerebral palsy	spastic diplegia	cerebral palsy	cerebral palsy	343	infantile cerebral palsy
cervical squamous cell carcinoma	cervical squamous cell carcinoma	cervical cancer	cervix carcinoma	cervical carcinoma	180	malignant neoplasm of cervix uteri
cervical intraepithelial neoplasia	cervix uteri carcinoma in situ	cervix uteri carcinoma in situ	high grade cervical intraepithelial neoplasia	cervical carcinoma	233	carcinoma in situ of breast and genitourinary system

chronic lymphocytic leukemia	chronic lymphocytic leukemia	chronic lymphocytic leukemia	leukemia, lymphocytic, chronic, b-cell	chronic lymphocytic leukemia	204	lymphoid leukemia
chronic obstructive pulmonary disease	chronic obstructive pulmonary disease	chronic obstructive pulmonary disease	severe chronic obstructive pulmonary disease	chronic obstructive pulmonary disease	496	chronic airway obstruction, not elsewhere classified
colorectal adenocarcinoma	colorectal cancer	colon cancer	colorectal cancer	colorectal adenocarcinoma	153	malignant neoplasm of colon
crohn's disease	crohn's disease	crohn's colitis	crohn disease	crohn's disease	555	regional enteritis
cystic fibrosis	cystic fibrosis	cystic fibrosis	cystic fibrosis	cystic fibrosis	277	other and unspecified disorders of metabolism
dengue fever	dengue disease	dengue disease	dengue	not found	61	dengue
dilated cardiomyopathy	dilated cardiomyopathy	dilated cardiomyopathy	cardiomyopathy, dilated	dilated cardiomyopathy	425	cardiomyopathy
discoid lupus erythematosus	discoid lupus erythematosus of eyelid	lupus erythematosus, discoid	lupus erythematosus, discoid	cutaneous lupus erythematosus	373	inflammation of eyelids
down syndrome	down syndrome	down syndrome	down syndrome	down syndrome	758	chromosomal anomalies
duchenne muscular dystrophy	duchenne muscular dystrophy	duchenne muscular dystrophy	muscular dystrophy, duchenne	duchenne muscular dystrophy	359	muscular dystrophies and other myopathies
emphysema	pulmonary emphysema	emphysema	pulmonary emphysema	emphysema	492	emphysema

endometrial carcinoma	endometrial carcinoma	endometrial carcinoma	endometrial carcinoma	endometrial neoplasm	182	malignant neoplasm of body of uterus
endometriosis	endometriosis	endometriosis of uterus	endometriosis	endometriosis	617	endometriosis
gastric adenocarcinoma	gastric adenocarcinoma	stomach cancer	stomach carcinoma	gastric adenocarcinoma	151	malignant neoplasm of stomach
head and neck squamous cell carcinoma	head and neck squamous cell carcinoma	carcinoma, squamous cell of head and neck	carcinoma, squamous cell of head and neck	head and neck squamous cell carcinoma	195	malignant neoplasm of other and ill-defined sites
chronic hepatitis b (carrier)	hepatitis b	hepatitis b	hepatitis b, chronic	chronic hepatitis b infection	70	viral hepatitis
chronic hepatitis c	hepatitis c	hepatitis c	hepatitis c, chronic	chronic hepatitis c infection	70	viral hepatitis
hepatocellular carcinoma	hepatocellular carcinoma	carcinoma, hepatocellular	adult primary hepatocellular carcinoma	hepatocellular carcinoma	155	malignant neoplasm of liver and intrahepatic bile ducts
hiv	human immunodeficiency virus infectious disease	human immunodeficiency virus infectious disease	hiv infections	hiv infection	42	human immunodeficiency virus [hiv] disease
hodgkin's lymphoma	hodgkin's lymphoma	hodgkin's lymphoma, lymphocytic-histiocytic predominance	classical hodgkin lymphoma	hodgkins lymphoma	201	hodgkin's disease
huntington's disease	huntington's disease	huntington's disease	huntington disease	huntington disease	333	other extrapyramidal disease and abnormal movement disorders

idiopathic pulmonary fibrosis	idiopathic pulmonary fibrosis	idiopathic pulmonary fibrosis	idiopathic pulmonary fibrosis	idiopathic pulmonary fibrosis	516	other alveolar and parietoalveolar pneumonopathy
influenza	influenza	influenza	influenza, human	influenza infection	487	influenza
irritable bowel syndrome	irritable bowel syndrome	irritable bowel syndrome	irritable bowel syndrome	irritable bowel syndrome	564	functional digestive disorders, not elsewhere classified
ischemic cardiomyopathy	cardiomyopathy	cardiomyopathy	ischemic cardiomyopathy	cardiomyopathy	414	other forms of chronic ischemic heart disease
ischemic stroke	cerebrovascular disease	cerebrovascular disease	ischemic stroke	stroke	434	occlusion of cerebral arteries
juvenile idiopathic arthritis	juvenile rheumatoid arthritis	rheumatoid arthritis, systemic juvenile	juvenile rheumatoid arthritis	chronic childhood arthritis	714	rheumatoid arthritis and other inflammatory polyarthropathies
leukopenia	leukopenia	leukopenia	leukopenia	not found	288	diseases of white blood cells
limb-girdle muscular dystrophy	limb-girdle muscular dystrophy	limb-girdle muscular dystrophy	muscular dystrophies, limb-girdle	not found	359	muscular dystrophies and other myopathies
non-small cell lung carcinoma	non-small cell lung carcinoma	non-small cell lung carcinoma	carcinoma, non-small-cell lung	non-small cell lung carcinoma	162	malignant neoplasm of trachea, bronchus, and lung
major depressive disorder	major depressive disorder	depressive disorder, major	depressive disorder, major	unipolar depression	296	episodic mood disorders

teratozoospermia	azoospermia	male infertility	teratospermia	azoospermia	792	nonspecific abnormal findings in other body substances
malignant pleural mesothelioma	malignant pleural mesothelioma	mesothelioma, malignant	pleural malignant mesothelioma	mesothelioma	163	malignant neoplasm of pleura
measles	measles	measles	measles	not found	55	measles
multiple sclerosis	multiple sclerosis	multiple sclerosis	multiple sclerosis	multiple sclerosis	340	multiple sclerosis
myocardial infarction	myocardial infarction	myocardial infarction	myocardial infarction	myocardial infarction	410	acute myocardial infarcation
nasopharyngeal carcinoma	nasopharynx carcinoma	nasopharyngeal carcinoma	nasopharyngeal carcinoma	nasopharyngeal neoplasm	147	malignant neoplasm of nasopharynx
obesity	obesity	obesity	obesity	obesity	278	overweight, obesity and other hyperalimentation
esophageal squamous cell carcinoma	esophagus squamous cell carcinoma	esophageal squamous cell carcinoma	esophageal squamous cell carcinoma	esophageal carcinoma	150	malignant neoplasm of esophagus
osteoarthritis	osteoarthritis	osteoarthritis	osteoarthritis	osteoarthritis	715	osteoarthrosis and allied disorders
osteoporosis	osteoporosis	osteoporosis	osteoporosis	osteoporosis	733	other disorders of bone and cartilage
pancreatic cancer	pancreatic cancer	pancreatic cancer	pancreatic carcinoma	pancreatic carcinoma	157	malignant neoplasm of pancreas
parkinson's disease	parkinson's disease	parkinson's disease	parkinson disease	parkinson's disease	332	parkinson's disease

