USING PATHWAY NETWORKS TO MODEL CONTEXT DEPENDENT CELLULAR FUNCTION

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE FACULTY OF SCIENCE AND ENGINEERING

2017

By

Ruth Alexandra Stoney

School of Computer Science

Contents

Word count: 47,746

List of Figures

List of Tables

List of Abbreviations

CPDB	ConsensusPathDB
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic acid

GC	Gene Cover
GI	Genetic Interaction
GO	Gene Ontology
GWAS	Genome-wide association study
HGNC	HUGO Gene Nomenclature Committee
HPD	Human Pathway Database
HPO	Human Phenotype Ontology
IEA	Inferred from Electronic Annotation
KEGG	Kyoto Encyclopedia of Genes and Genomes
Max_O	Maximum Overlap
MDS	Minimum Dominating Set
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
ORA	Over Representation Analysis
PDB	Protein Data Bank
PID	Pathway Interaction Database
PPI	Protein-Protein Interaction
ReCiPa	Redundancy Control in Pathway Databases
RNA	Ribonucleic acid
SNP	Single Nucleotide Polymorphism
SD	Standard devation

Glossary

Biological function – in the context of this thesis, functions are processes within the cell that the organism must fulfil to survive and show a normal phenotype, defined by the Gene Ontology. These intra-cellular processes are also referred to as **cellular functions**, **cellular processes** and **biological processes**.

Pathways – Sets of molecules than interact under a specific cellular context, here defined by ConsensusPathDB.

Molecular network – network comprised of nodes representing biological molecules such as proteins and genes. Edges represent interactions between individual molecules or genes.

Pathway network – network comprised of biological pathways. Edges represent functional similarity between pathways.

Functional module – sub-network cluster of nodes associated with a particular biological function

Functional pathway module – sub-network cluster of nodes (biological pathways) associated with a particular biological function

Disease module - network clusters of nodes associated with similar diseases

Disease pathway modules - network clusters of nodes (biological pathways) with associated shared disease

Dynamic functional network – networks of molecular interactions that incorporates information about the dynamic nature of the cell by mapping active modules, or generating sub-graphs active under particular conditions

Abstract

Using pathway data to model context dependent function
Ruth Alexandra Stoney
University of Manchester, Doctor of Philosophy, September 2017

Molecular networks are commonly used to explore cellular organisation and disease mechanisms. Function is studied using molecular interaction networks, such as protein-protein networks. Although much biological insight has been gained using these models of molecular function, they are hindered by their reliance on available experimental data and an inability to capture the complexity of biological processes.

Functional modules can be identified based on molecular network topology, making it essential that the edges accurately depict molecular interactions. However, these networks struggle to depict the temporal nature of interactions, giving the impression that all interactions are constant. This misrepresentation can result in functionally heterogeneous clusters. The notoriously inaccurate nature of experimental protein interaction data, along with variable conformity among network clusters and functional modules further impedes functional module extraction. Representation of genes by single nodes artificially merges the functions of pleiotropic genes, distorting the arrangement of function within molecular networks. This thesis therefore explores a more suitable model for representing function.

Pathways are composed of sets of proteins that are known to interact within a particular cellular context, corresponding to a discernible biological function. Their representation of context dependent cellular activity makes them ideal for use as nodes within a new pathway level model. Using combinatorial algorithms a reduced redundancy pathway set was produced to represent global cellular systems. Enrichment analysis provides reliable functional annotations for each pathway node, attributing independent functions to pleiotropic genes. Edges are based on functional semantic similarity, generating a network representation of functional organisation.

Both yeast and human biological systems are presented as functionally connected pathway networks. Pathway annotation and experimentation with semantic similarity measures provides insight into the cross-talk between biological processes. Pathway functional modules elucidate the intracellular implementation of processes. Disease modules highlight the effects of functional perturbations and disease mechanisms. The pathway model provides a complementary, high-level functional model that begins to bridge the gap between molecular data and phenotype. The utilisation of pathway data provides a large, well-validated data source, avoiding the inaccuracies inherent with molecular data. Pathway models better represent components of biological complexity such as pleiotropy and linear implementation of functions.

Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification of The University of Manchester or any other university or institute of learning.

Copyright statement

Acknowledgements

Firstly, I would like to thank my supervisors Prof Goran Nenadic, Dr Jean-Marc Schwartz and Prof David Robertson for their continued advice and support throughout my PhD. Being able to discuss my work from such different perspectives has been enormously helpful and contributed greatly to my experience. I would particularly like to thank Jean-Marc Schwartz for guiding me through my first international conference, as well as organising my collaboration with Prof José Nacher at Toho University. Also, a special thanks to José Nacher for his interesting and unique insight into the project. Further, I would like to thank Nikola Milosevic and Maksim Belovusov for continually helping me with maths and computer queries throughout my project. Thanks to Ryan Ames, who contributed and provided valuable guidance at the beginning of my PhD, Oyebode Oyeyemi, whose fascinating data set allowed me to explore the versatility of the methods I was developing, and Jamie Soul, who helped me validate the later stages of my work.

Additionally, I would also like to thank my parents for their endless encouragement and support. I would like to thank Thomas Ziesler for proof reading all of my work and generally being unreasonably supportive, and Catriona McGinn for ensuring that throughout all the work I always had fun.

Chapter 1

Introduction

Studying the intracellular organisation of function has enhanced our understanding of biology, offering insight into the relationships between molecular interactions that make up complex phenotypes. We show that molecular interaction models have facilitated great progress in the generation of functional maps, but are restricted both by limitations within the interaction data and the ability of molecular networks to represent biological complexity. We therefore propose a new model, using pathways as entities, to overcome these barriers and generate more representative biological networks.

1.1 Function

In order to fulfil their biological roles, cells must perform a multitude of highly interdependent functions. Some processes are continuous, such as glucose metabolism; others are cyclic, such as cell cycle division; and others are intermittent, such as responses to stimuli. Biological function has a hierarchical organisation, starting with very general functions involving large numbers of genes, which are sub-divided into increasingly specific processes. This hierarchical organisation is observed in the physical organisation of function in the cell (Barabási & Oltvai 2004) and reflected in conceptual formalisations such as the Gene Ontology (GO) (Ashburner et al. 2000).

1.2 Systems biology approaches to modelling function using molecular networks

Systems biology is a holistic approach, which aims to understand the complexity of biological systems. It is based on the understanding that biochemical entities, such as proteins, do not work alone, but instead collaborate to perform synchronised cellular functions. Graphs are often used to represent the interactions and relationships between biological molecules, allowing entire cellular systems to be depicted in a single model. The use of molecular networks to decipher the organisation of cellular function has been widely adopted (Barabási & Oltvai 2004; Spirin & Mirny 2003; Sharan et al. 2007; Ravasz et al. 2002), with functional maps being generated for many organisms including *Saccharomyces cerevisiae* (Janjić et al. 2014; Spirin & Mirny 2003; Zhu et al. 2008), *Mus musculus* (Guan et al. 2008), and *Homo sapiens* (Wu et al. 2010; Huttlin et al. 2015). These networks can incorporate interactions between many types of biological entities including proteins, DNA and metabolites. Protein interactions are traditionally viewed as forming the fundamental infrastructure of cellular function; correspondingly the data for protein-protein interactions is extensive and widely used. Functional networks based on protein-protein interactions have improved our understanding of disease (Barabasi et al. 2011), infection (Vidal et al. 2011) and drug pharmacodynamics (Suthram et al. 2010).

Traditional network representations ignore the temporal context-dependent nature of cellular interactions. The shortcomings of this approach were addressed by the development of dynamic networks, which show considerable improvements in their representation of cellular function. However, the incorporation of biological complexity into molecular models is hindered by problems with data accuracy and completeness (Hart et al. 2006; Sprinzak et al. 2003), as well as an inability to represent context dependent interactions (Ideker & Krogan 2012). An additional disadvantage to the study of molecular networks is their inability to represent pleiotropic gene function. By representing genes using single nodes the multiple functions of pleiotropic genes cannot be independently portrayed. The aim of this thesis is to explore an alternative approach to model function, based on biological pathways.

1.3 Using pathways to model function

Biological pathways are sets of interacting molecules, whose delineation and scope are used to model biomolecular relationships within a cellular context (Petri, Jayaraman, et al. 2014). Pathways are active under particular cellular conditions and contribute towards discernable cellular functions. The use of pathways rather than molecules as biological units overcomes two major problems of molecular networks. The first issue faced by molecular networks is the effect that inaccuracy within interaction data has on network topology, affecting all network analysis. Pathways reliably represent sets of interacting proteins, which are robust against inaccuracies within individual molecular interactions, avoiding this issue. The second problem limiting molecular networks is the representation of context dependent interactions. Within molecular interaction networks all interactions appear constant and static, therefore visualisation of their dynamic nature requires the generation of multiple sub-networks. Pathway data naturally incorporates cellular context, since interactions that occur in different contexts are partitioned into separate pathways. Molecular networks also tend to represent genes with single nodes; therefore the multiple functions of pleiotropic genes cannot be easily separated, whereas within the pathway network proteins with pleiotropic functions are present in multiple pathways, allowing their function to be judged independently in each instance. Finally, the pathway network also provides a method for studying function at a higher level, which has the potential to be developed into a multilayer approach to function.

1.4 Hypothesis and research questions

The hypothesis of this project is to demonstrate that pathway data can be used to create an alternative model of biological function that overcomes the existing issues of current molecular functional networks.

The thesis specifically addresses the following research questions:

1. How should the pathway network be constructed? The pathway network is a novel model therefore the best method of preparing, functionally annotating and linking the pathways into a network will have to be established.

1.1 Can a set of pathways appropriately represent each biological system? The pathway set will have to provide maximised coverage of the organism's gene set, with minimal redundancy. Heavily overlapping pathways are likely to complicate the interpretation of functional modules. The size of pathways should also be controlled, since excessively large pathways lack functional specificity.

1.2 What is the most appropriate metric to assess functional relatedness for the pathway edges? Many methods have been developed for calculating the similarities of GO term sets between pairs of genes, however less is known about linking functionally annotated pathways. Existing methods will be assessed on their suitability to form edges within the pathway network.

2. Is the organisation of function portrayed by the pathway network biologically meaningful? Can external data sources be used to support the validity of the network?

3. How do functional boundaries relate to pathway boundaries? What advantages do pathway models have in representing function compared to molecular models? Does the literature support the portrayal of function within and between pathways?

4. Are pathways suited to capture biological complexity such as context dependent and temporal gene function? Can the presence of pleiotropic genes be demonstrated within the network?

5. Does the pathway network provide novel insights into disease mechanisms? Can disease pathways be linked into connected modules? Does the distribution of disease pathways within the network reflect properties of the disease pathology or phenotype?

1.5 Contributions

The main contribution of this thesis is an alternative functional model based on pathways, capable of overcoming the issues faced by molecule-centric functional networks (see Section 2.3). The presented pathway network offers a more representative portrayal of function within the cell. The representation of gene pleiotropy is discussed and novel insight to functional cross talk is presented (see Section 4.4.8).

Disease pathway modules are identified within the network, indicating the functional perturbations that give rise to disease phenotypes (see Section 6.4.5). The utility of the network is demonstrated using a cancer case study, in which the distribution of nodes related to different types of cancer is used to make inferences regarding the characteristics of each cancer type. All data regarding the human network and disease modules is available at https://data.mendeley.com/datasets/3pbwkxjxg9/1.

Additional contributions include the development of new set theory methods for reducing redundancy in pathway and GO enrichment data (see Section Chapter 5). The methods developed are available for use at https://github.com/RuthStoney/set-cover-and-set-packing-to-reduce-redundancy-in-pathway-data. The thesis also contains analysis of semantic similarity measures and an evaluation of their suitability for pathway data.

1.6 Thesis outline

To overcome the difficulties faced by molecular functional networks, a model using biological pathways rather than molecular entities as network nodes is proposed. The network edges are based on shared functionality, revealing the relationships between pathways and creating a biologically intuitive network. We first developed the project in yeast, since it presents a small, well-studied biological system, before progressing onto human data. Figure 1 shows the progression of network development through the three stages of the project. The following section provides a summary of each chapter of the thesis.

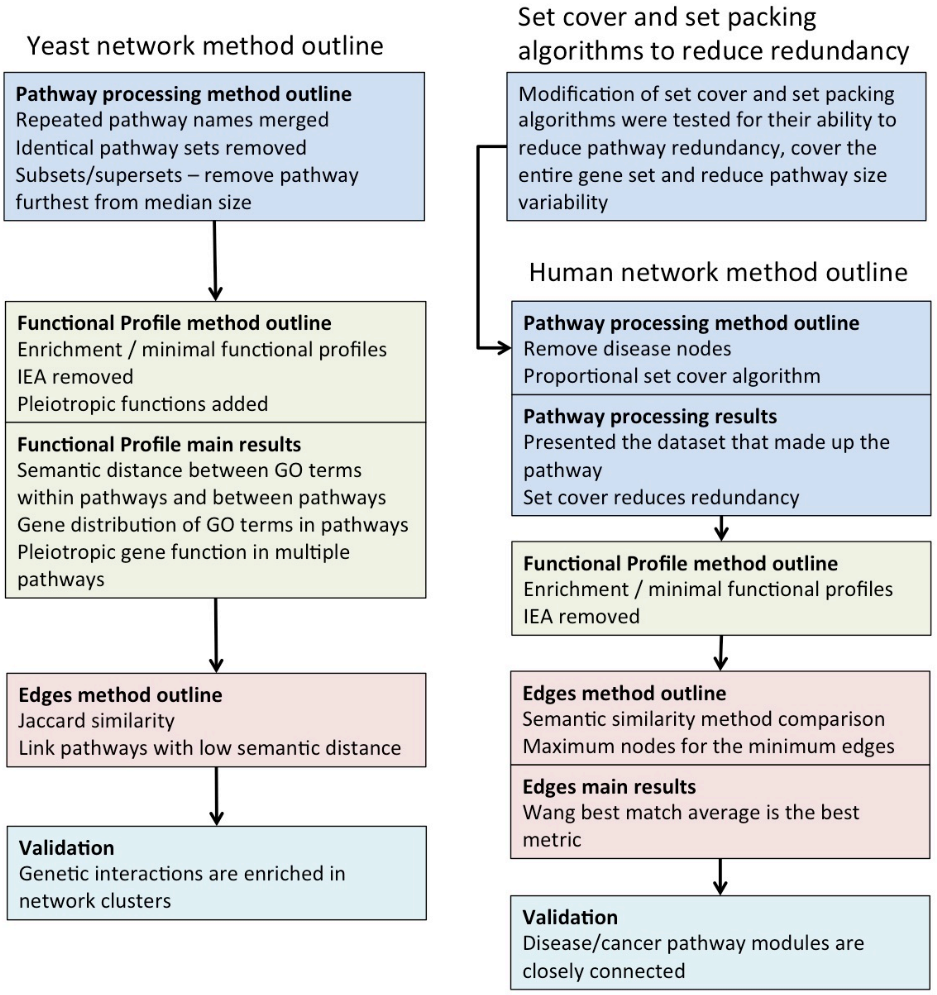

Yeast network method outline
Pathway processing method outline
Repeated pathway names merged
Identical pathway sets removed
Subsets/supersets – remove pathway furthest from median size
Functional Profile method outline
Enrichment / minimal functional profiles
IEA removed
Pleiotropic functions added
Functional Profile main results
Semantic distance between GO terms within pathways and between pathways
Gene distribution of GO terms in pathways
Pleiotropic gene function in multiple pathways
Edges method outline
Jaccard similarity
Link pathways with low semantic distance
Validation
Genetic interactions are enriched in network clusters
Set cover and set packing algorithms to reduce redundancy
Modification of set cover and set packing algorithms were tested for their ability to reduce pathway redundancy, cover the entire gene set and reduce pathway size variability
Human network method outline
Pathway processing method outline
Remove disease nodes
Proportional set cover algorithm
Pathway processing results
Presented the dataset that made up the pathway
Set cover reduces redundancy
Functional Profile method outline
Enrichment / minimal functional profiles
IEA removed
Edges method outline
Semantic similarity method comparison
Maximum nodes for the minimum edges
Edges main results
Wang best match average is the best metric
Validation
Disease/cancer pathway modules are closely connected

Figure 1: Outline of the thesis

1.6.1 Literature review of models of biological complexity - Chapter 2

Firstly, we review the use of molecular interaction networks to generate functional modules and explore functional organisation, with a focus on protein-protein interaction (PPI) networks. The issue of representing dynamic and temporal interactions within static networks is then explained, leading to an evaluation of dynamic network methods. We initially focus on the use of expression data to provide cellular interaction context, which is represented using active sub-networks. We then review the range of molecular interaction data types available. Next we evaluate the limitations of molecular data, exploring the effects that data incompleteness and inaccuracy can have on network topology, which in turn affects the identification of functional modules. The next section reviews gene pleiotropy and the problem of representing multiple gene functions in molecular networks in which each gene is represented by a single node. The idea that pleiotropic nodes may also distort the arrangement of function, by bringing together functions which are separate in the cell, is presented. A further section introduces disease modules, explaining their relationship to functional modules and how functional networks can be useful in exploring disease mechanisms. Finally, we review existing pathway networks, showing how the Pathway Ontology and VisaANT organise pathways into metagraphs based on shared biological concepts.

1.6.2 Review of methods and resources – Chapter 3

This chapter provides a comprehensive review of the methods and resources used within this thesis, starting with an explanation of the data used. The pathway data was retrieved from ConsensusPathDB (CPDB) and the functional annotation data came from the Gene Ontology (GO). The disease annotations used to generate disease pathways and modules were retrieved from the Human Phenotype Ontology and the genetic interactions used to validate the yeast network were taken from BioGrid.

In order to link the pathways into a network based on shared functionality, the function of each pathway was ascertained. We explain how enrichment analysis can be used to identify pathways associated with GO terms, as well as disease annotations. The use of semantic similarity measures to indicate the similarity of GO terms is described. We describe the Resnik and Wang methods to measure distances between individual GO terms, along with the best match average and pairwise average approaches to measure distances between sets of GO terms. These semantic similarities form the basis of the network edges.

Next we address the issue of redundancy within the pathway data set and the enriched pathway GO terms. We look at existing methods that use pathway merging to reduce overlap between pathways. We then introduce set theory algorithms to develop a method for generating a pathway subset with reduced redundancy and controlled size variability. Finally, we describe some software and statistical methods used within the thesis.

1.6.3 Development and analysis of the Yeast Network – Chapter 4

A pathway network was first generated in yeast to determine: whether the pathways formed a cohesive network of function; whether the independent functions of pleiotropic genes were represented within multiple pathways; and to determine the distribution of functions within and between pathways.

To determine the feasibility of the thesis, a set of pathway nodes were allocated enriched GO function. Careful allocation of functional annotations was particularly important in the yeast model, since functional links between pathways were generated using set methods. We developed a method to assign the most specific GO terms available to annotate pathways, later referred to as the enrichment set cover algorithm. Network edges were primarily generated using the Jaccard coefficient for shared GO terms. In the yeast model the need to link functionally similar terms was met by generating edges for a small proportion of highly similar GO terms.

Following construction of the network, its topology and functional organisation were explored, and the main functional modules were observed. The ability of the network to facilitate the multiple functions of pleiotropic genes was also demonstrated. Functional analysis showed that GO terms within pathways are more similar than GO terms between pathways, although some pathways contained semantically diverse annotations. Finally, validation was performed by degrading the network into clusters and establishing the high prevalence of genetic interactions (GI) within these clusters.

1.6.4 Reducing pathway redundancy through set theory algorithms – Chapter 5

Direct application of the previous method to human data was hindered by the high levels of redundancy within the human dataset. The requirement for a minimally redundant set of pathways capable of covering all of the genes within the data set lead to the exploration of the combinatorial algorithms set cover and set packing. Neither algorithm was suitable in its original form, since set cover preferentially selects large pathways with low functional specificity, while set packing tends to show poor coverage of the data set. Methods capable of reducing redundancy, maximising coverage of the dataset and reducing pathway size variability were therefore developed.

1.6.5 Exploring semantic similarity and disease modules within the human network – Chapter 6

Application of the modified set cover algorithm to the human data set produced a suitable set of pathways, with reduced redundancy and good coverage of the dataset. The methods developed in yeast were reapplied to identify the most significantly enriched set of GO terms capable of covering the genes in each pathway. Rather than using the Jaccard coefficient to generate edges between pathways, the ability of the Resnik and Wang methods (Resnik 1999; Wang et al. 2007), in conjunction with the pairwise average and best-match average, was explored. Semantic similarity methods were judged on their ability to assign shorter semantic distances to GO terms within pathways, than GO terms between pathways, based on the distribution of function in yeast and within the literature (Guo et al. 2006). The superior performance of the

Wang best-match average approach suggested that some combinations of functionally diverse GO terms might appear within pathways.

Following generation of the network, functional clusters were evaluated and then validated using disease data. Within systems biology, diseases are modelled as perturbed functions therefore, a relationship is expected between the distribution of disease nodes and the distribution of function. The tendency for diseases to cluster within the network was measured, then a case study linked different types of leukaemia and gastrointestinal cancers to pathway functions.

1.6.6 Discussion – Chapter 7

The discussion begins by outlining the main findings of the thesis, focusing on the representation of function within pathways and network topology. We first demonstrate that the network is biologically informative. The functional cohesiveness of the functional modules within the networks is validated by significant increases in genetic interactions within the modules. The relationship between disease modules and functional perturbations is examined and insights into cancer mechanisms, inferred from the distribution of cancer nodes within the network, are confirmed within the literature.

Next we discuss how application of the set cover algorithm reduced pathway redundancy by 60% and controlled pathway size, while covering 99.95% of the human gene set. We then explore the finding that functions are often covered by multiple pathways and pathways often cover multiple functions. Literature support for multifunctional pathways is reviewed, supporting our assertion that multifunctional pathways are responsible for functional cross-talk and co-regulation across the cell. This functional cross talk is captured within our cohesive functional network. We demonstrate that the network captures further biological complexity by facilitating context dependent gene function. We conclude this section by discussing the issues that motivated the developments which occurred in the method between the yeast and human networks.

Next, we illustrate the advantages of using pathways to study function. Firstly, we show that pathways provide a source of context dependent interactions capable of capturing gene pleiotropy, and are much more extensive and accessible than gene expression data. Next, we consider how functional modules within molecular networks are determined by network topology, and we discuss how incompleteness and inaccuracies within interaction data can affect these modules. We offer pathway networks as a solution to this problem since molecular interaction data is not incorporated into the network. In addition, we consider the finding that molecular network topology is most heavily influenced by cellular components, in contrast to pathways, which are most likely to be connected by shared biological processes. These findings suggest that pathway networks may be more capable of detecting functional modules than molecular networks.

The finding that the Wang best match average is the most suitable method for generating pathway edges contradicted many previous studies, which suggested that the Resnik method showed better performance. We suggest that this result may be related to pathway multi-functionality, as well as the inclusion of indirect similarity comparisons between genes in the same pathway that may not directly interact.

The following section addresses the application of the pathway network to the study of disease. We explore the benefits that linking pathways based on functionality can have for exploring disease mechanisms. Disease modules can be detected as closely linked sets of pathways affected by a particular disease. By visualising the functions connecting disease pathways as well as the additional functions in close network proximity, disease mechanisms can be uncovered. This process is impeded in molecular networks by difficulties in identifying disease modules. The understanding of pleiotropic disease genes could also benefit from the use of pathway networks, since perturbed functions can be traced to individual gene instances within particular pathway contexts. Finally, disease gene overlap and co-morbidity is discussed as the effect of functional perturbations spreading in different directions throughout the network to generate varying disease phenotypes.

The final sections of Chapter 8 outline the limitations of the thesis and discuss future work. Issues surrounding the variation in the output of the heuristic methods required to reduce redundancy are considered. The inclusion of ‘part-of’ GO edges has caused some controversy within the semantic similarity literature, which is assessed, along with issues in disease pathway detection. For future work, controllability analysis is suggested to analyse the flow of information between functions. We also suggest a metric to compare the molecular interaction networks' and the pathway networks' ability to predict gene functions. A second metric ascertains whether physical interactions are more tightly correlated between functionally related genes in the molecular network or functionally related pathways in the pathway network. In addition, we suggest that the issues in disease pathways detection could be resolved by applying genome-wide association studies to disease SNP data to enhance the sensitivity of disease pathway identification.

1.6.7 Conclusion – chapter 8

We conclude the thesis by stating that we have constructed two biologically informative, validated functional networks, using yeast and human data. The use of functional pathway networks overcomes many of the issues of molecular networks, including their inability to represent context dependent interactions and gene pleiotropy. An additional advantage is that functional modules are not affected by inaccuracies or incompleteness in molecular interaction data. As a result we suggest that pathway networks provide an excellent solution in many situations where molecular interaction networks may be inadequate. These include situations where molecular interaction data is scarce or data inaccuracy is suspected. Pathway networks can resolve situations in which gene pleiotropy has distorted the arrangement of function within molecular networks, or molecular networks have failed to detect functional modules based on network topology. Additionally, pathway networks may be used to provide context dependent interactions when contextual or temporal molecular data is not available. As well as overcoming the problems of molecular networks, pathway networks are also more concise and provide intuitive interpretation of functional organisation.

Chapter 2

Literature review of models of biological complexity

This chapter starts by describing the use of static and dynamic molecular interaction networks to study function. We review the limitations of these methods, most importantly Interaction data incompleteness and inaccuracy, as well as an inability to represent gene pleiotropy. Next, the use of disease modules to elucidate disease mechanisms is reviewed. Finally, we explore existing pathway networks.

2.1 Systems Biology

Molecular biology has traditionally focused on the study of individual biochemical entities; however, genes, proteins and metabolites do not work alone in the cell. By examining the interactions and relationships between biological entities, systems approaches are able to provide holistic insights into the inner working of the cell. Network approaches rely on biological entities being represented as nodes (Barabási & Oltvai 2004). Edges represent relationships such as physical interactions, GIs, or correlated expression. Networks typically have a scale-free distribution, meaning that a small number of nodes have a very high degree distribution (hubs) compared to the majority of the nodes in the network (Albert 2005; Barabási & Oltvai 2004). Scale-free topologies result in short paths between nodes, known as the small world property.

Network topologies also tend to show hierarchical modularity of cellular functions. Modularity refers to groups of nodes which show dense intra-component connectivity and sparse inter-component connectivity (Albert 2005). By studying network topology inferences are made about the roles of genes and the organisation of function within the cell.

Network edges can be directed or undirected depending on the type of data they represent. Protein-protein interaction (PPI) networks and genetic networks (I. Lee et al. 2004) are examples of undirected networks. Gene regulatory networks, which represent the binding of transcription factors to regulatory element, require directed edges to show the flow of control (Zhu et al. 2009). The pathway network presented is a new type of undirected network.

2.2 Functional Networks

Function is often studied using molecular interaction networks. These networks can be constructed using various types of molecular interaction data, however in this section we focus on PPI networks, with gene expression data used to indicate temporal or context dependent interactions.

2.2.1 Static protein-protein interaction functional networks

PPI networks have been extensively used to study the mechanisms by which proteins co-operate to perform cellular functions. In these biological models, nodes represent genes and the edges represent the protein-protein interactions of each gene's product. Biological functions rely on the combined actions of multiple proteins, therefore it is initiative that PPI networks show that proteins tend to interact with other functionally related proteins (Spirin & Mirny 2003; Albert 2005). This results in the appearance of functionally related sub-networks, with tight clusters representing protein complexes and functional modules. The arrangement of functionally related proteins into a modular hierarchical topology has been shown to occur across species. Applications of functional networks include predicting gene function (Song & Singh 2009; Wang et al. 2011) and elucidating disease mechanisms (Janjić & Pržulj 2012c; Barabasi et al. 2011).

2.2.2 Functional modules and protein complexes

Network clustering algorithms search for communities, which are defined as subgraphs with a higher density of connections than the surrounding network (Albert 2005; Sharan et al. 2007). These communities represent protein complexes and functional modules. Protein complexes occur when a set of proteins interact simultaneously to create a single multi-molecular entity (Spirin & Mirny 2003). This is subtly different to functional modules, which are also sets of interacting proteins working towards a common cellular objective; however the interactions in a functional module may occur separately in time and space (Spirin & Mirny 2003; Lu et al. 2006). Functional modules may contain protein complexes (Lu et al. 2006; Li et al. 2012). Both functional modules and protein complexes have a similar appearance in networks, being typically detected as highly connected clusters and both are considered to be functionally homogenous.

This thesis focuses on functional modules; however both functional modules and complexes can be derived from the same networks using similar techniques and many papers do not distinguish between the two entities types (Li et al. 2012; Chen et al. 2014; Lu et al. 2006).

The interactions that comprise protein complexes are considered more reliable than other protein-protein interactions, therefore some papers utilise protein complex interactions as gold standard interaction data (Zhu et al. 2008; Jianxin Wang et al. 2013). Studies may also specifically investigate the functional properties of protein complexes (Gavin et al. 2006; Tarassov et al. 2008). Since proteins are subcomponents of functional modules, their components tend to be highly functionally cohesive.

2.2.3 PPI networks misrepresent dynamic cellular interactions

Sets of interacting proteins do not form constant static entities within the cell, rather they assemble when they are needed to perform a function, then disassemble after use (Srihari & Leong 2012). The dynamic nature of these interactions is not shown in static maps, where these interactions appear constant. Moreover, some proteins participate in the formation of multiple-context dependent complexes (Li et al. 2012; Srihari & Leong 2012). In a study mapping complex formation during different stages of the yeast cell cycle, Srihari and Leong (2012) found that some proteins form different complexes at various stages of the cell cycle. In static networks, edges are assumed to be constant; however, if a protein participates in multiple separate complexes these transient interactions will form a single cluster.

This relates to previous work which uncovered two types of highly connected nodes (hubs) within PPI networks: “date” hubs and “party” hubs (Han et al. 2004; Wagner et al. 2007). Party hubs interact with all of their partners simultaneously, whereas date hubs form transient interactions with different neighbours at different times. These variable interactions are demonstrated by measuring the correlation between the gene expression patterns of hub nodes with the gene expression patterns of their neighbours. Party hubs show highly correlated gene expression with their neighbours, whereas date hubs show a low correlation. Date hubs interact with different partners at different times; therefore, linking all partners within a single network cluster is not biologically representative. Figure 2 (taken from Wagner et al. 2007) shows how date and party hubs link biological components.

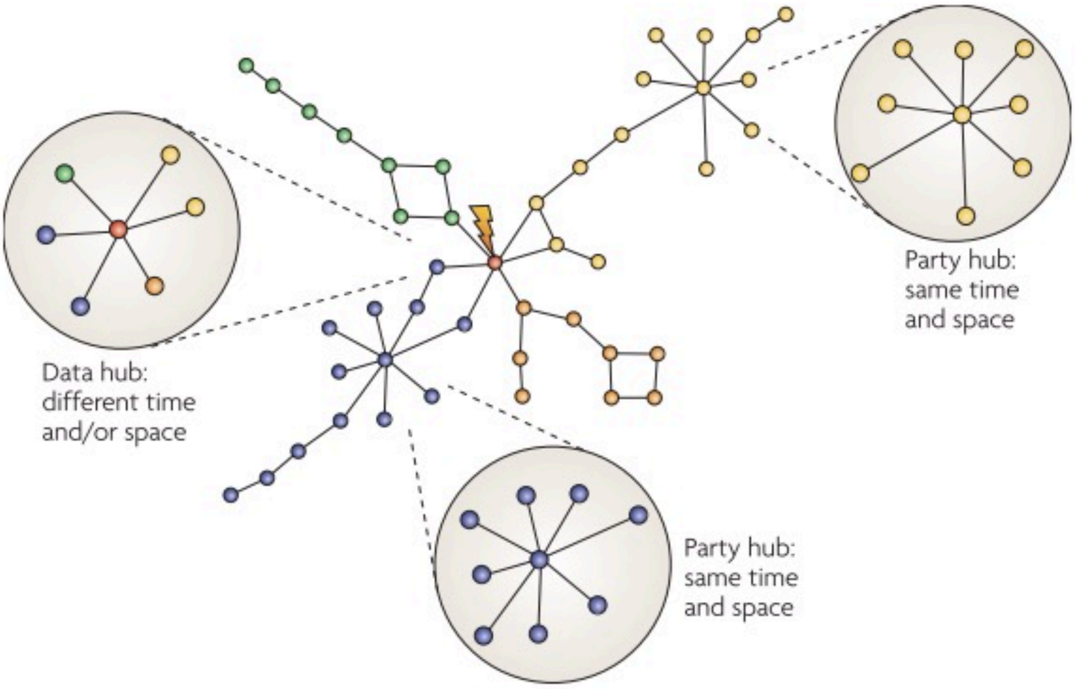


Figure 2: The functional roles of ‘date’ and ‘party’ hubs. ‘Party’ hubs interact simultaneously with many partners linking functional modules, whereas ‘date’ hubs interact with their partners at different times and/or locations, co-ordinating multiple functions. Colours indicate mRNA expression and functional similarity. (Wagner et al. 2007) nature publishing group

Tsai et al (2009) suggested that it is a misrepresentation to claim that a single hub protein binds with tens or hundreds of other proteins. They suggested that this is an artefact that arises from the representation of genes as single nodes, rather than multiple nodes representing different gene products, which may have different amino acid sequences and 3-D structures (Tsai et al. 2009). Alternative splicing is suggested as the dominant means of creating multiple gene products, which is known to alter protein function and binding properties. In PPI networks, each gene is represented by a single node, making differentiating between the interactions of splicing subtypes impossible.

When proteins are involved in different functions as a result of these different interactions, this multi-functionality is referred to as pleiotropy (see Section 2.4). This has a major effect on the mapping on functional modules (see Section 2.2.2) since these pleiotropic nodes can cluster together functionally diverse genes that would not interact in the cell.

2.2.4 Co-expression networks represent context dependent transcriptional states

A common approach used to address the issue of dynamic interactions is to generate networks using co-expression data to form edges between genes that show corresponding changes in expression under various cellular conditions (Segal et al. 2003; Stuart et al. 2003; Ihmels et al. 2002). Methods, such as the Pearson's correlation, can be used to identify pairs of genes that show significantly correlated changes of expression across the microarray datasets (Stuart et al. 2003). Microarray analysis provides a snapshot of the transcriptional responses invoked by the cellular condition being tested. This generates context-dependent data showing sets of genes transcribed within each condition tested. This enables sets of genes, transcribed under a particular subset of cellular conditions, to be revealed and linked within the generated network. Network modules can be extracted, corresponding to sets of genes that are co-regulated in response to cellular conditions. Modules show high levels of functional coherence and conservation across species (Stuart et al. 2003), making the use of gene expression data highly promising in the search for context-dependent interactions. Co-expression networks do not, however, reveal the physical interactions responsible for the execution of function in the cell.

2.2.5 Generating dynamic interaction networks through integration of PPI and gene expression data

Several papers have attempted to address the issues of dynamic PPI interactions by incorporating additional data to improve functional module detection. Protein interactions are assumed to be associated with co-expression; therefore, microarray data is commonly used.

Ideker et al. (2002) used gene expression data in conjunction with PPI data and protein-DNA data to identify sub-networks that showed changes in response to various cellular conditions. These were identified as connected regions of the PPI and protein-DNA network that showed altered gene expression under particular conditions (Ideker et al. 2002). They demonstrated the validity of this method by introducing mutations to the galactose-utilisation (GAL) pathway and identifying co-regulated networks that were affected by the perturbations. The differentially expressed genes were associated with seven sub-networks that showed altered expression under the conditions examined (Figure 3). Note that the sub-networks shown in Figure 3 (taken from Ideker et al 2002) do not correspond directly with obvious topological network clusters. The clusters are based on the differential expression of genes, allowing them to exclude many genes within topological clusters that showed unaltered gene expression.

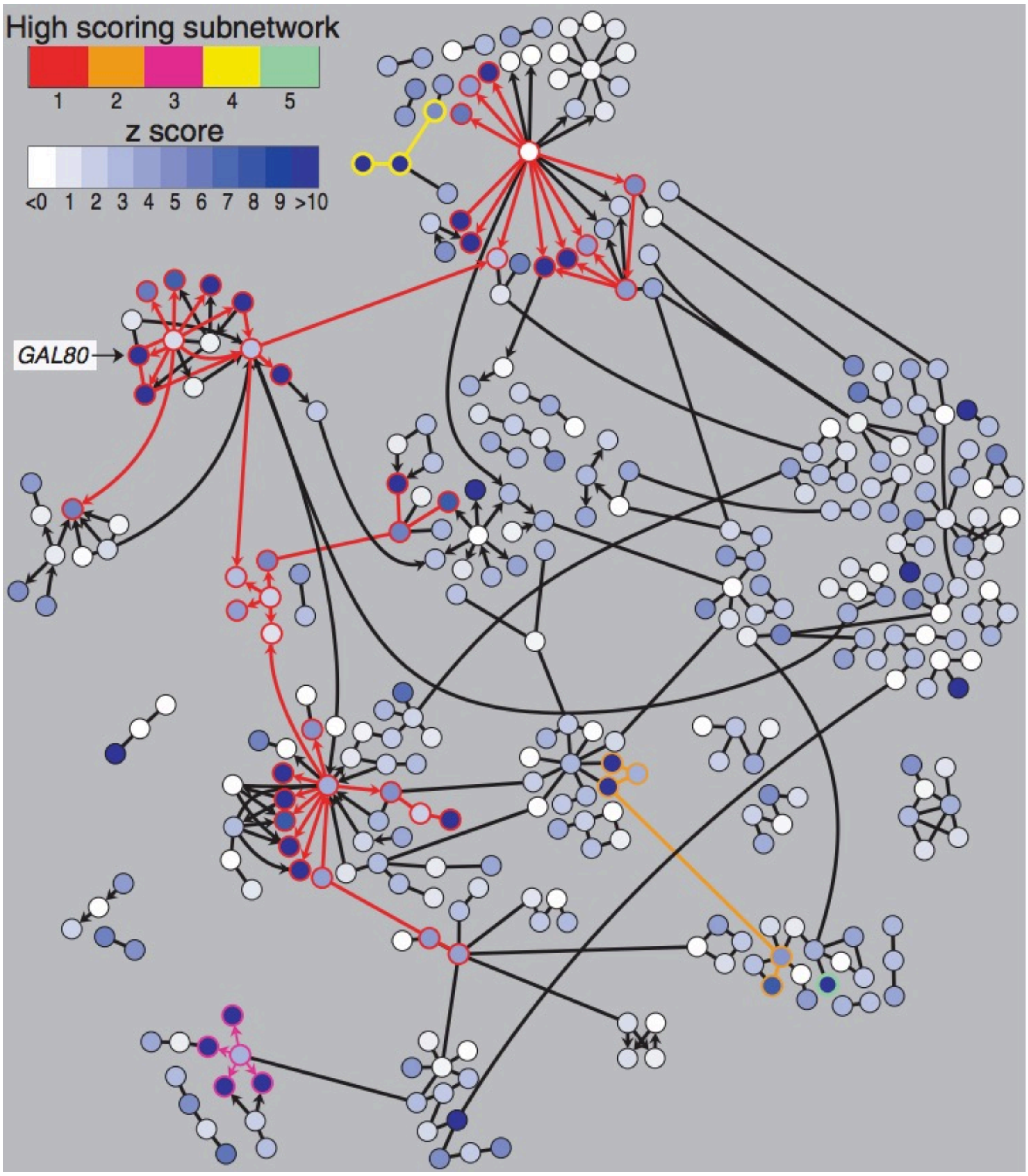


Figure 3: Altered sub-networks within a protein-protein and protein-DNA network. Directed edges protein-DNA interactions, indicating that the source node influences the transcription of the target node. Undirected edges represent protein-protein interactions. The node colours indicate differential expression in response to one or more gene-perturbations. (Ideker et al. 2002) nature publishing group

This demonstrates the complexity and cross talk of cellular responses. These sub-networks showed strong links to known regulatory mechanisms in the literature, for

Also note that multiple sub-network clusters, which are mostly connected, form rather than a single module. example a detected linear sub-network was found to correspond to the core of the galactose-induction circuit.

The concept of dynamic and static network communities was formalised by Komurov and White (2007). Using data from 272 microarray experiments they identified static modules and dynamic modules in the yeast PPI network. In this modular network, communities of dynamically co-regulated interactions allow the cell to carry out condition-related processes, while static modules provide sets of interactions that are always present and act as a structural core. This implies that some functions, such as general mRNA transcription and splicing, may be accurately within static PPI networks, while other functions, such as translation initiation, are represented poorly because their components are dynamic.

A set of time series networks have been constructed for several points in the cell cycle, composed using only highly expressed genes (Tang et al. 2011). Expression data from the yeast cell cycle was mapped onto a PPI network and sets of functional modules were generated, which were then compared to a manually curated protein complex set. The time series networks consistently showed greater overlap between the functional modules produced and the known complexes than comparative static networks. The biological significance of the functional modules generated was also assessed using GO enrichment analysis. Modules generated using time series data consistently showed higher numbers of significantly enriched GO terms, suggesting increased functional cohesiveness. Similar results have been produced using time series data focusing on yeast metabolism. Complexes and functional modules retrieved following the integration of expression data were shown to be more significantly enriched with GO terms, indicating biological validity (Li et al. 2012). These results were confirmed in a later study demonstrating the ability of metabolic time series data to improve the sensitivity and specificity of protein complex prediction (J Wang et al. 2013). A set of networks showing PPI interactions between genes expressed under various UV irradiation conditions has also been produced (Hegde et al. 2008). Approximately 40-60% of cellular genes were expressed in each condition, with variation between networks implicating different cellular response mechanisms. This example indicates that around half of the nodes in a comparative static network, would be present despite not being transcriptionally expressed. Similarly Komurov and White (2007) found that only 16.5% of genes where present within their static network and 21.2% were highly dynamic. These finding undermine the ability of static networks to generate functional modules.

2.2.6 Enhancing functional representation through integration of multiple data types

Further improvements to functional networks have been generated by combining multiple types of molecular data including: GI data; subcellular localisation data; and genetic control factors such as transcription factor binding sites and expression quantitative trait loci. The following sections review examples of each approach.

2.2.6.1 Genetic interaction data

GIs provide an additional approach to the generation of functional networks. A genome-scale GI map generated in yeast showed that genes with similar biological processes clustered into coherent modules (Costanzo et al. 2010). GIs are known to occur frequently between pairs or genes with similar functions within network modules, as a result of the cells' inability to cope with multiple perturbations to a single biological process (Kelley & Ideker 2005).

Significant overlap was found between modules extracted from PPI networks, co-expression networks and GI networks (Ames et al. 2013). Combining these modules into a single network, generated modules that were more functionally cohesive than those extracted from networks produced using single data types. Figure 4 (taken from Ames et al 2013) shows how the combination of a PPI network (blue nodes and edges) with a GI network (red nodes and edges) gives a more complete portrayal of the genes involved in DNA replication control, either individual network. Coverage of the Gene Ontology was also found to be more complete in modules extracted from the combined data network compared to modules from the individual networks. The improvements gained by combining data types demonstrates that none of the previous methods is capable of individually providing a complete biological representation of cellular function.

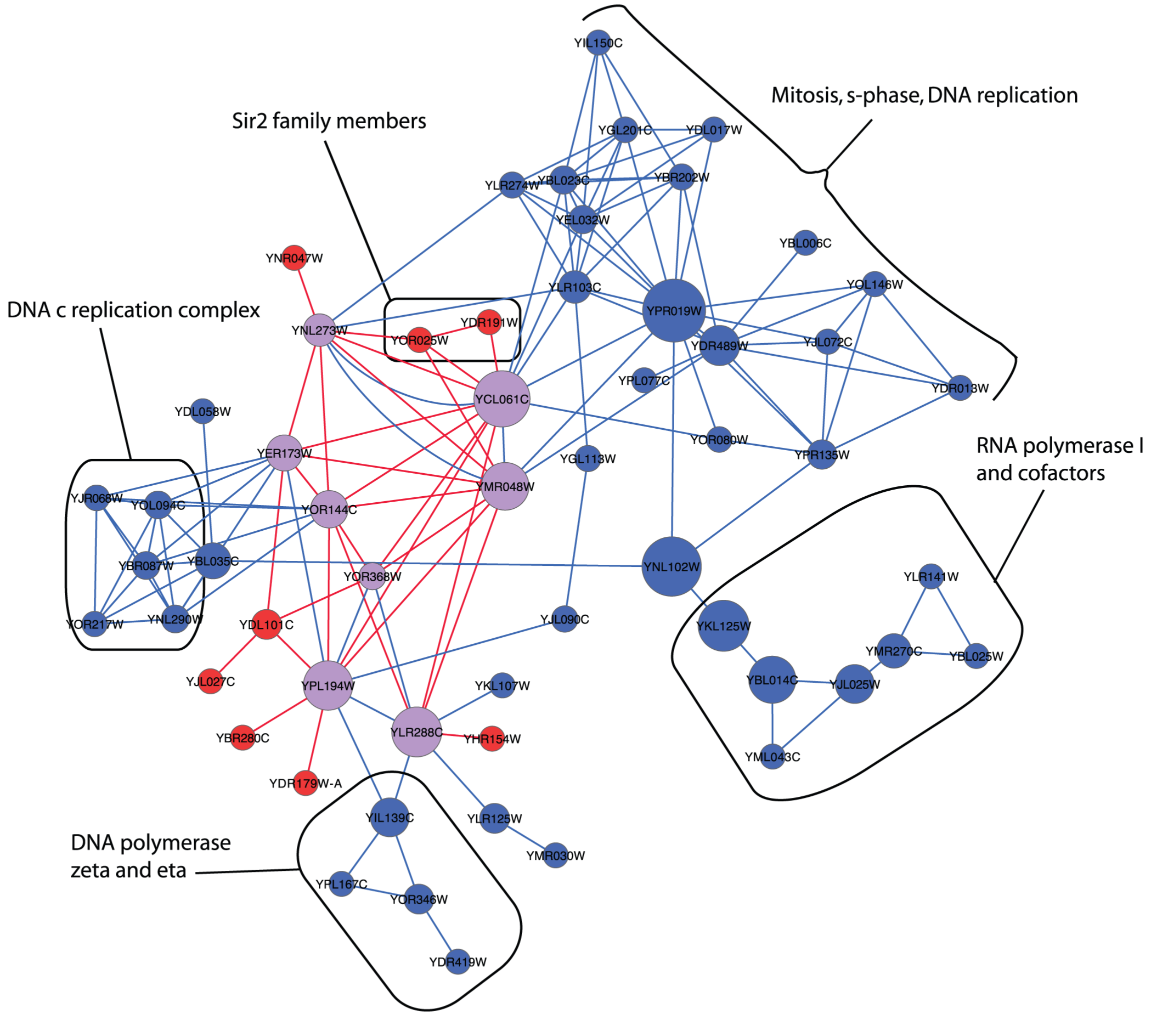


Figure 4: Consolidating functional clusters by combining PPI and GI data. DNA replication module in which the blue nodes represent genes from a PPI network cluster, the red nodes represent genes for a GI network cluster and the purple node are present in both PPI and GI network clusters. By merging the networks a more extensive portrayal of DNA replication is presented than possible using either network individually *(Ames et al. 2013)*.

2.2.6.2 Subcellular localisation data

Subcellular localisation data can be used with PPI and expression data to assist in the identification of functional modules (Lu et al. 2006). Network clusters which included subcellular localisation data were found to be more robust to false positive interactions and better highlight known protein complexes. As well as demonstrating the importance of combining multiple sources of biological data, Lu et al. (2006) illustrates the difference between protein complexes and functional modules by stating that protein complexes must be co-localised and co-expressed, but neither of these restrictions apply to functional modules. An example of a mitochondrial module is given, which includes proteins found within the mitochondrial ribosome and as well as the mitochondrial membranes. The membrane protein acts as a functional facilitator allowing transport into the mitochondria. Therefore although co-localisation data can be used to assist in the detection of protein complexes and ensure that co-localised protein groups are clustered within networks, its ability to detect functional modules is limited.

2.2.6.3 Genetic control factors

Genetic control factors are another data source that can be integrated into molecular interaction databases. A yeast co-expression network was enhanced through the addition of PPI, transcription factor binding site and expression quantitative trait loci (eQTL) hot spot data (Zhu et al. 2008). By integrating different data types, the generated network showed an increased ability to predict the effects of gene knock out perturbations. The transcription factor binding site and eQTL data were also used to make inferences about the functional control of the network.

Interestingly a low correlation was detected between the co-expression and protein-interaction data (Zhu et al. 2008). This may genuinely reflect the different information being represented by these datasets or it may be due to high levels of false positive results in the PPI data. A low correlation between mRNA levels and protein levels has reported in many papers (see section 2.3.2) indicating that these networks reflect different data. However, a high confidence set of protein interactions generated from protein complex data showed increased correlation with the co-expression data, suggesting that inaccuracy in the PPI data may be a contributing factor. This increased correlation may also reflect the tendency for components of complexes to be co-expressed, while interacting protein molecules are less likely to be co-expressed.

2.2.7 Genome wide proteomics assays detect context dependent protein interactions

Genome wide protein-fragment complementation assays provide an opportunity to view *in vivo* protein interactions and may provide a more reliable data source of context dependent protein interactions (Tarassov et al. 2008). This method allows detection of context dependent interactions, presenting a faithful representation of gene expression timing and subcellular localisation. Protein complexes with similar functional annotations interacted more often than would be expected to occur by chance, although interactions between proteins with different GO annotations were not uncommon. Interactions between proteins from different subcellular compartments were also common, reflecting the complex organisation of cellular function. Further use of protein-fragment complementation assays in a range of cellular conditions will eventually lead to highly detailed accurate functional maps, however its use is currently limited. This may be because probe instability makes large-scale screening difficult (Ohmuro-Matsuyama et al. 2013). Additionally false positives may occur due to the florescent intensity of the reconstituted reporter fragments used (Snider et al. 2015).

Networks constructed using protein profiling data from mass-spectrometry based technologies were compared to mRNA co-expression networks (Wang et al. 2017) for three types of cancer. The networks showed a marked difference in topology with protein co-expression being primarily driven by functional similarity, while mRNA co-expression was largely determined by chromosomal co-localisation resulting in less functionally coherent modules. Network-based gene function prediction showed that the protein co-expression network had better predictive performance for up to 92% of GO terms. This demonstrates the ability of this approach to significantly enhance molecular networks, however, it is limited by the availability of data. Data sets must be produced for each cellular condition studied, for example breast cancer, colorectal cancer and ovarian cancer in this study.

2.3 Limitations of molecular networks

PPI networks and network generated from combined molecular data types are the standard models used for studying the organisation of function within cells. However, data inaccuracy and incompleteness, as well as the weak correlation between protein interactions and protein expression undermines the used of molecular interaction models. The following sections explore these shortcomings motivating the development of a new model structure.

2.3.1 PPI data inaccuracy, incompleteness and bias

The accuracy of PPI networks is often criticised due to the high number of false positives and data incompleteness (Karagoz & Arga 2013; Hakes et al. 2008; Hart et al. 2006; Snider et al. 2015; Sprinzak et al. 2003). This inaccuracy stems from issues in the experimental procedures used to acquire the interaction data. Protein-protein interactions can be mapped in small studies, focusing on a single protein of interest, a panel of proteins or in an unbiased global fashion (Brückner et al. 2009)(Brückner et al. 2009). Two commonly used experimental approaches are yeast-2-hybrid (Y2H) screening and mass spectrometry approaches, such as affinity purification mass spectrometry (AP-MS).

In the Y2H system the gene on interest (the bait) is fused to the DNA binding domain of a transcription factor for a reporter gene (Brückner et al. 2009). Proteins suspected to be interaction partners (the prey) are fused to the transcription factor’s activation domain. Yeast cells are transformed with vectors containing the modified bait and prey genes under the control of yeast promoters. If the two proteins interact, then the reporter gene is expressed.

Classic Y2H systems only detect interacting proteins within the nucleus and are therefore unsuitable for membrane associated proteins, integral membrane proteins, cytosolic proteins or proteins localised in other cellular compartments (Brückner et al. 2009). Truncated proteins have previously been used to circumvent this issue, however, such proteins are prone to misfolding, leading to high rates of false negatives. Modified systems such as the split-ubiquitin system (Johnsson & Varshavsky 1994), are able to detect interactions outside of the nucleus. False negatives may also arise if the fused yeast reporter proteins or anchors impede protein interactions (Brückner et al. 2009). Additionally, yeast may lack the ability to add posttranscriptional modifications to eukaryotic genes that are required for protein interaction. Transient interactions are also often missed. False positives also arise in Y2H screens and may be caused by high expression of bait and prey proteins within cellular compartments that do not correspond to their natural environment.

AP-MS is another popular method used to detect interactions. Proteins are purified, along with their binding partners during affinity purification, then identified using mass spectrometry. Affinity purification describes the capture of biological material using a ligand attached to a solid support (Dunham et al. 2012). There are two commonly used methods of affinity purification, using antibodies or epitopes. Antibodies are generated to bind to a protein of interest (the bait) and then attached to beads. The soluble fraction of the cell lysate is then run past the beads, binding the protein of interest along with any interacting partners. Alternatively an epitope-tag may be fused to the C or N terminal of the protein. The tagged protein is then purified using an affinity matrix that recognises the epitope. This method of epitope tagging allows highly efficient protein purification and since multiple proteins can be tagged with the same epitope, background contaminants should be consistent across all purifications. Identification of the purified proteins is carried out using a mass spectrometer, which measures the mass to charge (*m/z*) ratio of charged molecules. In tandem MS the *m/z* ratio intact ionized protein is first measured, then the protein is fragmented and the *m/z* ratio of the fragments is determined. This process allows the amino acid sequence of the peptide to be deduced.

An advantage of immunoprecipitation, compared to epitope tagging and Y2H, is that it avoids modifying cellular proteins, which can disrupt protein binding. However, false negatives may still occur due to antibodies disrupting protein-protein interactions or proteins losing affinity for their target protein following posttranslational modifications (Dunham et al. 2012). False positives may also arise due to cross-reactivity of specific antibodies. If the epitope tagging method is used, the fusion of the tag onto the protein may interrupt protein-protein interactions, resulting in false negatives. False positives and false negatives may also arise as a result of the overexpressed or problematically tagged proteins misfolding, or being misregulated or mislocalised. In both immunoprecipitation and epitope tagging approaches cell lysis may bring proteins from different cellular compartments into proximity, causing false positives if the proteins bind. An advantage of the Y2H system is that it avoids these issues, since the method does not involve cell lysis. Although MS is a highly effective method of identifying proteins false positives can arise, particularly in situations where proteins are identified by low numbers of peptides and the confidence scores generated for these peptides are low.

To control inaccuracies from high throughput data, Gene Ontology data has been used to assess the reliability of PPIs (Karagoz & Arga 2013). A high confidence PPI network was generated by scoring interactions based on the GO semantic similarity (see Section 3.4) of their components using the biological process, cellular component and molecular function ontologies. The resulting network had a sensitivity of 86% and specificity of 68% and coverage of 72%, based on a positive gold standard data set consisting on very high confidence STRING interactions, the MIPS dataset which is acquired through small scale experiments, and the core network of DIP. The gold standard negative dataset consisted of proteins from different subcellular compartments under the assumption that they could not interact, however, as demonstrated by Lu et al. 2006 there are exceptions to this rule. As the specificity of Karagoz and Arga 2013's network is 68%, even within this high confidence dataset almost one third of the interactions are expected to be false positives. Techniques incorporating graph topology (Kuchaiev et al. 2009) have also been applied to estimate the confidence of interactions, which show promising results predicting interactions based on older datasets, which have since been added to updated versions. However, lack of a gold standard PPI network to compare the predicted confidence scores to makes interpretation of results.

To counteract the effects of false positive PPIs, studies may restrict interactions to those reported in small literature-curated studies (Cusick et al. 2009; Karagoz & Arga 2013). A weakness of this approach is the inherent selective bias regarding the genes that biologists choose to study, which shows a large preference for essential genes and disease genes (Hakes et al. 2008). The reliability of literature curated studies is further undermined by the finding that only 25% of interactions are found in multiple studies

(Cusick et al. 2009). Comparisons between databases that collect literature curated interactions show low levels of overlap implying that individual databases are far from comprehensive.

Estimates regarding the completeness of PPI networks have reported that yeast maps are 50% complete and human networks are only 10% complete (Hart et al. 2006). The effect of sampling bias in the data is also considered to have a highly distorting effect on networks; for example resulting in the apparent repression of hub-hub interactions, which are present when data bias is controlled (Hakes et al. 2008). These findings carry serious implications for the current utilisation of PPI network topology.

2.3.2 Gene co-expression is a poor predictor of protein interactions

Expression data has been used to indicate whether pairs of proteins are present under particular conditions allowing them to interact. However, the low correlation between gene expression and protein abundance is well documented (Ghaemmaghami et al. 2003; Gygi et al. 1999). Cellular protein levels are controlled through processing and degradation of mRNAs, followed by translation, localisation, post-translational modifications and degradation (Vogel & Marcotte 2012). Recent reviews have suggested that post-transcriptional, translational and degradation regulation are at least as important in determining protein concentration as transcriptional control (Maier et al. 2009a; Vogel & Marcotte 2012).

Reported correlations between mRNA and protein levels are widely variable, ranging from 0.35 to 0.75 (Vogel & Marcotte 2012; Schwanhausser 2011; Maier et al. 2009b; Ideker et al. 2001) and even r = 0.01 when the correlation is measured within single cells. The relationship between mRNA levels and protein levels varies widely depending on the organism and biological properties of the proteins (Vogel & Marcotte 2012). Some genes show highly variable mRNA levels, such as genes involved in transcriptional regulation and chromatin organisation; while others, such as ribosomal genes show very stable mRNA levels. In yeast, mRNAs with highly variable expression were found to be tightly correlated with protein levels (r=0.89), while genes with stable mRNA levels showed little or no correlation with protein levels (Greenbaum et al. 2003).

Translation, rather than transcription, was found to have dominant control over protein levels in mouse cells, with wide variations in the number of proteins produced by each mRNA per hour (Schwanhausser 2011). Transcription did have some influence with the correlations between mRNA levels and protein levels reported as R^2 = 0.41, however considering the translation rate boosts the correlation to R^2 = 0.95. The variability of protein half-life was also found to be a controlling factor. Biological processes involving fast response to stimulus, such as transcription factors and chromatin modifying enzymes, require proteins with short half-lives. Housekeeping genes such as those involved in central metabolism produced stable proteins, with some half-lives exceeding 200 hours. The finding that some functions require consistent protein levels while others require variable protein levels is in agreement with Komurov and White's proposal of static and dynamic modules (Komurov & White 2007).

2.4 Pleiotropy

The ability of a gene to affect multiple phenotypic traits is referred to as pleiotropy (He & Zhang 2006a). Over 50% of the genes in yeast been suggested to be pleiotropic (Dudley et al. 2005). Pleiotropy can result from genes having more than one molecular function or a single function having more than one cellular effect. Pleiotropy has been correlated with the number of interactions a gene participates in (Promislow 2004; He & Zhang 2006a). Structural flexibility is known to be a factor by allowing proteins to bind to multiple partners, for example *p53* achieves this using disordered domains (Oldfield et al. 2008). The representation of pleiotropic genes within GI networks is problematic, since genes are only represented by a single node.

The ability of proteins to have multiple catalytic functions is referred to as protein promiscuity. Promiscuous proteins participate in secondary reactions while retaining the ability to perform their primary function (Nobeli et al. 2009). This gain of additional functions may stem from proteins binding to additional partners, or enzymes performing multiple chemical reactions.

Various factors within the cellular environment may determine which functional role a protein adopts. Protein function may be controlled through gene expression at different times or in different cellular locations. Differential expression is commonly used by viruses to generate multiple functions from their limited gene set. For example, the Epstein-Barr virus only has one serine/theonine protein kinase which assumes various roles, phosphorylating different proteins, during the different stages of virus replication (Wang et al. 2005).

Proteins may also switch functionality depending on the presence of ligands within the cell. For example, mammalian cytosolic aconitase acts as an enzyme that interconverts citrate and isocitrate in the absence of iron, but acts as an RNA binding protein when iron is present (Philpott et al. 1994). Protein function can also be affected by factors such as pH and temperature. An example is thymidine kinase from *Thermotoga maritima* which shows high substrate specificity at 82°C, but gains the ability to bind to different partners 37°C, due to a conformational changes (Lutz et al. 2007).

Proteins are described as moonlighting if, as well as performing their primary enzymatic role, they also have a secondary non-enzymatic function, which may be structural or regulatory (Copley 2003). The first moonlighting proteins to be described were crystallins, which have enzymatic capabilities and are also structural proteins in the vertebrate eye (Piatigorsky & Wistow 1989). It is possible for a protein to have several moonlighting functions (Copley 2012). For example, GAPDH's primary enzymatic role is in energy metabolism, however, it is also involved in apoptosis, vesicular transport, nuclear tRNA transport and as a crystanllin in eye lenses.

Previous sections (2.2.3) of the thesis have discussed context dependent interactions, and the dependency between network topology and functional module identification. Gene pleiotropy builds on this issue by stating that the function of some genes (which are represented by a single node) varies depending on the interactions they engage in. The function of the gene's interaction partners will, therefore, also be different. This can distort the distribution of functions within the network, since functions that are independent in the cell, separated by the cellular conditions mentioned, are artificially brought together in the network by the pleiotropic node. Therefore, if a pleiotropic node performs different functions by binding to different partners at different cellular time points, all of these functionally diverse partners will be brought into close proximity in the network. This is particularly problematic since one of the major principles guiding functional networks is “guilt-by-association”, meaning that proteins are likely to be involved in the same functions as their interaction partners (Oliver 2000). Gene pleiotropy can lead to false positives since a partnering gene may have multiple pleiotropic functions, which will not all be applicable (Gillis & Pavlidis 2011).

2.5 Disease modules

Proteins associated with similar disorders are likely to physically interact, leading to the formation of disease modules within interaction networks (Goh et al. 2007; Janjić & Pržulj 2012a; Vidal et al. 2011). Disease modules may overlap with functional modules indicating that disease phenotypes result from perturbed cellular functions. For example, the Fanconi anaemia module overlaps with the DNA repair module, indicating a causative relationship (Goh et al. 2007). Modelling disorders as perturbed cellular functions offers insight into complex and polygenic diseases (Mitra et al. 2013; del Sol et al. 2010). The multiple genes associated with polygenic diseases reflect the sets of genes within a functional module, which if disrupted will result in perturbed module functionality. Complex diseases may involve perturbations within multiple functional modules. This is well documented in complex diseases such as cancer, in which many different mutations can result in a similar disease phenotype. Generation disease modules along with functional modules on molecular interaction networks can therefore be useful in elucidating disease mechanics (Barabasi et al. 2011).

The formation of disease modules allows identification of candidate disease genes by the 'guilt-by-association' principle which states that if a gene product interacts with many known disease genes it is likely to also be implicated in the disease (Barabasi et al. 2011). This approach can assist in the identification of candidate disease genes. For example, sickle cell anaemia is characterised by a single point mutation but patients can present with a range of phenotypes. This implies that other additional disease modifying genes are affecting the phenotypes, in the surrounding disease module. Highly informative disease modules can be generated by combining related disease phenotypes. For example, by generating a PPI sub-network of proteins involved in multiple types of inherited cerebellar ataxia and their adjacent interaction partners, a disease module was generated in which the 18 of the 23 disease genes interacted (Lim et al. 2006). Many of the shared interaction partners of the ataxia genes have been found to modify neurodegeneration in animal models, providing direct support for the guilt-by-association hypothesis. The data used by Barabasi et al. (2011) was specially generated using yeast two-hybrid screens targeted towards the ataxia causing genes. Of the interactions identified, 96% were novel, indicating the incompleteness of PPI data.

Proteins can be involved in multiple disease modules meaning that disease modules can overlap (Goh et al. 2007). Relationships between diseases have been made particularly explicit in studies linking diseases by shared genes (human disease network) and genes by shared diseases (disease gene network). In each network diseases and disease genes cluster into disorder groups such as cancers, renal disorders, neurological disorders and haematological disorders (see Figure 5 taken from Goh *et al* 2007). The disease gene network highlights the tendency for common disorders to be polygenic, supporting the hypothesis that different combinations of genes from a disease module may give rise to a single disorder. The ability of genes to cause multiple disorders is illustrated by the human disease network, suggesting that cellular context and the actions of multiple genes are important for determining the phenotype. Diseases with shared genes were found to have increased comorbidities in many cases, particularly in causes where diseases share multiple genes (Park et al. 2009). Co-morbidities are more likely if the mutations occur in the same functional domain in each disease case, supporting the idea that the proteins ability to fulfil a particular cellular role is being disrupted. Examining the multiple disease phenotypes caused by genes gives insight into their functions and helps to develop understanding of disease mechanisms.

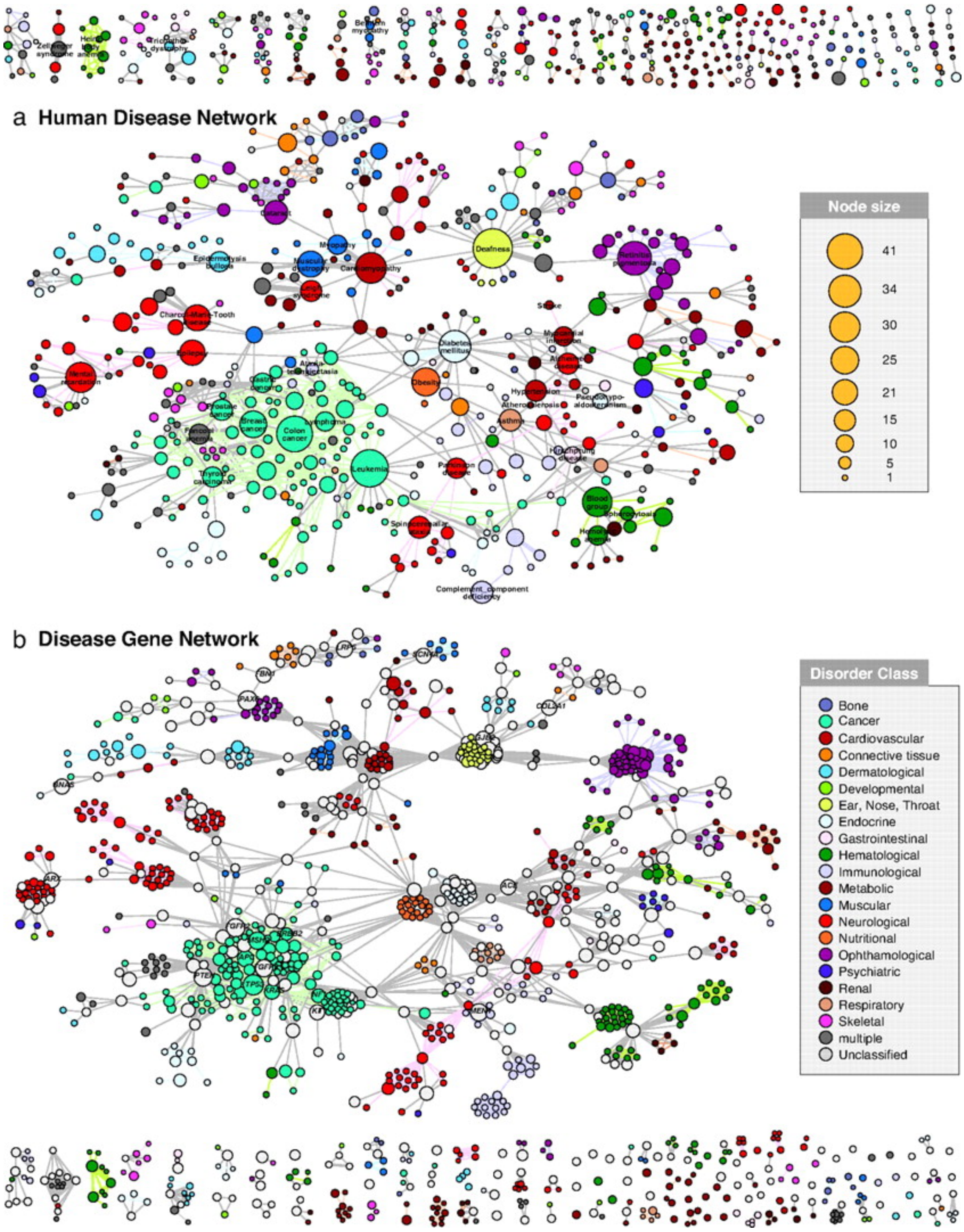


Figure 5: The human disease network and the disease gene network (Goh et al. 2007) "Copyright 2007 National Academy of Sciences." a) In the human disease network nodes represent diseases and edges represent shared genes. Node colours indicate disease classes and node sizes reflect the number of genes involved in the disease. b) In the disease gene network each node represents a disease gene and nodes are linked if they are associated with a shared disease. The size of each node indicates the number of disorders the gene is associated with (Goh et al. 2007).

2.6 Existing pathway networks

Previous studies have arranged pathways based on shared function and molecular interactions. These studies use nested network structures to identify pathways that are subsets of larger pathways, as well as enabling the user to identify closely related sets of pathways. They do not, however, reduce and flatten hierarchical pathway redundancy or attribute full sets of function to each pathway gene set. The relationships between pathways are therefore restricted.

2.6.1 Pathway Ontology

The pathway ontology uses a directed acyclic graph to capture relationships between pathways (Petri, Jayaraman, et al. 2014). The ontology contains five types of nodes, representing different types of pathways are present within the ontology: metabolic nodes, regulatory nodes, signalling nodes, disease nodes and drug nodes. The structure of the Pathway Ontology is based on the Gene Ontology with “is_a” and “part_of” relationships forming directed edges pathways to increasingly general pathway concepts. For example, Figure 6 shows a screenshot from the Pathway Ontology (hosted by the Rat Genome Database), visualising the parent nodes of chromatin remodelling. The chromatin-remodelling pathway is connected to eight parent nodes, each containing more general pathway concepts linking to increasing numbers of pathways. The database includes pathways from the PID and KEGG, providing pathway annotations for over 49,000 genes and has a depth of up to 10 nodes (Petri, Jayaraman, et al. 2014). Pathway nodes are not disjoint, allowing pathway overlap, which is particularly likely between closely related pathways. Many pathway names link closely to GO terms for example the ‘fatty acid biosynthetic process’ in the Gene Ontology and the ‘fatty acid biosynthesic pathway’. Pathway suites are also generated linking pathways by common concepts, for example the ‘Glucose Homeostatis Pathway Suite Network’ brings together pathways involved in glucose metabolism with regulatory and signalling pathways (Figure 7).

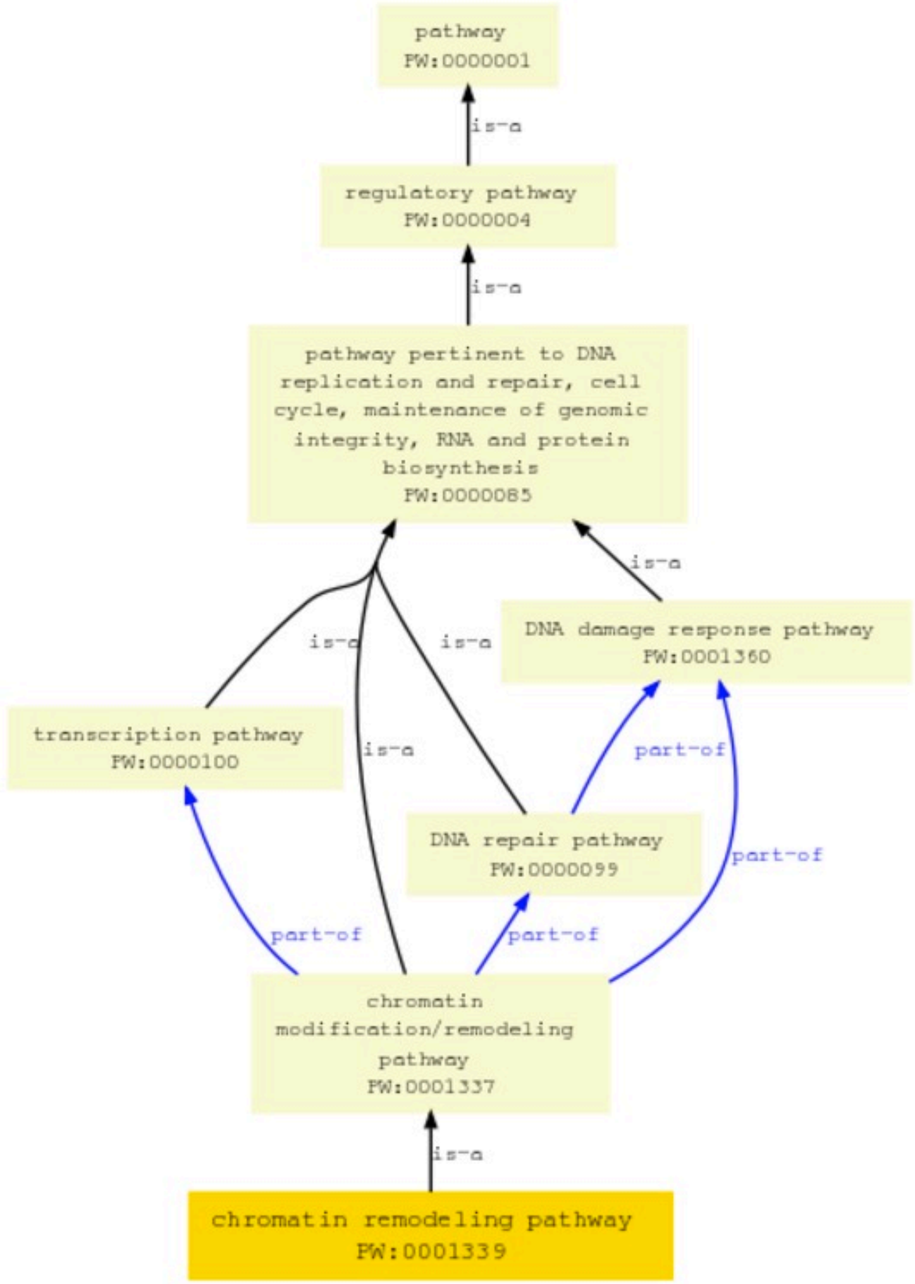

pathway
PW:0000001
is-a
regulatory pathway
PW:0000004
is-a
pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis
PW:0000085
is-a
is-a
is-a
is-a
DNA damage response pathway
PW:0001360
transcription pathway
PW:0000100
is-a
part-of
DNA repair pathway
PW:0000099
part-of
part-of
part-of
chromatin modification/remodeling pathway
PW:0001337
is-a
chromatin remodeling pathway
PW:0001339

Figure 6: Pathway Ontology diagram Screenshot taken from http://rgd.mcw.edu/rgdweb/pathway/pathwayRecord.html?acc_id=PW:0001339 on 02/08/2017 showing the parent nodes of ‘chromatin remodelling pathway’

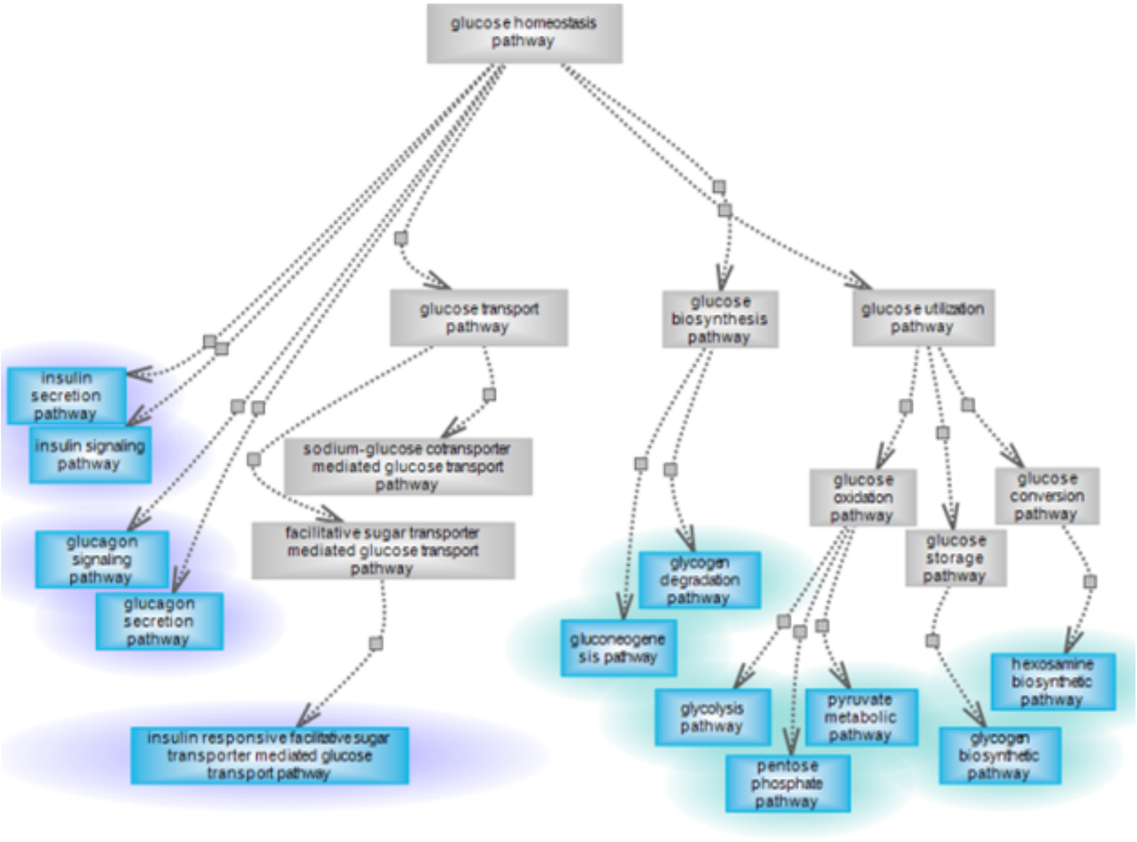

glucose homeostasis pathway
glucose transport pathway
glucose biosynthesis pathway
glucose utilization pathway
insulin secretion pathway
insulin signaling pathway
sodium-glucose cotransporter mediated glucose transport pathway
facilitative sugar transporter mediated glucose transport pathway
glucose oxidation pathway
glucose conversion pathway
glucose storage pathway
glucagon signaling pathway
glucagon secretion pathway
glycogen degradation pathway
gluconeogene sis pathway
hexosamine biosynthetic pathway
glycolysis pathway
pyruvate metabolic pathway
pentose phosphate pathway
glycogen biosynthetic pathway
insulin responsive facilitative sugar transporter mediated glucose transport pathway

Figure 7: Pathway Ontology glucose homeostatis pathway suite. Taken from http://rgd.mcw.edu/wg/pathway/glucose-homeostasis-pathway-suite on 22nd August 2017

A disease pathway is a pathway in which defects in one or more components affect its normal functioning, contributing towards a disease phenotype. The Pathway Ontology provides an alternative system through which known disease pathways can be visualised in their altered state (Petri, Jayaraman, et al. 2014). By linking pathways that are functionally similar, the Pathway Ontology is able to decipher disease mechanisms and view pathway cross talk. Connections between risk factors, pathway perturbation and drug treatments can also be observed (Petri, Hayman, et al. 2014). For example the connections between Bisphenol A, cancer, the anti-estrogen drug tamoxifen and aromatase inhibitors were mapped using the ‘Estrogen Pathway suite’. This collection of estrogen related pathways brings together estrogen signalling, estradiol biosynthesis, biosynthesis of cholesterol, and lipid homeostasis, mapping the biological connections between the risk factor, disease and drug. The pathway ontology differs from the approach used in this project in that it uses known molecular connections to link pathways, making it more reliant on detailed molecular knowledge. Pathways were also clustered on a small number of disease related concepts however; the list is largely limited to cancer processes and does not provide extensive insight into disease.

2.6.2 VisaANT pathway metagraph

Another approach to represent the hierarchical nature of proteins, pathways and modules is by utilising graphs that allow the nesting of nodes within nodes (Hu, Mellor, et al. 2007). Compound graphs gain this capability using metanodes. Metanodes may be represented in two states, an expanded state in which reveals the sub-graph inside the metanode and a contracted state in which the metanode can be treated as a simple node (Hu, Ng, et al. 2007). The subgraphs within metanodes are connected by standard edges (which could represent protein interactions for example). If two nodes that were components of different metanodes were linked by an edge, then a metaedge could be generate to link the two metanodes. Metagraphs are similar to compound graphs, with the additional function of allowing nodes to exist in multiple metanodes, which may or may not be hierarchically nested. This is particularly useful for pathway data since it allows the representation of genes in multiple pathways, as well as functional modules and expression clusters.

VisANT is a multidimentional map, which represents metabolic pathways within a metagraph, to allow exploratory pathway analysis and multi-scale visualisation of multiple pathways (Hu, Mellor, et al. 2007). Within the generated metagraph 'semantic zooming', allows the user to zoom into large pathways, viewing smaller sub-pathways, protein complexes and individual genes. Figure 8 shows three depictions of the same events within G1 phase of the yeast cell cycle, at different levels of the metagraph. In the top left image all metanodes are expanded, to show proteins (pink) nested within protein complexes. The orange complexes are also nested within larger turquoise complexes. Note that the genes *SWI6* and *CDC28* are both present within multiple turquoise complexes. Physical interactions between nodes are represented as edges. In the top right image the orange complexes and in the bottom right images all metanodes have been contracted. Each time a metanode is contracted, edges that had linked it's components to nodes within other metanodes must be generalised to metaedges. Each metaedge inferers a relationships between the complete metanode and the node or metanode adjacent to it. Metanodes were also generated to show the shared components between the protein complexes.The development of metaedges allows the generation of new networks, for example the cellular network of yeast complexes has been mapped based on protein interactions between the components of yeast complexes. A similar map has also been generated based on shared components between complexes.

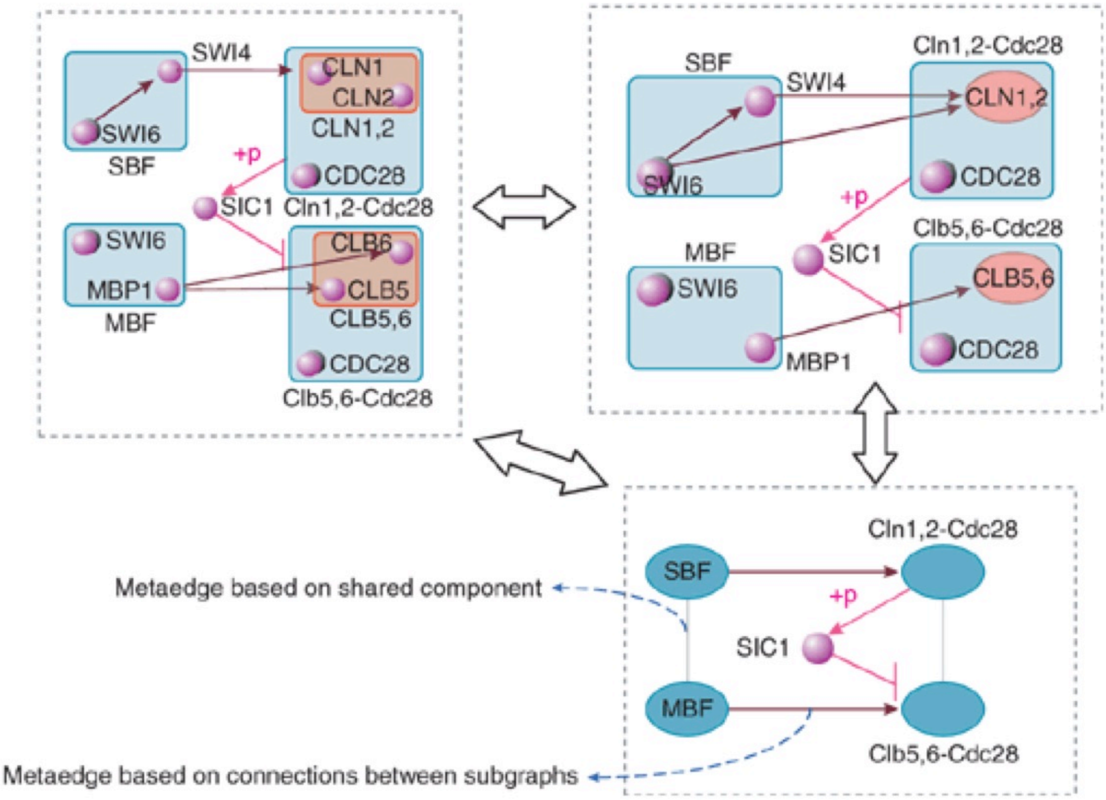


Figure 8: Metagraphs represent multiple network levels. All three depictions show the same graph. All of the metanodes are full expanded in the top left image. The orange protein compound metanodes have been contracted in the top right image. In the bottom image the larger turquoise protein complex metanodes have also been contracted. When metanodes are contracted, metaedges are constructed to replace the previous edges. Links depicting physical interactions between simple nodes are replaced with metaedges linking the metanodes. Additional edges can be constructed to indicate that metanodes contain a shared component *(Hu, Mellor, et al. 2007)*.

Gene Ontology structure can also be visualised within the zoomable multiscale visualisation (Hu et al. 2009). GO functions can in integrated into the metagraph (Figure 9a), by using GO terms as nested metanodes. VisANT is capable of generating clusters from molecular interaction networks (see Section 2.2.2) and representing the generating modules as metanotes. The generation of metaedges also provides some new analysis options (Hu, Mellor, et al. 2007). Edges representing physical interactions between proteins (Figure 9b), are used to generate metaedges between GO terms. This provides a useful tool for clustering functions based on known interactions and showing functional crosstalk. For example Figure 9 shows that genes involved most of the sub-processes of sequence-specific DNA binding physically interact, excepts for genes involved in ribosomal DNA binding. The inverse graph could be produced, where the nodes were pathways and metanodes connect pairs of pathways if any of their component genes share a GO term.

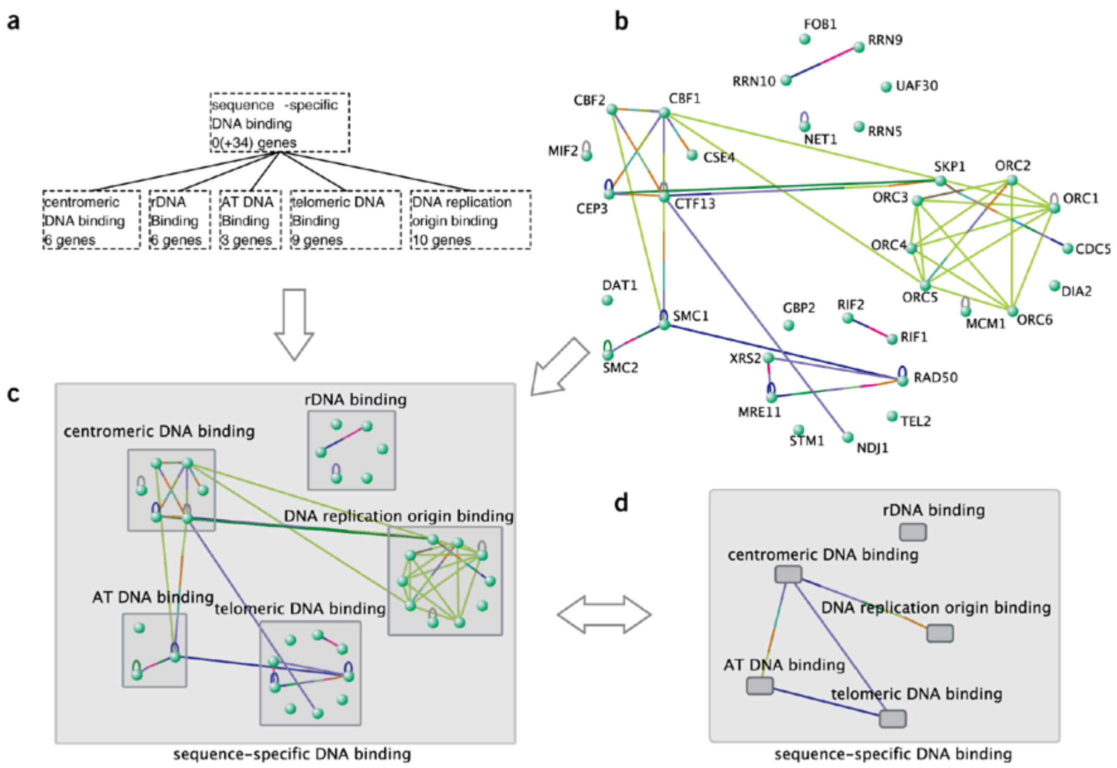


Figure 9: Integrating GO data and PPIs in a metagraph. hierarchy of GO terms covering 34 genes B) PPI interactions between the 34 genes (green circles) C) PPIs of the genes clustered into GO terms, represented hierarchically within the grey boxes D) graph with the metanodes compacted *(Hu, Mellor, et al. 2007)*

The zoomable nature of metagraphs provides interesting options for representing context dependent interactions within pathways and generating hierarchical functional modules (Hu, Mellor, et al. 2007). Generating functional metaedges between pathway metanodes based on overlapping gene annotations, provides an approach superficially similar to the method presented in this thesis, however without the use of enrichment analysis the pathway context of gene function is lost. Positioning a gene within a pathway will ensure that its physical interactions are context dependent (since physical interactions are used to generate the pathways incorporated into VisANT), however it will not effect the gene's GO annotations. For gene annotations to be context dependent, enrichment analysis is required. The use of enrichment analysis assigns a set of GO terms to a pathway that is different to the combined annotations of its' component gene. The generation of this new pathway property violates the standard metagraph structure, which was why it was not used in this project. In addition, linking functions by shared gene interactions, while interesting, is limited by the accuracy and completeness of molecular data (see Section 2.3). Allowing nodes to exist in multiple functional metanodes facilitates pleiotropy within the nodes, however, without knowledge of context dependent interactions, the metaedges linking nodes would also struggle to facilitate context dependent interactions.

Chapter 3

Review of methods and resources

This chapter first reviews the pathway data and gene annotation sources used, then describes the enrichment methods used to assign properties such as function to the pathways. Next, methods to deal with redundancy in the pathway datasets and enrichment results are discussed. Finally, additional software and network analysis methods are outlined.