placental malaria	malaria	malaria	malaria	malaria	84	malaria
polycystic ovary syndrome	polycystic ovary syndrome	polycystic ovary syndrome	polycystic ovary syndrome	polycystic ovary syndrome	256	ovarian dysfunction
preeclampsia	pre-eclampsia	pre-eclampsia	pre-eclampsia	preeclampsia	642	hypertension complicating pregnancy, childbirth, and the puerperium
prolactinoma	prolactinoma	prolactinoma	prolactinoma	hyperprolactinemia	253	disorders of the pituitary gland and its hypothalamic control
prostate cancer	prostate cancer	prostate cancer	prostate carcinoma	prostate carcinoma	185	malignant neoplasm of prostate
psoriasis	psoriasis	parapsoriasis	psoriasis	psoriasis	696	psoriasis and similar disorders
pulmonary arterial hypertension	pulmonary hypertension	pulmonary hypertension	pulmonary arterial hypertension	pulmonary hypertension	416	chronic pulmonary heart disease
clear-cell renal cell carcinoma	renal clear cell carcinoma	clear-cell metastatic renal cell carcinoma	non-hereditary clear cell renal cell carcinoma	clear cell renal carcinoma	189	malignant neoplasm of kidney and other and unspecified urinary organs
rheumatoid arthritis	rheumatoid arthritis	rheumatoid arthritis	arthritis, rheumatoid	rheumatoid arthritis	714	rheumatoid arthritis and other inflammatory polyarthropathies

rosacea	rosacea	rosacea	rosacea	rosacea	695	erythematous conditions
sarcoidosis	sarcoidosis	sarcoidosis	sarcoidosis	sarcoidosis	135	sarcoidosis
sepsis	disease by infectious agent	sepsis	sepsis	sepsis	995	certain adverse effects not elsewhere classified
sickle cell disease	sickle cell anemia	sickle cell anemia	anemia, sickle cell	sickle cell anemia	282	hereditary hemolytic anemias
systemic lupus erythematosus	systemic lupus erythematosus	systemic lupus erythematosus	lupus erythematosus, systemic	systemic lupus erythematosus	710	diffuse diseases of connective tissue
essential thrombocythemia	essential thrombocythemia	essential thrombocythemia	thrombocythemia, essential	essential thrombocythemia	238	neoplasm of uncertain behavior of other and unspecified sites and tissues
thyroid carcinoma	thyroid carcinoma	thyroid cancer	thyroid carcinoma	thyroid carcinoma	193	malignant neoplasm of thyroid gland
type 1 diabetes mellitus	type 1 diabetes mellitus	type 1 diabetes mellitus	diabetes mellitus, type 1	type i diabetes mellitus	250	diabetes mellitus
type 2 diabetes mellitus	type 2 diabetes mellitus	type 2 diabetes mellitus	diabetes mellitus, type 2	type ii diabetes mellitus	250	diabetes mellitus
ulcerative colitis	ulcerative colitis	ulcerative colitis	colitis, ulcerative	ulcerative colitis	556	ulcerative colitis
vulvar intraepithelial neoplasia	vulva cancer	vulva cancer	vulvar intraepithelial neoplasia, usual type	vulvar intraepithelial neoplasia	233	carcinoma in situ of breast and genitourinary system

APPENDIX E: SIGNIFICANCE THRESHOLDS OF CONCORDANCE FOR DIFFERENT SUBGROUP SIZES

95th percentile values of the distribution of average correlation over randomly selected subgroups of PD studies. For a correlation in any (real) subgroup to be considered significant, it must be greater than or equal to the 95th percentile value of random subgroup correlation for its size. Smaller subgroups are more likely to show higher correlation through chance alone, and therefore smaller subgroups need higher average concordance to be considered significant.

Subgroup size	**95th percentile**	**Subgroup size**	**95th percentile**
3	0.26	17	0.09
4	0.20	18	0.09
5	0.18	19	0.09
6	0.16	20	0.08
7	0.15	21	0.08
8	0.13	22	0.08
9	0.13	23	0.08
10	0.12	26	0.08
11	0.11	27	0.07
12	0.11	28	0.07
13	0.11	29	0.07
14	0.10	30	0.07
15	0.10	31	0.06
16	0.09	32	0.06

APPENDIX F: CONCORDANCE OVER BASE VS SUBSET SHARED GENES

Concordance calculated over the base set of genes, consisting of the 2,513 genes shared between all studies, vs concordance calculated over the larger sets of genes shared between subsets of studies. In most cases the result is not substantially different; concordance over the base set was reported in the main text in order not to bias the results due to changing geneset sizes.

Subset	**Subset size**	**Score over 2,372 genes**	**Score over shared genes**	**Number of shared genes**
All PD studies	33	0.05	-	2,513
SN	8	0.30	0.30	2,976
Striatum	9	0.07	0.06	6,153
Human	19	0.08	0.09	4,776
Human, in vivo	15	0.15	0.15	5,082
Mice	9	0.03	0.04	7,609

APPENDIX G: PATHWAY ENRICHMENT RESULTS FOR PARKINSON'S DISEASE STUDIES

Reactome pathways identified by gene set enrichment analysis of differential gene expression in human Parkinson's disease studies

Upregulated pathways (NES>0, FDR<0.25)	**Number of human studies in which pathway is in top 50 pathways by abs(NES):**
Attenuation phase	4
Interferon alpha/beta signaling	4
Diseases associated with the TLR signaling cascade	3
Diseases of Immune System	3
Cellular response to heat stress	3
HSF1-dependent transactivation	3
Initial triggering of complement	3
Eukaryotic Translation Elongation	3
Viral mRNA Translation	3
Peptide chain elongation	3
Eukaryotic Translation Termination	3
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	3
Formation of a pool of free 40S subunits	3
Influenza Viral RNA Transcription and Replication	3
SRP-dependent cotranslational protein targeting to membrane	3
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	3
Nonsense-Mediated Decay (NMD)	3
Influenza Life Cycle	3
3' -UTR-mediated translational regulation	3
L13a-mediated translational silencing of Ceruloplasmin expression	3
Interferon gamma signaling	3
GTP hydrolysis and joining of the 60S ribosomal subunit	3
Influenza Infection	3
Cap-dependent Translation Initiation	3
Eukaryotic Translation Initiation	3
Translation	3
Regulation of Complement cascade	3
Downregulated pathways (NES <0, FDR<0.25)	
The citric acid (TCA) cycle and respiratory electron transport	10
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	9

Respiratory electron transport	9
Vpu mediated degradation of CD4	8
Vif-mediated degradation of APOBEC3G	7
Antigen processing: Ubiquitination & Proteasome degradation	7
Ubiquitin-dependent degradation of Cyclin D	7
Ubiquitin-dependent degradation of Cyclin D1	7
GLI3 is processed to GLI3R by the proteasome	7
Degradation of GLI2 by the proteasome	7
SCF-beta-TrCP mediated degradation of Emi1	7
Dectin-1 mediated noncanonical NF-kB signaling	6
Cross-presentation of soluble exogenous antigens (endosomes)	6
Autodegradation of the E3 ubiquitin ligase COP1	6
Regulation of activated PAK-2p34 by proteasome mediated degradation	6
Regulation of Apoptosis	6
degradation of AXIN	6
SCF(Skp2)-mediated degradation of p27/p21	6
CDK-mediated phosphorylation and removal of Cdc6	6
Stabilization of p53	6
Autodegradation of Cdh1 by Cdh1:APC/C	6
Degradation of GLI1 by the proteasome	6
CDT1 association with the CDC6:ORC:origin complex	6
AUF1 (hnRNP D0) destabilizes mRNA	5
p53-Independent DNA Damage Response	5
p53-Independent G1/S DNA damage checkpoint	5
Transmission across Chemical Synapses	5
APC/C:Cdc20 mediated degradation of Securin	5
Hedgehog ligand biogenesis	5
Cdc20:Phospho-APC/C mediated degradation of Cyclin A	5
APC:Cdc20 mediated degradation of cell cycle proteins prior to satisfation of the cell cycle checkpoint	5
Hh mutants abrogate ligand secretion	5
Hh mutants that don't undergo autocatalytic processing are degraded by ERAD	5
Assembly of the pre-replicative complex	5
Degradation of beta-catenin by the destruction complex	5
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	4
Regulation of ornithine decarboxylase (ODC)	4
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	4
Cyclin A:Cdk2-associated events at S phase entry	4
Cyclin E associated events during G1/S transition	4
Hedgehog 'off' state	4
APC/C:Cdc20 mediated degradation of mitotic proteins	4

Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins	4
Orc1 removal from chromatin	4
degradation of DVL	4
Asymmetric localization of PCP proteins	4
Host Interactions of HIV factors	3
Neuronal System	3
Dopamine Neurotransmitter Release Cycle	3
Serotonin Neurotransmitter Release Cycle	3
Na+/Cl- dependent neurotransmitter transporters	3
Norepinephrine Neurotransmitter Release Cycle	3
HS-GAG degradation	3
Protein folding	3
Chaperonin-mediated protein folding	3
Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding	3
Regulation of APC/C activators between G1/S and early anaphase	3
APC/C-mediated degradation of cell cycle proteins	3
Regulation of mitotic cell cycle	3
Mitochondrial translation initiation	3
Switching of origins to a post-replicative state	3

Reactome pathways identified by gene set enrichment analysis of differential gene expression in Parkinson's disease studies: humans and animal models

	Number of times pathway is in top 10 pathways by abs(NES) in:		
	All studies	**Human studies**	**Substantia nigra studies**
Upregulated pathways (NES > 0, FDR<0.25)			
Eukaryotic Translation Elongation	6	3	3
Viral mRNA Translation	6	3	3
Eukaryotic Translation Termination	6	2	2
Peptide chain elongation	6	3	3
Formation of a pool of free 40S subunits	4	1	1
3' -UTR-mediated translational regulation	4	2	3
GTP hydrolysis and joining of the 60S ribosomal subunit	3	1	2
L13a-mediated translational silencing of Ceruloplasmin expression	3	1	2
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	3	2	1
Downregulated pathways (NES <0, FDR<0.25)			
The citric acid (TCA) cycle and respiratory electron transport	7	6	3
Respiratory electron transport	6	5	3
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	6	5	3
Antigen processing: Ubiquitination & Proteasome degradation	3	2	0
Phosphorylation of CD3 and TCR zeta chains	3	2	1
Vif-mediated degradation of APOBEC3G	3	3	1
Degradation of GLI2 by the proteasome	3	3	2
Hh mutants abrogate ligand secretion	3	3	2
Amyloids	3	1	0
Cholesterol biosynthesis	3	0	0

APPENDIX H: UNION OF TOP 10 GENES ACROSS ALL 33 PARKINSON'S DISEASE STUDIES

Reading across each successive row of the table, this list corresponds to the rows of the heatmap in Figure 3.7.

AMH	MAP3K6	SH2D2A	SLC7A8	WNK1	SLC6A3	CFLAR	FLNB
GLI2	ZFP36	CRK	CXCL2	CYP1B1	CH25H	ADM	NFKBIA
SGK1	ST3GAL6	ALB	CDKN1A	NFE2L2	TFCP2	MAFF	TTR
TYMS	CD74	OXT	CAPN3	FOXO3	SPP1	CX3CR1	HLA-DRB1
UGT8	AZGP1	CXCR4	ANGPT2	RCN1	AQP4	RAB3IL1	EEF1D
ITGAM	SERPINB1	MTF2	TIA1	PIM1	TINF2	GADD45B	IER3
ATF3	TULP4	SAT1	ANXA3	PRSS23	INHBA	SSTR2	FOS
EGR1	TH	NR4A2	TBL1X	UIMC1	RND2	SLC6A8	BOK
CD22	ABCA1	MAP4K4	MLLT10	VIM	LPL	CD14	C1QB
PTPRC	SPARC	RPS11	COTL1	BAG3	HSPB1	AXL	CASP4
CD4	VCAN	FMOD	PCBD1	SLC6A13	SMAD5	NFIX	SLC9A3R2
COPE	CRLF1	IGFBP2	JUP	DUSP1	JUNB	IDH1	GPX2
AGT	MID1	CLTA	GFAP	CD44	CHI3L1	CASP1	VAMP8
CEBPG	ANXA4	LGALS1	CAV1	CSRP1	KLKB1	RAB31	CP
CTSC	UBL5	GPX3	PFKFB2	PPARGC1A	NOS1	VEGFC	GRIK1
ADCYAP1	DDX6	XIAP	ZFX	TSPAN12	ARHGAP5	MAN2A1	CDH1
MAP7	SLC4A7	SDC1	SMARCA4	ACOT7	ACTN1	DPYSL4	PTK2B
MMP9	LAMB3	BCL2L1	ATP1B2	HTR4	SEC14L1	ITM2C	EPHX1
PLEC	TMEM176A	PHGDH	GUSB	SLC4A1	GABBR1	ACTC1	FAM65B
LPP	BAX	CYTH3	DRD2	SEPT9	SMPD1	HS3ST1	CA4
DUSP6	MARK2	GFRA2	KIF5C	GNAS	ZNF148	TCF4	SPOCK1
ZFPM2	AP1S2	ENC1	YWHAZ	LAPTM4B	MAPK10	GABRG2	YWHAH
SNCB	VAMP2	CIRBP	TPBG	DDC	DLK1	PAX6	FABP7
PPAT	RIOK3	GOSR1	CDKN1B	RAD23B	TPD52	EIF1AX	TFPI
HSD17B11	SLC16A1	NAMPT	ME1	CFH	OLR1	PDK4	DNAJB6
RAD21	SDHD	PKIA	SMAD1	NAP1L4	RGS2	PDPK1	RAB1A
RC3H2	SLC6A1	KCNB1	PSMB4	EIF2AK1	HAGH	FECH	BPGM
RPL15	PDE1A	MAP1B	TTC3	FUT9	INPP4A	NOV	ITGB1BP1
FEZ2	FGF9	CADPS	AIF1	GMFB	PIP4K2A	ST13	SLC11A2
TRIP12	RABEP1	RGS4	STMN2	GAP43	SNAP25	SYT1	PRKACB
RCN2	PRKCB	YWHAB	SNX10	NSF	CHGB	TPPP3	MAP2
SCG5	TAGLN3	NEFL	THY1	CCK	SNCA	SERPINI1	CMAS
GHITM	TOMM20						

APPENDIX I: UNION OF TOP 10 GENES ACROSS ALL 33 STUDIES PLUS ALZHEIMER'S DISEASE AND TUMOUR STUDIES

Reading across each successive row of the table, this list corresponds to the rows of the heatmap in Figure 3.9.