3.1 Pathway databases

Biological pathways represent collections of genes and interactions depicting a physical component of a biological function. They can include metabolic pathways, genetic regulatory pathways and signalling pathways (Chowbina et al. 2009). Pathway data is distributed across many databases, each with its own biological focus. For example, the Reactome database primarily focuses on signalling pathways (Fabregat et al. 2016), while KEGG is known for metabolic pathways (Kanehisa et al. 2014). The fragmentation of data across multiple sources impedes researchers from performing comprehensive searches utilising all known data (Belinky et al. 2015). Several metadatabases consolidate existing databases into a more comprehensive resource, providing a single searchable format which overcomes many incompatibility issues (Chowbina et al. 2009; Kamburov et al. 2009; Cerami et al. 2011). For example, the range of gene identifiers in current use (HGNC symbols, SwissProt IDs and KEGG IDs).

In addition, different databases may contain pathways with same name but different gene content, due to variable pathway boundaries (Belinky et al. 2015). Databases may also contain pathways with identical content and different names.

Existing resources have so far struggled to overcome the issue of pathway redundancy, which largely arises from the arbitrary nature of pathway boundaries. Overlap between pathways is common, in addition to large pathways subsuming smaller pathways. Vivar et al. (2013) showed that pathway redundancy was present within KEGG, Biocarta, PID, Reactome and the Chemical and Genomic Perturbation database which is a component of the molecular signatures database (Liberzon et al. 2011). This problem increases when databases are combined. The hierarchical redundancy of pathways is visualised within the Pathway Ontology, a directed acyclic graph mapping the tendency of large pathways to subsume smaller pathways (see Section 2.6.1).

ConsensusPathDB (CPDB) is an example of a metadatabase, as it integrates humans, mouse and yeast data from 30 heterogeneous data sources, including KEGG (Kanehisa et al. 2017), WikiPathways (Kutmon et al. 2016), Pathway Interaction Database (PID) (Schaefer et al. 2009), Reactome (Croft et al. 2014) and Biogrid (Chatr-Aryamontri et al. 2015). A recent report showed that of the 161,396 interactions covered by CPDB, 75% were only present in one of the source databases (Kamburov et al. 2013). The number of pathways present in multiple databases decreased rapidly, with very few pathways appearing in more than five databases (see Figure 10). This highlights the highly complementary nature of the databases used. However, the appearance of 25% of interactions in multiple databases still indicates a high level of redundancy.

These findings were confirmed in another study, which examined gene overlap between four large databases: QIAGEN (http://www.qiagen.com/geneglobe/), KEGG, Reactome and Wikipathways (Belinky et al. 2015). Of the 10,770 genes contained within these sources, more than 4,000 were unique to a single database and 1,413 were found in all four databases. Again, these findings show that integrating databases is essential to generate a comprehensive view of cellular pathways; however, overlap merging resources results in high levels of redundancy. Pathway redundancy still remains a major obstacle to database integration.

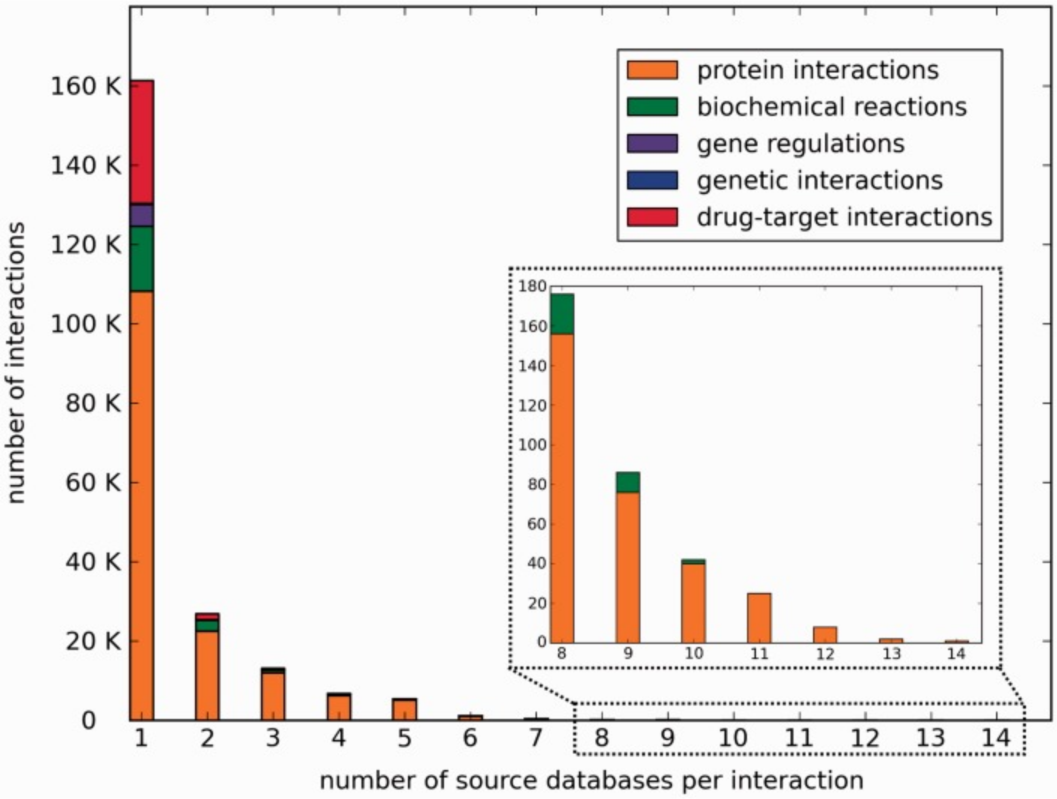


Figure 10: Histogram showing the number of source databases per interaction in ConsensusPathDB *(Kamburov et al. 2013)*

The Human Pathway Database (HPD) integrates pathway data from the PID, Reactome, Biocarter (Nishimura 2001), KEGG and the Protein Lounge (http://www.proteinlounge.com). Within 999 human pathways this data base covers over 59,000 human molecular entities (Chowbina et al. 2009). Users are able to search for pathways related to protein names or search terms, using a maximum overlap threshold to limit redundancy within the results. Overlap between pathway results may also be visualised as a network, representing pathways as nodes and overlapping biological entities as edges.

Figure 11 (taken from Chowbina *et al* 2009) shows overlap between breast cancer pathways. Some pathways show high levels of redundancy, for example, the 'DNA Repair Mechanism' pathway shares 49 biological entities with 'Chks in Checkpoint Regulation', as well as 38 biological entities with 'p53 signalling'. The 'Molecular Mechanisms of Cancer' pathway shares 196 biological entities with 'JAK/STAT Pathway'. A second network may be constructed showing regulatory relationships are present between biological entities in different pathways.

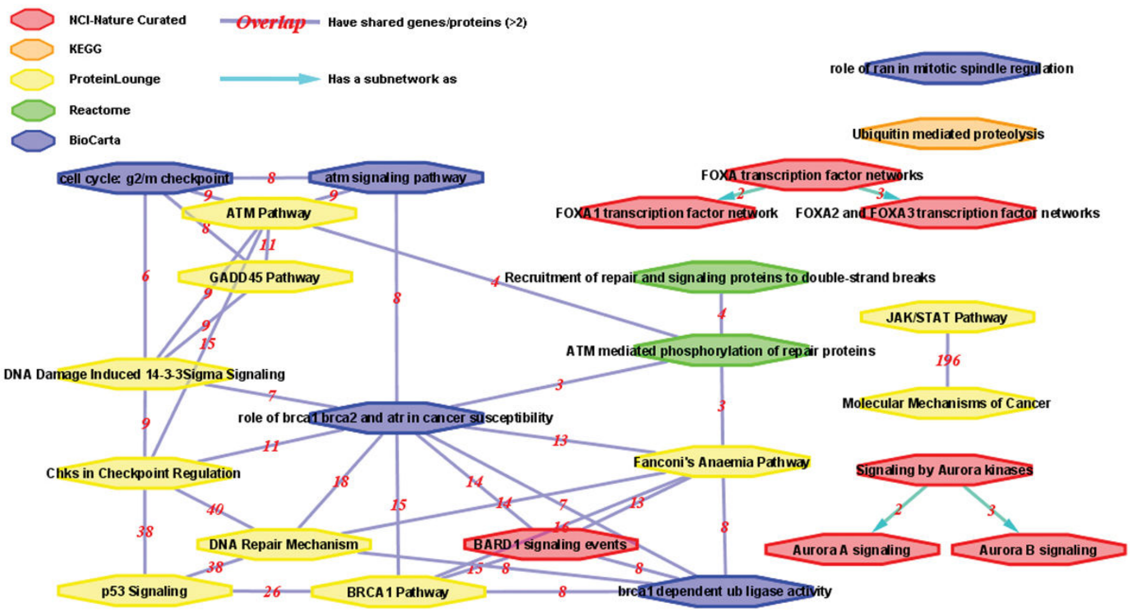

NCI-Nature Curated
KEGG
ProteinLounge
Reactome
BioCarta
Overlap
Have shared genes/proteins (>2)
Has a subnetwork as
role of ran in mitotic spindle regulation
Ubiquitin mediated proteolysis
FOXA transcription factor networks
FOXA1 transcription factor network
FOXA2 and FOXA3 transcription factor networks
cell cycle: g2/m checkpoint
atm signaling pathway
ATM Pathway
GADD45 Pathway
Recruitment of repair and signaling proteins to double-strand breaks
JAK/STAT Pathway
ATM mediated phosphorylation of repair proteins
Molecular Mechanisms of Cancer
DNA Damage Induced 14-3-3Sigma Signaling
role of brca1 brca2 and atr in cancer susceptibility
Chks in Checkpoint Regulation
Fanconi's Anaemia Pathway
Signaling by Aurora kinases
DNA Repair Mechanism
BARD1 signaling events
Aurora A signaling
Aurora B signaling
p53 Signaling
BRCA1 Pathway
brca1 dependent ub ligase activity
8
9
9
8
11
6
4
8
9
9
15
4
196
7
3
3
9
11
13
18
14
14
7
13
15
16
40
38
38
8
13
8
26
8
2
3
2
3

Figure 11: Pathway overlap in the Human Pathway Database *(Chowbina et al. 2009)*. Pathway similarity network showing 25 breast cancer pathways. The node colours indicate the pathway source. Edges indicate overlap, with the number of shared biological entities indicated by the red numbers.

Human integrated Pathway DataBase (hiPathDB) collated data from PID, Reactome, BioCarta and KEGG, to generate 1661 pathways covering 8,976 biological entities. Pathways were intergrated into a single non-redundant superpathway, onto which individual pathway boundaries are mapped (N. Yu et al. 2012). To facilitate the construction of the superpathway, all interactions were reduced to binary pairwise interactions and molecular details of different gene products were lost, resulting in the loss of some contextual data. Reformatted pathways were then mapped into this network.

3.1.1 ConsensusPathDB

In this study, CPDB was used to provide an extensive set of pathways covering a diverse range of biological functions. CPDB was selected as it covers both yeast and human pathways, and includes a wider range of resources than the other meta-databases. Despite the wide coverage of this resource, a disadvantage to it's use is the limited gene coverage, with the CPDB including 2,114 yeast genes and 11,196 human genes within pathways at the time of download (see Section 4.3.1). A complication in the use of metadatabases is the high levels of redundancy through pathway overlap, especially when pathway resources are merged (see previous section). This complication is addressed in Section 3.5.1.

3.2 Annotation Databases

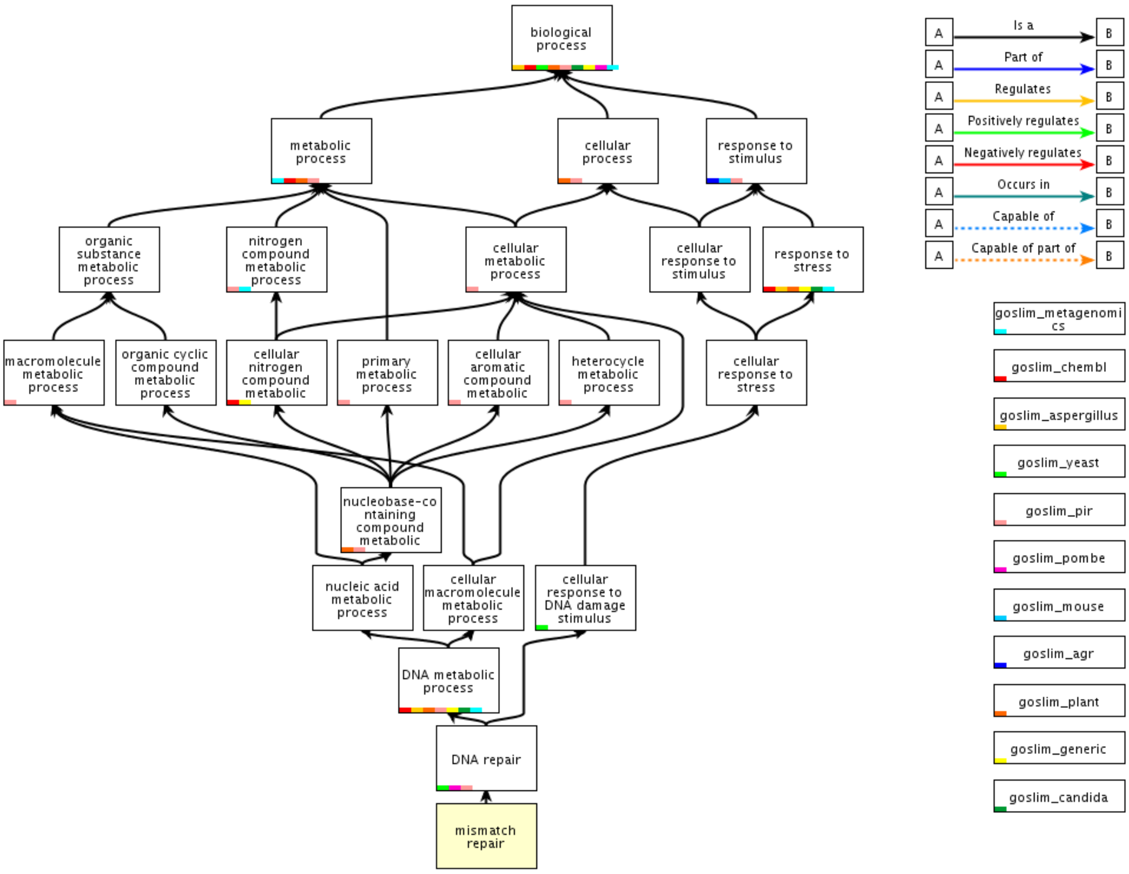
This section reviews the Gene Ontology, which was used to assign function to pathways, and the Human Phenotype Ontology which provided disease annotations. We also discuss Biogrid, which provides the genetic interactions used to validate disease clusters.

3.2.1 The Gene Ontology

The Gene Ontology is a widely used controlled vocabulary to describe the properties of genes (Ashburner et al. 2000). Of the three independent ontologies describing the biological processes, molecular functions and cellular components of genes, the biological process ontology was the most appropriate for this study. The biological processes ontology is the larges ontology containing 27,284 terms (Blake et al. 2015). By performing enrichment analysis each pathway, the pathway's function was ascertained.

The Gene Ontology is a directed acyclic graph, in which parent nodes represent general terms and child nodes give more specific information (see Figure 12). Within the biological function ontology, relationships between terms can be describes as 'is a' or 'part of'. Both of these relationships are transitive; therefore, for an annotation to be assigned to a gene all of the annotation's parent terms must also be applicable (Yon Rhee et al. 2008), a standard sometimes described as the 'true path rule' (Pesquita et al. 2009). The 'is a' relationship is intuitive, describing general terms in increasing levels of detail (for example a mitochondrion is an intracellular organelle is an organelle). The 'part of' relationships describe situations where the specific term describes a subcomponent of the general term (for example a replication fork is part of a chromosome). A chromosome can exist without a replication fork, but a replication fork cannot exist without a chromosome.

The Gene Ontology provides evidence codes for each annotation indicating the type of evidence used to generate the annotation (Yon Rhee et al. 2008). These annotations disclose whether the annotation is based on experimental or computational information, and whether it was manually curated. Annotations 'inferred from electronic annotation' (IEA) are computationally derived, usually through inference from sequence similarity to model organisms (Pesquita et al. 2009), and are not manually curated. The use of IEA can increase the rate of false positives (Mathur & Dinakarpandian 2012). However, removing them reduces coverage as they comprise the majority of annotations (du Plessis et al. 2011). Studies that use the same data sources as those used to generate IEA annotations should omit IEA annotations to avoid circularity (Pesquita et al. 2009). Since physical interactions are the basis of pathways, IEA terms were omitted from the thesis (as described in Section 6.3.1.2). The effects of removing IAE GO terms had little effect on the ability of various semantic similarity measures to distinguish between true and negative PPIs (Jain & Bader 2010), therefore, suggesting that removal of these terms will not effect later methods (see Section 3.4).


biological process
metabolic process
cellular process
response to stimulus
organic substance metabolic process
nitrogen compound metabolic process
cellular metabolic process
cellular response to stimulus
response to stress
macromolecule metabolic process
organic cyclic compound metabolic process
cellular nitrogen compound metabolic
primary metabolic process
cellular aromatic compound metabolic
heterocycle metabolic process
cellular response to stress
nucleobase-containing compound metabolic
nucleic acid metabolic process
cellular macromolecule metabolic process
cellular response to DNA damage stimulus
DNA metabolic process
DNA repair
mismatch repair
A Is a B
A Part of B
A Regulates B
A Positively regulates B
A Negatively regulates B
A Occurs in B
A Capable of B
A Capable of part of B
goslim_metagenomics
goslim_chembl
goslim_aspergillus
goslim_yeast
goslim_pir
goslim_pombe
goslim_mouse
goslim_agr
goslim_plant
goslim_generic
goslim_candida

QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 12 Gene Ontology hierarchy. Nodes represent the parent terms for mismatch repair and edges represent relationships between terms. Figure taken from http://amigo.geneontology.org/amigo/term/GO:0006298 on 18/08/2017.

3.2.2 Human Phenotype Ontology

The Human Phenotype Ontology (HPO) is a widely used resource for describing phenotypic abnormalities (Köhler et al. 2017). It contains >10,000 phenotypic terms (such as 'abnormality of the nervous system') which are linked to 123,724 rare diseases and 132,620 common diseases. The disease annotations are retrieved from Online Mendelian Inheritance in Man (McKusic 2009), Orphanet (Rath et al. 2012) and DECIPHER (Firth et al. 2009). HPO has extremely broad coverage of phenotypes, compared to the Unified Medical Language System which only covers 54% of the concepts included in the HPO, while SNOMED only covers 30%. HPO was used to assign disease annotations to pathways.

3.2.3 BioGrid

The Biological General Repository for Interaction Datasets (BioGRID) is an extensive open-source resource for GI and PPI data (Stark et al. 2006). BioGRID collects and annotates GI data from the published literature for all major model organisms and humans, through text mining and manual annotation. BioGRID releases updated interaction data from yeast on a monthly basis (Chatr-Aryamontri et al. 2015). This thesis used the yeast GI dataset, containing 207,188, interactions covering 5,674 genes. GIs reveal functional relationships between and within regulatory modules. Within this thesis, GIs were used to provide an independent dataset to validate the functional relatedness of genes.

3.3 Annotating pathways through enrichment analysis

This section describes enrichment analysis which is a statistical approach used to identify pathways associated with particular GO terms or diseases.

3.3.1 Functional pathway enrichment

Different biological concepts are used to interpret and describe the unifying concepts of large gene lists (Khatri et al. 2012). Statistical enrichment is used to identify the characteristics common across the gene set that provide functional insight into the list. Over-Representation Analysis (ORA) can be used with biological pathways, the Gene Ontology, or other gene set collections (Tarca et al. 2013). Each concept must have a predefined list of genes, which is compared to the experimenter's gene list. If more genes from the experimenter's gene list are found in the predefined gene list than expected at random, the experimenter's gene list is enriched for the particular concept (Doderer et al. 2012). The significance of the enrichment is quantifiably measured using a p-value. Various algorithms can be used including Chi-square, Fisher's exact test, Binomial probability and Hypergeometric distribution (Huang, Sherman, et al. 2009). It is common for ORA analysis to produce highly redundant results, for example large numbers of closely related GO terms, hindering interpretation by researchers. Therefore, several studies have proposed methods to reduce redundancy in enrichment analysis results (see Sections 3.5.2).

This thesis uses ORA to assign function to pathways and to identify pathways enriched with disease terms. The size of pathways is important for ORA, as pathways that are too large lack functional specificity (Belinky et al. 2015). Also large pathways may have disproportionately high statistical power, resulting in spurious annotations (Glass & Girvan 2014). In addition pathways with fewer than four genes are too small for enrichment analysis (Kamburov et al. 2013) and were not included in the study presented in this thesis (see Sections 4.3.5 and 6.3.1.3). In order to reduce pathway size variability a modified set cover algorithm was used to reduce pathway redundancy (see Section 3.6.2).

3.3.2 Pathway disease enrichment

Analysis of the gene mutations that give rise to disease phenotypes is an essential method for elucidating gene functionality (Robinson et al. 2008). Mutations in functionally similar genes give rise to phenotypically similar diseases which can be mapped to disease families. Phenotypic analysis offers insights into the pathophysiology of cellular networks, by revealing pathways or modules in which perturbation produces similar phenotypic consequences. Information regarding the mechanisms of disease progression can inferred from affected pathways and genes within these pathways can be identified as potential drug targets (Yu et al. 2007). Section 2.5 discusses the mapping of diseases onto interaction networks. Enrichment analysis can also be used to map diseases onto pathways. ORA can also be used to map diseases onto pathways, since diseases cluster within pathways as they interact with partners with similar functions (Liu & Chance 2013).

An alternative approach to the application of ORA to gene lists is to assess which functions are being disrupted to produce a gene phenotype. In these instances, lists of disease-associated genes are compared to genes associated with GO terms. For example, expression data extracted from different brain regions of Alzheimer's patients showed differences in enriched GO terms, corresponding to different pathways (Miller et al. 2013). Similarly expression data extracted from breast cancer samples was also used to identify GO processes associated with tumour metastasis (Yu et al. 2007). The most enriched terms were shown to be different for oestrogen receptor-positive and oestrogen receptor-negative cancers, demonstrating the ability of the method to detect different mechanistic processes. This method provides an interesting alternative to the methods based of molecular data (ORA and network disease modules), however, since it excludes physical cellular components form the analysis, it does not provide direct insight into disease mechanisms. It is therefore less suitable for evaluating the pathway network, which is concerned with organising physical biological pathways in a meaningful way.

3.4 Semantic similarities between GO terms

To measure the functional relatedness of each pair of pathways, the semantic similarity of the GO terms assigned to them must be quantified. This requires the similarity of GO term pairs to be calculated first (see Section 3.4.1), before the similarity of GO terms sets is assessed (see Section 3.4.2).

3.4.1 Semantic similarities between pairs of GO terms

There are various methods available to measure the semantic similarity between GO terms (Resnik 1999; Lin 1998; Wang et al. 2007). The methods available to measure the similarity between a pair of GO terms can be grouped into node-based methods and edge-based methods (Pesquita et al. 2009). Node-based methods consider the properties of the nodes such as the information content. This is a measure of how specific a term is, based on the number of genes it applies to. General terms apply to large numbers of genes, while more specific terms will apply to fewer genes. The information content of the most informative common ancestor is indicative of the similarity of two terms. This common approach is used in methods such as Resnik (Resnik 1999) and Lin (Lin 1998). A similar approach was used to generate semantic distances by subtracting the information content of each test term from the doubled information content of the most informative common ancestor (Hakes et al. 2007). In situations where test terms are semantically close, a semantic distance of around zero would be produced, providing a defined reference point. An alternative to the most informative common ancestor is the disjoint common ancestors used in GraSM (Couto et al. 2005). This measure considers all disjoint common ancestors (the set of common ancestors in which no terms subsume other terms).

Edge-based methods use the structure of the gene ontology to measure the distance between a pair of GO terms on the GO graph (Pesquita et al. 2009). Distance can either be measured as the shortest path between two nodes or the average of all paths, if multiple paths exist. The Wang method is a commonly used edge based method, since it takes the types of edges in the Gene Ontology into consideration (Wang et al. 2007)

Semantic similarities between pathways are not commonly measured. The contrast in the methodological approaches of the Wang and Resnik methods, along with their high performance in other studies (Sevilla et al. 2005; Guo et al. 2006; Mistry & Pavlidis 2008; Guzzi et al. 2012; Jain & Bader 2010; Wang et al. 2007), makes them a good choice to explore the range of available techniques.

3.4.1.1 Resnik method

The Resnik method is an node based approach based on methods based on WordNet (Resnik 1999). It has been found to outperform other approaches in several studies (Sevilla et al. 2005; Guo et al. 2006; Mistry & Pavlidis 2008; Guzzi et al. 2012; Jain & Bader 2010). In this method, similarity is calculated using the information content of the lowest common ancestor (see Equation 1). This measure indicates the probability of two terms sharing a particular parent term (Pesquita et al. 2009; Lord et al. 2003b). For a pair of GO terms (*A, B*), all parent terms common to both are identified, denoted as $T_A \cap T_B$. The proportion of available genes attributed to each parent term is then calculated, shown as *p(t)*. The parent term with the fewest genes attributed to it is selected as the most informative common ancestor. The proportion of available genes annotated with the most informative common ancestor is denoted as P_{ms} (*probability of the minimum subsumer*).

$$P_{ms}(A,B) = \ {}_{t \in T_A \cap T_B}^{min} \{p(t)\} \qquad (1)$$

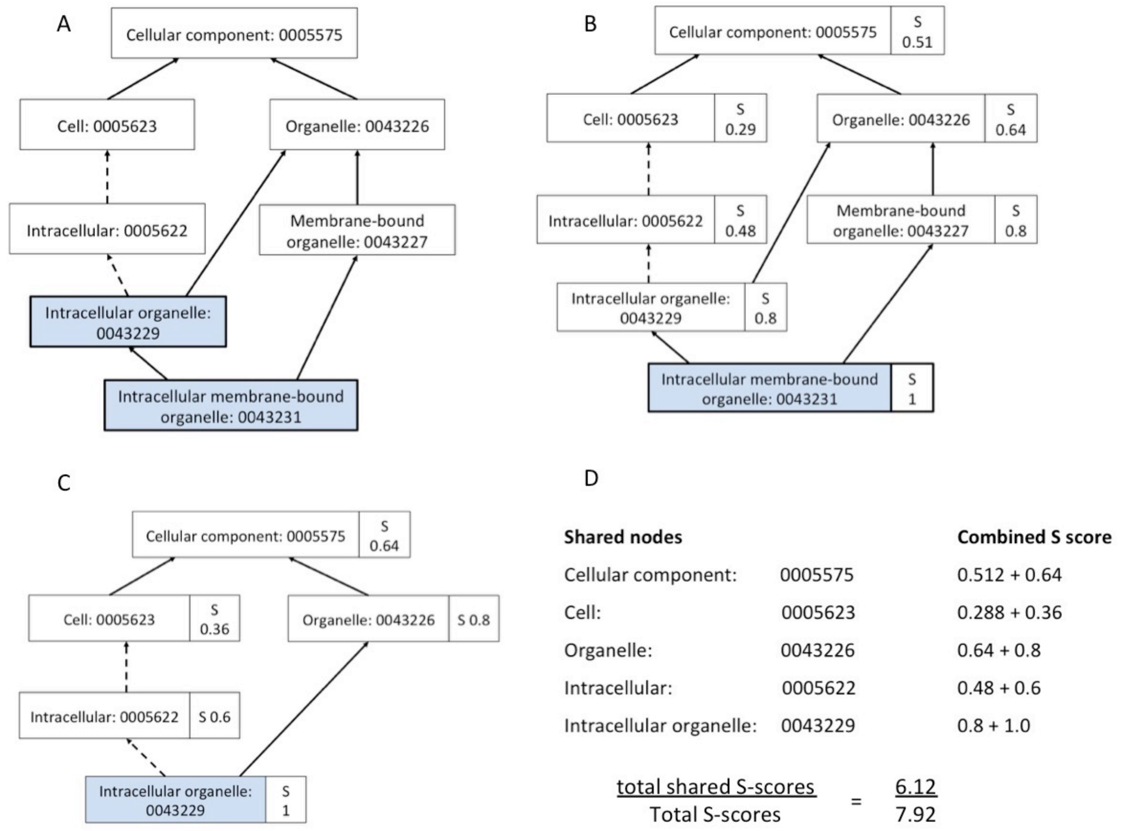
The information content of the most informative common ancestor is generated as the negative log of the probability of the minimum subsumer (Equation 2). This is the score used to describe the similarity between test nodes *A* and *B* (*sim(A,B)*).

$$sim(A, B) = -\log\left(P_{ms}(A, B)\right) \qquad (2)$$

3.4.1.2 Wang method

The Wang method is a commonly used edge-based approach (Wang et al. 2007). It uses the directed acyclic graph structure of the Gene Ontology to measure the overlap between the parent nodes of GO test terms (*A, B*). Parent nodes that are closer to a given term in the graph are considered to contribute more towards the test terms semantics. To illustrate this, Figure 13 shows the calculation of semantic distance between the terms ‘Intracellular organelle’ and ‘Intracellular membrane-bound organelle’. The term ‘organelle’ contributes more to the meaning of both terms than the term ‘cellular component’ and is therefore given greater weight when calculating similarity between the two terms.

To calculate the semantic similarity between the terms ‘Intracellular organelle’ and ‘Intracellular membrane-bound organelle’ (Figure 13A) sub-graphs containing each term and the term’s parents are generated (Figure 13B and Figure 13C). Each GO term is assigned an S-score, independently within each sub-graph, indicating the extent of the parent term’s contribution towards the semantics of the test term. Each test term gets an S-score of 1. The S-score of each parent term is calculated, moving from the bottom of the graph to the top, by multiplying the previous term’s S-score by 0.8 for an ‘is-a’ relationship and 0.6 of a ‘part-of’ relationship (see Figure 13B). In this way, the S-scores become smaller as parent terms become more distant from each test term. To calculate the semantic similarity, the product of the S-scores from all nodes present in the intersection of both sub-graphs is calculated (Figure 13D). This shared score is divided by the total of all of the S-scores in both sub-graphs.


A
Cellular component: 0005575
Cell: 0005623
Organelle: 0043226
Intracellular: 0005622
Membrane-bound organelle: 0043227
Intracellular organelle: 0043229
Intracellular membrane-bound organelle: 0043231
B
Cellular component: 0005575 S 0.51
Cell: 0005623 S 0.29
Organelle: 0043226 S 0.64
Intracellular: 0005622 S 0.48
Membrane-bound organelle: 0043227 S 0.8
Intracellular organelle: 0043229 S 0.8
Intracellular membrane-bound organelle: 0043231 S 1
C
Cellular component: 0005575 S 0.64
Cell: 0005623 S 0.36
Organelle: 0043226 S 0.8
Intracellular: 0005622 S 0.6
Intracellular organelle: 0043229 S 1
D

Shared nodes		Combined S score
Cellular component:	0005575	0.512 + 0.64
Cell:	0005623	0.288 + 0.36
Organelle:	0043226	0.64 + 0.8
Intracellular:	0005622	0.48 + 0.6
Intracellular organelle:	0043229	0.8 + 1.0

$$\frac{\text{total shared S-scores}}{\text{Total S-scores}} = \frac{6.12}{7.92}$$

Figure 13: Wang semantic similarity example. Example of semantic similarities between GO terms taken from *(Wang et al. 2007)*. The dashed arrows indicate a ‘part-of’ relationship between term and the solid arrows indicate an ‘is-a’ relationship. S denotes S-score, signifying how much each term contributes to the test term. Table D gives the S score calculation of each term, calculated by summing the scores for each term indicated an B and C.

Note that in Figure 13B there are two available paths to generate the S-score for ‘cellular component’. The highest score was generated using the path: ‘Intracellular membrane-bound organelle’, ‘Intracellular organelle’, ‘Organelle’. A lower score of 0.23 could have been generated using the alternative path: ‘Intracellular membrane-bound organelle', 'Intracellular organelle', 'Intracellular', 'Cell'. The path that generates the highest score will be selected, to contribute to the overall similarity.

The Wang method is presented in Equation 3 in which T_A and T_B are once again the parent terms of *A* and *B* respectively. $S_A(t)$ and $S_B(t)$ are the S-scores of parent terms related to *A* and *B*, and *SV(A)* and *SV(B)* are the combined scores of the parent subgraphs generated for *A* and *B* respectively.

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B}(S_A(t) + S_B(t))}{SV(A) + SV(B)} \qquad (3)$$

3.4.1.3 Comparison of semantic similarity approaches

The Resnik method has been shown to be highly effective; however it does have some notable limitations. The specificity of a GO term is determined by its position in the graph and its semantic meaning is determined by all of its ancestors (Wang et al. 2007). If two test terms share a parent term close to the ontology root, they should have a greater semantic similarity if each test term is also close to the root than if both test terms were positioned close to the ontology tips. However, in both described cases, the similarity score would be the same, since the distance between the test terms and the lowest common ancestor is not considered. In addition, test terms close to the root of the ontology will always receive a low similarity score, since the information content of the most informative common ancestor will always be low. Solutions to these limitation are presented by the Jiang method (Jiang & Conrath 1997) and Lin method (Lin 1998), both of which consider the information content of the minimum subsumer and the information content of the test terms. However, because these methods consider either absolute difference or the ratio of these measures, if the test terms are close to the ontology root the semantic similarity will consistently appear very high. This is due to the similar information content of shallow annotations, which are less meaningful than if the test terms and shared ancestor had been closer to the root. This can be problematic in the context of gene similarity as poorly annotated genes which are associated with such general terms can be very functionally different (Sevilla et al. 2005). An additional limitation of using information content is that it makes the output highly depend on the data set. If two studies use different sets of annotated genes, they will generate different results, undermining the objectivity of the method (Wang et al. 2007).

The limitations of node based methods provided motivation for the development of the Wang method. By using the Gene Ontology topology rather than information content, topological distance between the test terms is incorporated into the similarity measures (Wang et al. 2007). The output is also unaffected by the proximity of the test terms to the ontology root.

Limitations of edge-based methods are that they tend to assume that nodes and edges in the Gene Ontology are uniformly distributed, and that all edges imply comparable differences in semantic similarity (Pesquita et al. 2009). In the case of the Gene Ontology, this assumption is undermined by the tendency of study bias to affect the density of areas of the ontology graph. Node-based methods are not affected by the structure of the Gene Ontology, however higher numbers of genes are likely to be assigned GO terms in heavily studied functional areas than less studied functions, which in turn will affect information content scoring.

3.4.2 Semantic similarity between groups of GO terms

The methods described in sections 3.4.1 to 3.4.1.3 measure semantic distance between pairs of GO terms. However, enrichment analysis produces sets of GO terms rather than single terms. Semantic similarity methods are most commonly measured between sets of GO terms attributed to pairs of proteins (Resnik 1999; Wang et al. 2007; Xu et al. 2008) , however they can be applied to pairs of annotated pathways (Mathur & Dinakarpandian 2012). There are three groups of methods used to compare sets of GO terms: the 'pairwise average/maximum', the 'best match average' and groupwise methods.

3.4.2.1 Pairwise average/maximum

The pairwise average calculates the mean semantic distance between every pair of GO terms across two sets (Figure 14A). The first step in the calculation is to measure the semantic distance between each GO term in the first set and every GO term in the second set. The mean of all the semantic distances calculated is then generated. Lord et al. 2003a showed that the pairwise average in conjunction with the Resnik method produced results correlated with sequence similarity.

The pairwise maximum approach links sets of GO terms by their most similar GO term pair (Figure 14B). This approach has been used with several semantic similarity measures, including the Resnik measure, and has been shown to correlate the semantic similarity of gene pairs and shared gene expression (Sevilla et al. 2005; Jain & Bader 2010).

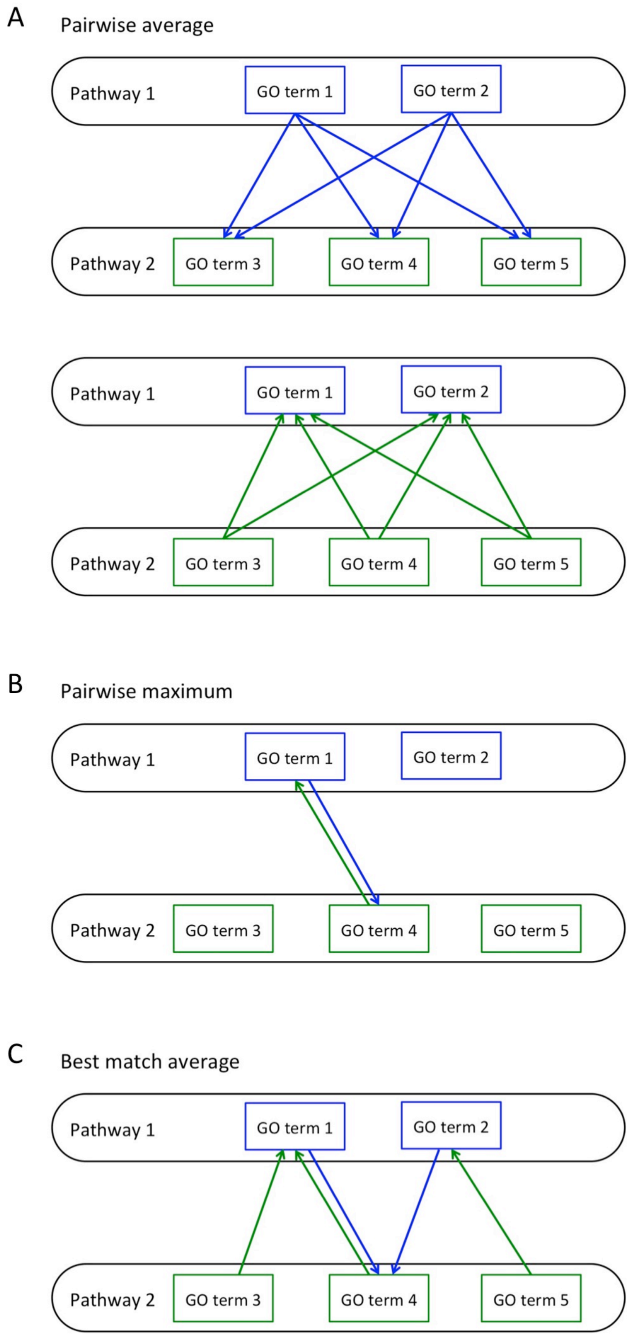


Figure 14: Semantic similarity measures. Arrows show the similarity measures incorporated into each metric

The pairwise average and the pairwise maximum vary in the way they facilitate gene pleiotropy. If the gene sets being compared only contain one GO term, or contain semantically similar GO terms, they will produce similar results. However, if genes have semantically variable terms, these methods will produce different results. If a protein has multiple context-dependent functions, then using the pairwise maximum approach can be beneficial, if only the most similar functions are relevant. However, using the pairwise maximum approach discards information since GO terms that are not included in the best match are ignored. Lord *et al*. 2003a argued that the function of a biological entity is comprised from all of its functions; therefore basing semantic similarities of a single function is detrimental. Using the pairwise average, all GO terms in both sets are considered in the calculation. Using the pairwise maximum genes are linked by their most similar pair of annotations allowing unrelated pleiotropic terms to be discarded. Since the study presented in this thesis utilises enriched GO term sets from biological pathways the pairwise average was selected as most appropriate. Functions attributed to pathways through enrichment analysis are by their nature context-dependent, therefore, any pleiotropic functions their proteins may have had will have been filtered out. However, the pathways themselves may have multiple functions (Guo et al. 2006), therefore effectiveness of both the pairwise average and the best match average (see below) were measured.

3.4.2.2 Best match average

The best match average calculates the average semantic distance between each GO term in one set and the closest GO term in the other set (Pesquita et al. 2009). Therefore, each GO each term will be matched to the most similar term in the other GO set (

Figure 14C). The approach combines properties of the previous two techniques, linking GO terms by their closest partner without discarding any GO terms in the set. This approach is especially suited to situations in which biological entities, such as proteins or pathways, are attributed with pairs of semantically diverse GO terms, which may cluster within the cell. If a pair of genes each had an identical set of two semantically distinct annotations, the best match average measure would assign them a semantic similarity of one. The pairwise average would assign a low semantic similarity because it would include distances between semantically diverse pairs. The pairwise maximum would also assign a semantic similarity of one, however, it would be unable to distinguish between two identical sets of semantically diverse GO terms and two sets of GO terms with one matching pair. The best match average was the approach originally suggested for use with the Wang method (Wang et al. 2007). The best match average has therefore been used in many studies (Jain & Bader 2010; Couto et al. 2005; Lord et al. 2003b) and has been shown to outperform the pairwise average (Pesquita et al. 2007).

3.4.2.3 Groupwise methods

Groupwise methods do not consider similarities between individual GO terms but instead look at the proportion of shared GO terms shared between biological entities. These approaches have different implementation to the previous approaches but show conceptual overlap. Set similarity or graph techniques, based on the entire sets of gene annotations may be used (Pesquita et al. 2009). An early measure was term overlap, which simply defined the similarity between two gene products as the number of GO terms they shared (H. K. Lee et al. 2004). The Jaccard coefficient provides a normalised version of this method (Gentleman 2005). In this approach, the Jaccard coefficient *J(A,B)* is calculated as the number of terms shared between two sets (A and B), divided by the total set of terms in both sets (see Equation 4).

$$J(A,B) = \frac{|A \cap B|}{|A \ \cup B|} \tag{4}$$

Other approaches such as simLP measure the depth of the longest shared path from the ontology root (Gentleman 2005), sharing base concepts with edge based techniques. Information content can also be incorporated into this set of techniques, giving similar results to Resnik's measure (Sheehan et al. 2008).

The disadvantage of these methods is that they are greatly dependent on the annotation density of the Gene Ontology. Well-studied areas of the Gene Ontology have more functionally similar terms than less studied areas. Therefore, two functionally similar genes could have a low number of shared terms if they were positioned in an understudied area of the Gene Ontology. In contrast, two dissimilar genes could share high numbers of very general terms from a better-studied area of the ontology. As a result the pathway network topology could be shaped by current trends in research rather than cellular properties.

3.5 Existing methods to reduce pathway data redundancy

The arbitrary nature of pathway boundaries results in high levels of redundancy within pathway datasets, particularly when multiple databases are combined (see Section 3.1). To address this issue and produce a non-redundant set of pathways capable of representing human biological function, methods for reducing redundancy were explored. We also discuss methods to reduce redundancy within pathway and GO enrichment results.

3.5.1 Reducing pathway redundancy

Various methods to reduce pathway redundancy by merging pathways have been developed (Vivar et al. 2013; Belinky et al. 2015). Pathways are treated as sets and therefore when pathways are merged overlapping genes are only represented once within the new pathway.

A simple approach to reducing pathway redundancy is to merge pairs of pathways if they overlap by more than a given threshold. This approach is used by Redundancy Control in Pathway Databases (ReCiPa), which allows users to define the 'max_overlap' threshold (Vivar et al. 2013). They demonstrated that merging pathways reduced pathway redundancy and made the refined data set more suitable for pathway enrichment analysis.

Pathway distiller (Doderer et al. 2012) provides a hierarchical clustering approach for merging data sets based on gene membership, GO similarity or shared protein interactions. The Jaccard coefficient is used to map similarity between each pair of pathways based on shared genes or interactions, or between GoSlim terms assigned to genes. Hierarchical clustering is used to generate the number of pathway clusters defined by the user (see Figure 15). As seen in Figure 15 hierarchical clustering can lead to large numbers of pathways being clustered into a single group, which can result in detrimentally large superpathways.

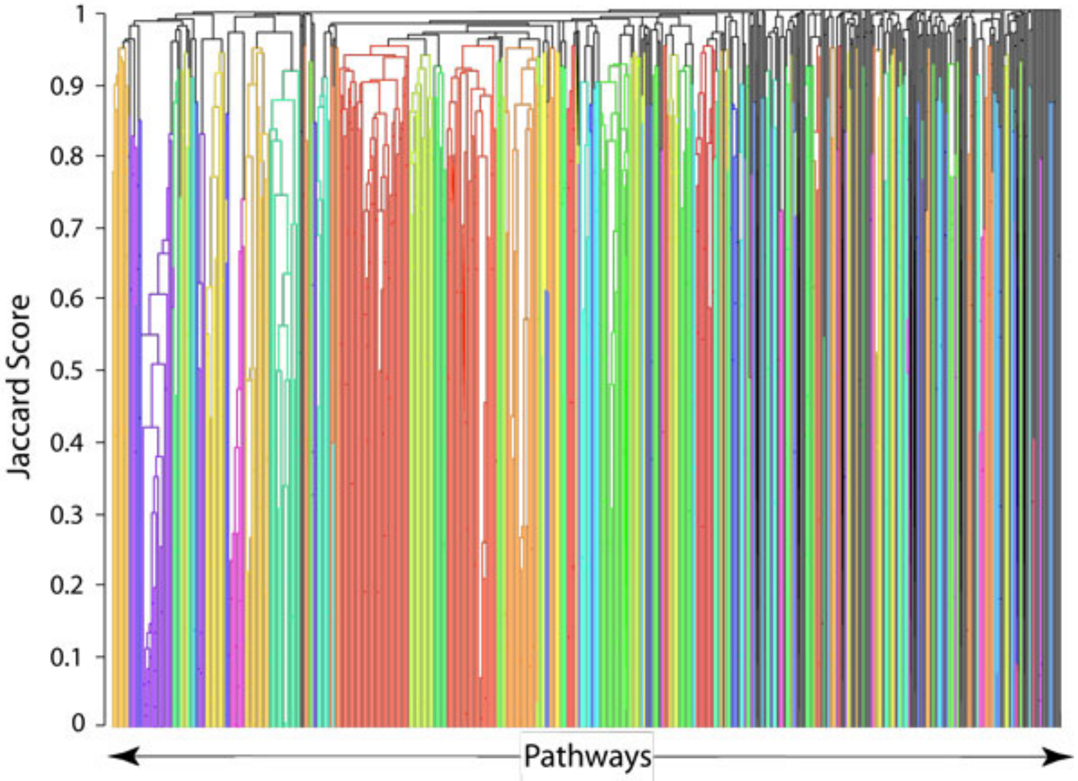


Figure 15 Pathway Distiller hierarchical clustering of pathways. Semantic similarities were generated between 2,462 pathways, which were clustered into 250 groups. Colours show the hierarchical clustering *(Doderer et al. 2012)*.

PathCards used a multistep procedure to reduce pathway redundancy, while protecting pathway informativeness by preventing pathways from becoming too large (Belinky et al. 2015). Small pathways (<200 genes) were first merged with larger pathways. Two thresholds for pathway merging were generated: a high threshold for the initial clustering, then a lower threshold to increase network cohesiveness. Hierarchical clustering was performed first, merging hierarchical pathway pairs with a Jaccard coefficient higher than the first threshold. A neighbour joining approach was then used to join each pathway to the adjacent pathway with the highest Jaccard coefficient, provided that it was higher than the lower threshold. Belinky et al (2015) intergrated twelve pathway sources including KEGG, Reactome, WikiPathways, and PharmGKB (Hewett et al. 2002) into 3,215 pathways, containing 11,478 genes. This allowed a notable reduction in redundancy with a modest increase in pathway sizes.

The presented methods have a number of issues. A limitation of the ReCiPa is that overlap between pathways can only be removed if it exceeds the given threshold. The threshold must be high enough to remove sufficient overlap, but must not be so high that pathway merging reduces functional specificity. This is a problem if pathway overlap is not uniform across the data set. Heavily studied areas of biology may have many overlapping pathways, requiring a higher threshold to prevent high numbers of pathways being merged into a “giant” uninformative set. These high thresholds would however prevent redundancy from being removed from the rest of the data set. PathCards addressed this limitation by having two thresholds. The higher threshold merges pathway pairs with very high overlap. This threshold can be set high as the second threshold is available to reduce lower level overlap. The second threshold is used in conjunction with neighbour joining allowing each pathway to merge with only one other pathway (unless two pathway pairs show equal overlap). This avoids large numbers of pathways becoming conjoined into giant superpathways. It is, however, still possible for giant superpathways to form if multiple pathways are to merge with a single pathway or if pathways merge in long chains.

A further limitation is that merging algorithms can only account for redundancy between pathway pairs. It is possible for all of the genes in a pathway to be covered by two or more pathways; however, if individual pathway pairs do not exceed the threshold(s), this redundancy cannot be removed.

Finally, when using the merged pathways produced by these methods, it is important to consider that these approaches have altered the original data. By merging pathways from multiple resources into superpathways, PathCards suggested many new molecular interactions. They demonstrated that many of these newly generated gene-gene connections are supported by high numbers of literature co-mentions compared to random gene pairs. Gene pairs in superpathways were also significantly more likely to be associated with protein-protein interactions than random pairs. However, literature co-mentions and experimental interaction scores did not support the majority of the new interactions. Furthermore existing knowledge regarding the unaltered pathways cannot be assumed to apply to the merged pathway. Also, gene overlap between pathways does not guarantee that they will be active in the same cellular context. For these reasons the study presented in this thesis chose to reduce redundancy without merging pathways (see Section 3.6).

3.5.2 Reducing redundancy from pathway enrichment results

The results of enrichment analysis are also prone to high levels of redundancy, if gene sets or pathways return large numbers of highly similar GO terms. Pathway Distiller offers two algorithms designed to merge pathways from enrichment analysis output (Doderer et al. 2012). The first algorithm clusters pathways based on the subset of differentially expressed/user defined genes that they contain. This approach uses iterative rounds of enrichment analysis, removing the differentially expressed genes in the most enriched pathway each time. The pathways that drop out of the enrichment output in each round are merged. This interesting approach avoids the use of thresholds, although it necessitates repeated enrichment testing. The second approach measures gene overlap and predicted interactions between the genes of a pathway pair. Clusters are formed using a neighbour joining approach to link each pathway to the pathway most similar to it. Again, this avoids thresholds but does not guarantee comparable levels of similarity between clusters, since even highly distinct pathways must be linked to a partner pathway. These approaches were used to reduce the number of pathways retrieved by enrichment analysis from over 1000 to between 12 and 318, depending on the methods and parameters used. This makes result analysis significantly easier. The merged pathways showed large increases in PPI interactions between gene members compared to randomised gene groupings.

Pathway Distiller also designed an algorithm capable of removing redundancy from entire pathway sets (see Section 3.5.1). The advantage of removing redundancy prior to enrichment analysis is that pathway enrichment methods assume independence between pathways. Pathway overlap results in interdependency and/or redundancy capable of skewing statistical assessment leading to incorrect pathway associations. ReCipa also demonstrated that reducing pathway redundancy lead to enhanced results within enrichment analysis (Vivar et al. 2013). They used pathway merging to reduce redundancy from the Molecular Signatures Database (MSigDB) and demonstrated that pathways ranked based on their association with obesity showed higher disease significance and less overlap when the reduced redundancy dataset was used.

3.5.3 Reducing redundancy from Gene Ontology enrichment results

Two algorithms, *elim* and *weight,* have previously been introduced to deal with the issue of redundancy in enrichment results (Alexa et al. 2006). The *elim* method acts on the principle that GO terms are less specific than their children and more specific that their ancestors. The most specific GO terms available to describe the genes in a set are preferentially selected based on their position within the GO topology. More general ancestor terms are only selected if they are required to cover all the enriched genes in the set. Alexa et al. (2006) acknowledged that in some instances parent terms have higher p-values than their child nodes and therefore introduced the *weight* method. In this method if an ancestor node has higher significance than a child node, then the significance of the child nodes will be decreased, preventing the child nodes from being reported.

The *weight* method is likely to produce results similar to the set cover for enrichment algorithm. Both select the most significant terms available to describe the enriched genes, however the set cover for enrichment algorithm does not take the topology of the GO into account. The gene sets covered by each term serve as a proxy for the ontology, based on the true path rule (see Section 3.2.1). In instances were gene sets are enriched with functionally diverse GO terms the weight method would apply both terms and the set cover enrichment algorithm would only apply the most significantly enriched. Therefore the *weight* method is better equipped to detect multifunctionality but is also more likely to detect spurious results. The algorithm used in the *elim* method is similar to the set cover for enrichment algorithm however they rank GO terms based on GO topology and enrichment significance respectively. The *weight* method outperformed the *elim* method, since the *elim* method has a higher risk of discarding relevant terms. Despite this finding they suggested that the *elim* method should be preferred given its simplicity.

3.6 Set cover and set packing

In this section an explanation of the algorithms that will be used to reduce redundancy within the pathway and GO enrichment data is provided.

3.6.1 Combinatorial set problems

Set cover and set packing algorithms are abstract, combinatorial optimisation problems (Kordalewski 2013). Sets represent groupings of unordered items, in which items may only be represented once. Set cover and set packing deal with sets of sets (*I*), which in our study represent the pathway datasets. Each set (s_i) contains elements, representing genes within each pathway. The set of all of the elements is collectively known as the universe, defined as the union of all the sets $\mathbb{U} = \cup_{i \in I} s_i$. Each set has an associated value v_i. The aim of set cover and set packing is to identify the selection of sets that cover the all the elements universe. The universe has been covered by subset H if $\cup_{i \in H} = \mathbb{U}$.

Another way to describe these problem is the following: given a set $\mathbb{U}$ of n elements and a collection of $S_1 \ldots S_m$ of subsets of $\mathbb{U}$, find the minimal collection of sets whose union is equal to $\mathbb{U}$ (Klienberg & Tardos 2003). In some instances, there may be a specified maximum number of sets, within which it is required to cover the universe or maximum cost.

Set cover and set packing are NP-hard problems and therefore they are typically approached using heuristic algorithms. This means that rather than finding the exact solution, which is not possible in polynomial time, an acceptable solution is found within a reasonable time frame (Kordalewski 2013). This is often done using greedy algorithms which generate solutions through a series of decisions, making each decision by selecting the lowest cost available at each step (Klienberg & Tardos 2003). The solution produced may not be unique and some degree of randomness is unavoidable, as explained in Section 3.6.6. Each method is described in more detail within the following sections.

3.6.2 Set cover

The aim of set cover is to generate the smallest subset H, that covers all the elements in the universe $\cup_{i \in H} = \mathbb{U}$ (Kordalewski 2013). Set overlap is permitted although it should ideally be minimal. The smallest subset is typically defined as the smallest number of sets (Skiena 2008), but could refer to the smallest total number of elements. A commonly applied greedy algorithm is to iteratively select the set with the largest number of uncovered elements $v_i = |s_i \cap R|$, where R represents elements that have not already been covered (Figure 16). Weights can be added, as costs or values for each set, to influence set selection (see Section 3.6.3).

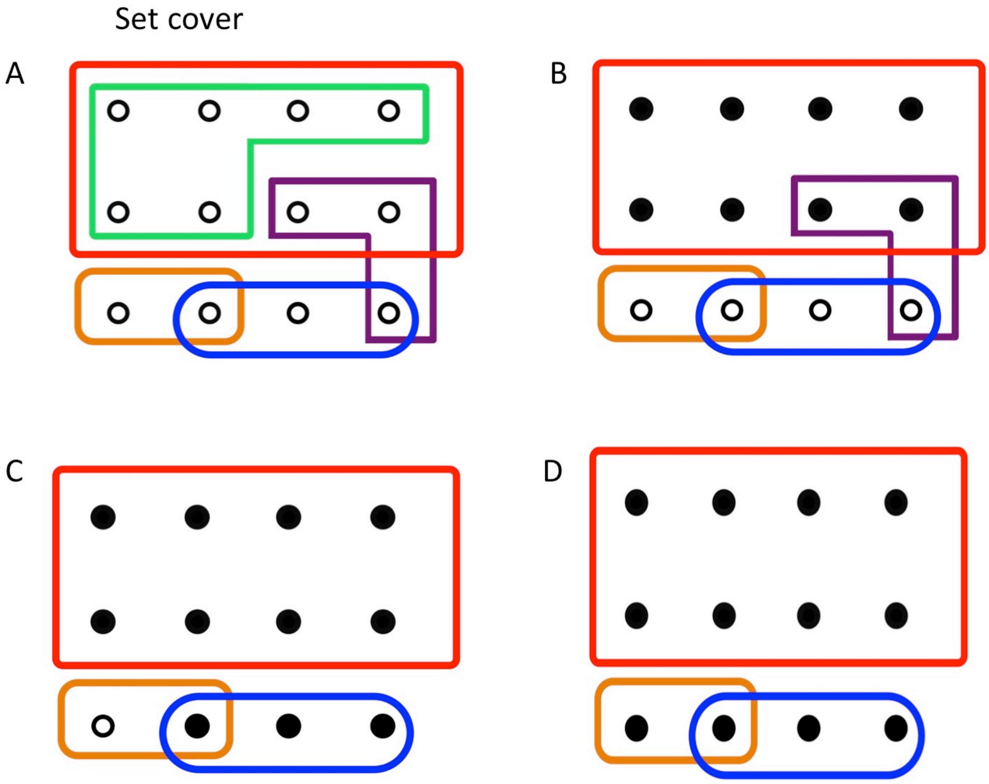


Figure 16: Set coverIs a simple set of overlapping sets B) The red set with 8 uncovered elements is selected first C) the blue set with 3 elements is selected second D) The orange set covers all the elements in the universe.

Set cover was used in other biomedical areas such as the assembly of RNA splice variants into full length transcripts (Song & Florea 2013). The tool CLASS (Constraint-based Local Assembly and Selection of Splice Variants) assembles transcripts from short RNA-seq reads to a reference genome. In this situation, the exons are the sets and the transcripts corresponding to each exon are the elements. The aim is to find the best combination of exons that can explain all of the splice variants. Set cover is suitable for this problem since the ability of sets to overlap reflects the ability of transcripts to cover multiple exons.

3.6.3 Hitting Set

The set cover problem can be reformulated into a bipartite graph problem, known as the hitting set, in which the sets represent one class of nodes and the elements represent a second class of nodes (Figure 17). The set nodes are linked to all the element edges that they contain. The aim is to generate the smallest subset of set nodes, which are connected to all element nodes.

In this representation, it is clear that some elements are covered by many sets, while other elements are covered by fewer sets. Since all the elements must be covered, sets that include elements that are not covered by many other sets should be assigned a higher value and preferentially selected. The value of a set is calculated as the sum of the value of its' elements. Elements in fewer pathways have more value and elements in more pathways have less value. Once an element has been covered by a set, there is nothing to be gained by covering it again in additional sets; therefore, the value of elements that have been covered previously is not added to the set value.

The hitting set algorithm has been applied to find a minimum number of transcription factors capable of covering a set of differentially expressed genes (Lu & Lu 2013). A bipartite graph was generated linking transcription factors to the set of genes they control and the minimum number of transcription factors needed to control each gene was varied from one to four, revealing cooperative transcription factor sets.

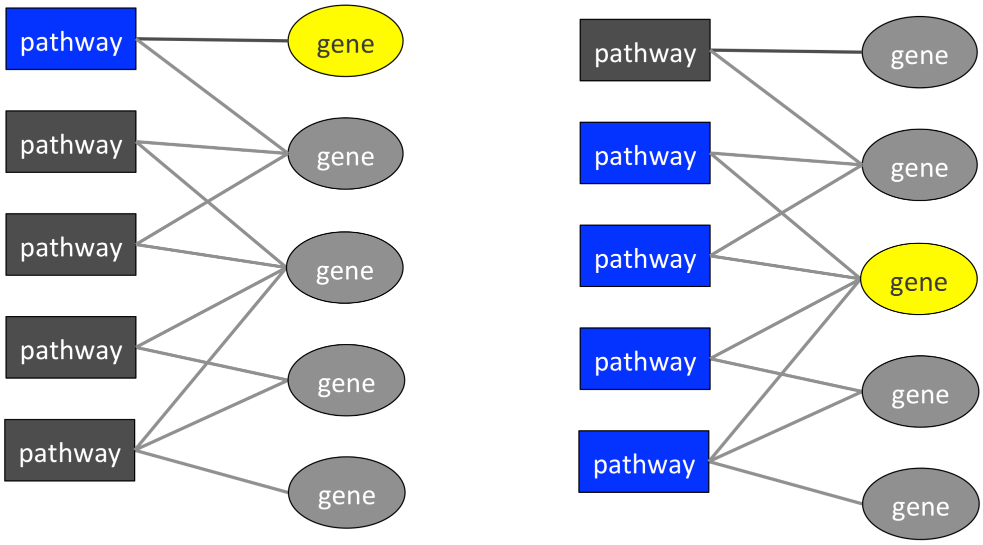


Figure 17: Hitting set bipartite graph to demonstrate covering gene sets using pathway

3.6.4 Set packing

Set packing aims to find a subset of disjoint sets that cover all of the elements in $\mathbb{U}$. Set packing disallows overlap between the selected sets, making it distinct from set cover (Klienberg & Tardos 2003). As a result, in some instances it may not be possible to generate a packing that includes all of the elements in the universe, especially when using heuristic methods. A heuristic method commonly used to perform set packing is explained in terms of an equivalent problem maximum independent set in Section 3.6.5.

A classic application of set packing is to generate schedules. For example, in a hospital setting set packing is used to allocate resources such as operation rooms and staff (Velásquez & Melo 2005). Resource use should be maximised, but no resources overlap (staff members cannot be in two places at once, rooms cannot be booked for two operations). In this situation, surgeries are the sets and resources are the elements.

3.6.5 Maximum independent set

The set packing problem is equivalent to a graph theory problem of maximum independent set (Kordalewski 2013). Given a graph *G = (V,E)*, nodes (*V*) are equivalent to sets. Edges are formed between nodes (*i,j*) if the corresponding sets have elements in common ($s_i \cap s_j \neq \emptyset$). An independent set is a set of nodes in which no pair of nodes are connected by an edge (Klienberg & Tardos 2003). The maximum independent set is the largest independent set that can be generated from the graph.

2. A common greedy heuristic for weighted maximum independent sets is to select the node with the lowest number of adjacent edges (Kordalewski 2013). Figure 18 shows the relationship between maximum independent sets and set packing. In weighted maximum independent sets, the heuristic is adjusted to include the

number of uncovered elements associated with each node. The value of a node would be calculated as

$$v_i = \frac{1}{\{j \in I|s_j \cap s_i \neq \emptyset\}|}| \qquad (5)$$

where each pathway's value (v_i) is calculated as the inverse of the number of overlapping pathways it has in the dataset. v_i is the value of pathway, *I* is the dataset, $\boldsymbol{s_i}$ is the pathway whose score you're calculating and $\boldsymbol{s_j}$ is each other pathway (Kordalewski 2013). Any sets that have elements in common with the selected set are then removed from the possible uncovered sets *R*, since overlap is not permitted. The value of all remaining sets is recalculated, and then the process is continued until only disjoint sets remain (Figure 18).

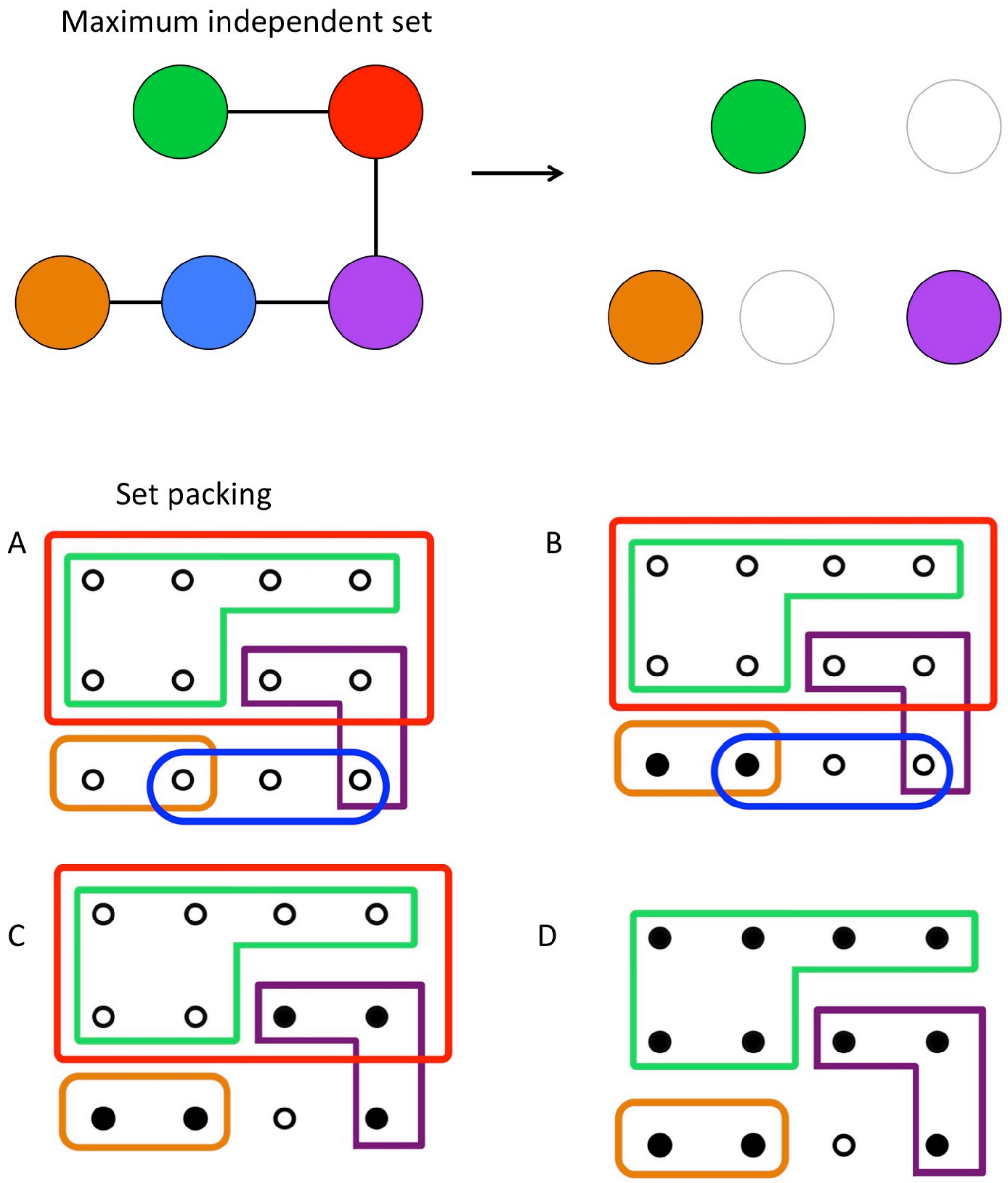


Figure 18: Relationship between maximum independent sets and set packing. Maximum independent set shows how removing the nodes with the highest degree generates an independent set of nodes. Set packing A) Is an simple set of overlapping sets B) The orange set, that overlaps with one other set, is selected first C) the blue set is removed because it overlapped with the orange set. The purple set, which overlaps with one other set, is next D) the red set is removed because it overlapped with the purple set. The green set covers all remaining available elements.

3.6.6 Randomness within set cover/packing solutions

Greedy algorithms generate solutions through a series of decisions, where each decision affects the input of the next. Weights or costs are used as the basis of each decision, however it is possible for multiple sets to have equal values. In these situations, an arbitrary decision is made that may affect the end results. Therefore, it is possible that multiple solutions for a given data set may be possible. In these situations, it is possible to use randomisation procedures to explore the variation possible and select the best outcome (Kordalewski 2013). This was not done within this thesis due, since all set cover and set packing outputs reduce dataset redundancy and are therefore considered acceptable for use.

3.7 DAVID gene functional classification tool

The DAVID Gene Functional Classification Tool clusters genes by function without using networks providing and alterative method for generating functional models (D. W. Huang et al. 2007), which is not dependent on molecular data. Designed to handle lists of up to 3,000 genes from high throughput genomic or proteomic studies, DAVID clusters genes into functionally related groups based on shared GO annotations. A fuzzy agglomeration method was developed to allow genes to participate in more than one functional group, allowing the method to incorporate gene pleiotropy. Genes associated with multiple terms can be visualised. Omitting any genes that do not fit easy into a cluster, such as orphan genes, enhances the functional heterogeneity of clusters. The algorithm is able to place genes from well defined protein families into appropriate clusters with 98-99% sensitivity and 95-99% specificity. Functional groups are ranked base on the group members' participation in the enriched term.

DAVID provides a valuable resource for generating functional clusters, however it is designed to deal with enriched gene lists rather than global gene lists, therefore the number of genes that can be analysed is restricted(Huang, Lempicki, et al. 2009). In addition it assumes that all genes in the set show differential transcription or translation under a particular set of conditions, which is not the case in global gene sets.

The clustering of enrichment output results is enables users to comprehensively pool all terms related to a single biological concept, without manually summarizing related terms. This prevents the functional diversity of enriched terms being lost, due to large numbers of highly significant redundant terms. Since most tools do not emphasise the inter-relationships between biological terms approaches to deal with this redundancy are required. In a study examining expression changes in peripheral blood mononuclear cells induced by the introduction of an HIV envelope protein, identified 16 functional groups from hundreds of enriched terms (D. W. Huang et al. 2007). Clearer, more concise results are produced by condensing redundant terms and summarizing interrelationships.

3.8 Network statistics

This section outlines network analysis approaches used to measure distances between nodes and describe node centrality within the network. We also review neighbour joining, which is used to show the relative similarity between sets.

3.8.1 Shortest paths

The shortest path is a graph measure used to indicate how connected node pairs are within a network (Newman 2001). The length of the shortest possible route between a pair of nodes is calculated. In a weighted network the edge weights are summed to calculate the distance between the node pairs. Within the hypothesis of network medicine the network parsimony principle states that the molecular pathway responsible for a disease phenotype can often be identified as the shortest molecular path between known disease components (Barabasi et al. 2011).

3.8.2 Betweenness centrality

Betweenness centrality is a related measure used to identify nodes that are influential in controlling the flow of information across the network (Newman 2001). The betweenness centrality of a node indicates the number of shortest paths that pass through the node. In protein interaction networks nodes with high betweenness centrality may be refered to as bottle necks and correlated with essentiality (Barabasi et al. 2011).

3.8.3 Hierarchical clustering

Neighbour joining and Unweighted Pair Group Method with Arithmetic mean (UPGMA) are distance-based methods for tree construction. They are both agglomerative hierarchical clustering methods commonly used for the creation of phylogenetic trees based on DNA sequence information. Within both methods a node represents each biological entity and distance matrices are generated to determine the relative similarity between each pair of nodes. Nodes are sequentially joined into clusters based on their similarity. Once multiple nodes have been joined into a cluster, they are treated as a combined entity in the calculation of remaining distances. This process continues until a single tree has been generated.

UPGMA and neighbour joining methods differ in the methods used to calculate distances between clusters and nodes or pairs of clusters. To generate the distance between a single node and a cluster the UPGMA approach would calculate the mean distance between the node and each item contained within the cluster (Sokal & Michener 1958). In contrast, neighbour joining links pairs of nodes using a newly generated ancestor node (Saitou & Nei 1987). The algorithm then calculates distances based on the ancestor node rather than it’s component entities. The following paragraph describes the steps implemented within the neighbour joining algorithm.

The first step in neighbour joining is to calculate the following measure of each node’s distance to all other nodes in the dataset (r) (Isaev 2006):

$$r_{i=\frac{1}{N-2}\sum_{k=1}^{N} d_{ik}} \tag{6}$$

where *d* is the starting distance matrix and *N* is the number of nodes .

Next the closest pair of nodes is identified by selecting the smallest value generated by the following algorithm

$$D_{ij} = d_{ij} - (r_i + r_j) \tag{7}$$

An ancestor node (*x*) is then generated to link the pair of nodes. The lengths of the vertices linking nodes d_i and d_j to *x* are calculated as follows

$$d_{xi} = \frac{1}{2}(d_{ij} + r_i - r_j) \tag{8}$$

$$d_{xj} = \frac{1}{2}(d_{ij} + r_j - r_i)$$

The distances between the ancestral node (*x*) and the other nodes are then calculated as

$$d_{x,m} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij}) \tag{9}$$

where *m* represents any node other than *i* and *j*. Having updated the distance matrix (*d*) with the ancestor term (*x*) the process can be repeated, until the tree is fully generated.

The trees generated from UPGMA are typically rooted, while those obtained through neighbour joining are not. The UPGMA method assumes that all the genes in the network are ultrametric, meaning that all nodes are equally distant from the root, shown by the equal branch lengths (Roling & Head 2005). This assumption is often not appropriate; therefore, the neighbour joining method is preferred. Similarly, in UPGMA neighbouring leaf nodes are considered to be equidistant from the branch joining them. In contrast neighbour joining is able to represent cases where nodes are not equidistant from the ancestor node that joins them, representing variation in evolutionary speed. Neighbour joining should ideally be applied to additive data matrices that satisfy the four-point condition, otherwise anomalies may occur, in the form of multiple trees being produced, negative branch lengths and tree distances that do not coincide with the original distance matrix. UPGMA are not useful in resolving these issue since a non-additive tree will not be ultrametric. Within this project proteins were linked based on functional similarity using the QuickTree Unweighted Pair Group Method with Arithmetic Mean joining method (Howe et al. 2002)

Chapter 4

Disentangling the multigenic and pleiotropic nature of molecular function

This chapter is directly adapted from my published work; full citation:

Stoney, R.A. et al., 2015. Disentangling the multigenic and pleiotropic nature of molecular function. *BMC Systems Biology*, 9(Suppl 6), p.S3[1]

4.1 Abstract

4.1.1 Background

Biological processes at the molecular level are usually represented by clusters in molecular interaction networks. Function is organised and modularity identified based on network topology, however, this approach often fails to account for the dynamic and multifunctional nature of molecular components. For example, a molecule engaging in spatially or temporally independent functions may be inappropriately clustered into a single functional module. To capture biologically meaningful sets of

[1] RA provided the Semantic similarity data and produced the results shown in Figure 30. RS performed the rest of the analysis and wrote the manuscript

interacting molecules, we use experimentally defined pathways as spatial/temporal units of molecular activity.

4.1.2 Results

We defined functional profiles of yeast, *Saccharomyces cerevisiae,* pathways based on a minimal set of Gene Ontology terms sufficient to represent each pathway's genes. The Gene Ontology terms were used to annotate 271 pathways, accounting for pathway multi-functionality and gene pleiotropy. Pathways were then arranged into a network, linked by shared functionality. Of the genes in our data set, 44% appeared in multiple pathways performing a diverse set of functions. Linking pathways by overlapping functionality revealed a modular network with energy metabolism forming a sparse centre, surrounded by several denser clusters comprised of genetic and metabolic pathways. Signalling pathways formed a relatively discrete cluster connected to the centre of the network. GIs were enriched within the clusters of pathways by a factor of 5.5, indicating the organisation of our pathway network is of biological significance.

4.1.3 Conclusions

Our representation of molecular function according to pathway relationships enables analysis of gene/protein activity in the context of specific functional roles, as an alternative to typical molecule-centric graph-based methods. The pathway network demonstrates the cooperation of multiple pathways to perform biological processes and organises pathways into functionally related clusters with interdependent outcomes.

4.2 Introduction

Biological functions must be carried out in a synchronised manner to ensure proper timing of processes like cell division and metabolism. Molecular functions arise from complicated sets of physical interactions between large numbers of proteins, RNAs and various regulatory pathways, which are extremely difficult to reconstruct, represent and analyse. In systems biology, molecular function is mapped using molecular interaction networks. Protein-protein interaction (PPI) networks are frequently used to map protein functionality (Janjić & Pržulj 2012b; Lee & Lee 2013; Schwikowski et al. 2000; Sharan et al. 2007; Yook et al. 2004). Within interaction networks, molecules are usually represented as single nodes connected by physical interactions. Functionally similar nodes tend to cluster together into dense subnetworks, referred to as functional modules (Vidal et al. 2011; Chen & Yuan 2006; Sharan et al. 2007) or “pathways” (Kelley & Ideker 2005), forming the basis of network analysis to study function (Schwikowski et al. 2000; Sharan et al. 2007; Yook et al. 2004).

One aim of identifying sub-networks is to illustrate the position and connectivity that molecules and functional modules have within the network (Chen & Yuan 2006). They are used to examine the organisation of different functions within the cell, showing how information is passed through physical interactions to enable the system to function as a whole. Many studies have used *Saccharomyces cerevisiae* to model functionality (Przulj et al. 2004; Kelley & Ideker 2005; Costanzo et al. 2010; Dutkowski et al. 2013) due to the availability of extensive PPI, GI and gene annotation data, making it an ideal organism for developing alternative methods of functional organisation.

A great deal of research has focused on computational methods used to identify clusters/sub-networks based on topological features (Blondel & Guillaume 2008; Song & Singh 2009; Wang et al. 2010). However, such networks tend to utilise the sum of a molecule's interactions, without accounting for the temporal and spatial nature of its interactions. Simply because two proteins can interact does not mean that they will interact in every context (Hyduke & Palsson 2010). Clustering approaches tend to treat spatial/temporal edges as if they are constant. These subnetworks, which represent functional modules, may as a result bring together functions that are unrelated in the cell. Evidence for this comes from discrepancies in community detection in networks created from different data types (Ames et al. 2013). The combination of different data types has been shown to improve the functional homogeneity of topological clusters.

To deal with the issue of spatial/temporal edges we propose a method using experimentally validated pathways as the units of cellular processes. In this context pathways represent groups of proteins shown to interact under specific experimental conditions. This differs from the definition used in Kelley (2005) (Kelley & Ideker 2005), in which clusters in PPI networks were described as pathways. In our approach proteins that participate in multiple, context dependent, interactions appear in multiple pathways, rather than being represented by a single highly connected node.

We use Gene Ontology (GO) annotations derived from experimental evidence or sequence homology to assign collective functionality to the pathways. Annotated pathways were then connected according to functional overlap. Linking pathways by shared functionality enables us to examine the flow of information among biological functions, giving insight into the organisation of function within the cell.

4.3 Methods

Gene annotation data was integrated with pathway data to produce a set of annotated pathways, which were assembled into a functional network and analysed. An outline of the methods is given in Figure 19.

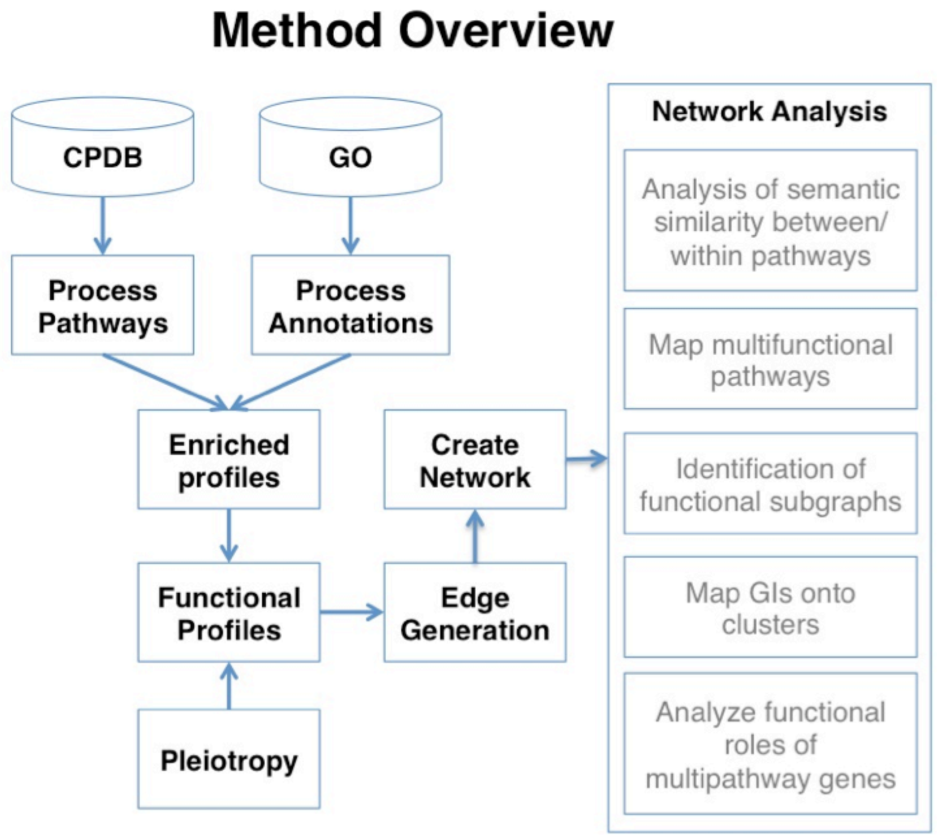

Method Overview
CPDB
GO
Process Pathways
Process Annotations
Enriched profiles
Create Network
Functional Profiles
Edge Generation
Pleiotropy
Network Analysis
Analysis of semantic similarity between/ within pathways
Map multifunctional pathways
Identification of functional subgraphs
Map GIs onto clusters
Analyze functional roles of multipathway genes

Figure 19: Outline of methods used in the construction and analysis of the yeast network

4.3.1 Pathway data

Saccharomyces cerevisiae pathway names and their constituent genes/proteins were downloaded from ConsensusPathDB (CPDB) (downloaded 17/01/2014) (Kamburov et al. 2009). Pathways were represented as sets of genes. The original data set consisted of 1050 pathways with 2114 genes.

CPBD collects pathway data from multiple databases, which results in a large degree of data duplication that needs to be removed. Three types of data duplication were identified: duplicated pathway names, duplicated gene sets, and small pathways that were subsets of larger pathways. Databases resourced by CPDB may assign slightly different gene sets to identical pathway names, as a result of varying pathway boundaries. Repeated pathway names were identified and amalgamated into single entities by merging the gene sets. Pathways with differing names but identical gene/protein sets were also present. Pathways with identical gene sets were identified and redundant pathways were removed.

Finally the gene/protein sets of some pathways were found to be subsets of larger pathways. Dealing with this form of data duplication is more complex, as the choice of which pathway to retain affects the final data set. The pathways retrieved from CPDB were also highly variable in their size (see Table 1). To reduce this variability and ensure pathways with high functional specificity were conserved, the pathway whose size was closest to the median pathway size was retained (min (|length of pathway 1| – median, |length of pathway 2| – median)).

Pathways containing less than three genes/proteins were considered too small for reliable statistical analysis of function and were removed. The effect that processing had on the data set is documented in Table 1. The final data therefore consisted of 271 pathways and 1433 genes, with a median of six genes/proteins per pathway.

Table 1: Transformation of data during processing.

	Original data	Duplicated names merged	Short (<3) pathways removed	Duplicate gene sets removed	Pathway subsets removed	Unannotated genes removed
Total pathways	1050	990	715	553	272	271
No. unique genes	2114	2114	2113	2113	1565	1433
Median GPP[2]	5	5	8	8	7	6
Mean GPP	11.9	12.2	16.3	17.6	11.4	10.2
SD	23.2	23.9	27.0	28.8	16.5	13.05

4.3.2 Generation of full set of GO identifiers for each gene

Functional gene annotations were downloaded from the Gene Ontology (release date 22/04/2014) (Ashburner et al. 2000). GO terms were assigned to the genes within each pathway. Only experimentally derived annotations or annotations generated using sequence orthologs were used, leaving 132 (9%) genes unannotated (Table 1). Unannotated genes were omitted from the data set. To increase annotation

[2] GPP indicates genes per pathway, SD indicates standard deviation

completeness, the GO hierarchy was downloaded and parent annotations were added to genes.

4.3.3 Removal of uninformative GO terms

The hierarchical nature of the Gene Ontology resulted in some annotations being too general and frequent to be considered informative. For this reason, and based on assessment of the GO annotation frequencies across the genes in the data set, annotations present in over 50% of genes were deleted; deleted annotations are listed in Supplementary Table 1 (Appendix A). These annotations are highly unlikely to be identified as enriched within a single pathway during later processing stages and removing them at this point reduces repeated testing.

4.3.4 Annotation of pathways

GO annotations associated with pathway genes were used to infer the function of the CPDB pathways. Only biological process annotations were used, molecular function and cellular component information was not incorporated. The Shapiro test (Shapiro & Wilk 1965) was performed to ensure that none of the GO terms were randomly distributed across the pathways ($p << 0.001$). Enrichment profiles were created to contain all the GO terms enriched within a pathway's genes. Functional profiles were then generated to show the most specific enriched GO terms capable of describing the gene set. Functional profiles should therefore be considered as describing the main functional roles of a pathway, at the highest level of specificity possible.

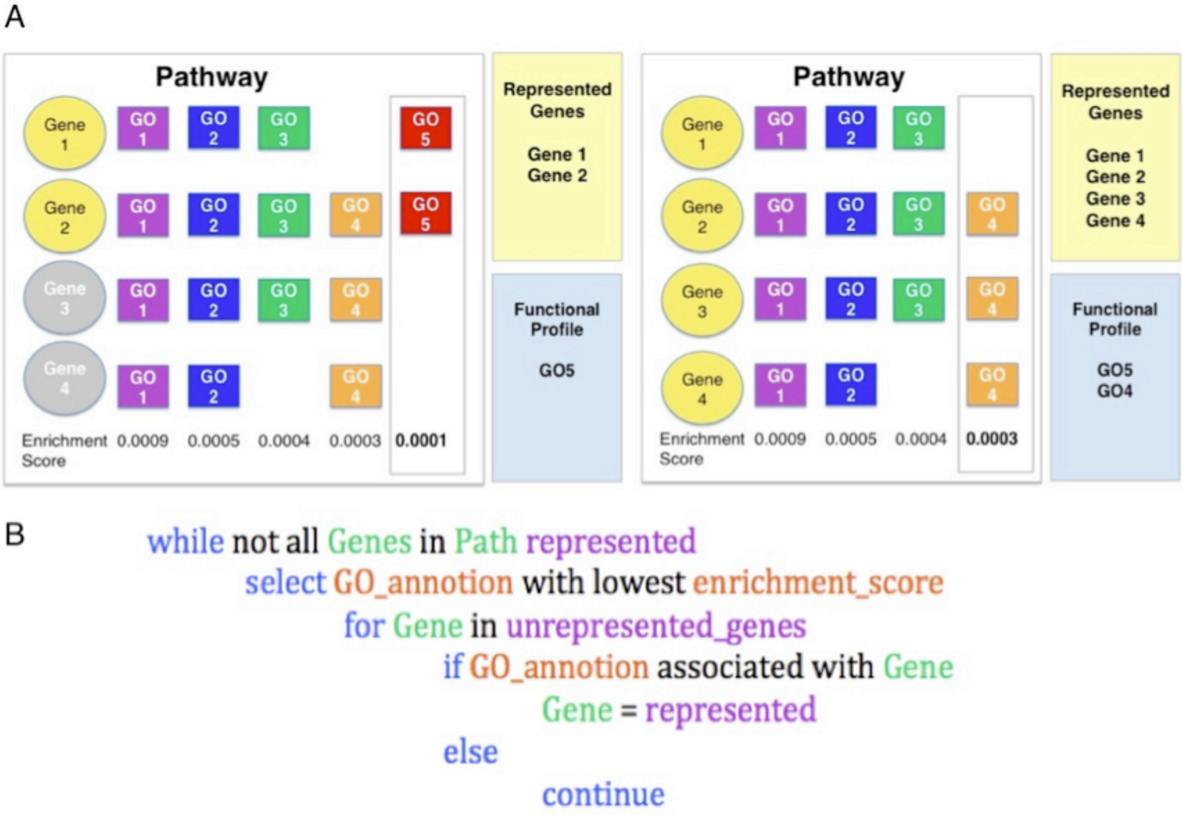
4.3.5 Enrichment profiles

Functional enrichment profiles were created using the Fisher's exact test to identify all annotations enriched to a p-value of 0.01, within the pathway's gene set. The parameters used were: instances of the GO annotation within the pathway (how many genes the annotation was attributed to), instances of other GO terms in the pathway, instances of the annotation outside the pathway, instances of all other GO terms outside the pathway. Using an enrichment score of 0.01 as the threshold for allocating GO terms, annotations are assigned at 99% specificity. Rather than correcting for multiple testing, false positive annotations are removed in later processing stages, designed to be flexible to the varying specificity of GO term-pathway relationships. P-values gained from Fisher's exact tests are therefore referred to as enrichment scores.

4.3.6 Functional profiles

The functional profile of a pathway is defined as a reduced set of enriched GO annotations that give maximum representation of a pathway's genes. Enriched annotations that were only present in one gene/protein within the pathway were excluded, as they are likely to be spurious and give a poor representation of the pathway's function.

The remaining annotations in each pathway's enrichment profile were considered for inclusion in the functional profile, in rank of their enrichment score (lowest enrichment scores first). The first (most highly enriched) GO term is selected and checked against the annotations of each gene/protein in the pathway (Figure 20). Genes associated with the annotation are considered represented. If all genes were not represented by the first GO term, GO terms associated with the remaining genes were considered. Any genes connected with this GO term were then considered represented. This process was continued until all genes were represented or until all the GO terms with significant enrichment scores were utilised. This resulted in a set of functional profiles with a median of two annotations per pathway.



```
while not all Genes in Path represented
    select GO_annotion with lowest enrichment_score
        for Gene in unrepresented_genes
            if GO_annotion associated with Gene
                Gene = represented
            else
                continue
```

Figure 20: Yeast functional profile creation. (A) The figure shows one pathway, with genes represented as circles and gene annotations shown in boxes to the right of each gene. The aim of the algorithm is to select the minimum number of GO annotations necessary to represent all the genes in the pathway, preferentially selecting annotations with low enrichment scores. In this example GO5 is the annotation with the lowest enrichment score and is therefore selected first. GO5 is associated with genes 1 and 2, therefore GO5 is sufficient to represent these genes. GO4 is selected next and represents genes 2, 3, and 4; therefore GO4 and GO5 represent all the genes in the pathway.

4.3.7 Pleiotropic genes within pathways

Pleiotropy describes situations where a gene contributes to more than one phenotype, implying that the gene/protein is involved in more than one function. This may be due to multiple instances of the gene/protein in different pathways, or the genes within a single pathway effecting multiple functions (He & Zhang 2006a), resulting in pathway multi-functionality. These additional functions may be missed in the initial formation of functional profiles, as only the most enriched annotations for each gene set are included. A second processing stage was added to capture pleiotropic annotations. Semantic distances between GO terms were taken from Ames et al. (2013). Semantic distances were available for 88% of GO annotations within the enriched profiles. Defining phenotypic pleiotropy is complex, as the distinction between different characters and multiple attributes of a single character is often unclear (Wagner & Zhang 2011). To ensure that the terms we add are truly pleiotropic we have chosen to use only terms that are very semantically different from existing terms in the functional profile.