VIM	TGFBI	IGFBP3	ABCA1	ANGPT2	CTSC	CD74	GPX3
CCL4	CASP1	VAMP8	LGALS1	CAV1	C1QB	CD14	CP
NNMT	ANXA1	CD44	CTSS	OLR1	AZGP1	PTPRC	CFH
AQP4	SERPINB1	PRSS23	TMEM176A	VCAN	CASP4	AXL	RAB31
SPP1	PDGFRA	CCND1	IL1RAP	ME1	DUSP6	EGR1	FOS
HCLS1	AIF1	CX3CR1	HLA-DRB1	ITGAM	EEF1D	CXCR4	INHBA
ANXA4	BAG3	GEM	CYR61	DUSP1	JUNB	ZFP36	SGK1
NFKBIA	RC3H2	ST3GAL6	SMAD1	FHL2	ADM	MMP9	HSPB1
TBL1X	SEC14L1	RND2	SPARC	COTL1	ACTN1	HS3ST1	FMOD
GFAP	AGT	ATF3	PIM1	TINF2	IER3	GADD45B	TFPI
CEBPG	BCAT1	IL13RA2	SLC16A1	HSD17B11	PDK3	RABEP1	NAMPT
CRK	GOSR1	SLC4A7	CFLAR	FLNB	GLI2	UBL5	ARNTL
TFRC	ALB	GPX2	RPS16	LTB	GLTSCR2	KLF2	LCK
CXCL2	DPYSL4	LAMB3	BCL2L1	CLTA	KLKB1	AQP1	TYMS
LPP	IDH1	ITGB4	SMAD5	NFIX	MYLK	GUSB	GRIK1
ACTC1	COPE	CYTH3	FOXO3	MLLT10	VEGFC	RAB3IL1	CA2
SLC9A3R2	PHGDH	SLC4A1	JUP	SLC6A13	FEZ2	TRIP12	LAPTM4B
MAP2	MAP1B	CADPS	ZNF148	TCF4	ZFPM2	JAK1	AP1S2
ENC1	YWHAH	TTC3	RCN2	YWHAB	YWHAZ	SMARCA4	SDC1
Sep-09	CIRBP	DDX6	RPL18	RPS11	UIMC1	PTEN	RAD21
PKIA	PDPK1	RAB1A	PPAT	RIOK3	EIF2AK1	BPGM	FECH
RAD23B	COPS8	PITPNA	PSMB4	PSMB7	PFKFB2	CDH1	NOS1
BCAN	PCBD1	BAX	RXRA	CD4	EPHX1	ATP1B2	IKZF2
SERPINB9	RARA	SLC6A8	MARK2	HTR4	GFRA2	CSRP1	CRLF1
PLEC	SMPD1	PDK4	DNAJB6	ACSL1	SPOCK1	NOV	GJA1
RGS2	GAD1	SNX10	THY1	PPP1R2	TPBG	DLK1	DDC
COL1A2	CA12	LPL	RCN1	MMP15	PRC1	IGFBP2	ELAVL1
MID1	FABP7	MTF2	GAP43	CYP1B1	CH25H	CDKN1A	FKBP5
IL1R2	NFE2L2	TFCP2	MAFF	CHI3L1	MAP3K6	AMH	WNK1
SH2D2A	SLC7A8	SLC6A3	TTR	STMN2	NEFL	SYT1	KIF5C
PVALB	SNAP25	PRKCB	CHGB	FGF9	GABRA1	SLC12A5	TIA1
PAX6	GNAS	HIVEP3	NNAT	SSTR2	TULP4	ANXA3	MAP4K4
SH3GL3	OXT	CAPN3	UGT8	GMFB	CTTN	MAN2A1	SLC11A2
PRKACB	TPD52	BOK	CD22	MAP7	CDKN1B	ARHGAP5	PIP4K2A
TSPAN12	ZFX	IVNS1ABP	NAP1L4	SDHD	XIAP	ST13	KCNB1
SLC6A1	FUT9	ITGB1BP1	PRNP	EIF1AX	SUMO1	EPS15	MAPK10
GHITM	CMAS	TH	NR4A2	CCK	RGS4	PDE1A	SNCA
GABRG2	SERPINI1	TPPP3	SCG5	TAGLN3	PTK2B	SYT5	DRD2
FAM65B	PPARGC1A	HAGH	GABBR1	INPP4A	CA4	ADCYAP1	ACOT7
NSF	SNCB	VAMP2					

APPENDIX J COMPARISON OF RESULTS USING HIPPIE AND OMNIPATH

HIPPIE, which is based on protein-protein interactions rather than the signalling interactions contained in OmniPath, is a slightly larger network, recording interactions for a median of 51% of measured genes compared to OmniPath's 30%. This means that the available network size (the size of the 'base' network) is also larger, at a median of 2,413 nodes compared to 1,408 nodes for OmniPath. Surprisingly, this does not result in much larger path-sets: HIPPIE path-sets contain a median of 90 nodes, compared to 82 for OmniPath.

This may be linked to the degree of nodes in path-sets: the degree distribution of the two graphs is similar overall, but interestingly, nodes in OmniPath path-sets have a higher average degree, at a median of 15.5 compared to the median degree of 10 for nodes in HIPPIE path-sets. Overall, the distribution of path-set network component size is similar in both networks: the median of both is 3 (path-sets contain many small components and only a few larger ones), with a maximum component size of 306 for OmniPath and 355 for HIPPIE path-sets. However, the mean component size is 11.6 for OmniPath and 7.0 for HIPPIE, suggesting that the major connected components in HIPPIE paths tend to be slightly smaller. These results are summarised in Table J.1.

Table J.1 Comparison of network and path-set topology in OmniPath and HIPPIE

	OmniPath	**HIPPIE**
Number of unique proteins	6,972	12,162
Number of usable interactions	43,963	62,615
Percentage of measured genes included in the network	30%	51%
Median available network size	1,408	2,413
Median path-set size	82	90
Median degree of nodes in path-sets	15.5	10
Mean network component size of path-sets	11.6	7.0
R^2 of relationship between number of path-sets and degree of a gene	0.47	0.45
R^2 of relationship between number of random-sets and degree of a gene	0.68	0.75

In general, the genes included in the path-sets are quite different between the two base networks. Figure J.1 shows an example of this, comparing the dysregulated path-set for asthma in OmniPath (as shown in Figure 4.3 of the main text) to that in HIPPIE. The difference between the two seems to be due to the low overlap between the signalling interactions included in OmniPath and the protein-protein interactions included in HIPPIE – of the 43,963 interactions in OmniPath and the 62,615 interactions in HIPPIE, only 8,966 are common to both (a Jaccard overlap of 0.09). The median Jaccard overlap of edges in both networks is correspondingly low at 0.05. Nevertheless, the overall properties (in terms of ability to capture KDGs and shared edges between diseases) are similar, as shown in Tables J.2, J.3, and J.4.

Table J.2 Known disease-associated genes and drug-interacting genes (KDGs) in path-sets in OmniPath and HIPPIE

Results in HIPPIE are overall similar to those in OmniPath, albeit slightly lower. One notable difference is that the random-path-sets perform comparably worse in HIPPIE. The relatively good performance of the random-path-sets in OmniPath was explored in Section 4.3.2 in relation to the tendency of path-set analysis to select genes of higher degree. Like in OmniPath, nodes in the HIPPIE random path-sets tend to have slightly higher degree than nodes in the real path-sets (median degree of 11.75 in the random path-sets compared to 10 in the real path-sets). However, unlike OmniPath, drug targets in HIPPIE do not have very significantly higher median degree than non-drug-targets (median degree of 4 for both, Wilcoxon p-value 0.012). This could explain the comparatively worse performance of the random-path-sets in HIPPIE.

		Path-set	**LFC-set**	**Random-path-set**	**Random-gene-set**
What proportion of sets contained at least one KDG?	OmniPath:	0.58	0.56	0.54	0.43
	HIPPIE:	0.55	0.52	0.46	0.38
How many KDGs were found per set on average?	OmniPath:	2.95	1.89	2.40	1.18
	HIPPIE:	2.06	1.70	1.50	0.85
What percentage of genes in the set were KDGs?	OmniPath:	3.2%	2.4%	2.5%	1.3%
	HIPPIE:	2.2%	1.9%	1.5%	0.9%

Figure J.1 Comparison of asthma path-sets in OmniPath (top) and HIPPIE (bottom)
The path-set based on HIPPIE is slightly larger (79 nodes) than the path-set based on OmniPath (58 nodes), but contains more short paths connecting only two genes. The HIPPIE path-set does not show the same hub-based structure as the OmniPath path-set – the key hub node MAPK1 is included in the path-set, but connects to only two other genes. Both sets include the asthma-associated gene SMAD3, but pick up different drug-interacting genes.

HIPPIE path-sets also yield similar numbers of disease pairs which share a significant number of edges, with 647 significant disease pairs of a median shared-edge-set size of 20 nodes, compared to 676 disease pairs in OmniPath which have a median shared-edge-set size of 21 nodes. Of these, 261 pairs are common to both. The pairs are overall similar in terms of their biological relevance (Table J.3) and the disease-associated genes and drug-targets contained in their shared edges (Table J.4).

Table J.3 Biological relevance of shared edges

Significant disease pairs in OmniPath and HIPPIE are similar in terms of their likelihood to be in the same Disease Ontology category or to share drugs, but disease pairs in HIPPIE are comparatively less likely to share disease-associated genes.

		Disease pairs with shared paths	**Disease pairs without shared paths**	**Fisher test *p*-value**
In the same Disease Ontology top-level class	OmniPath:	44.7%	26.3%	$<2.20e^{-16}$
	HIPPIE:	45.0%	26.3%	$<2.20e^{-16}$
In the same Disease Ontology sub-class	OmniPath:	18.1%	7.2%	$<2.20e^{-16}$
	HIPPIE:	17.8%	7.3%	$<2.20e^{-16}$
Share drugs (in Phase III clinical trials or approved)	OmniPath:	17.2%	8.0%	$1.33e^{-13}$
	HIPPIE:	14.5%	8.2%	$2.03e^{-7}$
Share disease-associated genes	OmniPath:	19.2%	10.8%	$5.27e^{-10}$
	HIPPIE:	14.4%	11.2%	0.015

Table J.4 Disease-associated and drug-interacting genes in shared edges

A slightly smaller percentage of disease pairs significant in HIPPIE include a disease-associated or drug-interacting gene than pairs significant in OmniPath; however, the drug-interacting genes tend to be slightly more likely to be relevant to both diseases.

		All significant disease pairs	**Top 100 most significant disease pairs**
Percentage of significant disease pairs which include a disease-associated gene for either disease in their shared-edge-set	OmniPath:	41.7%	62.2%
	HIPPIE:	35.2%	45%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	OmniPath:	6.4%	7.4%
	HIPPIE:	6.5%	6.4%
Percentage of significant disease pairs which include a drug-interacting gene for either disease in their shared-edge-set	OmniPath:	51.9%	74.5%
	HIPPIE:	47.8%	60%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	OmniPath:	9.8%	15.9%
	HIPPIE:	12.2%	16.9%

APPENDIX K: PATHWAY ENRICHMENT RESULTS FOR GENES IN MULTIPLE PATH-SETS

List of 54 genes in the dysregulated path-sets of 25 or more diseases

Gene	# Path-sets	Gene	# Path-sets	Gene	# Path-sets
EGFR	66	**AKT1**	30	**EGR1**	27
MAPK1	56	**EP300**	30	**ITGB4**	27
STAT1	55	**JUN**	30	**LYN**	27
STAT3	49	**PIK3R2**	30	**MYC**	27
SRC	48	**PXN**	30	**PLAUR**	27
CD44	46	**RAC1**	30	**SYK**	27
CTNNB1	37	**RELA**	30	**CDKN1A**	26
PRKCA	37	**CCND1**	29	**CRK**	26
AR	36	**CREBBP**	29	**FAS**	26
ESR1	36	**FOS**	29	**IGF1R**	26
MAPK14	36	**PRKCD**	29	**SMAD4**	26
PTPN11	36	**BCL2**	28	**CALM1**	25
SMAD3	36	**CASP8**	28	**CBL**	25
FYN	35	**FGFR1**	28	**CTNND1**	25
TP53	33	**IRS1**	28	**JUP**	25
RAF1	32	**PLCG1**	28	**MDM2**	25
MAPK8	31	**SMAD2**	28	**PAK1**	25
SNCA	31	**CDK2**	27	**VCAN**	25

Pathway enrichment results for these 54 genes

The 8 terms in non-italic text are also enriched amongst genes in multiple random-path-sets.

	Background (all genes in network)	**Genes in multiple path-sets**			
GO-Slim Biological Process	**#**	**#**	**Expected**	**Fold Enrichment**	**FDR**
negative regulation of apoptotic process	55	6	1.34	4.49	4.44E-02
regulation of biological process	458	29	11.12	2.61	2.25E-05
biological regulation	556	31	13.5	2.3	7.83E-05
regulation of transcription from RNA polymerase II promoter	123	11	2.99	3.68	4.94E-03
metabolic process	964	37	23.41	1.58	1.22E-02
transcription from RNA polymerase II promoter	151	11	3.67	3	1.97E-02
transcription, DNA-dependent	171	11	4.15	2.65	4.42E-02
intracellular signal transduction	317	21	7.7	2.73	5.56E-04
signal transduction	535	32	12.99	2.46	2.22E-05
cell communication	590	32	14.33	2.23	9.43E-05
cellular process	1348	48	32.74	1.47	1.48E-03
cell surface receptor signaling pathway	242	16	5.88	2.72	4.93E-03
phosphate-containing compound metabolic process	336	18	8.16	2.21	1.84E-02
response to stimulus	539	27	13.09	2.06	3.08E-03
Unclassified	674	4	16.37	0.24	3.99E-03

APPENDIX L SHARED EDGES RESULTS AT A THRESHOLD OF TOP 100 PATHS

Using path-sets based on the stricter threshold of the 100 top paths (median path-set size of 72) to identify shared edges between disease pairs, instead of the more lenient 500 top paths threshold (median path-set size of 195), results in fewer significant disease pairs (287 vs 676) and a smaller median shared-edge-set size (9 as opposed to 21 nodes), as may be anticipated. This makes little difference to the percentage of disease pairs which are biologically relevant in terms of ontological class or disease/drug sharing (Table L.1), but the percentage of shared edges which include a disease-associated or drug-interacting gene at a threshold of 100 top paths are lower than those in the main text (Table L.2), as the smaller edge-sets mean less chance to capture a relevant gene. Interestingly, those genes that are found are also less likely to be associated with both diseases, suggesting that the smaller path-sets are less useful for capturing potential drug repurposing hypotheses.