Within functional profiles, the median semantic distance between pairs of GO terms was 6 and 95% of GO term pairs had semantic distances above 11.2 [3]. Therefore a

[3] Semantic similarities were taken from (Ames et al. 2013) using a method described in (Hakes et al. 2007). The method subtracted the information content of each test node from the the doubled information of the lowest common ancestor, resembling the Jiang & Conrath (1997) approach. This generates distances rather than similarities,

semantic distance of 11.2 was used as the measure of pleiotropy. To avoid false positive annotations, GO terms from enriched profiles were only considered for pleiotropy if they had an enrichment score below 0.0005. The semantic distance between each GO term in each pathway's enriched profile and all the GO terms in the functional profile was measured. Any enriched annotations that had a distance greater than 11.2 from all of the GO terms in the functional profile, were considered pleiotropic and added to the functional profile. Using these parameters 32 terms were added to 25 pathways.

A concern when adding pleiotropic terms was that large semantic distances may be more likely to arise in larger pathways with more genes, resulting in less specific pathway functions.

Figure 21 shows the number of GO annotations and the maximum semantic distances between annotations in enriched profiles. Examination of the point sizes, indicating the number of genes in a pathway, shows that although pathway size is linked to the number of enriched GO terms, it does not affect the maximum semantic distance between the terms. Terms in several small pathways' pass the threshold distance of 11.2, indicating that small pathways can contain semantically diverse enriched terms,

---

therefore GO terms with a score higher than 11.2 must by be two highly specific GO terms (low information content), with a general lowest common ancestor (high information content).

which if omitted from the functional profile, could result in useful information being lost.

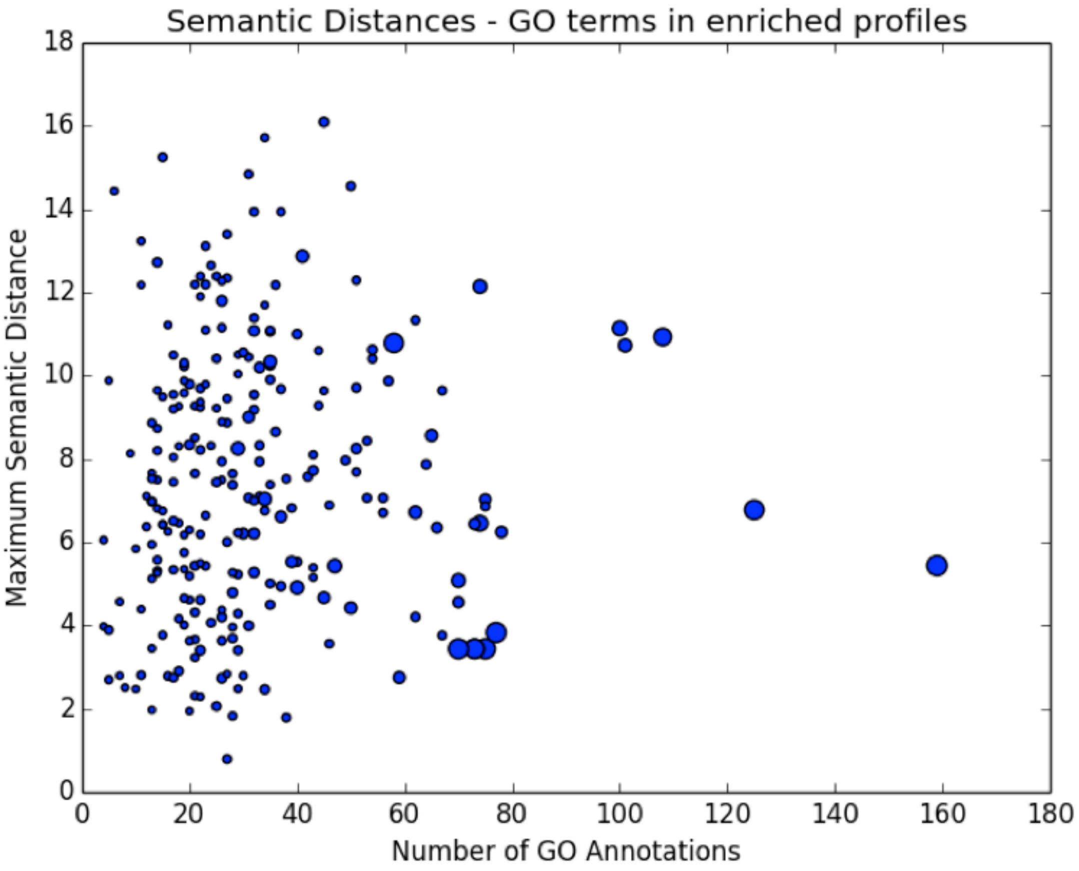


Figure 21: Annotation variability within different sized pathways. The Y-axis represents the maximum semantic distance between GO terms in each pathway's enriched profile. The X-axis represents the number of GO terms in the pathway's enriched profile. Circle size indicates the number of genes in the pathway. Pathways with low numbers of genes and annotations can be seen achieving high maximum semantic distances.

4.3.8 Network generation

The annotated pathways were used as nodes and linked by shared functionality into a network. Edges were created using the Jaccard coefficient to measure proportional overlap between pairs of pathway annotations (Equation 4). Jaccard coefficient scores were used to weight the edges in an undirected network.[4] The degree distribution of the network structure was analysed using the R igraph package version 1.0.0 (Csardi & Nepusz 2006).

4.3.9 Linking functionally similar annotations

Due to the size and hierarchical nature of GO it is possible that multiple annotations may describe very similar cellular functions. Pathways with different annotations describing highly similar functions would not be linked, therefore the network would fail to represent the pathways' functional similarity. To overcome this issue, links have been created between nodes with semantically similar annotations below a threshold (*T*) of 0.8 (Equation 6). We calculated pairwise similarity scores (S^{ab}) between GO terms (*a* and *b*) of pairs of pathways (A and B), retaining only scores below *T*. The retained similarities were normalised, then summed to give a value (V_{AB}) expressing the total similarity between the annotations in both pathways (if a GO annotation

[4] Cytoscape 2.8.3 was used to visualise the network (Shannon et al. 2003), due to its compatibility with the MultiColouredNodes v.2.54 plugin (Warsow et al. 2010). The MultiColouredNodes v.2.54 plugin was used to visualise multiple attributes to single nodes with pie charts.

appeared in both functional profiles it was not compared to itself). The resulting value was then divided by the number of possible pathway pairs, to obtain the edge weight (W_{AB}). GO term pairs with scores below the threshold of 0.8 represent the most extreme cases of semantic similarity (<0.1% of semantic distances), ensuring that the majority of the edges in the network represent identical shared annotations (74%).

$$V_{AB} = \sum_{a \in A, b \in B} \frac{T - S^{ab}}{T} \qquad W_{AB} = \frac{V_{AB}}{|A| * |B| - |A \cap B|} \tag{10}$$

4.3.10 Genetic interaction analysis

GIs frequently occur between genes/proteins in pathways that share functions (Costanzo et al. 2010). Based on this knowledge it is expected that topological clusters[5] (see supplementary Figure 1, in Appendix A) in the network will be enriched for GIs. This was tested using a set of GIs from BIOGRID (data set BIOGRID-ORGANISM-Saccharomyces_cerevisiae-3.2.119, downloaded 27/11/2014) (Stark et al. 2006). Excluding GIs involving genes that were absent from the data set resulted in a list of 29309 GIs. For each GI, the set of pathways that each gene/protein participates in was retrieved, and all pathway combinations were examined: if both genes/proteins appeared in a single pathway, a within-pathway GI was recorded; if each gene/protein appeared in a different pathway but the pathways were in the same cluster, a within-cluster GI was recorded (Figure 22). GIs linking pathways from different clusters or involving unclustered pathways were recorded as uncharacterised.

[5] Clusters were generated using the Cytoscape ClusterONE plugin.

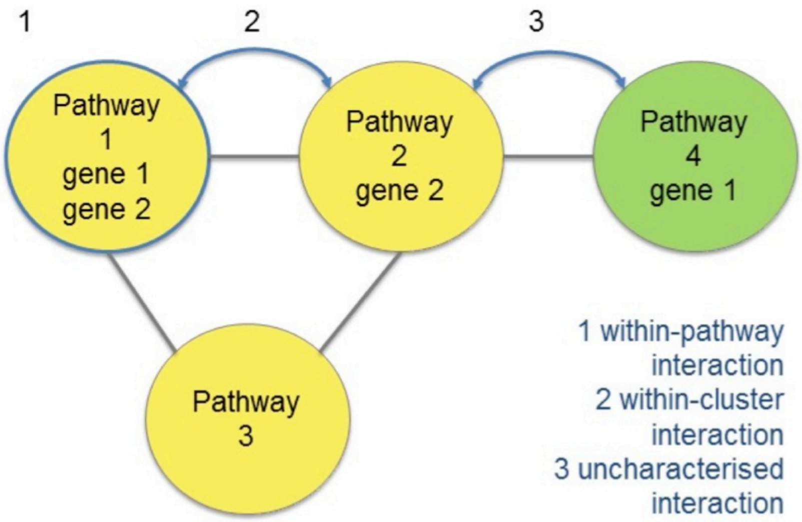

1
2
3
Pathway 1 gene 1 gene 2
Pathway 2 gene 2
Pathway 4 gene 1
Pathway 3
1 within-pathway interaction
2 within-cluster interaction
3 uncharacterised interaction

Figure 22: Genetic interactions within pathways and pathway clusters. GIs are classified depending on whether the genes are in the same pathway, within two pathways in the same network cluster, or uncharacterised (between two pathways in different clusters or involving unclustered pathways). Genes present in multiple pathways will result in GIs appearing in many pathway pairs. In these situations all pathway pairs are classified separately. The yellow nodes show 3 pathways in a single cluster. The green node represents a pathway in a separate cluster. All possible ways of connecting gene1 and gene2 across all pathways are explored

4.3.11 Characterising the profiles of multi-pathway genes/proteins

To establish whether genes/proteins acting in multiple pathways are performing different roles, we performed pairwise comparisons of semantic distances between the annotations in multiple functional profiles. The sum of the semantic distances was divided by the number of genes in the profiles' union.

4.3.12 Generation of the gene/protein overlap heat map

Many proteins were present in multiple pathways. To examine the relatedness of these pathways' functions, a heat map was created to compare gene/protein overlap against functional similarity. Pathways were arranged into a tree based on functional similarity, shown in both axes. This was calculated by carrying out pairwise comparisons of all GO terms between functional profiles, and taking the mean semantic distances. The tree structure was created by QuickTree using the Unweighted Pair Group Method with Arithmetic Mean joining method (Howe et al. 2002). The heat map was created by calculating the percentage of gene/protein overlap between pathways and colouring cells accordingly.

4.4 Results and discussion

We produced a set of functionally annotated pathways, which were assembled into a network to show functional organisation. The major functional subgraphs are identified and the relationship between functions is discussed. The functional variability of genes/proteins that participate in multiple pathways is evaluated. GI enrichment within network clusters was measured.

The vast majority of biological functions require the cooperation of multiple genes and proteins. However, functional representations associated with individual genes/proteins are derived from the curation of scientific papers (Camon et al. 2004), making them highly idiosyncratic and often failing to capture the cooperative aspect of biological function. In order to create systems-wide models that are more suitable to biological interpretation and understanding, new representations are needed that better reflect the cooperative nature of function. Biological pathways are a suitable candidate for higher-level representation of biological function, since they group genes and proteins that interact to produce a specific cellular or physiological outcome.

4.4.1 Generation of a functionally representative set of pathways

A set of 1050 *Saccharomyces cerevisiae* pathways was obtained from CPDB and processed to remove data duplication and reduce pathway size variability (pathway sizes in the original data set ranged from 1 to 310). Removal of duplicated pathway names and gene sets, as well as pathways containing fewer than 3 genes, reduced the number of pathways in the data set to 553 (Table 1). Further processing of duplicated data selectively removed pathways whose size deviated from the median, helping to reduce the standard deviation from 23.2 in the original data set to 13.1 in the final data set. The largest pathway in the original data set was 'Metabolism' containing 310 genes, which would have dominated much of the network; however the largest pathway in the final data set was 'Protein processing in endoplasmic reticulum' with a more comparable 78 genes.

4.4.2 Assignment of Gene Ontology terms to genes

Annotations were available for 92% of genes in the data. Adding parent annotations to the GO terms initially assigned to the genes increased the median number annotations from two to 38 and the maximum from eight to 149.

Removing highly frequent, uninformative annotations from the data set reduced the median number of annotations per gene from 38 to 31. Within this final data set the range of annotations assigned to genes was large, ranging from 1 to 208; 75% of genes had between 14 and 66 annotations. This variability may be due to genes being attributed GO terms with large numbers of parent annotations or gene/protein multi-functionality.

4.4.3 Generation of functional profiles of pathways

The Fishers exact test produced large numbers of overrepresented GO terms for each pathway (median 26, range 1-159). This is in part related to the hierarchical nature of the Gene Ontology, implying that many of these annotations are describing a small number of functions at various levels of detail. Functional profiles were created to give a succinct representation of each pathway's specific functions, by selecting a reduced set of GO terms to describe the maximum number of genes/proteins inside each pathway (median 2, range 1-9). Only 35% of pathways were described by a single GO term, demonstrating that functions defined by the Gene Ontology cannot be directly mapped onto pathways, as the relationship is more complex. A moderate correlation was found between the number of genes/proteins in a pathway and the number of GO terms in its functional profile (coefficient 0.5). The majority of pathways had unique functional profiles, however 13% of functional profiles were not unique to a pathway indicating that some GO functions may be shared by discrete groups of pathways.

4.4.4 Improved functional profile comprehensiveness through incorporation of gene pleiotropy

The functional profile algorithm (see Section 4.3.6) selects the most enriched annotations for genes/proteins within the context of each pathway; however, multiple functions performed by genes/proteins may be missed. As a result of incorporating pleiotropic terms, 32 additional annotations were added to 25 pathways, with each pathway retrieving between 1 and 3 terms. Examples of the information added by including pleiotropic terms are given in Table 2. For the full set of pleiotropic and non-pleiotropic GO terms associated with each gene within the pathway dataset, please see Supplementary Data 1.

Table 2: Examples of the data added through the inclusion of pleiotropic genes. The annotation overlap across the pathways illustrates the functional overlap of these pathways. Sucrose is degraded into fructose and trehalose is degraded into glucose, prior to cellular import *(Gagiano et al. 2002)*. Mannose and fructose are both transported into the cell by hexose transporters and degraded into Fructose-6-phosphate.

Pathway	**Original Annotations**	**Pleiotropic Annotations**
sucrose degradation	cellular carbohydrate catabolic process	fructose import
trehalose degradation II	cellular carbohydrate catabolic process	glucose import
mannose degradation	fructose import	fructose metabolic process

We analysed the semantic distance between GO terms co-occurring within functional profiles (Figure 23). The distribution of semantic distances indicates that functional profiles have a much higher proportion of close GO terms than enriched profiles. The most frequent (mode) semantic distance between GO terms in functional profiles is 4 (median 6.1), which is notably lower than in enriched profiles (mode 6, median 6.2). Merging the GO terms from within functional profiles and within enriched profiles gives the distribution of semantic distances between random pairs of annotations, accounting for annotation frequency. Both functional and enriched profile sets contain many more semantically close genes/proteins than expected from chance (modes 9 and 7 respectively). Although most functional profiles contain semantically similar annotations, some are functionally diverse, as shown by the tail of the functional profile distribution (Figure 23). The spike in frequency seen at the semantic distance of 11-12, is due to the addition of pleiotropic annotations. A peak is also seen at a semantic distance of 8-9, corresponding to the mode distance in combined enriched profiles. This indicates that the pathways may incorporate a second cellular function, possibly acting as functional bridges, facilitating cellular coordination.

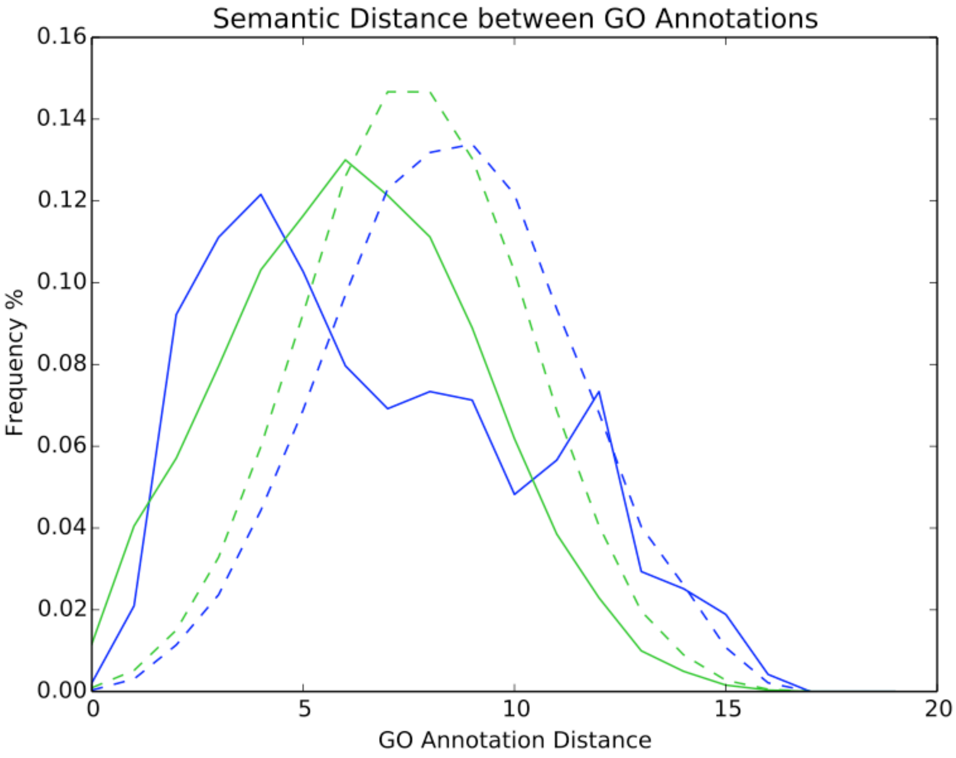


Figure 23: Semantic similarity of GO annotations within/between functional and enriched profiles. The solid blue line shows the frequency of distances between pairs of GO terms within each pathway's functional profile. The solid green line shows the frequency of distances between pairs of GO terms within each pathway's enriched profile. Annotations in functional profiles were merged and distance frequencies are shown by the dashed blue line. This process was repeated for the enriched profiles to create the dashed green line. Merging profiles gives the random expected distance between annotations, controlling for annotation frequencies. When merging profiles annotations appear multiple times, however, annotations were not compared to themselves.

4.4.5 Functional diversity of pathways

Multiple functions can be distributed across the genes/proteins within a pathway in three ways. Functional profile annotations are either distributed across overlapping, disjoint (discrete) or pleiotropic sets of genes within the pathway (Figure 24 A, B & C respectively). The majority of pathways (84%) had all of their functional profile annotations distributed across overlapping gene/protein sets. This overlap of functions illustrates how information is passed from one function to the next, connecting cellular functions. Instances where a pathway's genes are split into discrete functional groups, may indicate that the boundaries of pathway are in discord with the functional boundaries presented by the Gene Ontology. This discrete distribution of function occurs in 26 pathways, many of which are positioned in areas of the network involved with energy production and amino acid metabolism. These pathways have a median of three GO terms and semantic distances between GO terms are higher than those observed within other pathways (median 10). Pleiotropic annotation distributions were created by the addition of pleotropic terms following initial functional profile creation, present in 25 pathways.

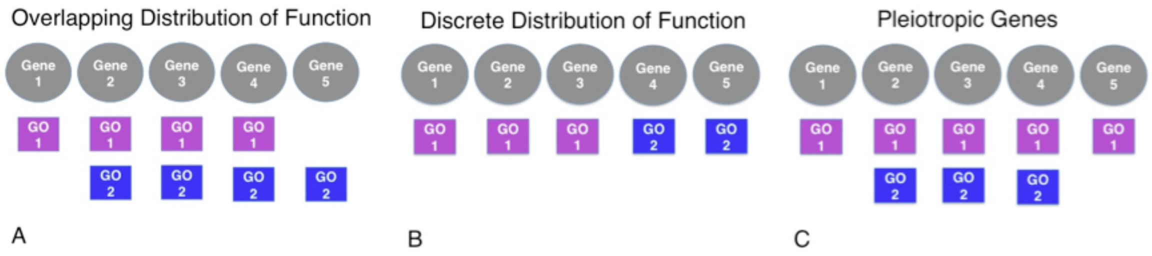


Figure 24: Distribution of multiple functions across genes within pathways. Functionality may be distributed across a pathway's genes in the following ways: pathways may have multiple functions distributed across overlapping genes (A); multiple functions may be divided into discrete (disjoint) sets of genes (B); or pleiotropic genes may have multiple layers of functionality (C).

Pleiotropic annotation distributions were created by the addition of pleotropic terms following initial functional profile creation, present in 25 pathways. Of these pleiotropic pathways 22 were connected within the network. These pathways have been labelled in Figure 25. The pathways omitted were 'Transport to the Golgi and subsequent modification', 'Other types of Oglycan biosynthesis' and 'Unwinding of DNA'. The GO terms associated with the 'Transport to the Golgi and subsequent modification' pathway were 'COPII-coated vesicle budding', 'regulation of vesicle targeting, to, from or within Golgi' and 'purine ribonucleoside triphosphate catabolic process' (pleiotropic), which were not found within any other pathways and were not semantically similar enough to any GO terms associated with other pathways to become linked (see Section 4.3.9). This situation also applies to the 'Other types of Oglycan biosynthesis' pathway which was associated with the GO terms 'protein O-linked glycosylation', 'regulation of response to stress' (pleiotropic) and 'protein exit from endoplasmic reticulum' (pleiotropic); and to the ‘Unwinding of DNA’ pathway which has the GO terms 'double-strand break repair via break-induced replication' and 'DNA duplex unwinding' (pleiotropic).

The addition of pleiotropic terms can give a fuller picture of the processes carried out by pathways. The pathway ‘synthesis of UDPNacetylglucosamine’ demonstrates the ability of pleiotropic terms to give insight into the function of pathways, which would otherwise be missed. The GO term ‘amino sugar biosynthetic process’ was primarily selected to represent the ‘synthesis of UDPNacetylglucosamine’ pathway with the terms ‘nucleotide-sugar biosynthetic process’ and ‘fungal-type cell wall biogenesis’ added as pleiotropic terms. The term ‘fungal-type cell wall biogenesis’ reflects the cell’s ability to convert of UDPN acetyl-glucosamine to UDP-GalNAc, which is used to generate cell walls (Milewski et al. 2006). The pathway 'Other types of Oglycan biosynthesis' is described by its allocated GO term 'protein O-linked glycosylation', but the pleiotropic term 'protein exit from endoplasmic reticulum' highlights the location of O-linked glycosylation within the Golgi (Alberts et al. 2002). Another pleiotropic term associated with this process is ‘regulation of response to stress’, appropriate since changes in glycosylation can induce or be induced by endoplasmic reticulum stress (Gerlach et al. 2012). A further example is the pathway 'Transport to the Golgi and subsequent modification', which has the terms 'COPII-coated vesicle budding' and 'regulation of vesicle targeting to from or within Golgi', and addition to the pleiotropic term 'purine ribonucleoside triphosphate catabolic process', which captures the use of ribonucleotides for glycosylation within the Golgi (Stanley 2011).

The cluster of pathways ‘thioredoxin system’, ‘sulfur relay system’ and ‘glutathioneglutaredoxin system’ are all linked by the GO term ‘cellular response to oxidative stress’. Since cellular stress can induce a range of changes within the cell different pleiotropic terms are seen in each pathway. The ‘glutathioneglutaredoxin system’ pathway is associated with the highest number of associated pleiotropic terms; ‘transition metal ion homeostasis’ is implicated since glutathione is involved in addressing metal induced oxidative stress (Jozefczak et al. 2012) ‘negative regulation of binding’ may correspond to glutaredoxin’s negative regulation of ASK1 in oxidative stress situations (Song et al. 2002) and ‘protein glutathionylation’ which is a well established response to oxidative stress (Niwa 2007). In contrast the ‘thioredoxin system’ pathway had the pleiotropic term ‘protein deglutathionylation’, however glutathionylation and de glutathionylation are both used to regulate mitochondrial NADP(+)-dependent isocitrate dehydrogenase during oxidative stress (Niwa 2007). Finally the ‘sulfur relay system’ pathway was associated with the pleiotropic term ‘adenosine metabolic process’, which may be because adenosine 5’-phosphate reductase has been found to differentially control sulphur flux during stress conditions (Scheerer et al. 2010).

The pathways ‘trehalose degradation II trehalase’, ‘mannose degradation’, and ‘sucrose degradation’ are linked by the shared term ‘frucose import’ which is pleiotropic to ‘trehalose degradation II trehalase’ and ‘sucrose degradation’. Although none of these pathways deal directly with fructose, sucrose is degraded into fructose prior to cellular import (Gagiano et al. 2002). Mannose and fructose are both transported into the cell by glucose (hexose) transporters and degraded into Fructose-6-phosphate. Trehalose is degraded into glucose extracellularly, then imported using the same transporter. Similarly the ‘glucose transport‘ pathway associated with the terms ‘glucose import’ and ‘single organism transport’, as well as the pleiotropic term ‘mannose metabolic process’ representing the ability of glucose transporters to import mannose into the cell (Rodrıguez et al. 2005).

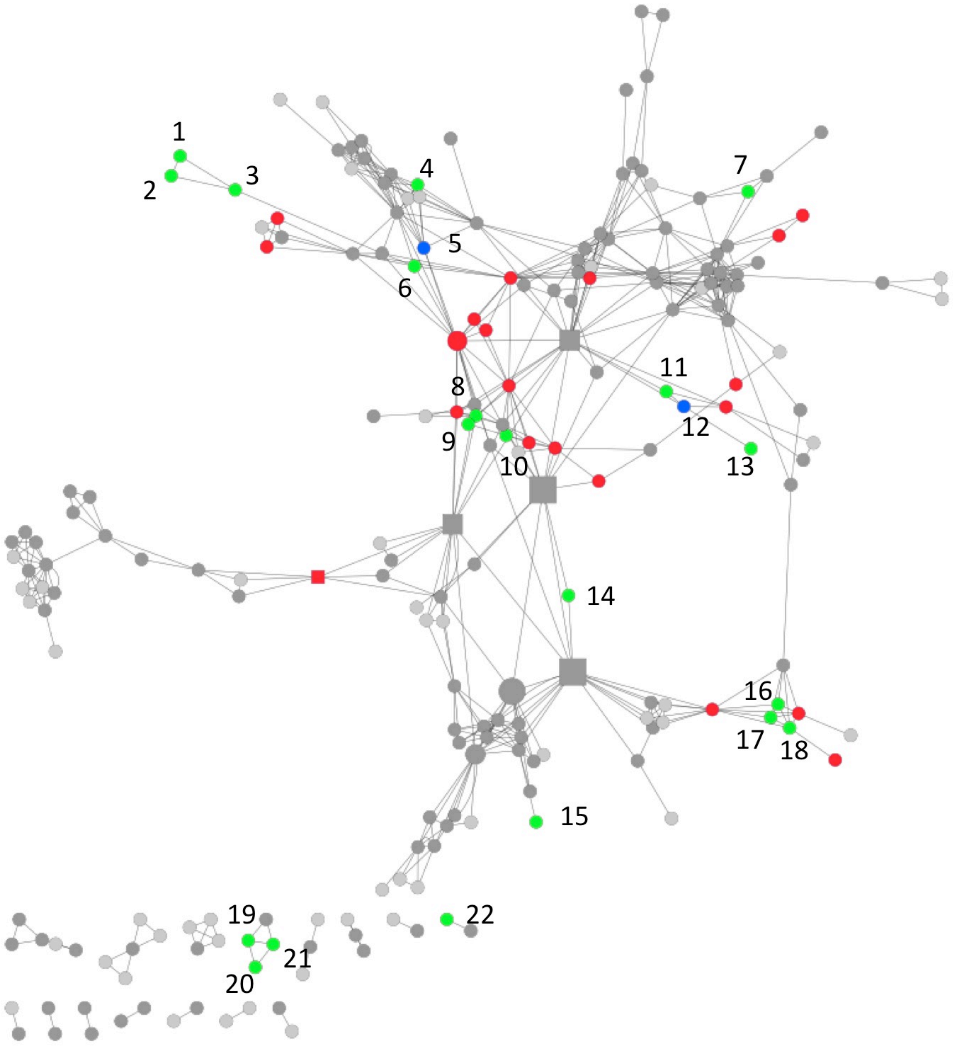
The ‘xylose monophosphate cycle’ pathway was attributed the terms ‘single-organism carbohydrate metabolic process’, ‘single-organism carbohydrate catabolic process’ and 'glycerol catabolic process' reflecting it's general role is sugar metabolism, while the pathway's pleiotropic term 'response to toxin' captures the pathway's ability to trap free formaldehyde, while is a cytotoxic compound (Yurimoto et al. 2005).

Other pleiotropic terms add additional information regarding pathway function, for example the 'acetaldehyde biosynthesis' pathway has the GO term 'amino acid catabolic process to alcohol via Ehrlich pathway' and the pleiotropic term 'fermentation'. The 'methylglyoxal catabolism' pathway has the pleiotropic term 'methylglyoxal metabolic process' which gives a more precise description of function than its other allocated terms, 'lactate metabolic process' and 'single-organism metabolic process'. Similarly, the pathway 'Unwinding of DNA' pathway is best described by the pleiotropic term 'DNA duplex unwinding', compared to its primary term 'double-strand break repair via break-induced replication'. The pleiotropic term associated with 'glycogen breakdown glycogenolysis' is 'glucose 1-phosphate metabolic process', which provides additional information to the terms 'single-organism catabolic process' and 'energy reserve metabolic process'.

The addition of plieotropic terms may be essential for the formation clusters within the network, such as the 'lactose degradation', 'trehalose Anabolism' and 'dolichyl glucosyl phosphate biosynthesis' cluster. The GO terms originally allocated to these pathways were 'galactose catabolic process' and 'UDP-glucose metabolic process'; 'oligosaccharide biosynthetic process'; and 'UDP-glucose metabolic process', respectively. Based on these initial functional allocations the 'lactose degradation' and ‘dolichyl glucosyl phosphate biosynthesis’ pathways would be linked by the shared term ‘UDP-glucose metabolic process’. However, it is the addition of the pleiotropic term ‘nucleotide-sugar metabolic process’ to each pathway that generates the cluster. The terms is appropriate since lactose and trehalose are degraded into glucose (Gagiano et al. 2002), and UDP-glucose is a nucleotide sugar and trehalose is generated using NDP-glucose (which is the collective term for nucleotide sugars including glucose) (Jules et al. 2008).

The pleiotropic pathways ‘fatty acid betaoxidation I’ and ‘fatty AcylCoA Biosynthesis’ are linked by the general term ‘fatty acid metabolic process’, however, the pleiotropic terms enhance the similarity of these pathways. The pathways ‘fatty acid betaoxidation I’ and ‘fatty AcylCoA Biosynthesis’ were also assigned the terms ‘long-chain fatty acid transport’ and ‘fatty-acyl-CoA metabolic process’ (pleiotropic); and ‘organic acid transport’ and ‘long-chain fatty-acyl-CoA metabolic process’ (pleiotropic), respectively. Shared details regarding the length of the fatty acids and the specific generation of AcylCoA would be lost without the inclusion of the pleiotropic terms.

The pathways ‘pyruvate dehydrogenase complex’ and ‘glutamate degradation III’ are linked by the plieotropic term ‘thioester metabolic process’. The pathways also include the GO terms ‘acetyl-CoA biosynthetic process from pyruvate’ and ‘succinyl-CoA metabolic process’, respectively, referring to the thioesters of coenzyme A included in each pathway (Slenter et al. 2018; Kutmon et al. 2016; Kelder et al. 2018).



1 Trehalose Anabolism
2 dolichyl glucosyl phosphate biosynthesis
3 Lactose degradation
4 fatty acid betaoxidation I
5 fatty AcylCoA Biosynthesis
6 Glycogen breakdown glycogenolysis
7 acetaldehyde biosynthesis
8 methylglyoxal catabolism
9 pyruvate dehydrogenase complex
10 glutamate degradation III
11 superoxide radicals degradation
12 ethanol degradation IV
13 Phase 1 Functionalization of compounds
14 Glucose transport
15 Synthesis of UDPNacetylglucosamine
16 thioredoxin system
17 Sulfur relay system
18 glutathioneglutaredoxin system
19 trehalose degradation II trehalase
20 mannose degradation
21 sucrose degradation
22 Xylulosemonophosphate cycle

Figure 25: Pathway multi-functionality. This figure shows a network of pathways linked by shared functionality. Red nodes represent pathways those functional annotations are divided into two or more discrete gene sets. Green nodes represent pathways with pleiotropic genes. Blue Nodes have both discrete functionality and pleotropic genes. Dark grey nodes represent pathways with overlapping multifunctional distributions and light grey nodes represent pathways with a single function. Node sizes indicate the number of GO terms attributed to pathways. Pathways with high betweenness centralities are shown with square nodes.

Figure 25 shows the position of pathways with overlapping functionality (dark grey), discrete functionality (red) and pleiotropic genes (green). Nodes showing discrete functionality and pleiotropic genes tend to form clusters, and are particularly frequent within energy metabolism. Pathways with both discrete and pleiotropic function (blue) are seen linking discrete and pleiotropic pathways. The size of the nodes indicates the number of GO annotations attributed to each pathway. Pathways with as few as 2 annotations have discretely distributed functionality or pleotropic genes.

Square nodes (Figure 25) show the five pathways with the highest betweenness centrality (range 0.33 to 0.19), indicating that these pathways are important for the transfer of information across the cell. Betweenness centrality was highest in: endocytosis; glyoxylate and carboxylate metabolism; mitochondrial protein import; toll-like receptor cascades; and adenosine ribonucleotides ide novo biosynthesis. Endocytosis is the process by which the cell imports proteins and lipids from the cell surface and can be seen linking the cell membrane and signalling pathways to the metabolic pathways (Carroll et al. 2012). Glyoxylate and carboxylate metabolism is necessary for the cell to grow on fatty acids and C2-compounds such as ethanol and can be seen in centre of the network between lipid metabolism and energy metabolism (Kunze et al. 2006). Mitochondria participate in several metabolic processes, however the majority of their proteins must be imported. The proteins required depend on the metabolic process taking place, therefore mitochondrial protein import connects many cellular functions and can be seen in the centre of the network (Dudek et al. 2013). Toll-like receptor cascades are essential for the cell to respond to pathogens (Akira & Takeda 2004). Within the network this pathway connects cell membrane and signalling pathways to the main body of the network. Adenosine ribonucleotides ide novo biosynthesis is necessary for transcription, DNA repair and replication. This pathway can be seen linking gene expression to nucleotide biosynthesis.

4.4.6 Functional network subgraphs

By mapping the most frequent GO terms onto the network of pathways, functional groups of pathways are clearly observed. Groups of pathways are formed involving genetic processes, metabolic processes and signalling (Figure 26). Energy metabolism appears in the centre of the network, reflecting its necessity of energy to all biological functions. Transcription and nucleotide processes dominate one side of the network, with protein and lipid metabolism at the other.

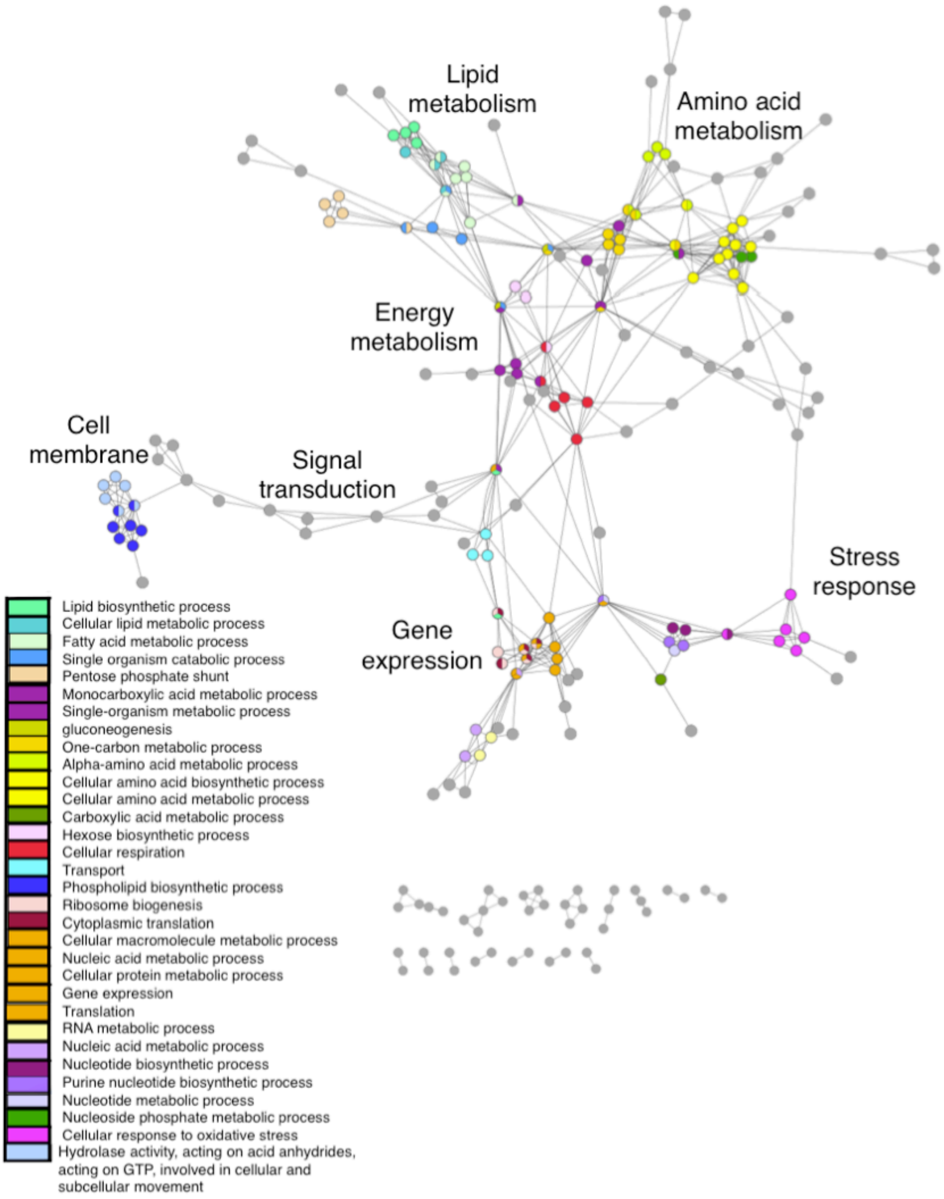

Lipid metabolism
Amino acid metabolism
Energy metabolism
Cell membrane
Signal transduction
Gene expression
Stress response
Lipid biosynthetic process
Cellular lipid metabolic process
Fatty acid metabolic process
Single organism catabolic process
Pentose phosphate shunt
Monocarboxylic acid metabolic process
Single-organism metabolic process
gluconeogenesis
One-carbon metabolic process
Alpha-amino acid metabolic process
Cellular amino acid biosynthetic process
Cellular amino acid metabolic process
Carboxylic acid metabolic process
Hexose biosynthetic process
Cellular respiration
Transport
Phospholipid biosynthetic process
Ribosome biogenesis
Cytoplasmic translation
Cellular macromolecule metabolic process
Nucleic acid metabolic process
Cellular protein metabolic process
Gene expression
Translation
RNA metabolic process
Nucleic acid metabolic process
Nucleotide biosynthetic process
Purine nucleotide biosynthetic process
Nucleotide metabolic process
Nucleoside phosphate metabolic process
Cellular response to oxidative stress
Hydrolase activity, acting on acid anhydrides, acting on GTP, involved in cellular and subcellular movement

Figure 26: Frequent GO terms in yeast functional profiles. This figure depicts the same network as shown in Figure 25, portraying the major functional groups. Colours represent frequent GO terms within the network. Pathways with less frequent GO annotations are shown in grey. GO similar terms have been attributed the same colour. Labels show the major functional communities.

Cell signalling forms a highly detached branch attached to the main body of the network by cellular transport processes. Functional maps created using yeast PPI data also found that cellular communication and signal transduction were highly segregated from the rest of the network (Yook et al. 2004). However, several differences in functional organisation are also observed. The network constructed by Yook et al.

(2004) placed cellular organisation and transcription rather than energy metabolism at the network core. Protein synthesis was found to be the least connected functional module, whereas in our network protein synthesis pathways are found within the main body of the network.

A further difference between our network and PPI networks, is that PPI networks tend to be hub-based networks, meaning that the network topology is dominated by small number of highly connected hub proteins with scale-free properties (Yook et al. 2004; Albert 2005; Winterbach et al. 2013). Scale free distributions, characterised as having a power law degree distribution of P(k) $\sim k^{-\gamma}$ where γ is typically between 2 and 3 are common in both biological and non-biological contexts(Barabási & Oltvai 2004). Within our network hub nodes would be expected to appear as highly multifunctional pathways. However, application of the Kolmogorov-Smirnov test revealed that the degree distribution did not follow a power law distribution ($p<<0.05$).

4.4.7 Co-occurrence with genetic interactions

It is known that GIs tend to occur within pathways and between functionally similar genes (Kelley & Ideker 2005; Costanzo et al. 2010). It is therefore expected that the pathway-clusters identified here will be enriched for GIs. To test this he proportion of GIs that occurred within pathways and within clusters (Figure 22) was compared to output from randomised GI data (Table 3). GIs within pathways were increased by a factor of 6.5 compared to randomised data and within-cluster GIs were enriched by a factor of 5.5. The topological network clusters are shown in supplementary figure 1 (Appendix A).

Table 3: Enrichment of GIs within pathways and network clusters. Percentages of within-pathway or within-cluster GIs, compared to randomised interactions.

	Genetic interaction data	Randomised data
Within-pathway	5.45%	0.840%
Within-cluster	4.37%	0.800%

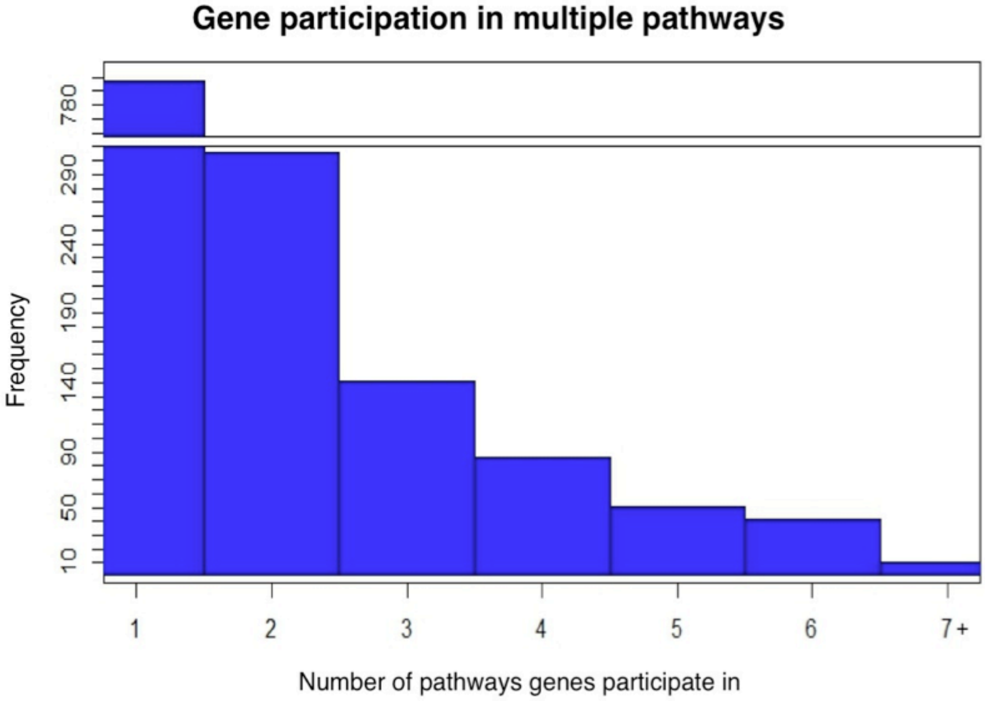


Figure 27: Frequency of gene participation in multiple pathways. Of the 1433 genes in the network, 797 (56%) of genes were found in 1 pathway, 304 (21%) of genes were found in 2 pathways, 332(23%) genes were found in 3 or more pathways.