Table L.1 Biological relevance of shared edges

		Disease pairs with shared paths	**Disease pairs without shared paths**	**Fisher test *p*-value**
In the same Disease Ontology top-level class	500 threshold:	44.7%	26.3%	$<2.20e^{-16}$
	100 threshold:	49.1%	26.9%	$3.55e^{-15}$
In the same Disease Ontology sub-class	500 threshold:	18.1%	7.2%	$<2.20e^{-16}$
	100 threshold:	18.1%	7.7%	$1.33e^{-8}$
Share drugs (in Phase III clinical trials or approved)	500 threshold:	17.2%	8.0%	$1.33e^{-13}$
	100 threshold:	20.2%	8.2%	$5.20e^{-10}$
Share disease-associated genes	500 threshold:	19.2%	10.8%	$5.27e^{-10}$
	100 threshold:	18.5%	11.2%	$3.06e^{-4}$

Table L.2 Disease-associated and drug-interacting genes in shared edges

		All significant disease pairs	**Top 100 most significant disease pairs**
Percentage of significant disease pairs which include a disease-associated gene for either disease in their shared-edge-set	500 threshold:	41.7%	62.2%
	100 threshold:	27.5%	30.8%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	500 threshold:	6.4%	7.4%
	100 threshold:	0.9%	1.8%
Percentage of significant disease pairs which include a drug-interacting gene for either disease in their shared-edge-set	500 threshold:	51.9%	74.5%
	100 threshold:	28.2%	47.7%
Percentage of these genes associated with both diseases (excluding same-disease pairs)	500 threshold:	9.8%	15.9%
	100 threshold:	7.5%	11.1%

APPENDIX M: RESULTS AT DIFFERENT FEATURE SET SIZES

To test the dependence of the results on the chosen feature set size, different feature set sizes of 20, 50, and 200 were tested were tested for ontological, literature co-occurrence, genetic, and transcriptomic feature spaces (phenotypic and drug feature spaces being of fixed size).

As expected, the proportion of links that were significantly greater than those observed in random maps (and therefore the number of links in each disease map) varied depending on the feature set size, with larger feature set sizes resulting in greater differentiation from random maps (Table M.1). Related to this, the mean Jaccard overlap of drugs shared by diseases linked in the map increases at smaller feature set sizes, where only the highest similarity links pass the random significance threshold and are included in the map (Table M.1).

Evaluating the full similarity matrices (in terms of the proportion of links which share drugs, and the ability to predict DO categories) produced similar results at different feature set sizes, with some minor variations in the performance of individual feature spaces (Figures M.1 and M.2). Finally, at all feature set sizes, literature co-occurrence, phenotype, and ontological spaces were the most highly correlated to the fused space.

Table M.1 Comparison of disease maps at different feature set sizes

	Feature set size			
	20	50	100	200
Percentage of links in full similarity matrix which are significant (included in map)	2.98	4.76	6.91	9.12
Percentage of links in disease map classed as novel	16.2	15.6	15.3	16.3
Mean Jaccard overlap of drugs shared by diseases linked in the disease map, approved and Phase III/approved only	0.099/0.107	0.082/0.081	0.069/0.069	0.061/0.059
Mean Jaccard overlap of drugs shared by novel links in the disease map, approved and Phase III/approved only	0.031/0.062	0.029/0.049	0.025/0.04	0.022/0.034

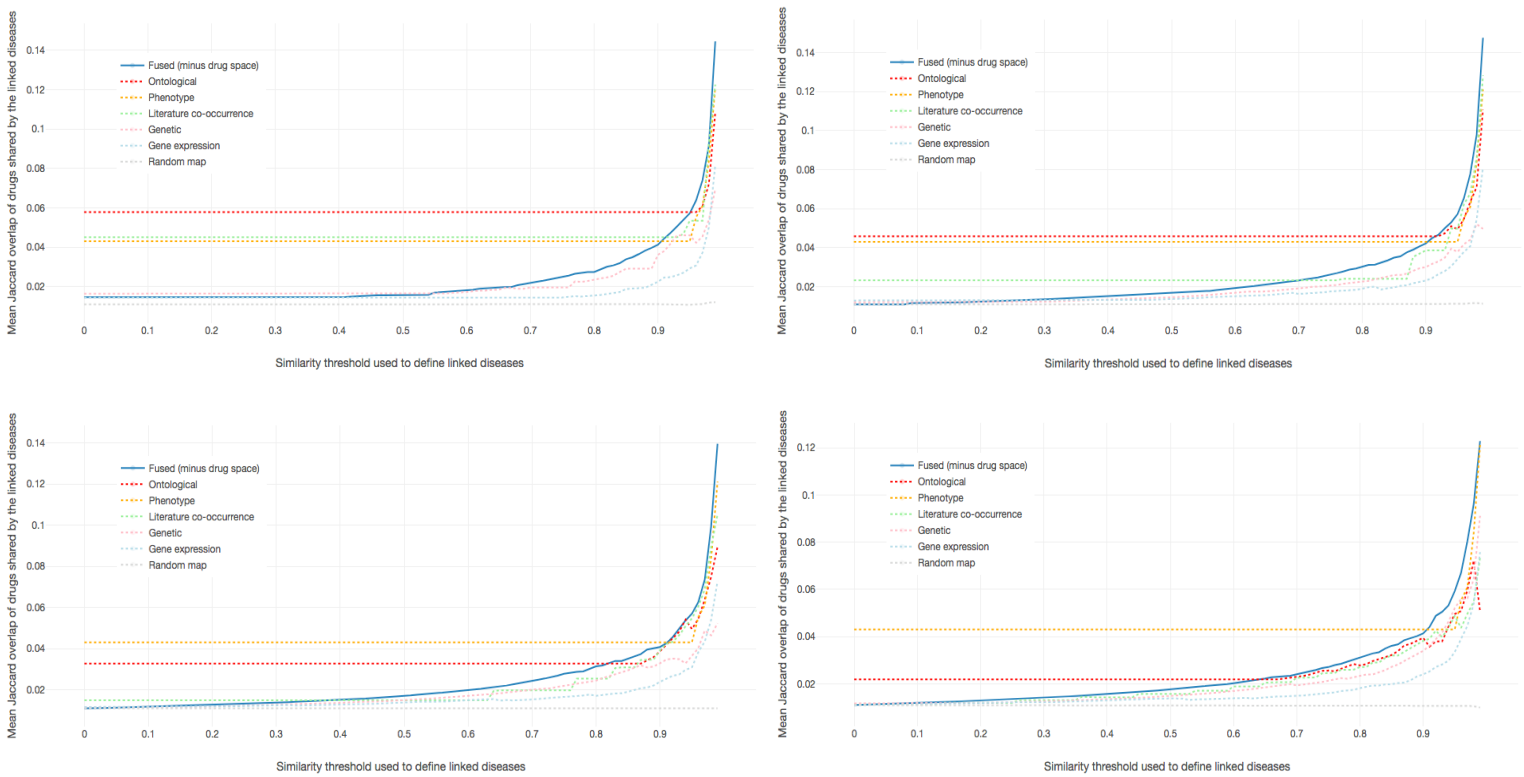


Figure M.1 Comparing the drug overlap (approved or in Phase III clinical trials) of links in disease networks at different thresholds of similarity, and different feature vector sizes.

Similar overall results are obtained for feature vector sizes of 20 (top left), 50 (top right), 100 (bottom left), and 200 (bottom right), with some minor variation in the performance of individual spaces. In particular, at a feature vector size of 20, literature space and ontological space marginally outperform the fused space at certain thresholds. The effect of different feature vector sizes on sparsity can be observed here, particularly for the ontological space, which shows higher sparsity at smaller feature set sizes. This indicates that smaller feature sets are insufficient to capture much overlap between diseases in this space (compared to the size of the feature universe).

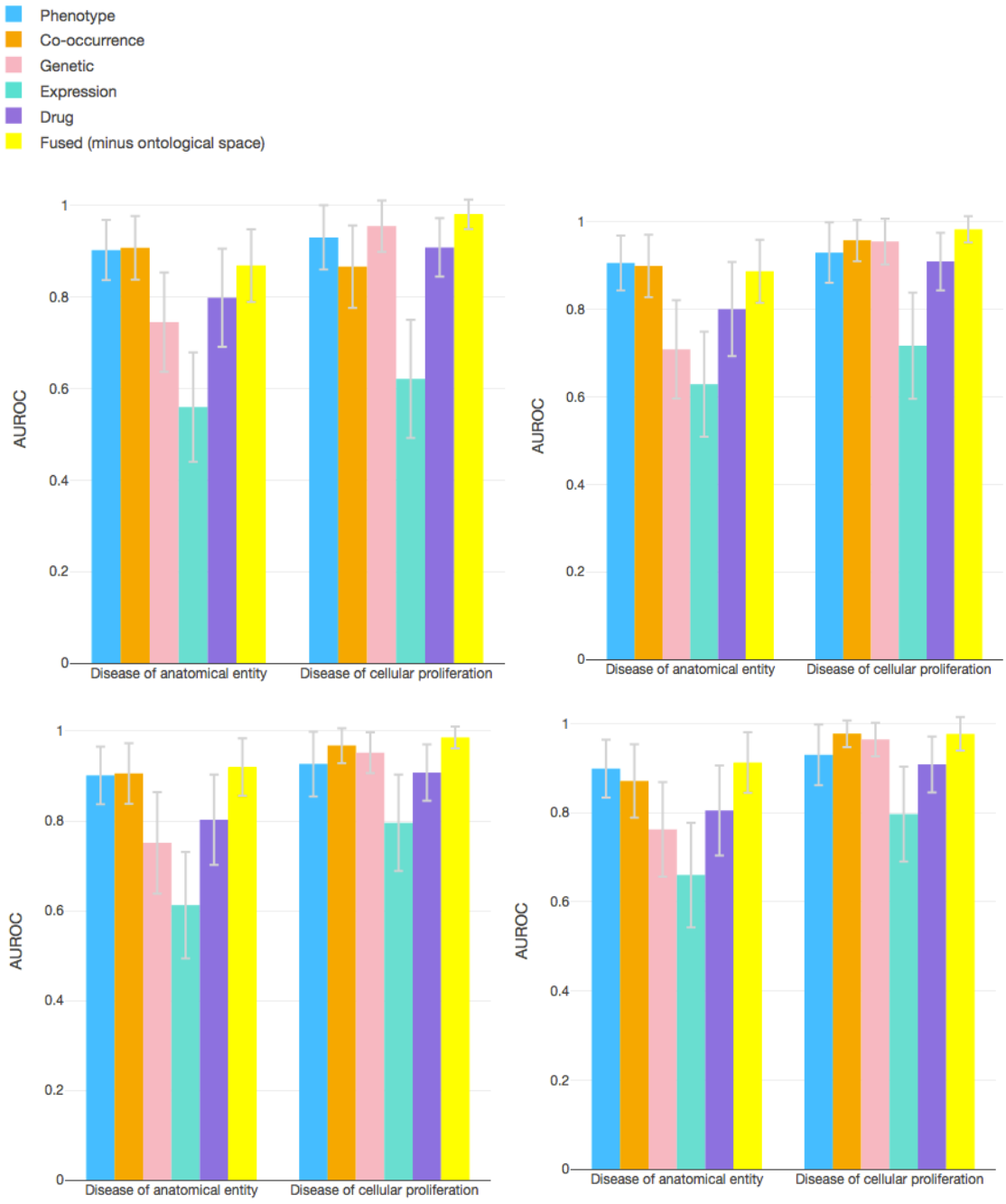


Figure M.2 Random Forest AUROC scores for individual and fused disease similarities at different feature vector sizes.

Overall, results are similar at feature vector sizes of 20 (top left), 50 (top right), 100 (bottom left), and 200 (bottom right), and improve with increases in feature vector size. For the class disease of anatomical entity, *AUROC scores for the fused disease map (minus the ontological space) improve from 0.869 for length 20, to 0.913 for length 200, but literature co-occurrence and phenotypic spaces outperform the fused space at feature set sizes 20 and 50. For the class* disease of cellular proliferation, *the fused map outperforms any of the individual maps at feature set sizes of 20, 50, and 100; at a feature set size of 200 it performs equally to literature co-occurrence (with AUROC scores of 0.9769 and 0.9768 respectively.*