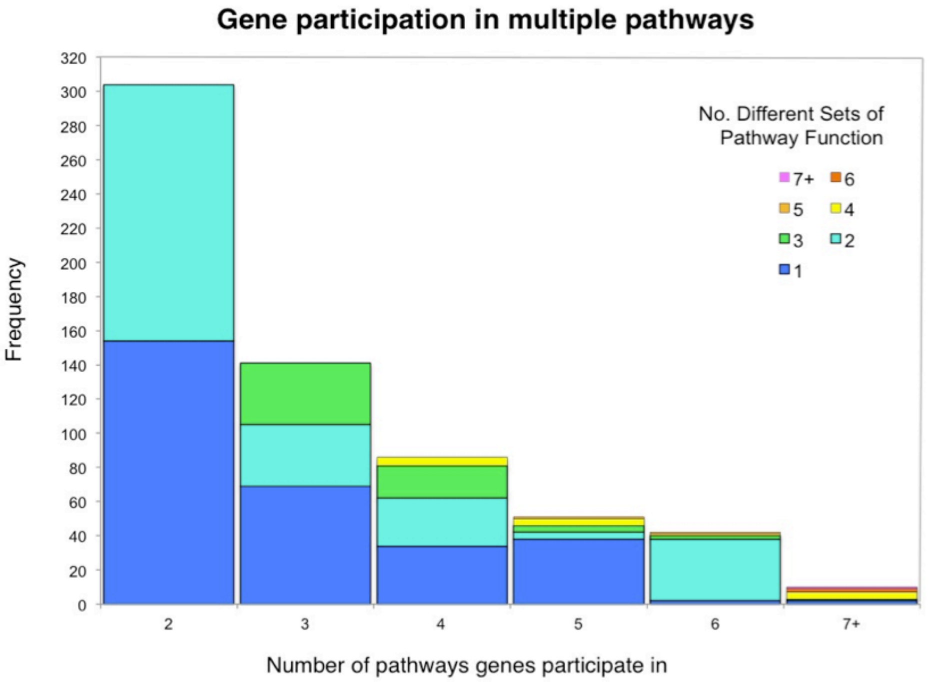


Figure 28: Functional variability of multi-pathway genes. Bars indicate the number of pathways that multi-pathway genes participate in. Bar colours indicate the number of disjoint functional profiles associated with genes' pathways

4.4.8 Pathway dependent gene/protein multi-functionality

Of the 1433 genes/proteins in the data set, 44% were found in multiple pathways, with the maximum number of pathways a gene/protein appeared in being 11 (gene AAT2)*(*Figure 27). If genes/proteins perform different functions in the context of different pathways then the functional profiles of these pathways will be different. Within our dataset, 83% of multi-pathway genes have a distinct functional profiles for each pathway they participate in. Pathway profiles were considered distinct if they did not contain identical sets of GO terms, however overlapping annotations were allowed. Annotation overlap between functional profiles is expected to be partially due to physical overlap between the pathways. Figure 28 shows the number of disjoint (non-overlapping) sets of functional annotations found in genes participating in multiple pathways. Two or more disjoint gene sets are frequently observed indicting that the genes are participating in distinct, context dependent pathways.

To further explore the possibility that genes/proteins acting in different pathways have different functions, we measured semantic distances between the functional profiles of multi-pathway genes/proteins (Figure 29 orange line). The mode distance between functional profiles is 4 showing that many of these pathways have highly similar profiles. These are likely to represent physically overlapping pathways. However, semantically distant GO terms (scoring between 5 and 11) were much more common between functional profiles than within functional profiles (blue line). This indicates that the pathway dependent functions of multi-pathway genes/proteins are frequently very different and the peak in frequency at a semantic distance of 8 indicates that these pathways may be as functionally unrelated as two pathways selected at chance.

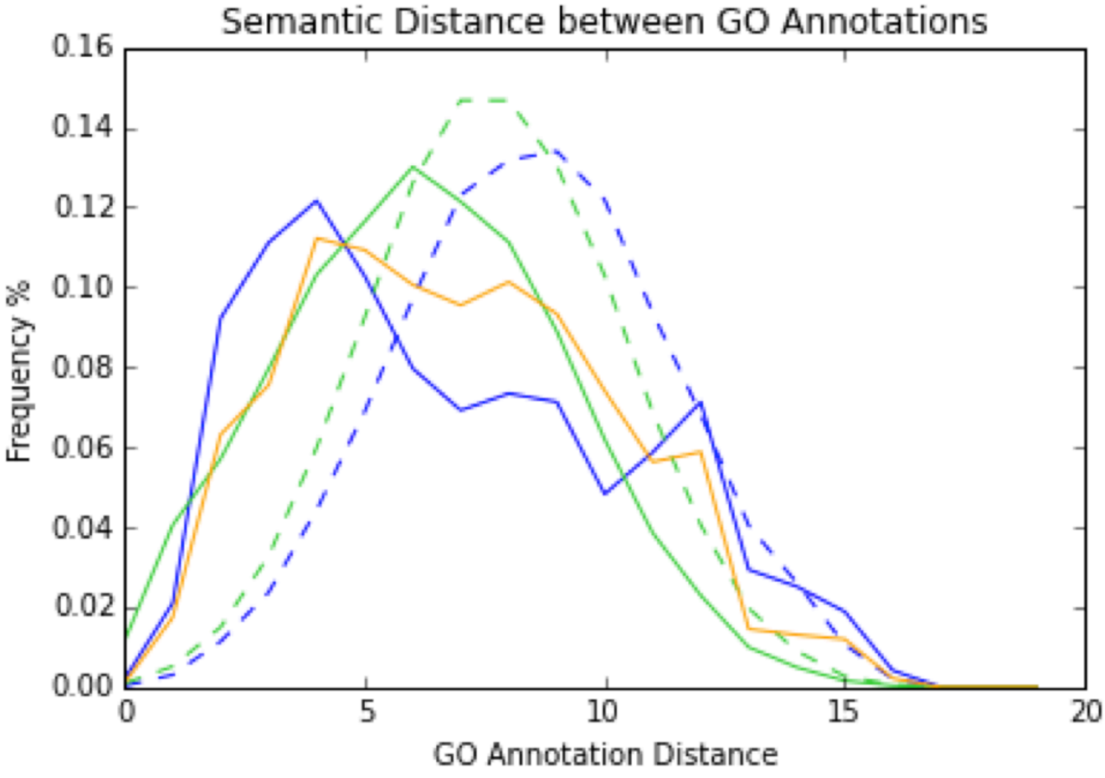


Figure 29: Semantic distance between multi-pathway genes' functional profiles. The solid blue and green lines show the frequency of distances between pairs of GO terms, within each of the pathway profiles. The dashed lines show the frequency of semantic distances between random annotation pairs. The orange line shows pairwise GO annotation distances between functional profiles of pathways sharing a multi-pathway gene. Comparison to the solid blue line shows increased semantic distances between the functional profiles of multi-pathway genes

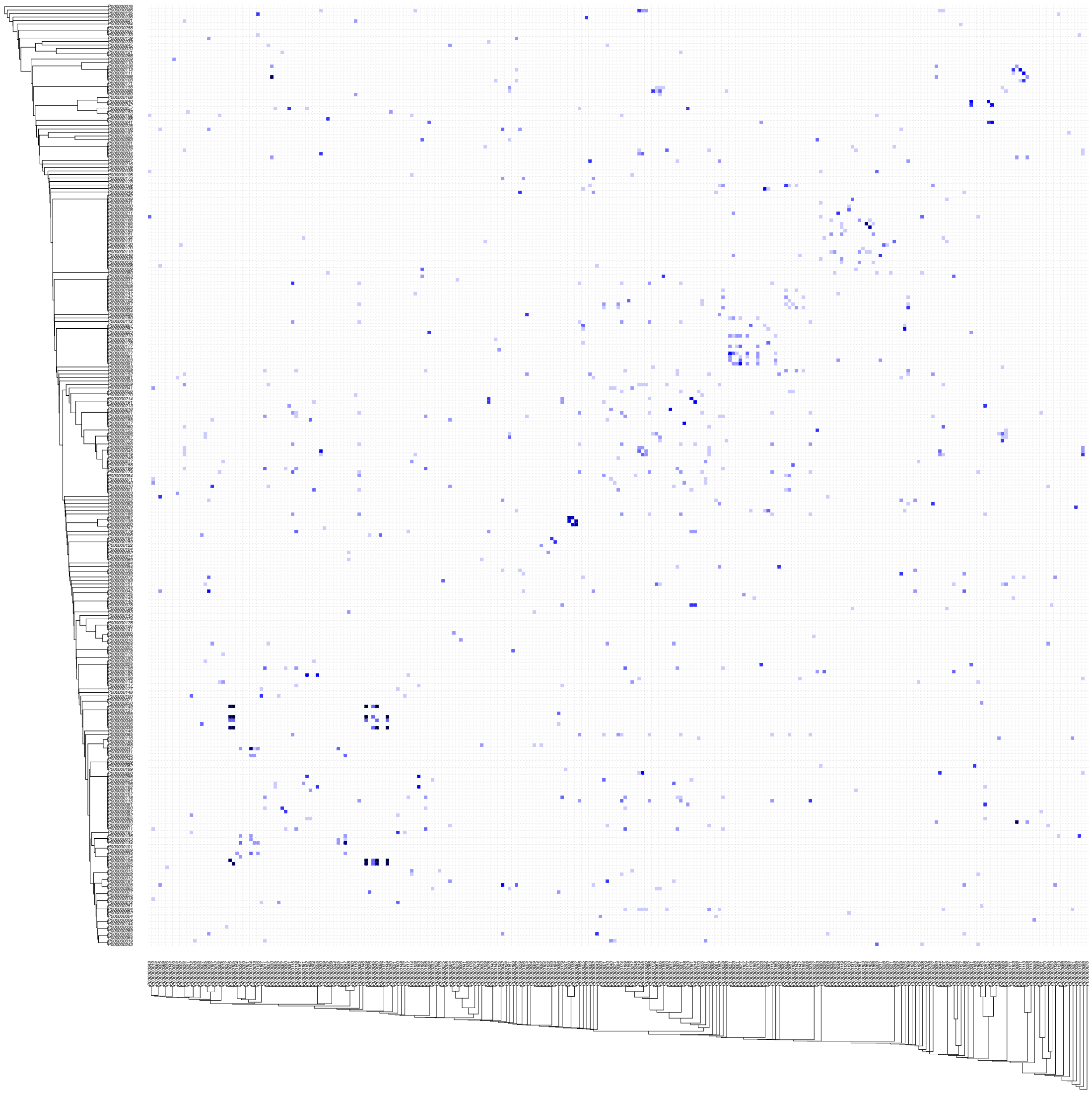


Figure 30: Relationship between pathway functionality and gene/protein overlap. Trees show pathways clustered by functional similarity. The heatmap shows gene/protein overlap. Dark cells along diagonal show that similar pathways are likely to share genes, showing some degree of pathway overlap in the data set. Dark cells positioned away from the diagonal show functionally unrelated pathways sharing genes.

Finally we examined the relationship between gene/protein overlap and pathway function by organizing pathways based on functionality then considering gene/protein overlap (Figure 30). Functionally divergent pathways can be seen sharing genes, indicating genes performing different roles depending on pathway context.

4.4.9 Comparison to Over Representation Analysis

To validate our results we compared them to DAVID (Huang, Sherman, et al. 2009; D. W. Huang et al. 2007), a tool commonly used for over representation analysis (ORA). We used DAVID to group genes based on GO annotation similarity. We then measured the number of shared annotations within each gene's functional profile, from pairs of genes in the same or different ORA groups. Gene pairs within the same DAVID groupings shared a mean of 3.9 annotations (variance 56.6, n 63,143) while genes in different DAVID clusters shared a mean of 0.6 (variance 4.7, n 764,398), indicating that the edges within our network are supported by DAVID functional groupings (T-test p=0.0). This result may, however be effected by a small number of gene pairs in the same DAVID group sharing large numbers of GO terms. A phi correlation of 0.23 was discovered between gene pairs that share at least one GO term and gene pairs in the same DAVID cluster (Table 4). This indicates a moderate to weak relationship between gene pairs sharing at least one GO term and being in the same DAVD cluster. This demonstrates that the use pathways to incorporate gene context, produces different results to assigning functional annotations without context.

Table 4: Gene pairs with common GO terms and DAVID groupings

	Shared annotation	No shared annotations
Shared DAVID group	32,544	30,599

No shared DAVID groups	131,141	633,257

When shared pathways, rather than annotations, were compared to DAVID groupings the correlation decreases down to 0.05 (see Table 5). This reflects the complex relationship between gene functions and pathway boundaries (functions are comprised of many pathways, pathways can participate in multiple pathways). Both tables produced significant results using the chi-squared test (p<<0.01) [6].

Table 5: Gene pairs with common pathways and DAVID groupings

	Shared pathway	No shared pathways
Shared DAVID group	1,498	48,745
No shared DAVID groups	6,873	757,525

4.4.10 Limitations

This method produces pathway annotations from GO data and organises pathways into a network representation of cellular function. The network contains 271 pathways, coving a wide range of functions including metabolism, signal transduction, gene expression and DNA maintenance. Yeast contains 6604 genes of which 5151 are characterised (Cherry et al. 2012), therefore the 1433 annotated genes analysed within

[6] The t-test, phi correlation and chi-squared were calculated in R. The phi correlation was calculated using the “psych” package (Revelle 2017)

the pathways of this network should not be considered complete coverage. This method can however be adjusted to allow more genes and pathways into the final data set, or to study specific sets of pathways. The highly frequent GO terms in supplementary table 1 (Appendix A) highlights the bias towards metabolic pathways within this network.

4.5 Conclusion

We have developed a method for organising cellular processes based on function, which accounts for temporal interactions modelled through pathways and allows multifunctional genes to be portrayed independently in their different biological contexts. The network illustrates the physical structure of function, as multiple pathways co-operate to ensure cellular processes are coordinated. Pathway multi-functionality was examined, determining that pathways vary greatly in the number and diversity of GO functions they facilitate. The functional variability of genes within multiple pathways was also demonstrated. Appreciation of multi-functionality at the level of both genes and pathways is critical for understanding pleiotropic genes and their relationship to multiple phenotypes, interpreting GIs and considering the transfer of information within the cell. Our representation of cellular function will enable analysis of gene/protein activity in the context of their functional roles, instead of the typical molecule-centric approach. This method can be adapted to incorporate different data types into the network, such as expression data and GI data. Future work will include incorporation of expression data to create directed edges showing the information flow between nodes.

Chapter 5

Reducing pathway redundancy using set theory algorithms

A shortened version of this chapter was submitted to BMC Bioinformatics on 13th September 2017[7]. All tools and data are available at https://github.com/RuthStoney/set-cover-and-set-packing-to-reduce-redundancy-in-pathway-data.

5.1 Abstract

5.1.1 Background

Pathway databases, such as KEGG, Reactome and ConsensusPathDB, contain high levels of overlap which can create redundancy in pathway enrichment analysis. Attempts to reduce this redundancy have focused on merging pathways, allowing neater data representation and aiding the interpretation of enrichment results. Previous methods have approached this problem using pathway merging, however merging pathways reduces functional specificity. In addition these methods often

[7] The submitted manuscript contained all of the methods and results regarding set cover, but omitted set packing for simplicity. RS performed all analysis and wrote the manuscript.

require a ‘maximum overlap’ threshold, and redundancy below this threshold cannot be addressed.

5.1.2 Results

We propose an alternative method using the set cover algorithm, to reduce pathway redundancy without merging pathways. The proposed approach considers three objectives: removal of pathway redundancy; controlling pathway size; and coverage of the gene set. By applying set cover to the ConsensusPathDB we were able to produce a reduced set of pathways, representing 100% of the genes in the original data set with 74% less redundancy, or 95% of the genes with 88% less redundancy. We also analyzed a set of enriched osteoarthritis pathways, revealing that within the top ten pathways, five were redundant subsets of more enriched pathways. Applying set cover to the enrichment results removed these redundant pathways allowing more informative pathways to take their place.

5.1.3 Conclusion

Our method provides an improved approach for handling pathway redundancy, while ensuring that the pathways are of homogeneous size and gene coverage is maximised. Pathways are not altered from their original form, allowing knowledge regarding the data set to be directly applicable. The ability of the presented algorithms and parameters to meet the objectives are discussed, enabling users to prioritise redundancy reduction, pathway size control or gene set coverage. The application of set cover to upregulated gene sets selects the most informative pathways for biological interpretation.

5.2 Background

Pathways are sets of genes corresponding to functionally related interacting proteins. Pathway data is available from many databases dependent on biological focus. It is commonly used for pathways analysis, a method of reducing large sets of over-expressed genes into sets of enriched pathways (Khatri et al. 2012). Pathways can be used to study disease and drug interactions (Kanehisa et al. 2008) and have uses in systems biology (Stoney et al. 2015).

The fragmented nature of pathways across multiple databases makes it difficult to perform inclusive analysis of all known data. To address this issue, many attempts have been made to consolidate pathway databases such as ConsensusPathDB (CPDB) (Kamburov et al. 2009), PathwayCommons (Cerami et al. 2011), The Human Pathway Database (HPD) (Chowbina et al. 2009), Pathway Interaction Database (PID) (Schaefer et al. 2009), HiPath (N. Yu et al. 2012) and NCBI Biosystems (Geer et al. 2009). They consolidate pathways from different databases into a consistent searchable format, however the arbitrary nature of pathway boundaries results in overlap and redundancy. This redundancy forms an obstacle to the use and interpretation of pathway data. The HPD addresses the redundancy between pathways by visualizing overlap to the user (Chowbina et al. 2009).

Redundancy Control in Pathway Databases (ReCiPa) (Vivar et al. 2013) uses a pathway merging algorithm, to combine pathways with high levels of overlap. Pathways are treated as sets, therefore when pathways are merged, overlapping genes are only represented once within the new pathway. Users select a maximum overlap threshold and pathways pairs displaying greater levels of overlap are merged. Vivar et al. (2013) reduced overlap across five large databases (KEGG, Biocarta, CGP, NCI-PID, and Reactome), finding redundancy in all of them. They proceeded to merge pathways from the Molecular Signatures Database (MSigDB) whose overlap exceeded 75%, then demonstrated improved success performing enrichment analysis on the new dataset. Pathways ranked based on their association with obesity showed less overlap and genes within the pathways showed higher significance towards the disease.

Pathcards described a multistep procedure to reduce pathway redundancy, again by merging pathways (Belinky et al. 2015). Two thresholds were calculated with the aims of minimizing pathway overlap and preventing pathways from becoming too large to be informative. Pathway pairs with overlap greater than the first threshold were merged using hierarchical clustering, then remaining pathway pairs with overlap greater than the second threshold were merged using nearest neighbor joining. By merging pathways into super-pathways, Pathcards suggested many new molecular interactions. They demonstrated that many of these newly generated interactions are supported by high numbers of literature co-mentions and high experimental interactions scores according to STRING. However, the majority of the new interactions were not supported by literature co-mentions and experimental interaction scores, reducing the confidence of the super-pathway data.

A disadvantage to the threshold merging approach that Pathcards and ReCiPa use is that redundancy between pathway pairs can only be removed if the overlap exceeds the threshold. The threshold is restricted by the need to preserve the informativeness of the pathways and to prevent the data set from being consolidated into a small number of large pathways. Applying uniform thresholds to a data set with unequal coverage can be problematic. Heavily studied areas with many overlapping pathways require a higher threshold to prevent large numbers of pathways being merged into a giant uninformative set. Pathway Distiller avoids thresholds by merging each pathway to the pathway it overlaps with the most (Doderer et al. 2012), allowing varying levels of similarity within clusters, since even pathways with very little overlap must be linked to a partner pathway.

Pathway enrichment analysis and functional enrichment analysis are major applications of pathway data. The impeding effect that similar, hierarchical and redundant terms have on interpretation of enrichment analysis results is well documented, with tools such as DAVID clustering terms into related groups to simplify results for the user (D. W. Huang et al. 2007). Alexa et al. (2006) introduced two algorithms, *elim* and *weight*, which use the Gene Ontology topology to deal with the issue of redundancy in enrichment results (Alexa et al. 2006). The *elim* method preferentially selects the most specific enriched GO terms, by selecting the terms closest to the ontology tips. More general ancestor terms are only selected if they are required to cover all of the enriched genes in the set. They acknowledged that in some instances parent terms have higher p-values than their child nodes and therefore the *weight* method was introduced. In this method, if an ancestor node has higher significance than a child node, then the significance of the child nodes will be decreased, preventing the child nodes from being reported.

We propose an approach using set theory algorithms to identify the minimum subset of pathways required to cover the dataset, rather than merging pathways. Since pathways are not merged, database and literature information on existing pathways remains directly applicable and functional specificity is not lost through pathway size expansion. The proposed method also removes the risk of biologically distinct pathways being merged. The algorithm's ability to remove overlap is not limited by a minimum overlap threshold and it can consider redundancy between more than two pathways. We propose a collection of simple efficient algorithms suitable for use on large datasets.

We implemented a version of our algorithm capable of handling ranked pathway enrichment data and applied it to a set of enriched osteoarthritis pathways (Dunn et al. 2016). By selecting GO terms based on p-values, the *weight* method (Alexa et al. 2006) shares some conceptual similarity with the set cover method introduced. However, the set cover method relies on shared gene membership across pathways to indicate redundancy, rather than looking for related annotations within the gene ontology topology.

5.3 Approach

We downloaded pathway data from ConsensusPathDB (CPDB), an open source online collection of pathways, that incorporates 32 sources including KEGG, Wikipathways, PDB, Reactome. CPDB makes these resources available as a single download, which we acquired on 24/09/2015 containing 4,011 pathways. We applied set theory algorithms to the CPDB data set, analyzing their effectiveness at: reducing pathway overlap; reducing pathway size variability; and preserving the maximum number of genes in the data set. We considered two algorithms set packing and set cover. Set packing reduces redundancy by selecting the pathways that show minimal overlap and deleting overlapping pathways. In contrast set cover algorithms select the pathways with the most uncovered genes. This can lead to set cover algorithms selecting the largest pathways in the data set. The selection of large pathways is detrimental to pathway specificity and since set cover does not directly limit overlap, set packing appeared to be the most promising alternative.

Set packing is a well-defined algorithm in computer science for handling overlapping sets of sets, which is useful when overlap between the selected sets is not allowed (Kordalewski 2013). It was therefore used to schedule surgical procedures given a set of finite resources, where no resource may be simultaneously used in two procedures (Velásquez & Melo 2005). Set packing has also been used to examine genomic rearrangements between genomes (Chen et al. 2011).

Set cover is a related algorithm which has been used by CLASS, a bioinformatics program that maps RNA sequence data to transcripts (Song & Florea 2013). Set cover has also been used to predict protein-protein interactions based on binding domains (C. Huang et al. 2007), to reduce the complexity of single nucleotide polymorphism (SNP) sets (Ao et al. 2005) and to minimize the number of probes needed to analyze DNA (Borneman et al. 2001).

Both algorithms deal with elements and sets, which relate to genes and pathways respectively. All the unique genes in the data set are collectively referred to as the universe. The aim is to produce a reduced selection of sets (pathways), which collectively cover all the elements (genes) in the universe (dataset). The subset of the original data generated by set cover is called the cover set and the subset generated by set packing is called a packing (Kordalewski 2013). Each time a pathway is added to the cover set or packing the genes in the pathway become covered.

Application of the set packing algorithm lead to unacceptable gene loss and the set cover algorithm lead to large highly general pathways dominating the cover set. Therefore, modifications of both algorithms were implemented to better covering the dataset and control pathway size, while reducing redundancy.

When dealing with enrichment analysis data the aim is to reduce redundancy between pathways, while preserving the order of enrichment significance denoted by the p-values. We designed an algorithm that would select the set of enriched pathways with the lowest p-values capable of covering all the genes in the data-set. This helps ensure that the most enriched pathways represent as many over-expressed genes as possible. The filtered results will also return the most enriched pathways available for each gene.

5.4 Methods

5.4.1 Overlap score

To measure overlap across different algorithms we measured the mean number of pathways that each gene appears in. Within the raw data genes appeared in a mean of 12.4 pathways. We refer to this metric as the overlap score.

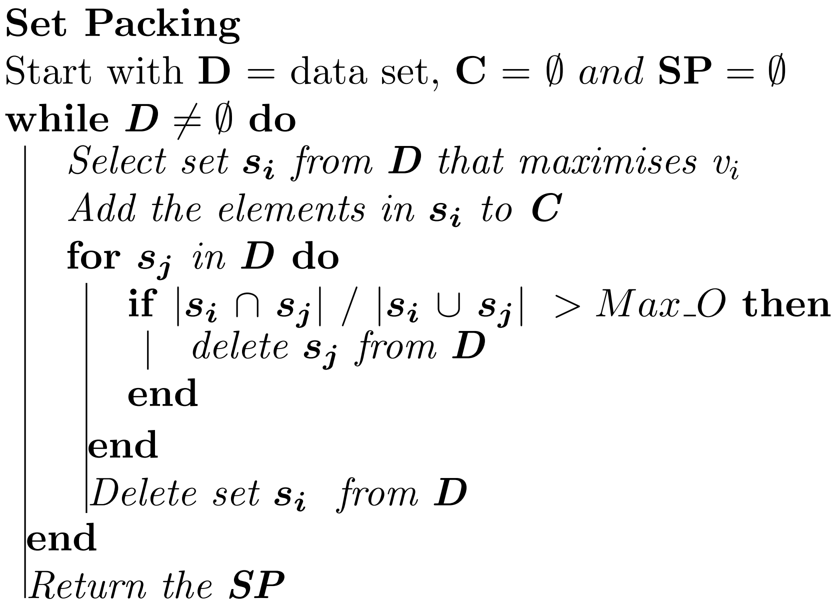
5.4.2 Set packing

Set packing generates a selection of sets (a packing) in which none of the sets overlap, which makes it an obvious starting point for reducing redundancy. It generates this discrete output by selecting pathways that overlap with the minimum number of other pathways, then deleting any pathways that overlap with the selected pathways (Supplementary Figure 2 in Appendix B). First, set packing values must be calculated for each pathway, corresponding to the number of pathways each pathway overlaps with (Equation 7).

$$v_i = \frac{1}{|\{j \in I \mid \boldsymbol{s_i} \cap \boldsymbol{s_j} \neq \emptyset\}|} \qquad (11)$$

where each pathway's value (*v_i*) is calculated as the inverse of the number of overlapping pathways it has in the dataset. *v_i* is the value of pathway, *I* is the dataset, ***s_i*** is the pathway whose score you're calculating and ***s_j*** is each other pathway (Kordalewski 2013).

The pathway with the highest value is selected first, then any overlapping pathways are deleted. Deleting pathways makes it necessary to recalculate the set packing values, before the next highest scoring pathway is selected. This process continues until all of the pathways have been added to the packing or discarded. Algorithm 1 shows a pseudocode depiction of the algorithm.



```
Set Packing
Start with D = data set, C = $\emptyset$ and SP = $\emptyset$
while $D \neq \emptyset$ do
    Select set $s_i$ from D that maximises $v_i$
    Add the elements in $s_i$ to C
    for $s_j$ in D do
        if $|s_i \cap s_j| \;/\; |s_i \cup s_j| > Max\_O$ then
            delete $s_j$ from D
        end
    end
    Delete set $s_i$ from D
end
Return the SP
```

Algorithm 1: Set packing

where ***D*** is the full set of pathways, ***C*** in the covered genes, ***SP*** is the set packing output, and ***s_i*** and ***s_j*** are pathways.

We found that the unmodified application of this algorithm successfully removed all overlap and did not inflate size variability, however 74% of the genes in the dataset were lost. To preserve a larger proportion of the data set, we modified the algorithm to allow pathways that overlapped extensively with the selected pathway to be removed, while pathways with only a slight overlap were retained. It was necessary to introduce a maximum overlap (*Max_O*) threshold, to allow overlapping pathways to be retained if the proportion of overlapping genes was less than the threshold. In the standard set packing algorithm, the *Max_O* is set to 0, indicating that 0% overlap is permitted. We experimented setting *Max_O* to 10, 20, 30, 40 and 60%, using the Jaccard cocoefficient to measure overlap. The Jaccard coefficient measures the number of overlapping genes, divided by the total number of genes in both pathways. In the resulting packing, no two pathways can overlap by more than the *Max_O* Increasing *Max_O* successfully increased the proportion of the data set conserved.

5.4.3 Set cover

We applied the set cover algorithm to the data set, which generates a selection of pathways called a cover set, in which all the genes in the data set are present or "covered". Set cover begins by first assigning values to each pathway (v_i). Set cover values correspond to the number of uncovered genes each pathway contains (Equation 8).

$$v_i = |\ \boldsymbol{s_i}\ \cap \boldsymbol{R}\ | \qquad (12)$$

where ($\boldsymbol{s}_i$) is the pathway's gene set and $\boldsymbol{R}$ is the set of all uncovered genes.

At the beginning of the algorithm all the genes in the dataset are uncovered so the algorithm selects the largest pathway. The genes from the selected pathway are then covered, so it is unnecessary to cover them again using additional pathways. The algorithm then recalculates how many uncovered genes each pathway contains and continues to add the pathway with the maximum value to the set cover until all genes in the data set are covered (Supplementary Figure 3).

```
Set Cover
Start with $\mathbf{R} = \mathbf{U}$, $\mathbf{C} = \emptyset$ and $\mathbf{SC} = \emptyset$
while $|\boldsymbol{C}|/|\boldsymbol{U}| * 100 < GC$ do
    Select set $\boldsymbol{s_i}$ that maximizes $\boldsymbol{v_i}$
    Add $\boldsymbol{s_i}$ to $\boldsymbol{SC}$
    Add the elements in $\boldsymbol{s_i}$ to $\boldsymbol{C}$
    Delete the elements in $\boldsymbol{s_i}$ from $\boldsymbol{R}$
end
Return $\boldsymbol{SC}$
```

Algorithm 2: Set cover

where ***R*** is the uncovered genes, ***U*** is all the genes in the dataset, ***C*** is the covered genes, ***SC*** is the set cover result, *GC* is the gene coverage (see Section 5.4.4) and $\boldsymbol{s_i}$ is a pathway.

Application of the set cover algorithm was effective in reducing overlap between the pathways; however, it selected very large pathways with reduced informativeness (maximum size 2320, standard deviation 160.1, almost double the standard deviation on the original dataset 86.9). We therefore explored methods that avoid preferential selection of large pathways.

5.4.4 Gene Set Coverage

As the set cover algorithm approaches completion and the final sets are added to the cover set, increases in data coverage are gained at the expense of redundancy reduction. This is because the final sets required to cover the few remaining genes tend to have the most overlap with other pathways already in the set cover. In addition, fewer pathways are available to cover the final few genes, restricting options to control pathway size. To allow a user-defined compromise between the gene coverage, pathway redundancy and pathway size we introduce the Gene Coverage (*GC*) parameter. Setting *GC* below 100% allows the algorithm to finish before the final elements have been covered. We experimented setting *GC* to 90, 95, 99 and 100% of the number of genes in the data set.

5.4.5 Proportional set cover

When reducing pathway redundancy there are three competing aims: reducing redundancy; controlling pathway size; and covering the entire gene set. The proportional set cover algorithm was generated to focus on controlling pathway size.

To control the size of the pathways we altered the scoring mechanism to rank pathways based on the proportion of uncovered genes they contained, rather than the absolute number (Equation 9). This works because larger pathways are more likely to have a proportion of their genes covered when other pathways are selected. Additionally this mechanism directly penalizes overlap, which the standard algorithm does not. At the beginning of the proportional set cover algorithm none of the genes are covered so the proportion of uncovered genes in every pathway is 1. This would result in the starting pathway being selected at random. To ensure that pathway size variability is controlled as strictly as possible, we implemented the second part of Equation 9, which ensures that pathways of mean pathway size are preferentially selected when multiple pathways with the same proportion of uncovered genes are available.

$$v_i = \frac{|\ \boldsymbol{s_i}\ \cap \boldsymbol{R}\ |}{|\boldsymbol{s_i}|} + \frac{1}{abs(\ |\boldsymbol{s_i}| - \overline{|\boldsymbol{s_\iota}|})\ * k} \tag{13}$$

where $\boldsymbol{s_i}$ is the pathway's gene set, $\overline{|\boldsymbol{s_\iota}|}$ is the mean pathway length, ***R*** is the uncovered genes set and *k* is a large constant to limit the influence of the second term (taken equal to 10,000).

5.4.6 Hitting set cover

The set-covering problem can be reformulated into the equivalent set-hitting problem. In this formulation genes and pathways are visualized as bi-partite graph in which the pathways are connected to the genes that they contain. In this depiction it is clear that some genes are only linked to a single pathway, which must be selected if the gene is to be covered. The importance of pathways can therefore be considered as a factor of how infrequent their genes are. The hitting set cover is therefore designed to reduce redundancy as much as possible without directly selecting for pathway size.

We calculated the frequency of each gene in the data set (*F*), then assigned the gene's value *gv(j)* as 1/*F*. We then assigned a value v_i to each pathway defined as the sum of each uncovered gene's scores divided by the number of genes in the pathway (Equation 10).

$$gv(j) = 1/F(j) \tag{14}$$

$$v_i = \frac{\sum_{j \epsilon \boldsymbol{s_i} \cap \boldsymbol{R}} gv(j)}{|\, \boldsymbol{s_i} \,|}$$

where ***gv(j)*** is the value of a gene, *F(j)* is the number of pathways a gene is in, $j\epsilon\boldsymbol{s_i} \cap \boldsymbol{R}$ means for each uncovered gene in the pathway and $|\boldsymbol{s_i}|$ is the length of the pathway.

5.4.7 Set cover for pathway enrichment analysis

Pathway analysis is a frequently used method; therefore a modified set cover algorithm to address this situation could be highly useful. The universe represents up-regulated genes and the sets are enriched pathways. Enrichment analysis output provides a unique circumstance in which the sets are already scored (p-values). We wish to reduce redundancy (gene overlap) between enriched pathways and it is essential that the pathways with the lowest possible p-values are selected. Equation 11 allows pathways with the lowest p-values to be selected, unless all of their genes are covered by pathways with even lower p-values.

$$\boldsymbol{s_i} \cap \boldsymbol{R} = \theta \rightarrow b = 0, \qquad \boldsymbol{s_i} \cap \boldsymbol{R} \neq \theta \rightarrow b = 1 \tag{15}$$

$$v_i = (1 - pvalue_i) * \ b$$

where $\boldsymbol{s_i}$ is the pathway's gene set, ***R*** is the uncovered gene set, *b* is a binomial operator, $pvalue_i$ is the pathway's p-value and v_i is the pathway's set cover value.

5.5 Results

We started with the large, extensively redundant CPDB data set and used set theory algorithms to reduce pathway overlap, while controlling pathway size and seeking to cover as much of the data set as possible. The unmodified set packing algorithm only covered 25% of the gene set, rendering it unusable, therefore the *Max_O* parameter was implemented. The ability of this modified algorithm to meet the three objectives is first described. Next we describe the ability of the standard set cover algorithm and two modified set cover algorithms, in conjunction with the GC parameter, to meet the above objectives.

Set packing algorithm

Figure 31 shows the effectiveness of various *Max_O* values at reducing overlap and maximizing coverage of genes in the pathway set. When no overlap was allowed between pathways (*Max_O* = 0) each gene only appeared in one pathway; however, only 26% of genes were recovered from the data set. As the permitted overlap was increased, the proportion of the genes returned in the packing set also increased. This is expected as the *Max_O* dictates the maximum amount of overlap permitted between 2 pathways; for example, the threshold of 0.3 produced a set of pathways covering 97.1% of the data set, in which no two pathways could overlap by more than

30%. The overall overlap score was reduced to 5.8, less than half of the original overlap score (12.4). When *Max_O* was increased to 0.6, 99.7% of the genes in the data set were covered by the packing but the overlap score increased to 9.2.

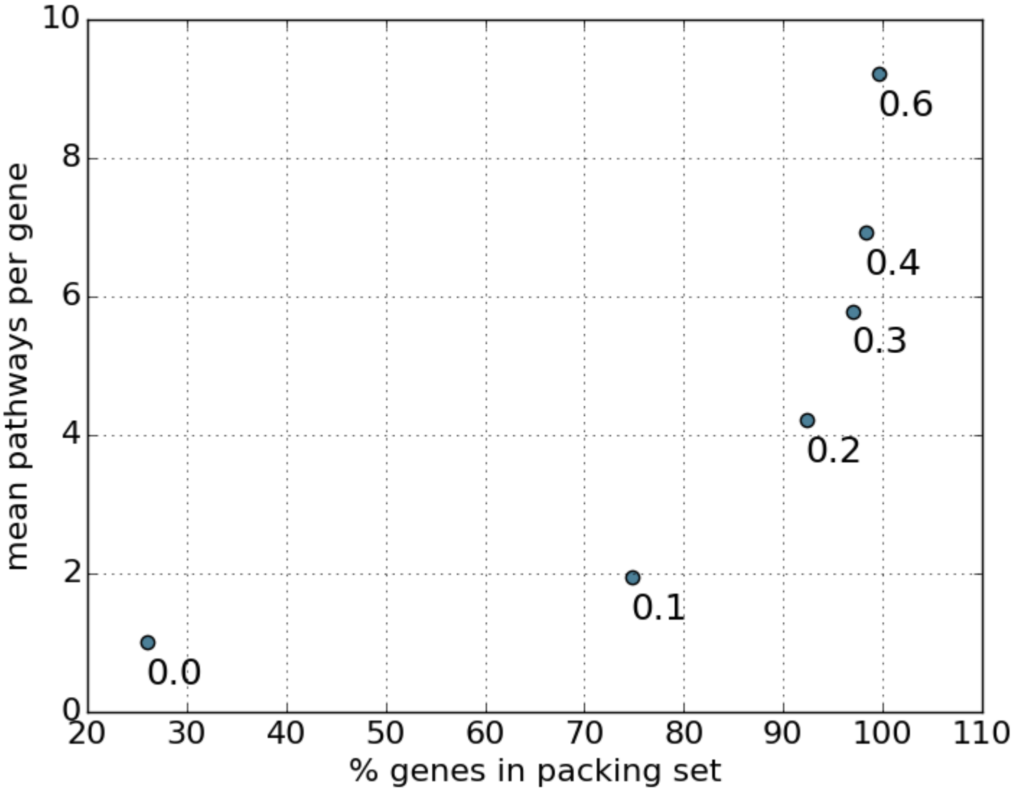


Figure 31: Set packing Gene Cover. Overlap score (mean pathways per gene) in the output of the set packing algorithm with Max_O set to a range of thresholds (0, 0.1, 0.2, 0.3, 0.4, 0.6). The Y-axis shows the overlap score, the X-axis shows the proportion of the gene set covered and the scatter points show different Max_O thresholds.

Pathways cannot overlap by more than the *Max_O* preventing high levels of overlap between pathways, however it's ability to reduce redundancy between pathway pairs that overlap by less than *Max_O* is limited. The algorithm will stop once all genes have been covered or deleted, preventing unnecessary pathways from being added, however the *Max_O* parameter shares disadvantages to the thresholds present in

ReCiPa and PathCards (Vivar et al. 2013; Belinky et al. 2015). To find a more satisfactory solution we moved onto the set cover algorithm.

5.5.1 Pathway redundancy varies between different set cover algorithms

The original pathway data set contained 11,196 genes and 3,305 pathways; the starting overlap score (see Section 5.4.1) was 12.4. The standard set cover algorithm reduced overall redundancy from 12.4 to 4.1, a 73% reduction (since the minimal possible overlap is 1). The overlap score for proportional set cover was 4.36, slightly higher than the standard set cover algorithm, but still representing a 70% reduction in overlap from the original data. The hitting set cover algorithm was designed to select pathways that contained rare genes within the data set, resulting in the greatest reduction in overlap (overlap score of 3.95 equivalent to a 74% reduction).

After application of the set cover algorithms the distribution of the remaining overlap varied greatly. Figure 32 shows the Jaccard coefficient between pairs of pathways, in the outputs produced by each of the three algorithms. The standard set cover algorithm produced the lowest maximum overlap (0.68) between the pathway pairs. However, compared to the original data, a higher proportion of pathway pairs in the set cover output showed between 10-30% overlap. Proportional set cover had the greatest maximum overlap at 0.93, out of all of the algorithms. The hitting set cover algorithm produced a maximum overlap between two pathways of 0.82, despite having the lowest overlap score.

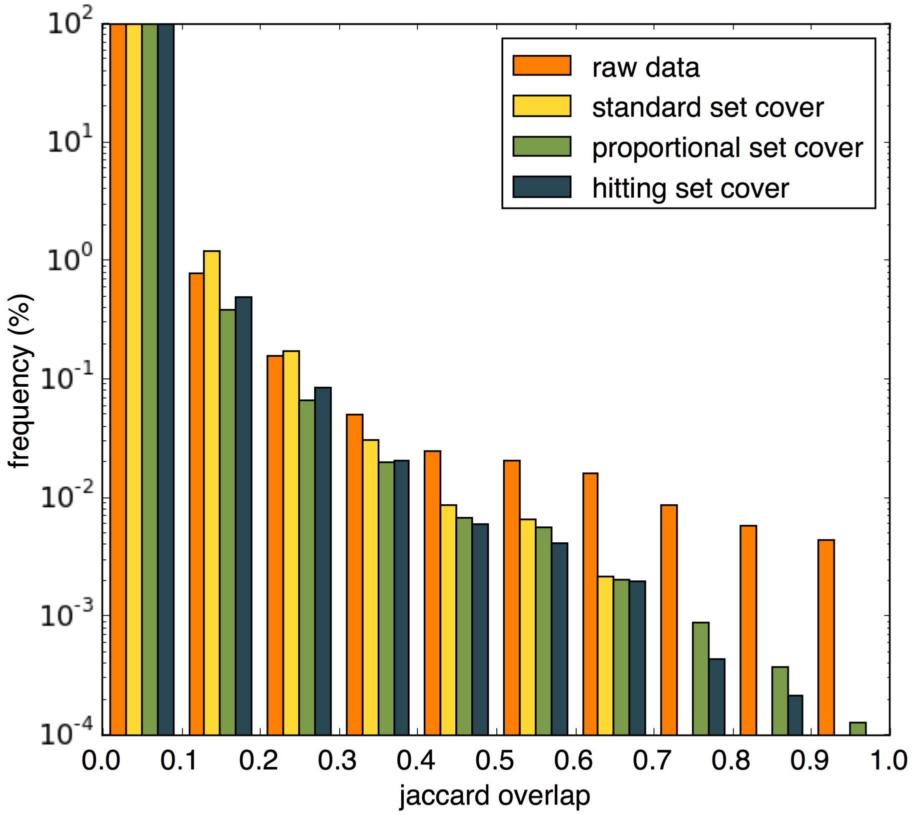


Figure 32: Jaccard coefficient between pathway pairs in the cover sets produced by each algorithm

5.5.2 Gene Coverage can be lowered to reduce redundancy

For each of the algorithms it is possible to use the *GC* parameter to prioritize reductions in redundancy over gene coverage by stopping any algorithm before all of the genes in the dataset have been covered. Figure 33 shows improved ability of the set cover algorithms to reduce pathway overlap for different values of *GC*. If 99% of the genes are required then the hitting set algorithm achieves the lowest overlap score of 3.24, equivalent to an 80% reduction in overlap. Redundancy can be further reduced if only 95% of the genes are covered, with the proportional and hitting set algorithms producing an overlap score of 2.41, equivalent to a 88% reduction in redundancy. Both the proportional set cover and the hitting set cover are more effective at reducing redundancy than the standard set cover if *GC* is set to less than 100%.

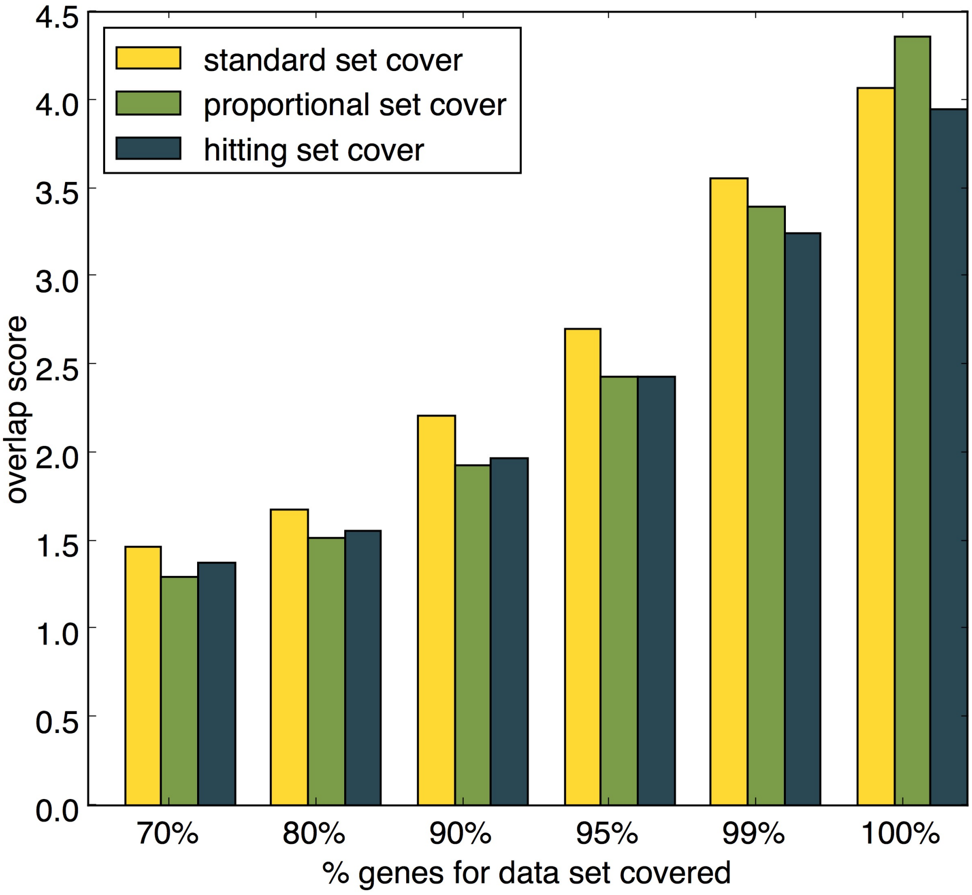


Figure 33: Redundancy in set cover outputs given different *GC* values

5.5.3 Pathway size is affected by the set cover algorithm and Gene Coverage setting

When *GC* was set to 100% the standard set cover algorithm represented all of the genes in the dataset using only 524 pathways (16% of the original pathway set). However, many of these were very large increasing the mean size to 87.2 (standard deviation 160). These pathways have reduced informativeness since functional specificity is lost.