APPENDIX N RESULTS OF WEIGHTED MAP

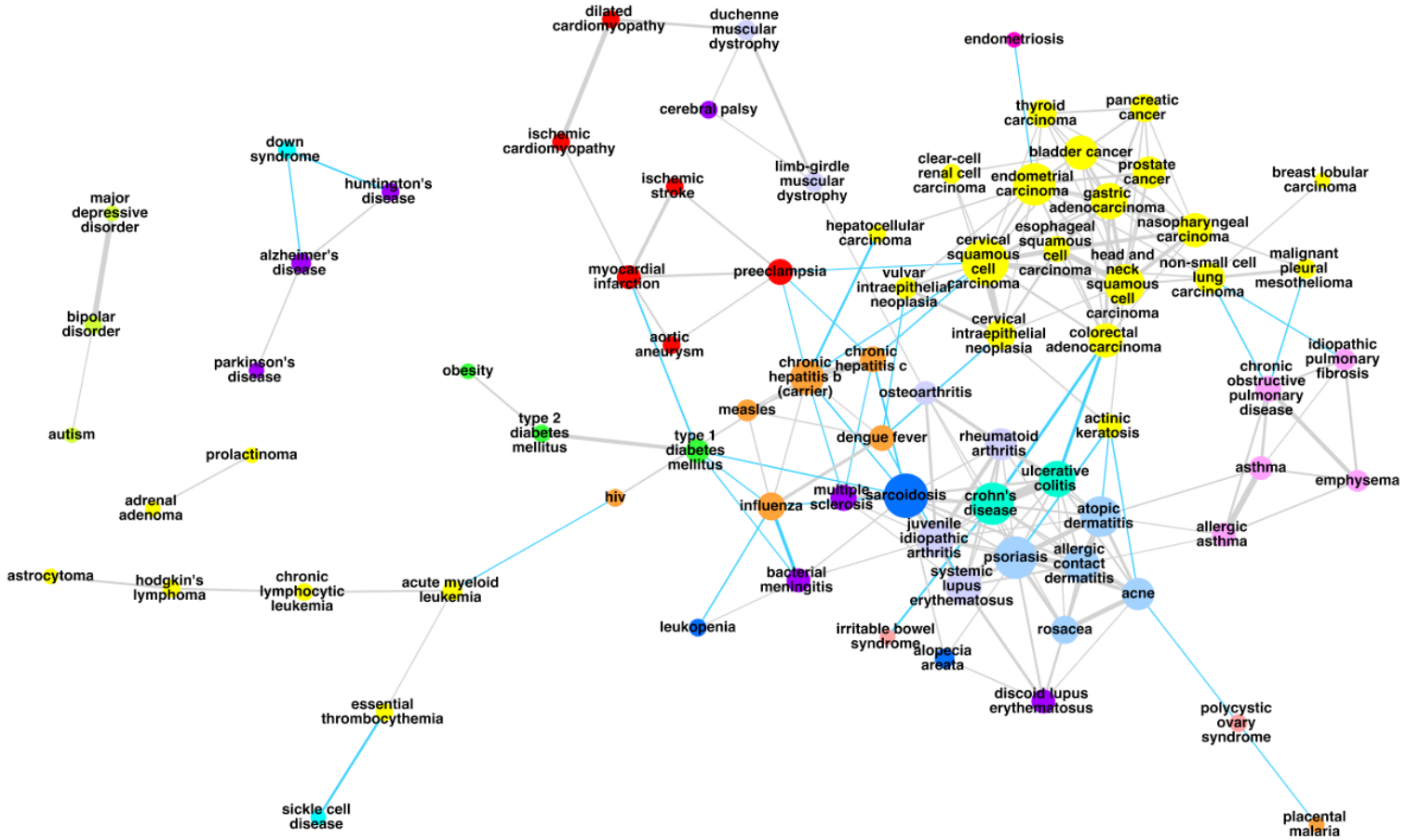


Figure N.1 Disease map resulting from weighted similarity fusion

The map shown here is constructed from a weighted mean of the individual spaces, so that the 'traditional' spaces (ontological, phenotypic, and literature co-occurrence similarity) together make up only a third (instead of a half) of the fused similarities (weighting scheme 1,1,1,2,2,2). The resulting disease map looks similar to the balanced (unweighted) disease map shown in Figure 5.2.

Table N.1 Comparison of balanced and weighted fused similarities

	Balanced	**Weighted**
Percentage of links in full similarity matrix classed as significant (and therefore included in the resulting disease map)	6.91 (242 links)	6.02 (211 links)
Percentage of links in disease map classed as novel	15.3 (37 links)	17.1 (36 links)
Mean Jaccard overlap of drugs shared by diseases linked in the disease map, approved and Phase III/approved only	0.069/0.069	0.075/0.077
Mean Jaccard overlap of drugs shared by novel links in the disease map, approved and Phase III/approved only	0.025/0.04	0.021/0.038
AUC of DO class prediction, disease of anatomical entity/disease of cellular proliferation	0.924/0.985	0.891/0.979

This table shows the results of adjusting the weights on the similarities so that the three 'traditional' spaces which show high similarity to each other (ontological, phenotype, and literature-based spaces) account for only a third (instead of half) of the fused similarities, i.e. the contribution of the other three spaces (genetic, expression, and transcriptomic) is doubled (weighting scheme 1, 1, 1, 2, 2, 2). Down-weighting these highly similar 'traditional' spaces means that they have less influence on the resulting disease map, however, the results shown here indicate that this down-weighting makes little difference to properties of the resulting disease map. For the DO class prediction, ontological space is excluded from the fused matrix, meaning that phenotypic and literature spaces together account for 25% of the fused similarities (weighting scheme 1, 1, 2, 2, 2). After this weighting, note that the fused space is no longer the best-performing space for the prediction of *disease of anatomical entity* (being slightly outperformed by phenotypic and literature co-occurrence similarities at 0.902 and 0.905 respectively). Note that the AUC quoted here for the balanced kernel varies slightly from that quoted in Section 5.3.5 as the classifier was re-run.

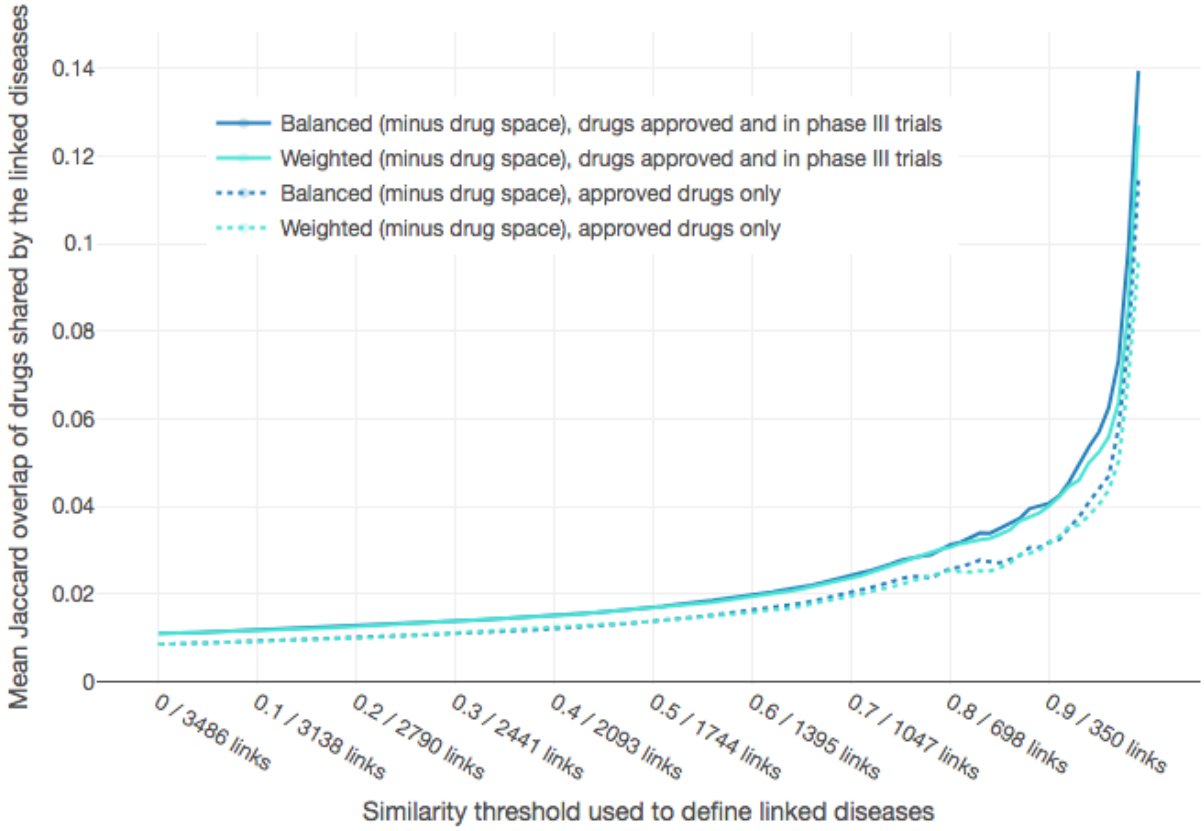


Figure N.2 Comparing the drug overlap of linked diseases in weighted and balanced spaces

The overlap of drugs for each link (mean Jaccard score) is very similar for balanced and weighted spaces. In this case, drug space is excluded from the evaluation, meaning that the contribution of the two remaining non-traditional spaces is tripled so that the traditional spaces contribute a third of the similarity (weighting scheme 1, 1, 1, 3, 3). Note however that in contrast to the balanced space, the weighted fused (minus drug) space is outperformed by literature space at the top 5% of similarities.