Figure 34A illustrates the tendency of this algorithm to select extremely large pathways.

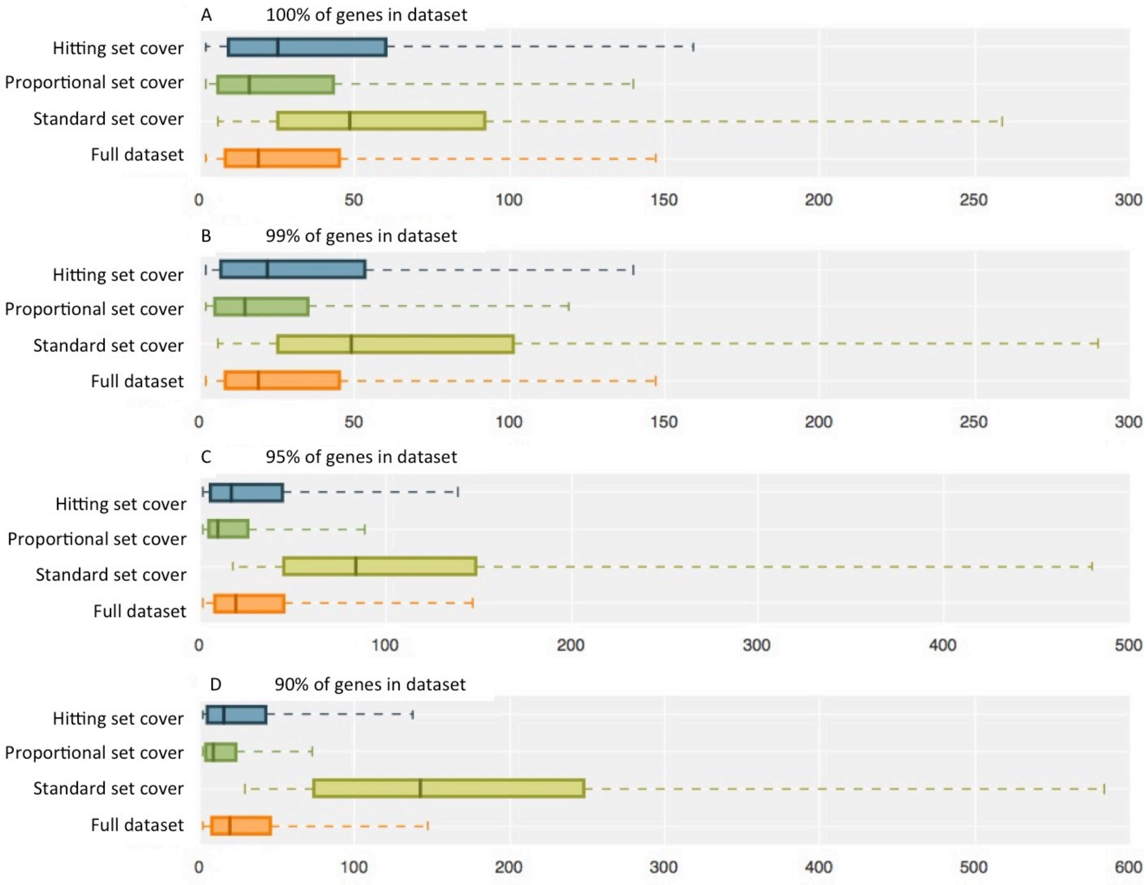


Figure 34: Affect of Gene Cover on pathway size. Pathway sizes in cover set when GC is set to A) 100%, B) 99%, C) 95% and D) 90%. The boxes indicate the 25th and 75th percentiles and the whiskers indicate the 5th and 95th percentiles.

The proportional set cover algorithm was designed to preferentially select moderately sized pathways. This returned a cover set of 1,336 pathways with controlled size variation (mean of 36.5, standard deviation 55.1) shown in

Figure 34A. The hitting set cover algorithm was less able to control pathway size than the proportional set cover algorithm, returning 957 pathways with a mean size of 46.2 (standard deviation 61.7).

Figure 34B – D show that as *GC* is reduced the tendency of the standard set cover to select very large pathways becomes more exaggerated. Decreasing *GC* also improves

the ability of the proportional set cover algorithm to select moderately sized pathways. The hitting set algorithm also tends to select smaller pathways when *GC* is reduced, since larger pathways often contain more frequent genes. Reducing *GC* affects pathway size since in the later stages of the algorithm, fewer pathways are available to cover the remaining genes, reducing the available options. Therefore lowering *GC* has the ability to help control pathway size when the proportional set cover and hitting set cover algorithms are used.

5.5.4 Reducing redundancy in pathway enrichment analysis

To demonstrate the ability of the set cover algorithm to handle enrichment data, we applied the enrichment set cover algorithm an osteoarthritis data set, retrieved from Dunn et al. (2016) (Dunn et al. 2016). From the osteoarthritis data set, 58.3% of the upregulated genes could be mapped to a CPDB pathway, which was a 17% improvement on the GOseq (Young et al. 2010) implemented data set. We retrieved 42 enriched pathways with a p-value lower than 0.05, following the Benjamini-Hochberg correction for multiple testing. Set cover for enrichment analysis reduced the number of pathways required to cover the upregulated genes to 23 (Supplementary Table 2). The heat-map in

Figure 35A shows the asymmetric overlap between the top ten pathways before application of the algorithm. The p-values from pathway enrichment determine the order in which pathways were considered for inclusion in the cover set. Pathways were omitted if all of their genes were covered by more enriched pathways. Note that overlap tends to be higher in the bottom left triangle as pathways added later were often smaller subcomponents of larger pathways.

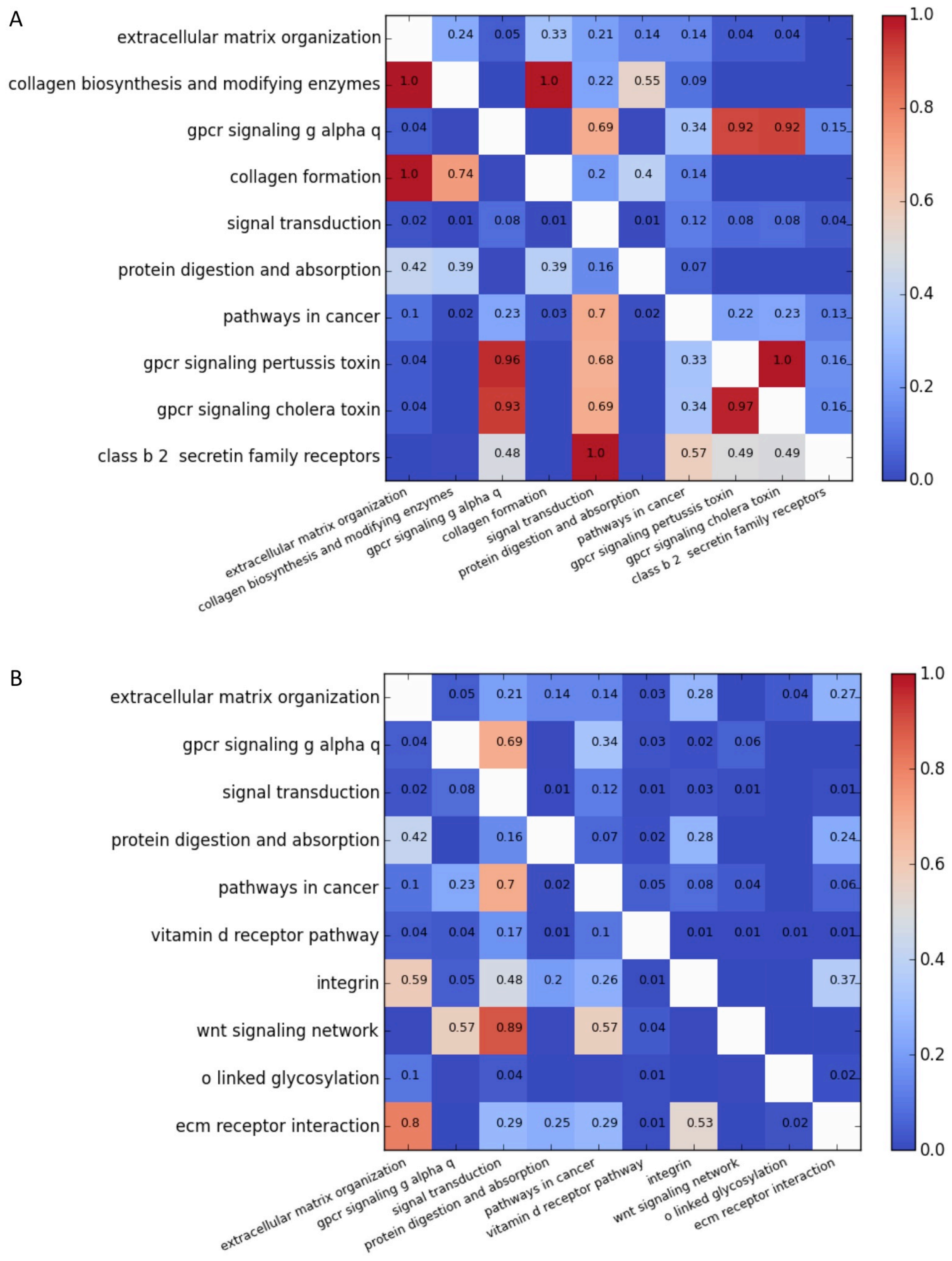


Figure 35: Pathway redundancy heat maps A) Pathway overlap for top ten enriched pathways. B) Pathway overlap for top ten enriched pathways after application of set cover. The values represent asymmetric overlap, i.e. for each pathway shown on the left axis values represent the proportion of genes included in the pathway shown on the bottom axis.

Some pathways in the enrichment set cover do still show significant levels of overlap, for example ’wnt signalling network’ is included despite 89% of its genes being covered by ’signal transduction’. This is acceptable because ’signal transduction’ is more highly enriched than ’wnt signalling network’, yet the ’wnt signalling network’ is worth including as it covers enriched genes missed by ’signal transduction’. The unmodified top ten enriched pathways only cover 78.0% of the enriched genes. Using the set cover enrichment algorithm increases this figure to 85.2% without disrupting the pathway order given by the enrichment p-values.

We can see that ’extracellular matrix organization’, the most enriched pathway, was placed in the cover set first. Next was ’collagen biosynthesis and modifying enzymes’; however, all of the genes in this pathway are contained in the larger pathway ’extracellular matrix organization’, as indicated by the red cell in the ’collagen biosynthesis and modifying enzymes’ row, ’extracellular matrix organization’ column. The corresponding cell in the ’extracellular matrix organization’ row reveals that 24% of the genes in ’extracellular matrix organization’ are also in ’collagen biosynthesis and modifying enzymes’. Figure 35B shows overlap between the top ten pathways after application of the enrichment set cover algorithm. Because ’collagen biosynthesis and modifying enzymes’ is a subset of ’extracellular matrix organization’, it is not returned in the cover set (

Figure 35B). The second item in this list then becomes ’GPCR signaling g alpha q’. In the enrichment set cover, ’collagen formation’ and ’class b 2 secretin family receptors’ are removed because they are subsets of ’extracellular matrix organization’ and ’signal transduction’ respectively. Additionally, ’GPCR signaling pertussis toxin’ and ’GPCR signaling cholera toxin’ are absent from the returned list, as all of their genes are found in ’GPCR signaling g alpha q’ or ’signal transduction’.

5.6 Discussion and conclusion

We described algorithms suitable for reducing overlap in large pathway data sets and enrichment analysis results. Set packing reduced redundancy without effecting pathway size and removed all redundancy, however it had unacceptably poor gene coverage. To improve coverage we implemented the *Max_O* threshold, however this limited the algorithms ability to reduce redundancy between pathways if the overlap was less than *Max_O*. These issues diminish the suitability of the set packing algorithm to reduce redundancy between pathways.

Implementation of the standard set cover significantly increased pathway size, therefore the proportional set cover and hitting set cover algorithms were developed to overcome this issue. The proportional set cover is the best algorithm for controlling pathway size and the hitting set cover is the preferred choice for covering all of the genes in the dataset with minimal pathway redundancy. We showed that reducing the GC parameter allows further reductions in pathway redundancy; for example, if only 95% of the genes in the CPDB dataset were covered redundancy can be reduced by up to 88%. In addition reducing GC increases pathway size control when the proportional set cover and hitting set cover algorithms are used.

For pathway enrichment analysis we aimed to reduce redundancy while selecting the most significant pathways based on p-values. As an application we used the modified set cover algorithm to reduce the results of enrichment analysis from a large osteoarthritis data set. We found that 5 out of the 10 top ranking pathways could be omitted as all of their genes were covered by more highly enriched pathways. Overlap between pathways returned from enrichment data is not always immediately obvious and requires further consideration. By reducing this redundancy, data interpretation is made more intuitive. Reducing redundancy also allows the user to explore substantially more of the data set using the same number of pathways.

Set cover uses greedy heuristic methods, which provide good approximations of the optimal solution in a time effective manner. These methods are extremely efficient and can be run in a matter of minutes, however it should be noted that they do not guarantee an optimal solution. This is particularly true for the proportional set cover algorithm where the randomness of early selections influences the result. However, all possible outcomes result in reduced redundancy. The enrichment set cover algorithm is exempt from these considerations unless multiple pathways have identical p-values.

We have provided a method to dramatically reduce redundancy in pathways facilitating cleaner analysis of cellular processes, while avoiding the issues introduced by pathway merging. Our algorithms are publicly available and have wide applicability to analysis of pathway datasets from any organism.

Chapter 6

Mapping biological process relationships and disease perturbations within a pathway network

A shortened version of this chapter was submitted to npj Systems Biology on the 4th September 2017[8]. All generated data and networks are available at https://data.mendeley.com/datasets/3pbwkxjxg9/1

6.1 Abstract

Molecular interaction networks are routinely used to map the organisation of cellular function. Edges represent interactions between genes, proteins or metabolites, however, in living cells, molecular interactions are dynamic, necessitating context-dependent models. Contextual information can be integrated into molecular interaction networks through the inclusion of additional molecular data, however there are concerns about completeness and relevance of this data.

[8] The submitted manuscript had a shortened introduction, omitted Figure 36 and Figure 40, and did not include a detailed discussion of the positioning of gastrointestinal and leukaemia nodes within the network . RS performed all analysis and wrote the manuscript.

We developed an approach for representing the organisation of biological processes using pathways as the nodes of a network. Pathways represent spatial and temporal sets of context-dependent interactions, generating a high level network, which incorporates contextual information without the need for molecular interaction data. Analysis of the pathway network revealed functionally linked communities, comparable to those found in molecular networks, including metabolism, signalling, immunity, and the cell cycle.

To examine the network's applications network we mapped a range of diseases onto the network. Pathways associated with diseases tend to be functionally connected, highlighting the perturbed functions producing disease phenotypes. Next we examined the distribution of different types of cancer. Cancer pathways tended to localise within the signalling, DNA processes and immune modules, with some cancer-associated nodes found in other network regions. Many pathways were common to multiple cancers, while pathways specific to different types of cancer reflected variations in genetic heterogeneity and risk factors.

By limiting our data sources to high quality pathway data and experimentally validated Gene Ontology annotations we generated a high confidence functional network, which avoids the shortcomings faced by conventional molecular models.

6.2 Introduction

Cellular processes are carried out by groups of interacting proteins (Barabási & Oltvai 2004). Understanding how these spatially and temporally organised sets of interactions lead to biological processes is fundamental to our comprehension of the cell. The traditional approach used to study function has been based on molecular interaction networks, which have improved our understanding of disease (Goh et al. 2007; Barabasi et al. 2011; Janjić & Pržulj 2012c), infection (Jiang et al. 2015), drug pharmacodynamics (Suthram et al. 2010) and evolution (Stuart et al. 2003). In this paper we describe data and networks as 'molecular' if they are concerned with interactions between individual biological molecules. This is in contrast to pathway-level representations, which represent pathway gene sets, with interactions between individual molecules omitted. Pathways are considered to collectively participate in biological process, the functions of individual genes or gene products are not considered.

There are various network approaches for studying biological processes using molecular interaction networks. Protein-protein interaction (PPI) data is frequently used to construct networks, in which proteins tend to interact with functionally related partners. This results in the emergence of functionally related sub-networks known as 'functional modules' (Barabasi et al. 2011). Modular organisation of function has been shown to exist across species, and is used to predict gene function (Song & Singh 2009; Wang et al. 2011). Similar networks have been generated using co-expression data

(Stuart et al. 2003), GI data (Costanzo et al. 2010), and by combining resources (Ames et al. 2013).

In PPI networks the edges link each protein to all of its known interacting partners; however this presents an over simplistic model. Protein interactions are dynamic, assembling when needed to perform a function, then disassembling (Srihari & Leong 2012; Przytycka et al. 2010; Tang et al. 2011). This is not captured in static networks, where interactions appear permanent in time. Clusters in static network often represent protein complexes, however some proteins participate in multiple temporally and spatially independent complexes (Li et al. 2012; Srihari & Leong 2012). In static networks the interactions representing these separate complexes will form a single cluster.

To capture the inherently temporal nature of molecular interactions, dynamic models incorporating additional data have been developed. For example, gene expression data have been mapped onto PPI networks to reflect the dynamic nature of protein interactions. Active sub-networks, defined as connected regions of the network that show altered gene expression under particular conditions, can then be identified (Ideker et al. 2002; Guo et al. 2007; Komurov & White 2007). Incorporation of expression data reveals a modular network, in which groups of dynamically co-regulated interactions perform condition-dependent processes (Ideker et al. 2002), while static modules provide the structural core (Komurov & White 2007).

Expression data have also been overlaid with PPI networks to study topics such as metabolic processes (Ideker et al. 2002), the cell cycle (Srihari & Leong 2012; Guo et al. 2007) and disease (Guo et al. 2007). Sub-networks of specific, condition-relevant interactions were constructed by limiting edges to interactions between co-expressed proteins. This made it possible to observe context dependent interactions (Srihari & Leong 2012) and the generated sub-networks showed increased functional coherence compared to static networks (Tang et al. 2011). The addition of intracellular localisation data can further refine these networks (Sprinzak et al. 2003). Additionally, longitudinally sampled data can be represented using a time series of networks (Srihari & Leong 2012). Refining the edges to those present at each time point produces modules that are smaller and more functionally specific (Tang et al. 2011).

Molecular interaction networks have provided great biological insight, however, they are undermined by the reliability of their data sources. We suggest that the utilisation of more reliable data could allow functional models to reach their full potential. The unreliability of PPI data has been well documented and leads to false positive and false negative interactions (Snider et al. 2015; Ji et al. 2014; Sprinzak et al. 2003). High throughput data tends to contain high numbers of false positives, while smaller studies are by definition limited in coverage. Algorithms have been developed to attempt to handle this inaccuracy within the data by including functional data, however this can add circularity in functional studies (Li et al. 2007).

Using expression data to filter networks can be undermined by interactions known to occur between proteins that do not have correlated expression (Snider et al. 2015). Further concern about the use of expression data to predict protein interactions comes from the notoriously weak correlation between gene expression and protein abundance (Ghaemmaghami et al. 2003; Gygi et al. 1999). Recent reviews have suggested that post-transcriptional, translational and degradation regulation is at least as important in determining protein concentration as transcriptional control (Maier et al. 2009a; Vogel & Marcotte 2012). Observed correlations between mRNA and protein levels have shown wide variations from r = ~0.35 to ~0.75 (Vogel & Marcotte 2012; Schwanhausser 2011; Maier et al. 2009b) and even as low as r = 0.01 when the correlation is measured within single cells. This variation may be attributed to various biological properties; correlations tend to be higher when mRNA expression is high and shows high variability (Gygi et al. 1999). In situations where expression is constant, protein levels may vary due to translational and post translational factors (Greenbaum et al. 2003). Difficulties in predicting protein levels using expression data undermines the assumption that co-expressed genes are more likely to interact, reducing the applicability of expression data to the generation of dynamic networks. The use of proteomic analysis can help identify context dependent interactions by identifying protein complexes (Altelaar et al. 2012), however protein complexes only constitute small subsections of functional modules making them insufficient for generating comprehensive functional maps.

In this work, we address these problems by introducing a representation of cellular function, which uses pathways, rather than genes, as the constitutive elements. Pathways are comprised of sets of proteins (and complexes) that interact with each other serially, for example, to form signalling or metabolic pathways. This allows us to group sets of proteins, known to interact under particular conditions, without requiring knowledge of the individual protein-protein interactions. This reduction in network complexity bypasses the issues of false positive and negative PPIs, since molecular interactions are not included in the network.

By using pathways we connect individual protein instances to their interaction partners under particular conditions (Stoney et al. 2015), resolving the problem of context and the multiple functions that a single protein can participate in (Ji et al. 2014). Proteins may be present in multiple pathways, allowing them to be represented independently in as many molecular instances as required. Since the function of each pathway is separately determined, pleiotropic gene function remains separate in the network. This approach also avoids gene expression data and the assumption that gene expression represents protein levels.

Finally, PPI networks are extremely large and complex, therefore reasonable time and computational constraints limit algorithm development (Ji et al. 2014) and reduce the accessibility of network analysis. By simplifying the network to a smaller number of pathways, computational analysis becomes less demanding and more accessible.

We present a human pathway network representing global biological function. By incorporating pathways from multiple data sources we aim to maximise functional coverage while minimising the overlap between pathways. To assess the ability of our network to interpret disease functions, we mapped a broad range of disorders onto the network, before focusing more specifically on cancer. Disease pathway 'modules' or clusters are known to form within molecular networks, showing overlap with functional modules (Liu et al. 2015; Barabasi et al. 2011; Goh et al. 2007). Cancer genes have been found to be especially highly connected within PPI networks (Goh et al. 2007), with different types of cancer forming highly connected overlapping modules (Janjić & Pržulj 2012a). Study of disease modules has helped elucidate mechanisms of complex diseases (Liu et al. 2015; Taylor et al. 2009; Wu et al. 2010). Our network provides a higher-level view of the pathways and functions affected by disease, without the inaccuracies inherent in molecular data. We examine the pathways and functions common to multiple cancer types, as well as the distinguishing pathways responsible for the phenotypic variation between different types of cancer.

6.3 Methods

To generate the data for the network, we selected a non-redundant set of pathways, representing healthy biological processes. We assigned function to the pathway nodes and generated edges based on functional similarity, to generate a model shown to biologically representative in yeast (Stoney et al. 2015). Finally we looked at the biological processes attributed to each area of the network and investigated the distribution of cancer pathways and other diseases. Figure 36 represents an outline for the method.

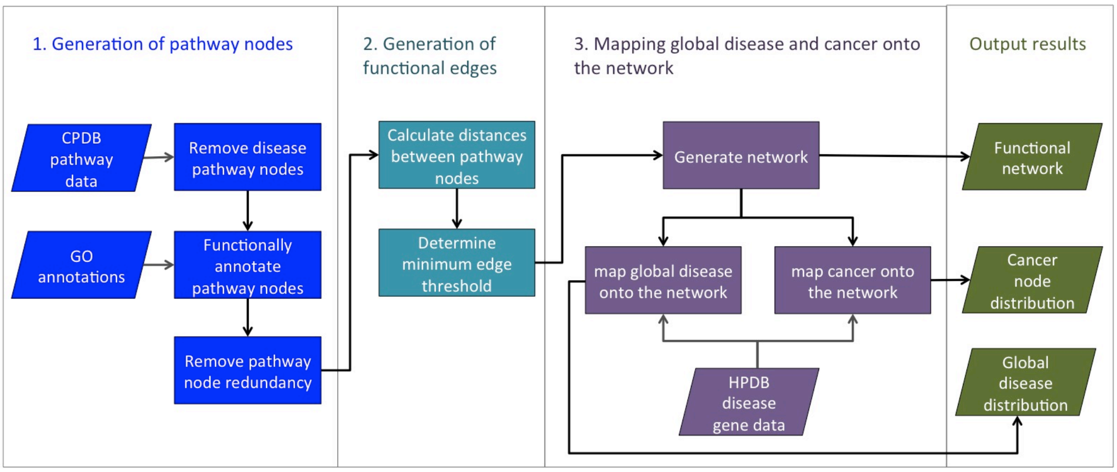

1. Generation of pathway nodes
2. Generation of functional edges
3. Mapping global disease and cancer onto the network
Output results
CPDB pathway data
Remove disease pathway nodes
GO annotations
Functionally annotate pathway nodes
Remove pathway node redundancy
Calculate distances between pathway nodes
Determine minimum edge threshold
Generate network
map global disease onto the network
map cancer onto the network
HPDB disease gene data
Functional network
Cancer node distribution
Global disease distribution

Figure 36: Workflow of the network construction

6.3.1 Generation of pathway nodes

Pathways were downloaded from ConsensusPathDB (CPDB) on 24th Sept 2015 (Kamburov et al. 2009), providing a dataset of 4,011 unique pathways containing 11,196 genes. CPDB collects and compiles data from major pathway databases such as KEGG, Reactome and WikiPathways. Of these pathways, 706 were exact duplicates and were removed. To be included in the network, pathways had to meet the following three requirements, they: represent the cell in a ‘normal’ state (so-called ‘disease pathways’ were removed, see Section 6.3.1.1); had high confidence enriched GO annotations (see Section 6.3.1.2); and belong to a reduced redundancy subset (see Section 6.3.1.4).

6.3.1.1 Removing disease pathway nodes

To generate the functional network, we identified a set of pathways representing ‘normal’ functions. We removed diseases by searching for disease terms within the pathway names (listed in Supplementary Table 3 in Appendix C), as they do not show the cell in a non-diseased state. This was considered necessary since in the later stages of the study, we mapped diseases onto the pathway network, to reveal function affected by particular diseases. The inclusion of disease pathway nodes would distort this distribution, as well as contributing to pathway redundancy.

6.3.1.2 Functional annotation of pathway nodes

To generate the network, we required functional profiles for each pathway node. We assigned high confidence GO terms to each gene, before using enrichment analysis to annotate pathways. Any pathway node that could not be functionally annotated was removed, as we could not calculate their similarity to other pathway nodes to establish network edges.

Functional annotation of pathway genes

The Gene Ontology provides Biological Process annotations for individual genes, along with information specifying how annotations are generated (Ashburner et al. 2000). We assigned high confidence Biological Process GO annotations to genes (downloaded 24th Sept 2016), discarding electronically annotated (IEA) terms as they are of lower confidence than experimentally validated terms. In addition, IEA terms are generated using bioinformatics techniques with similarities to those employed in functional networks, introducing circularity to the method (Pesquita et al. 2009).

We were able to assign high confidence, curated GO annotations to 88% of the genes in normal cellular pathways. We also added all non-IEA parent terms to the GO terms allocated to each gene, since for every GO term associated with a gene, all of the GO term's ancestors apply (Yon Rhee et al. 2008). To meet the minimum criteria for enrichment analysis, each pathway must contain at least four genes with Biological Process GO annotations (Kamburov et al. 2013). Any pathways that contained less than four annotated genes were removed.

Functional enrichment of pathway nodes

Functional enrichment analysis was carried out using the R package clusterProfiler (G. Yu et al. 2012). Enrichment analysis returned large sets of GO terms with p-values below 0.01 for pathway nodes (mean of 412.0 GO terms per pathway), using the Benjamini and Hochberg correction (Benjamini & Hochberg 1995) for multiple testing.

6.3.1.3 Minimisation of pathway functional profiles

We generated minimal sets of enriched high confidence GO terms to represent all of the genes in each pathway node. Later stages generating the network edges required the use of methods capable of measuring similarity between each pathway node's enriched GO sets. These methods are well established (Pesquita et al. 2009; Lord et al. 2003b; Resnik 1999; Wang et al. 2007), but they are not suitable for highly redundant sets of enriched GO terms. We therefore had to first remove similar enriched GO terms within each pathway.

We have previously described a set cover algorithm that reduced redundancy from enrichment analysis data (see Section 5.4.7), which we use here to remove redundancy from each pathway's enriched GO terms. For each pathway, the enrichment set cover algorithm selects the subset of the most enriched GO terms covering all of the genes. In this way, the most specific/enriched GO terms that describe the function of all the genes in each pathway are identified and retained. GO terms describing the same genes with a lower level of significance are discarded, resulting in a reduced functional profile. Note that only the non-IEA GO terms associated with each pathway's genes will be selected for inclusion in the minimal profile. The set cover algorithm used is a heuristic method; therefore it does not provide a unique solution. The solution provided reduces redundancy using a greedy algorithm, therefore multiple solutions are possible (see Section 7.3.1).

6.3.1.4 Reduction of redundancy between pathways

Following the removal of disease and functionally unannotated pathway nodes, all remaining pathway nodes were suitable for use in the network. However, because the data source used was highly inclusive, incorporating pathways from all areas of study, high levels of pathway overlap were present. An extensive effort was made to remove as much data duplication as possible, while preferentially selecting moderately sized pathways. Removal of redundancy was necessary since we aimed to generate a network in which linked nodes represent functional cooperation between distinct pathways. Otherwise we would risk generating groups of overlapping pathways, whose mutual function stemmed from a high proportion of shared genes.

We have previously described methods using set cover theory to reduce redundancy in pathway data sets (see Sections 5.4.3). These combinatorial optimisation algorithms identify subsets of pathways that cover all the genes in the dataset. The pathway set cover algorithms are different to the enrichment set cover algorithm (described above Section 6.3.1.3), as they are based on fundamentally different data types (proportional pathway overlap and enrichment annotations with p-values). As the data set contained pathways with up to 2,154 genes, controlling the pathway size was critical for preserving functional specificity. We therefore selected the proportional set cover algorithm (see Section 5.4.5) as it controls pathway size variability while minimising pathway overlap.

We note that significant improvements in the algorithm's ability to control pathway size variability have been observed when the algorithm was allowed to cover 'most' rather than all of the genes in the dataset (see Section 5.5.2). We found that allowing the set cover method to cover 99.95% rather than 100% of the genes in the dataset reduced the maximum pathway size from 2,154 to 426. Large reductions in pathway redundancy were also observed (see Results Section 6.4.1).

6.3.2 Generation of edges

To generate the edges in the network, we measured the semantic similarity of each pair of pathway nodes based on their associated GO terms in the minimised functional annotation profile (see Section 6.3.1.3). These values, between zero and one, formed the basis of the network edges.

6.3.2.1 Semantic similarities between pathways

To calculate the semantic similarity between pairs of pathways, we first needed to measure the similarity between pairs of GO terms. Since pathways are enriched with multiple GO terms, we established the most suitable method for comparing GO sets. Various measures are available for measuring the distance between GO terms and GO term sets (Resnik 1999; Wang et al. 2007; Lin 1998). We selected our method based on its ability to comply with the assumption that GO terms within pathways should be more closely related than GO terms between different pathways.

6.3.2.1.1 Measuring semantic distances between individual GO terms

Of the various methods available to measure the distance between two GO terms, the Resnik (Resnik 1999) and Wang (Wang et al. 2007) measures have been shown to outperform other methods in previous studies (Pesquita et al. 2009). We therefore implemented these methods using the R package GOSemSim (Yu et al. 2010).

6.3.2.1.2 Measuring the semantic distance between GO sets

To calculate the similarity between pathways, we tested two approaches: the method and the best-match average (Pesquita 2009). The pairwise method measures the similarity between every pair of GO terms between two pathways and then calculates the mean. The best-match average records the similarity between each GO term in the first pathway and the closest GO term in the second pathway. It then performs the symmetric calculation, before generating a mean distance based on both sets of scores. This produced a semantic distance between every pair of pathways generating a complete network. The complete network was impractical for global analysis, but useful for studying subsets of nodes within the network.

6.3.2.2 Pruning edges between pathway nodes

Our network links pathway nodes using weighted edges based on their similarity. Linking all nodes with a semantic distance greater than zero resulted in a highly connected network, limiting its suitability for network analysis and obscuring the network’s main structure. We aimed to reduce the number of edges in the network to show only the most significant functional links between the pathways. We generated a range of 50 thresholds between zero and one and calculated the proportion of nodes and edges retained by each. By subtracting the proportion of nodes retained by each threshold by the proportion of edges retained, we identified the threshold that linked the maximum number of nodes into the network using the fewest edges.

6.3.3 Mapping the distribution of biological function and disease onto the network

6.3.3.1 Mapping global diseases onto the network

In order to apply the pathway network to the study of disease we identified pathway nodes associated with a comprehensive range of diseases. To ensure that a broad range of diseases were covered we used the Human Phenotype Ontology (HPO) disease dataset, downloaded on the 30th of April 2016. This dataset contained 293,556 disease gene annotations for hereditary and non-hereditary disorders. This dataset includes both OMIM diseases such as 'migraine, familial hemiplegic, 1; FHM1' and disease phenotypes, such as 'visual hallucinations'. We used enrichment analysis to generate disease pathways. The genes (within our pathway dataset) associated with each disease or disease phenotype in the HPO data were identified, resulting in a set of 1,061 disease annotations connected to at least four genes (annotations with fewer than four genes were considered too small for enrichment analysis). We then used the Fisher's exact test to identify pathways associated with the disease annotations

6.3.3.2 Finding the shortest paths in disease sub-networks

To test whether diseases tended to cluster within the network, we measured the shortest paths between pathways associated with each disorder using NetworkX (Hagberg et al. 2008). This algorithm calculates the shortest path between two nodes. This measure conventionally uses distance rather than similarity. The shortest path is the combined weight of all the edges, linking the shortest route between two nodes.

We compared these results to sets of shortest paths generated from sets of random nodes. We selected randomised sets of nodes of equal size to the set of disease nodes. We repeated this method 100 times for each disease.

6.3.3.3 Measuring the largest connected component of disease-pathway modules

To identify whether diseases were forming linked sub-networks within the network, we selected subsets of edges connecting nodes related to each disease, excluding edges incident to non-disease nodes. We then measured the proportion of nodes, associated with each disease, that were connected within the largest single component. Connected components were generated using NetworkX (Hagberg et al. 2008). Furthermore, we used the full set of semantic similarities (prior to filtering the edges using the minimum threshold, see Section 6.3.2.1.2) to measure the semantic similarities between ‘disease pathways'. We used randomised disease sets (see Section 6.3.3.2) to assess the validity of these findings.

6.3.3.4 Mapping cancer onto the network

We selected cancers by searching for the terms: *cancer, tumour, tumor, melanoma, carcinoma, leukemia, lymphoma* and *sarcoma* in the set of HPO phenotypes enriched to a p-value of 0.01. We mapped the locations of 166 cancer related pathways onto the network and examined associations with biological processes. To measure the tendency of cancers to cluster within the network, we measured the shortest paths between pathway nodes with the same phenotype (see Section 6.3.3.2).

To provide a more in depth analysis, we focused on the distribution of various types of gastrointestinal cancer and leukaemia. We selected all cancers within these groups that were present in more than two pathways and examined how the network revealed properties of the cancer types.

6.4 Results and Discussion

6.4.1 Global functional organisation can be represented by a non-redundant set of 1,014 pathways

In order to generate a representation of biological processes based on pathways, we first selected a set of non-redundant, functionally annotated pathways. The original dataset contained 3,305 pathways with 11,196 genes (following removal of identical pathways, see Section 6.3). Figure 37shows the proportion of pathways that were removed at each stage of pathway preparation.

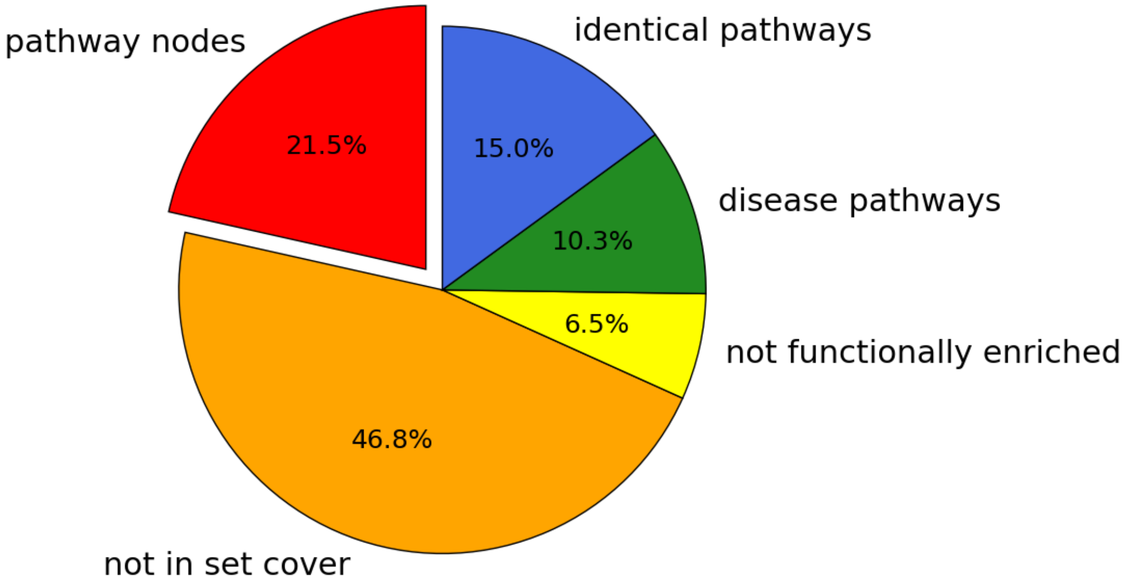


Figure 37: Pathway processing stages. Proportion of pathways that were removed from the initial data set because they had identical gene sets, were disease pathways, could not be functionally annotated, or were redundant (not in the set cover).

Disease pathways such as colorectal cancer, asthma and HIV infection were removed from the data set; as well as drug metabolic pathways such as doxorubicin and statin pathways; and addiction pathways such as cocaine addiction. Pathways involving naturally occurring substances such as vitamins (iron) and medically administered hormones (folic acid) were allowed to remain. We removed 484 disease pathways reducing the number of genes in the data set to 10,833.

The Gene Ontology (GO) (Ashburner et al. 2000) assigned a mean of 8.2 terms to each gene (median 5). Addition of parent terms increased the mean number of GO terms per gene to 75.3 (median 52). It was necessary to remove 1,263 genes, as they did not have experimentally validated GO annotations, resulting in a loss of two pathways. Of the unannotated genes, 4.0% had no biological process annotations and 7.6% only had Biological Process annotations inferred from electronic annotation (IEA), which are considered less reliable. We removed 298 pathways with fewer than four annotated genes, as they were too small for enrichment analysis. Enrichment analysis returned at least one high confidence enriched biological GO term (p-value <0.01) for 2,514 out of the 2,521 remaining pathways. Pathways without enriched GO terms were removed, as functional annotations were required to create edges in the network.

Between 1 and 3,459 enriched GO terms were assigned to each pathway (mean 411.8), using the p-value threshold of 0.01. These enriched GO terms varied greatly in their significance and included many similar terms and parent terms. We aimed to generate a network that linked pathways based on the similarity of their enriched GO terms; however, GO terms assigned with low significance had the potential to make spurious connections or link pathways based on highly general terms. To address these issues we selected the most specific set of GO terms available to represent the genes in the pathway. We used the set cover for enrichment analysis algorithm (see Section 5.4.7) to select the most significant GO terms capable of covering the genes in each pathway, reducing the mean number of GO terms from 411.8 to 4.7. These reduced functional profiles provide a precise representation of the pathways' function without large numbers of similar GO terms or parent terms.

The remaining data set still contained high levels of overlap. In addition to reducing redundancy, it was beneficial to reduce pathway size variability, with pathway sizes ranging from 4 to 2,154; the standard deviation was 89.4, approximately twice as large as the mean (46.0). The inclusion of such large pathways is unbeneficial, since they lack functional specificity and their statistical strength in enrichment analysis is disproportionately high (Khatri et al. 2012). We used the proportional set cover algorithm (see Section 5.4.5) to reduce redundancy while preferentially selecting pathways with sizes close to the median size of 23. We allowed the set cover algorithm to finish after 99.95% of the genes had been covered, reducing the number of pathways required from 2514 to 1014 (representing a 60% reduction). The only difference between this set cover and the set cover produced to cover 100% of genes was the absence of pathways ‘gene expression’ and ‘metabolism’. This reduced the maximum pathway size from 1442 (metabolism) to 426 (‘generic transcription pathways’), while resulting in the loss of only 4 genes.

Figure 38 demonstrates the ability of the set cover algorithm to reduce redundancy, by displaying the presence of genes in multiple pathways. Prior to redundancy reduction, genes appeared in a mean of 46.0 pathways, with many genes appearing in large numbers of pathways. After set cover, genes appeared in a mean of 4.2 pathways. Genuine cases of pleiotropy are preserved in the remaining overlap, as pathways with minor overlap are not removed. The use of this modified set cover algorithm enables us to use the combined data sources collated by CPDB without being undermined by excessive pathway overlap.

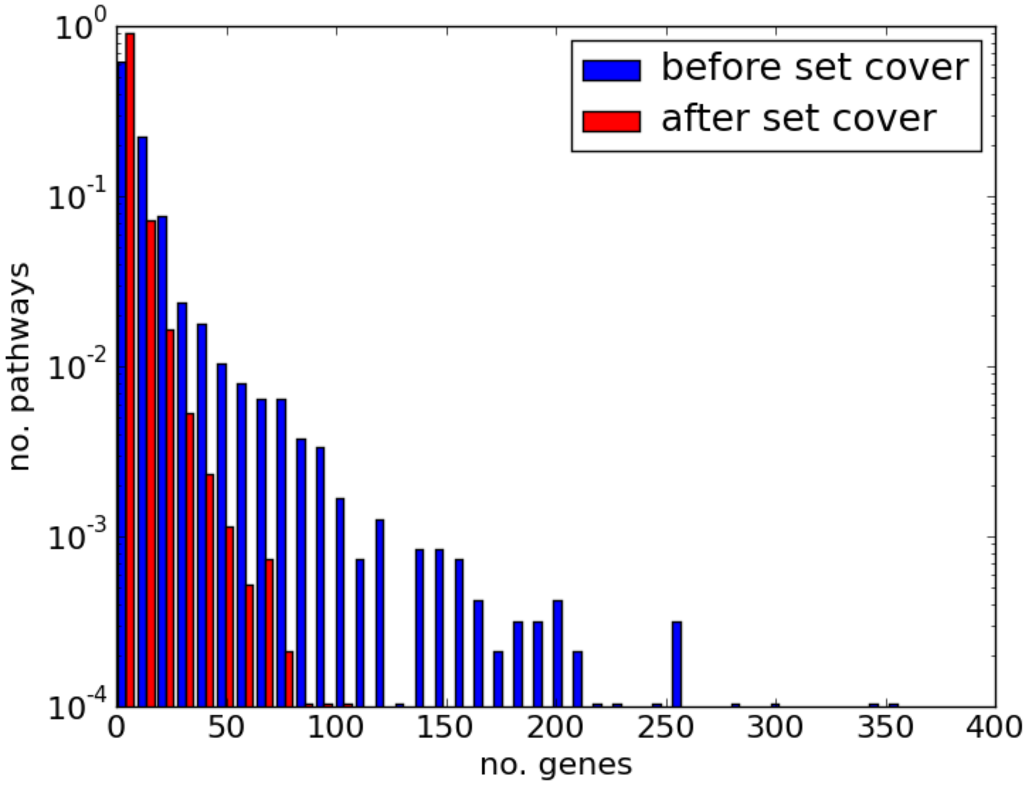


Figure 38: Genes in multiple pathways before and after applying the set cover algorithm.

Histogram showing the proportion of the genes in the data set that appear in multiple pathways (indicating redundancy), before and after set cover.

6.4.2 The Wang best-match average is the most suitable metric to measure the functional similarity of pathways

Pathways were linked to form a network based on the similarity of their shared GO terms. We compared the Wang and Resnik methods for measuring distances between GO term pairs (see Section 6.3.2.1.1). We then compared the pairwise and best-match average methods for measuring distances between sets of GO terms (see Section 6.3.2.1.2). To assess the suitability of each method, we identified the approach that gave the greatest difference between the semantic similarity of GO term pairs within pathways, compared to semantic similarities between different pathways. Figure 39A and B show the semantic distances between all pairs of GO terms within and between pathways, using the Resnik and Wang methods. Semantically similar GO pairs are consistently more frequent within pathways than between them, although the difference is small especially when using the Resnik method.

To generate the pairwise average measure, we calculated the mean similarity between GO terms within each pathway and between each pair of pathways. This increases the distinction between semantic similarities observed between pathway nodes and within pathways. The difference is clearer when distances between GO terms are generated using the Wang (Wang et al. 2007) measure (Figure 39D), rather than the Resnik measure (Figure 39C).

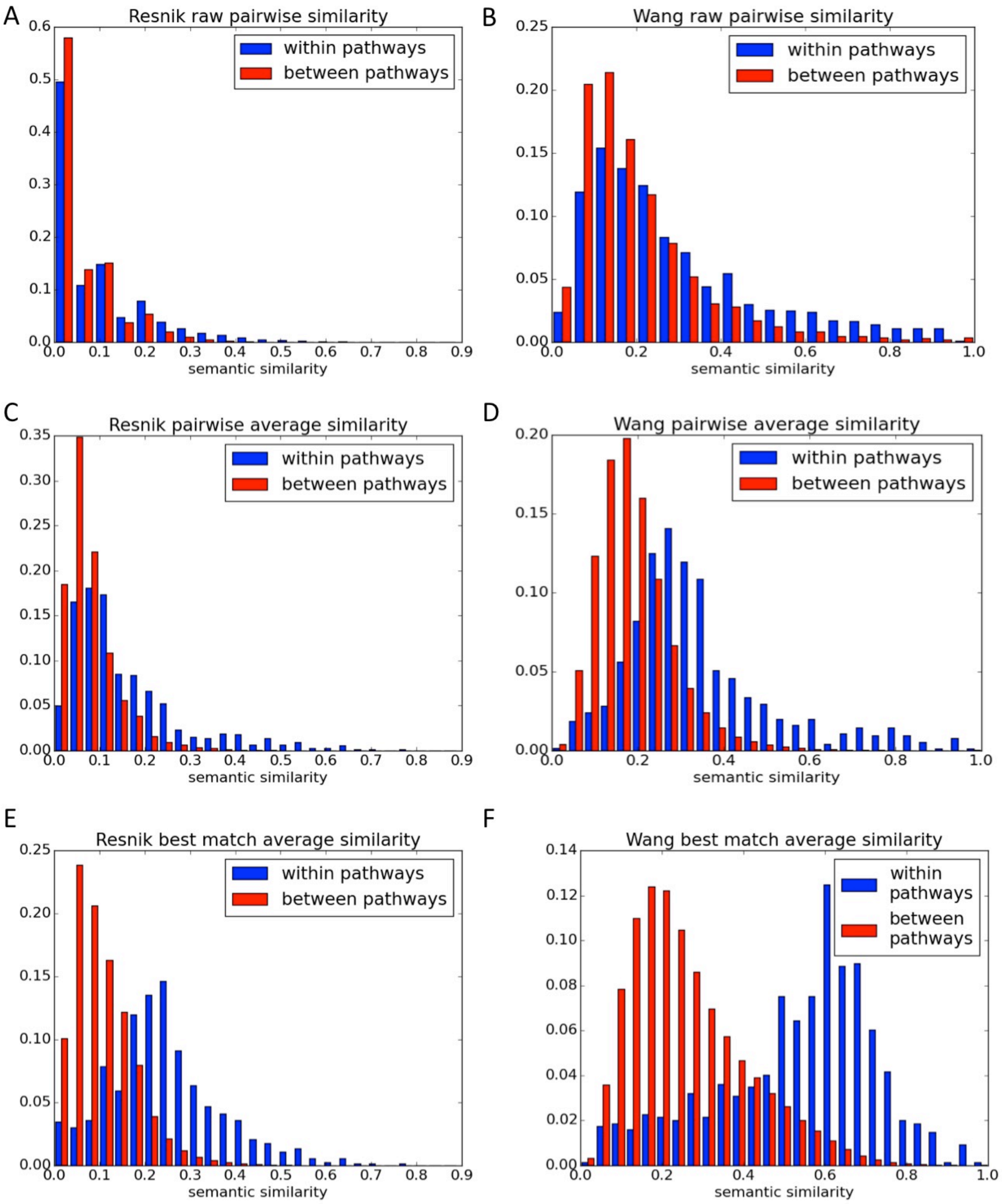


Figure 39: Pathway redundancy across cover sets. Semantic similarities between GO terms in the same pathway (red) and between pathways (blue). The y-axes show the proportion of GO term allocated different semantic distances. A and B are individual semantic similarity measures taken using the Resnik and Wang measures. C and D are pairwise average distances using the Resnik and Wang measures. E and F are best-match average distances using the Resnik and Wang measurements.

Figure 39E and F show the best-match average similarities between and within pathways. This enhances the distinction between semantic similarities within and between pathways, particularly when the Wang method is used to measure distances between GO terms.

The best-match average typically out-performs the pairwise method when unrelated annotations are allocated to the same pathway or gene (Pesquita et al. 2009). This is because rather than comparing each GO term to all available terms within each pathway or pathway pair, the best-match average is generated using the most similar GO term pairs. For example if ‘GO:1’ and ‘GO:2’ have a semantic similarity of 1, and are both allocated to ‘pathway x’ and ‘pathway y’, the pairwise methods will assign an average similarity of 0.5, despite the pathways having identical terms. The best-match average would assign a more intuitive score of 1. The finding that the Wang method outperforms Resnik indicates that pathways are not being assigned a single semantic function but instead are enriched with multiple semantically different GO terms. Clusters of pathways are formed within the network when pathways share at least one function.

The Wang method demonstrably out-performs the Resnik measure, in each recorded instance. To interpret these results, we note that the Resnik measure is based on the lowest common ancestor in the GO ontology capable of covering both GO terms. The score is calculated to describe the specificity of the lowest common ancestor, based on the number of genes associated with the term. A disadvantage of this approach is that it does not consider how far removed each GO term is from the common ancestor (Wang et al. 2007). Therefore two identical generic terms would receive the same score as two highly specific child terms of the generic ancestor, despite their increased difference. The Wang measure considers all ancestral terms shared by two GO terms and reduces the score if the shared ancestors are distantly removed from the terms being compared (Wang et al. 2007). In this way it is better able to distinguish between pairs of general GO terms and pairs of distantly removed GO terms. For these reasons we generated the network using the Wang method in conjunction with the best match average method to generate the network.

6.4.3 Pathways linked by shared functionality form a cohesive network

We linked the pathways into a network based on shared functionality, represented by semantic similarity between GO terms. We used the Wang method to calculate functional semantic similarities between each pair of pathways, in order to generate a set of weighted network edges. Inclusion of all the edges generates a highly dense network reflecting the cross-talk between all biological processes, which impedes analysis and structural visualisation of the network.

To reduce the number of edges while preserving the topological structure of the network, we removed weaker edges. To avoid disconnecting large numbers of pathway nodes from the network, we calculated the minimum edge weight threshold for reducing edges while retaining nodes. Figure 40 shows the proportion of pathway nodes and edges preserved using similarity thresholds between zero and one. Using the best-match average technique the optimum threshold to provide the highest number of nodes with the lowest number of edges, was 0.56, which conserved 987 nodes (97.1 %) and 20,642 edges (4.0%). We used the minimum edge threshold to select a set of edges to construct the network, with a density of 4.2%.

Figure 41shows the network with a sample of GO terms highlighted to illustrate some of the functions represented. Within the network two major functional pathway modules relating to metabolism and signalling can be observed. A DNA metabolic process module links transcription processes, chromatin organisation and mitotic cell cycle to metabolism. Immune responses are tightly clustered besides signalling and cellular responses to stimuli. Axon guidance has nodes in the immunity network region, reflecting its role in the primary immune response (Tordjman et al. 2002).

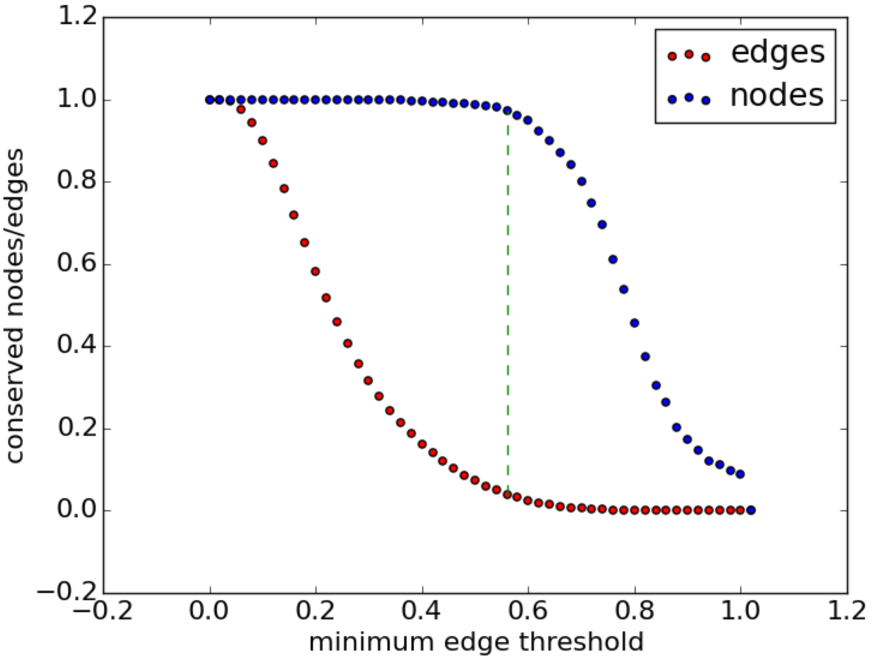


Figure 40: Minimum edge threshold. Linking all nodes with a semantic similarity >0 resulted in a highly dense network. To reduce the number of edges in the network, while minimising the loss of nodes, we experimented with thresholds between zero and one. Blue and red circles show the proportion of nodes and edges retained at each threshold. The optimum threshold is calculated as producing the greatest difference between the proportion of nodes retained and the proportion of edges retained. We used a threshold of 0.56 to select a set of edges to construct the network.

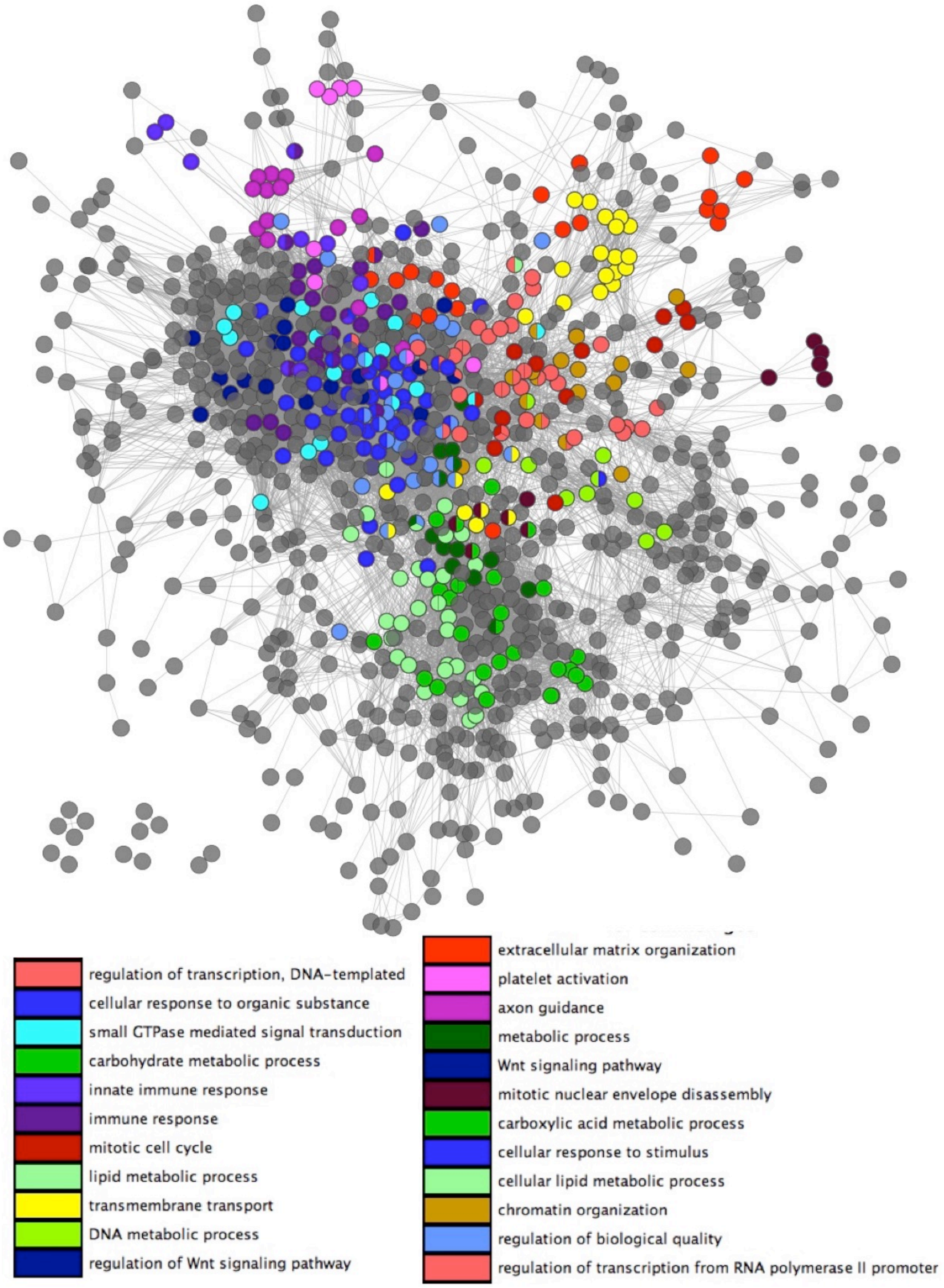

regulation of transcription, DNA-templated
cellular response to organic substance
small GTPase mediated signal transduction
carbohydrate metabolic process
innate immune response
immune response
mitotic cell cycle
lipid metabolic process
transmembrane transport
DNA metabolic process
regulation of Wnt signaling pathway
extracellular matrix organization
platelet activation
axon guidance
metabolic process
Wnt signaling pathway
mitotic nuclear envelope disassembly
carboxylic acid metabolic process
cellular response to stimulus
cellular lipid metabolic process
chromatin organization
regulation of biological quality
regulation of transcription from RNA polymerase II promoter

Figure 41 Major functional clusters in the human pathway network. Weighted network of pathways, linked by shared functionality. Edges were generated using the Wang best match average method to link pathways biased on their functional profiles, using a minimum weight cut off of 0.56.

6.4.4 The functional network enables identification of disease pathway modules.

We used enrichment analysis to assign 404 OMIM diseases to 219 pathways, using a p-value threshold of 0.01[9]. By focusing on diseases (e.g. cystic fibrosis) rather than phenotypes (e.g. chronic lung disease, elevated sweat chloride, hepatomegaly) we capture the range of symptoms induced by disorders.

We examined whether disease pathway nodes could form connected modules, excluding edges incident to non-disease nodes. The proportion of disease pathway nodes linked into a single connected component was higher than expected at random (Figure 42). 18% of disease pathway node sets could be linked into a single disease pathway module and 29% of disease pathway node sets have at least two connected nodes. In comparison only 2% of randomised pathway node sets formed a single connected module and 6% had at least two connected nodes.

The majority of disease pathways did not, however, form a single disease pathway module. The most likely explanation is that the subset of functional links selected to generate the network edges do not closely reflect the functional relationship between the disease pathway nodes.

[9] The Benjamini Hochburg correction was tested, using a threshold of $p<0.05$. Unfortunately this reduced the mean number of pathways per disease to 2.5 (median 2, range 1-7). Disease modules are less meaningful with such small sets of pathway nodes, therefore multiple testing was not applied (see Section 7.3.4)

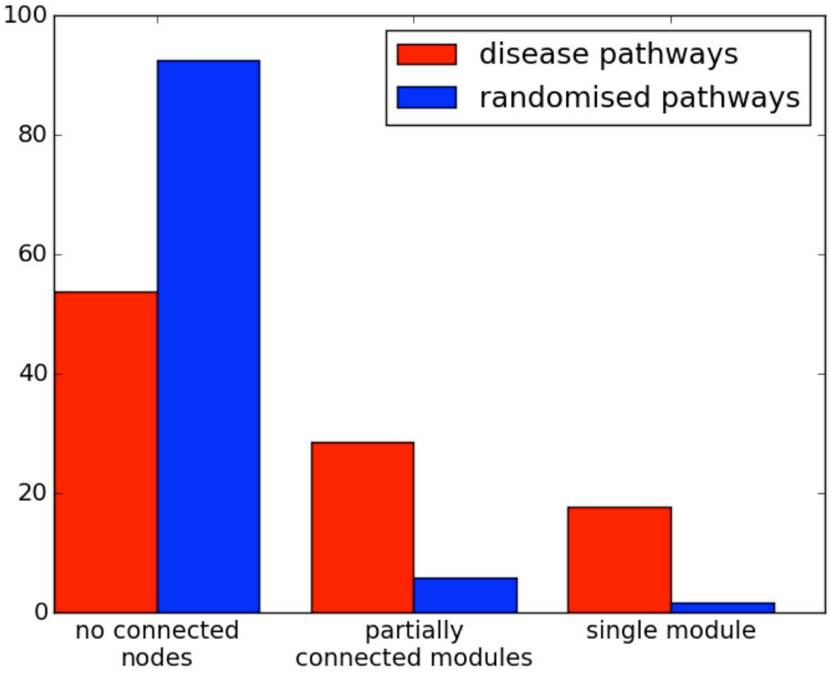


Figure 42: Proportion of disease nodes forming connected components

To test the hypothesis that disease nodes have close proximity within the network, generating disease pathway modules (despite not being fully connected), we measured the shortest paths between disease nodes. Figure 43A shows the distances between nodes with shared diseases, compared to an equal number of random pathways. Shortest paths between randomized disease nodes formed a roughly normal distribution, whereas shortest paths between disease nodes tended to be shorter, indicating that disease nodes are close within the network. To confirm the significance of the distributions we performed a one sample Kolmogorov-Smirnov test, which returned a p-value of << 0.01.

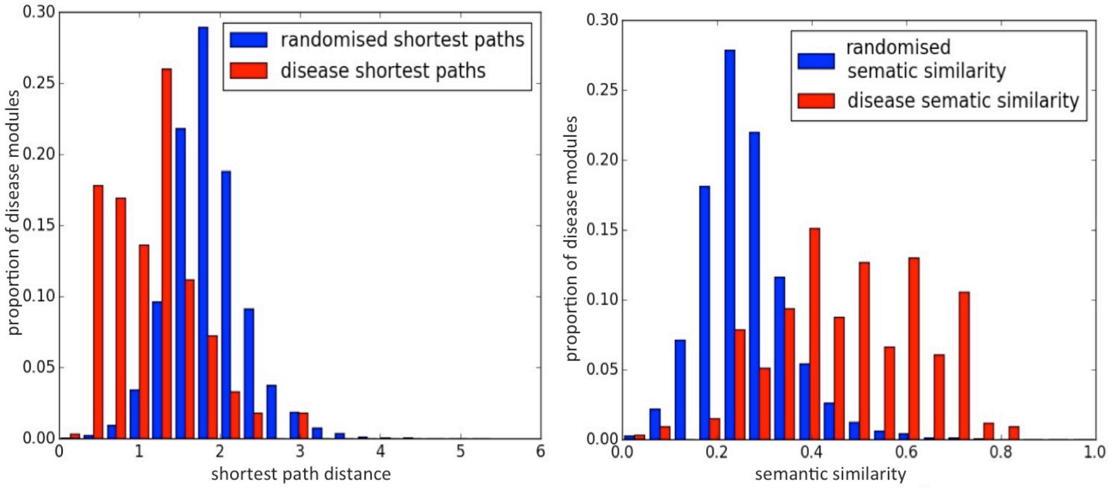


Figure 43: Disease module connectivity. A) Shortest paths between nodes enriched for the same disease and randomised disease nodes. B) Semantic distances between nodes enriched for the same disease and randomised disease nodes

Biological processes are known to be hierarchical with general functions covering multiple specific functions (Barabási & Oltvai 2004; Albert 2005). In addition cross-talk between functions co-ordinates the actions of the cell. To generate the network we selected a subset of high scoring edges to provide the main structure of biological process organisation, while controlling the density of the network. Selecting a higher number of edges would have allowed us to capture more detail regarding specific functions, as well as interactions between functions; however it would complicate the network, obscuring the major functional groupings.

To explore the idea that disease nodes may be linked by functions that are not represented as edges in the network, we looked at the distances between nodes in the full semantic similarity set (before filtering using the minimum edge threshold, see Section 6.3.2.1.2). This data set contains the direct semantic similarity measures between all pairs of nodes. We found that semantic similarity between the disease pathway nodes is greater than semantic similarity between randomised nodes, demonstrating that diseases cluster within this dataset (Figure 43B). We also confirmed that the set of shortest paths generated between disease nodes using the full set of semantic distances produced more significant results (p=3.1e-134) than the set of shortest paths generated using only network edges (Figure 43A, p=1.1e-22). This indicates that the full set of semantic similarities best captures disease pathway modules. These links, which may not be represented by edges, should be considered when observing distribution of disease nodes on the network.

6.4.5 Comparison between cancer pathway modules

We identified 166 pathways enriched with cancer genes at a p-value of <0.01. These were comprised of 39 types of cancer affecting a range of cell types. Many pathways were enriched for multiple cancer phenotypes (mean 3.3). The pathway associated with the most cancer types (17) was ‘extracellular vesicle mediated signalling in recipient cells’, which contains cancer causing genes including *WNT, EGFR, RAF, NRAS* and *KRAS*, and is upstream of pivotal cancer pathways (Vader et al. 2014). Other pathways associated with high numbers of cancers were the ‘*RAC1 PAC1 P38 MMP* 2 pathway’ containing *MAPK, ERK, KRAS, RAC, RAS* genes and ‘copper homeostatis’ which has been found to be relevant to multiple tumour types and is being trialled as a chemotherapy target (Denoyer et al. 2015).

To assess the claim that cancers cluster within particular network regions, we measured the shortest paths between cancer nodes within the network (Figure 44A).

The Kolmogorov–Smirnov test was applied to confirm the significance of the observed cancer clusters (p-value <<0.01). We also compared the semantic similarity of nodes within the cancer module Figure 44B and found that semantic similarity best captures the relationships within disease modules.

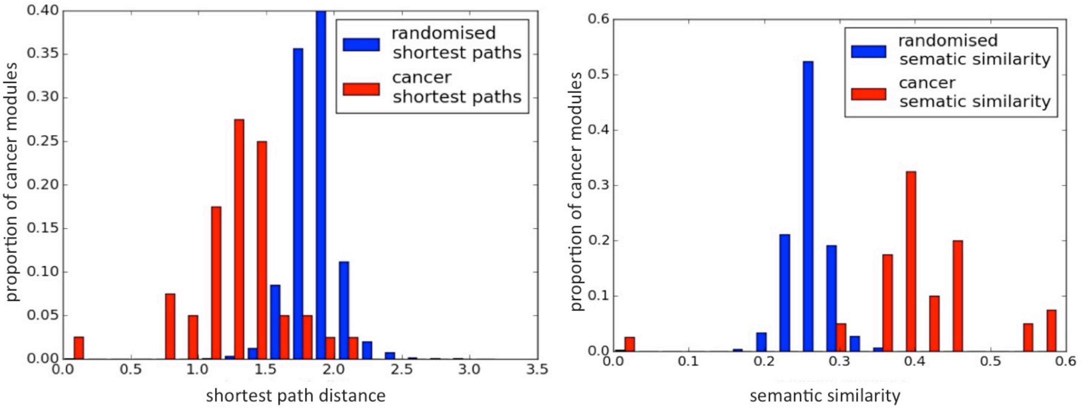


Figure 44: Cancer module connectivity A) Shortest paths between cancer nodes and randomised nodes. B) Semantic similarities between cancer nodes and randomised nodes.

We examined the distribution of cancer within the network. Figure 45 shows the topological position of a sample of cancers affecting high numbers of pathways in the dataset. Cancer pathways can be seen clustering primarily within the signalling, immune response and DNA process network regions.

The signalling and immune network region is the most densely populated with cancer nodes, including sarcoma pathways, juvenile leukaemia, and neurofibrosarcoma. Cancer nodes also cluster in the region concerned with DNA metabolism, response to stimulus and transcriptional control. Several breast cancer and nephroblastoma pathways are prevalent in this region.

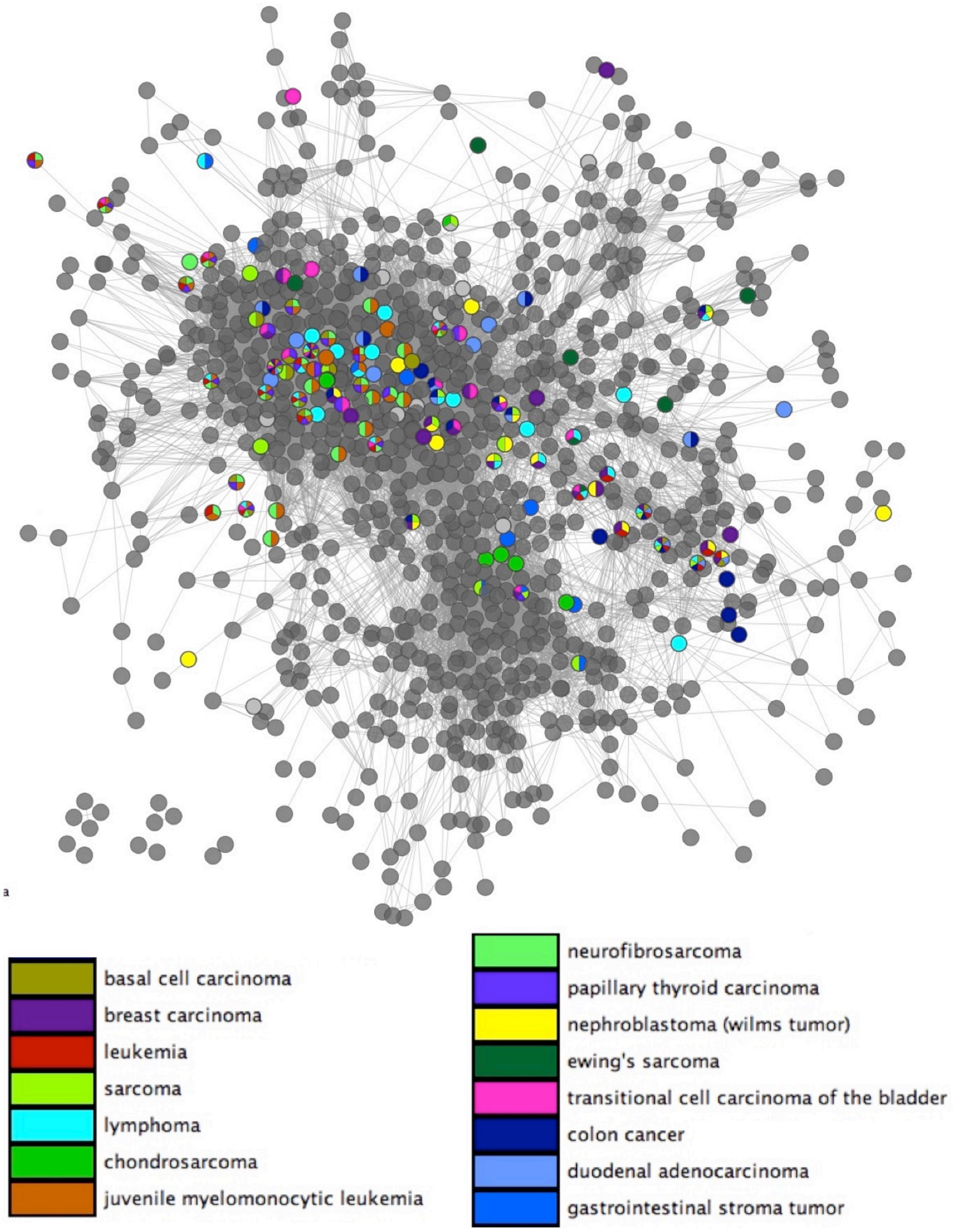

a
basal cell carcinoma
breast carcinoma
leukemia
sarcoma
lymphoma
chondrosarcoma
juvenile myelomonocytic leukemia
neurofibrosarcoma
papillary thyroid carcinoma
nephroblastoma (wilms tumor)
ewing's sarcoma
transitional cell carcinoma of the bladder
colon cancer
duodenal adenocarcinoma
gastrointestinal stroma tumor

Figure 45: Distribution of cancer pathways. Functional pathway network showing the distribution of pathways associated with common cancer types (in the data set).

To demonstrate the value of mapping individual diseases within the network, we examined the distribution of leukaemia and gastrointestinal tumours.

Figure 46A shows disease nodes associated with leukaemia widely distributed across the cell cycle, DNA metabolism and signalling regions of the network, while juvenile myelomonocytic leukaemia nodes are restricted to the signalling area of the network. Adult chronic leukaemia is a highly heterogeneous disease with wide variations in disease aggressiveness and most gene mutations occurring in less than 5% of patients (Ghamlouch et al. 2017). Disease mutations are most commonly linked to cell cycle, DNA repair, immune and RNA pathways. Acute leukaemia is also highly genetically diverse, with the most frequent mutations occurring in the *NPM1* gene, a phosphoprotein involved in a range of functions and *NOTCH1* gene, which regulates development (Rowe 2016). This is reflected by the highly distributed arrangement of most types of leukaemia within the network.

In contrast, juvenile myelomonocytic leukaemia is largely characterised by *RAS* signalling pathway mutations, found in 85% of patients (Sethi et al. 2013). The *RAS* pathway is an upstream signalling pathway controlling cell proliferation, survival and phenotypic transformation (Downward 2003). The distribution of nodes within the network neatly illustrates these differences in disease subtype heterogeneity, while highlighting their overlap.

Additionally, we examined the distribution of gastrointestinal cancers within the network (
Figure 46B). Gastrointestinal stroma tumour nodes appear around the metabolic regions of the network, within electron transport chain and tricarboxylic acid pathways. This corresponds to alterations in the *SDH* gene leading to electron

transport chain complex II dysfunction, which is a risk factor for gastrointestinal stromal cancer (Janeway et al. 2011). Tyrosine metabolism nodes are also detected in the network, reflecting the frequent *KIT* mutations in gastrointestinal stromal cancer and the subsequent use of tyrosine kinase inhibitors as an established treatment (Din & Woll 2008).

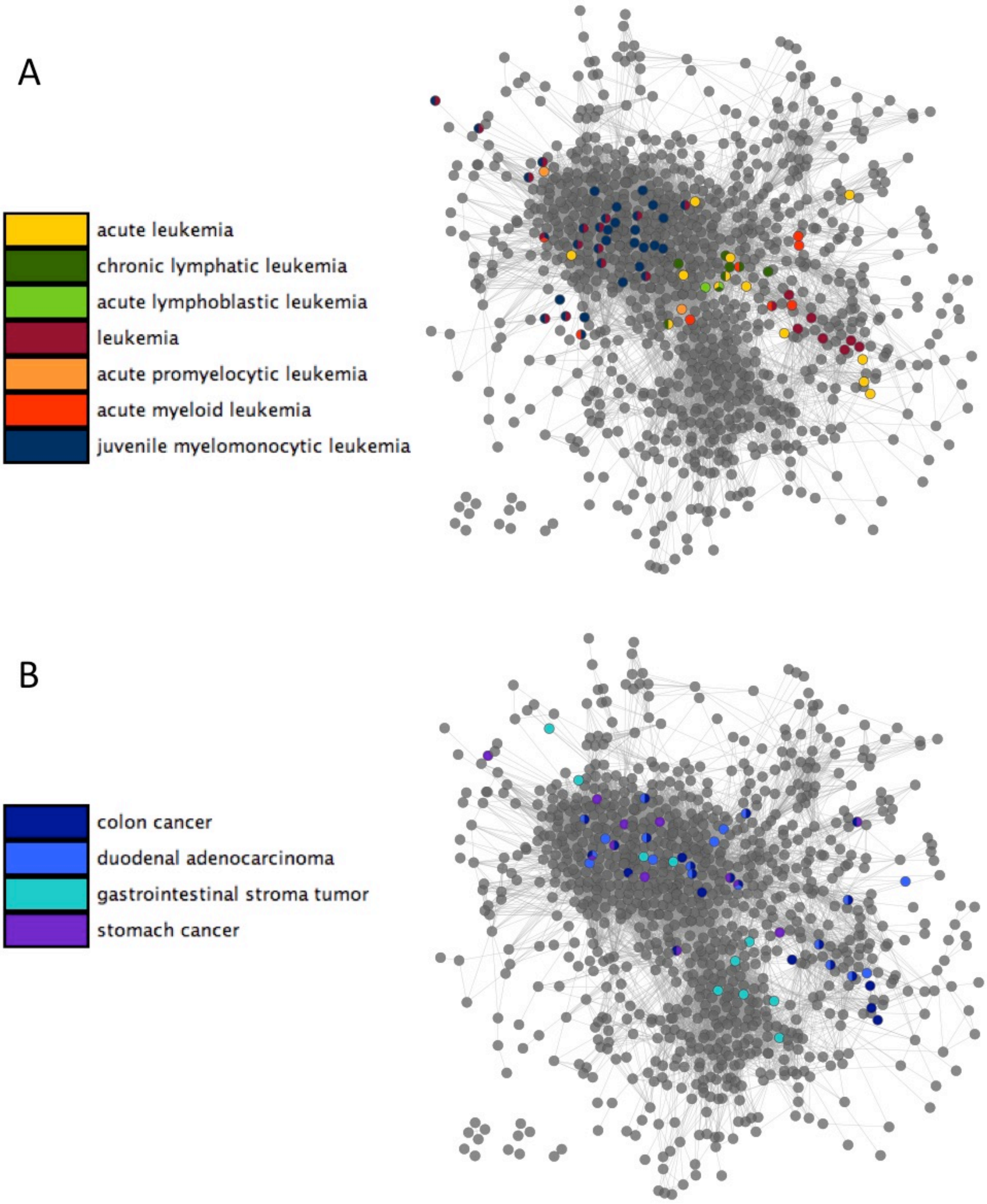

A
acute leukemia
chronic lymphatic leukemia
acute lymphoblastic leukemia
leukemia
acute promyelocytic leukemia
acute myeloid leukemia
juvenile myelomonocytic leukemia
B
colon cancer
duodenal adenocarcinoma
gastrointestinal stroma tumor
stomach cancer

Figure 46: Leukaemia and gastrointestinal cancers. The colon cancer and acute leukaemia pathways cluster within the ribosomal pathway nodes, by DNA metabolism and genetic processes. Ribosomal gene mutations are relatively frequent in these cancer types and impaired ribosome biogenesis is a risk factor for initiation of these cancers (Goudarzi & Lindström 2016).

Some pathway overlap between gastrointestinal stroma tumour, colon cancer and duodenal adenocarcinoma is observed, with shared pathways corresponding to common cancer processes and risk factors. The common risk factors of duodenal adenocarcinoma and colon cancer are gastrointestinal polyps and chromic inflammatory bowel disease (Amersi et al. 2005; Raghav & Overman 2013). Correspondingly, within the network both cancers are found to be enriched in *BMP* signalling pathways, which have been shown to affect gastric inflammation (Takabayashi et al. 2014). DNA repair, cell cycle, extracellular vascular mediated signalling and RAF activation pathways were frequently shared by gastrointestinal and leukaemia cancer types.

The distribution of cancer pathways is dependent on established disease gene annotations; therefore the distribution of disease pathways within the network is unlikely to uncover novel functions affected by cancers. It does however present an effective way of organising diseases such as cancers, based on affected functions, rather than anatomical sites, pathogens or mutations.

6.5 Discussion and conclusion

The use of molecular networks to study biological processes has been highly insightful; however, limitations within molecular data and issues representing multi-functional genes make the development of alternative methods highly desirable. We have constructed a functional network from existing pathway data and biological process annotations, which avoids the issues faced by traditional molecular networks. The pathway network portrays a high-level representation of the organisation of biological processes, composed of functional pathway modules. Clustering methods used in molecular networks identify specific relationships in which each node shows a high density of interactions with all of the other nodes in the cluster. These methods are less suitable for identifying linear relationships, captured by data types such as pathways. Other studies have also approached the issue that network structures, other than clusters, may represent functional modules (Pinkert et al. 2010). Pathways are sets of interactions, which were manually curated to adopt the most appropriate shape for the data, therefore they represent coherent functions independently of the molecular topology.

Mapping diseases onto molecular interaction networks has contributed towards the elucidation of disease mechanisms (Janjić & Pržulj 2012c), identification of new disease-associated genes (Barabasi et al. 2011) and indication of potential drug targets (Yu et al. 2007). However, gene mutations can be phenotypically diverse, such as *AKT1,* which is associated with schizophrenia, colorectal cancer, ovarian cancer and breast cancer (Chavali et al. 2010). Further evidence suggesting that diseases may act independently within different pathways comes from the finding that many disease pairs with shared genes do not show significant co-morbidity (Park et al. 2009). Phenotypically diverse genes may also interact with different partners in different tissues, for example *AKT1* participates in a range of interactions dependent on tissue type(Chavali et al. 2010), further supporting the hypothesis that the results arise from the gene acting in different pathways. This shows that pleiotropy allows genes to be involved in multiple disorders in different contexts, demonstrating that pathways are better suited than molecular networks to map functional perturbations occurring in diseases.

By mapping cancer pathways we were able to visualise the functional regions known to be fundamental to cancer, highlighting the similarities and differences between cancer types, and correlating the network distribution of cancer nodes with their genomic and phenotypic heterogeneity. This method can be generalised to facilitate understanding of any group of disorders or phenotypes.

Examining the similarities and differences between diseases is necessary to assess the shared applicability of knowledge and drugs. Our map makes these relationships immediately obvious. This method can be generalised to facilitate understanding of any group of disorders or phenotypes.

Chapter 7

Discussion

Within this chapter the main findings regarding the representation of function within pathways and the validation of the network topology are presented. Next we discuss the implications of this work for existing approaches in systems biology, emphasising the advantages gained from the pathway networks' independence from molecular data and their ability to capture functional modules that may not be identifiable using molecular networks. Next we discuss the application of pathway networks to disease studies, examining the ability of pathway networks to elucidate disease mechanisms. Finally, we discuss the limitations of the thesis and provide suggestions for future work, including further network analysis, improved edge construction and enhanced disease pathway detection.

7.1 Summary of main findings

The main finding of this project is that functionally annotated pathways can be used as entities within a biologically informative, validated network. We first discuss the methods used to validate the structure of the network and demonstrate its utility. We then examine the portrayal of function within pathways and the ability of the methods developed to incorporate pathway multi-functionality and gene pleiotropy. Next we discuss how, despite the modular nature of biological function, the arrangement of functions within pathways generates a large interlinked component. Finally, we compare the methodological differences between the yeast and human networks and how they affect the final topology.

7.1.1 The organisation of functional pathway maps is biologically informative

The main output of this project was two biologically informative networks (yeast and human pathways) with network regions corresponding to major cellular functions such as metabolism and signalling processes (see Figure 26 and Figure 41). External data sources were used to validate the topology of each network. Genetic interactions indicate functional relatedness between pairs of genes and therefore occur frequently within pathways (see Section 4.4.7). The increase in GIs found within the yeast network clusters was only slightly less than the increase in interactions observed within pathways, indicating extremely high functional cohesiveness within clusters (Table 3). Disease data was used to validate the human network (see Section 6.4.4), since mutations in functionally related genes are likely to generate similar phenotypes (Goh et al. 2007). Pathways enriched for disease genes were shown to be positioned within close proximity within the network Figure 43A. Disease pathways were also more likely to be directly adjacent within the network Figure 42. Highlighting the pathways affected by diseases, as well as the biological functions connecting them, could be applied to studying the mechanistic processes through which disease phenotypes arise. By focusing on gastrointestinal cancer and leukaemia we were able to explore the relationship between the positioning of disease pathways and disease pathogenesis, which was supported in the literature (see Section 6.4.5). We found that the distribution of the diseases studied seemed to be linked to their genetic heterogeneity; for example, the genetically heterogeneous adult leukaemia's tended to be widely distributed across the network, while juvenile leukaemia which is less genetically diverse is more localised within the network (Figure 45). We also found an example of pathways associated with multiple disorders indicating common disease mechanisms (duodenal adenocarcinoma and colon cancer share the BMP signalling pathway, implicated in gastric inflammation which is a risk factor for both conditions).

7.1.2 Minimally-redundant pathway nodes can portray complex arrangements of biological function

Chapter 5 introduced a method to generate a minimally redundant set of pathways covering global biological function. This set of pathways provided extensive coverage of the human dataset, allowing the complexity of biological function to be represented within a framework of context dependent cellular activity. The methods presented avoid pathway merging (Vivar et al. 2013; Doderer et al. 2012; Belinky et al. 2015), which is advantageous since the merged pathways may not be biologically representative. Merging also increases pathway size, reducing functional specificity, while the proposed set cover method controls pathway size (see Section 3.5.1).

The generated pathway set represents a novel resource for the representation of cellular function. Within yeast, 1,443 genes were covered representing 68% of the genes in the dataset. Development of set cover methods allowed the genes in annotated human pathways to be represented, with minimal redundancy. The refined dataset represented 10,833 genes within a set of 1,014 pathways, reducing the number of pathways in the original dataset by 60%. This demonstrates notable progress in the development of methods capable of preserving the scope and inclusiveness of pathway data.

The overlap between the Gene Ontology functions and pathway boundaries was explored (see Sections 4.4.3 - 4.4.5). Genes within pathways have been shown to have more related with the biological processes ontology than expected by chance (Guo et al. 2006). This shows that pathways tend to group genes based on their biological process. However, despite the high functional homogeneity found within pathways, GO terms did not conform neatly to single pathways. Pathways were enriched with multiple GO terms and GO terms were frequently allocated to multiple pathways, suggesting that the Gene Ontology and pathways datasets represent different dimensions of biology. This exemplifies the underlying principle of systems biology that single genes and pathways should not be studied in isolation, since physical units such as pathways collaborate with each other to implement function and encompass multiple biological processes to generate broad reaching effects on the cell.

The representation of pleiotropic gene function is also an important requirement for the pathway network. The use of pathways to represent context-dependent sets of interactions allows the multiple functions associated with some genes to be independently portrayed within in separate cellular contexts. In Section 4.4.8 the tendancy of genes to be assigned different GO annotations through enrichment analysis within different pathways, represents a demonstration of context-dependent function.

7.1.3 Functional pathway networks form cohesive maps of global cellular processes

Following the allocation of GO terms to pathways, edges were generated using semantic similarities to construct a cohesive functional network (see Sections 4.3.8 and 4.4.8). Since previous literature showed that function is highly modular, it was not clear in advance that the construction of a cohesive network would be possible (Han et al. 2004; Wagner et al. 2007; Mitra et al. 2013; Ryan et al. 2012).

The human network generated had a clustering coefficient of 0.59. This shows that although clusters are present the network remains connected, at a density of only 4.2%. Many methods to measure semantic similarity were available to generate the network edges (see Section 3.4). Within the human network we showed that the Wang best-match average most appropriately linked functionally similar pathways, accounting for pathway multi-functionality (see Section 6.4.2). The majority (79%) of yeast pathways were connected to at least one other pathway, with 63% being positioned within the largest connected component. These numbers increased in the human network, with 97% of pathways being linked within the network and 96% appearing in the largest connected component. This demonstrates that the majority of pathways participate in functions that are identical or semantically similar to a function performed within one or more other pathways. Multifunctional pathways form physical links between diverse cellular processes, generating a cohesive network. These pathways represent functional cross talk within the cell, which is visualised within the network. The interconnected nature of the Gene Ontology, which supplies the basis for all edges in the network, also helped ensure the network was well connected. If the pathway network representation of function had been more modular then multiple smaller sub-networks would have been formed, generating a disconnected map.

7.1.4 Development of methods between the yeast and human networks

Comparison of the yeast and human networks reveals some similarities regarding the major clusters including metabolisms, gene expression, immunity, and signalling. There are however, differences. For example, the yeast network has a sparser, more modular structure. This reflects the original aim of the yeast network, which was to show the major functional clusters through a modular network. Edges were primarily generated using the Jaccard coefficient to match GO terms (see Section 4.3.8), which does not account for similarity between different GO terms. In a second processing stage, only the minority of semantically similar GO terms were linked (0.1%) (see Section 4.3.9). The semantic similarity data used was generated based on normalised information content (Hakes et al. 2007). In contrast, within the human dataset semantic similarity methods were used to generate edges. This creates a more inclusive depiction of functional representation, which is not based on the assumption that a modular network best represents function. Instead, the human network was generated to create greater semantic similarity within the GO terms assigned to a pathway than the GO terms between pathways. The network is therefore denser and less modular as cross talk between functions is better represented.

The addition of multiple, less significant functions into pathway profiles (see Section 4.3.7) was also omitted from the human network. This is because the additional GO terms added an extra layer of complexity, but could potentially introduce spurious annotations. Although pathways may have multiple diverse functions, we decided to omit this stage from the processing of human papers following discussion of the yeast paper (Stoney et al. 2015) . Removing this stage make the process simpler, cleaner and less prone to error, although some biological complexity is lost.

7.2 Implications of this thesis for biological networks

This Section discusses the wider implications of the work covered within this thesis, starting with the assertion that pathways are a valuable, underused resource for studying function and that functionally enriched pathways may be better suited than molecular interaction networks to mapping the intracellular organisation of function. We attribute the high performance of the Wang best match average when generating edges to the multifunctional nature of pathways. Finally, we discuss the implications and possibilities generated by the pathway network for studying disease. The clusters of cellular activity captured by pathways within the network provide new opportunities to examine disease mechanisms and identify candidate disease genes. The hypothesis that pleiotropic genes cause multiple diseases by acting within multiple pathways is compared to the possibility that their diverse phenotypes arise from functional crosstalk within the cell.

7.2.1 Pathway data is an invaluable but neglected resource for modelling context dependent function

Within Section 4.4.8 we have demonstrated that some genes have different functions within the context of different pathways. This is an issue that molecular functional networks struggle to address. Dynamic networks accept that different interactions are active at different times and therefore use active modules (Guo et al. 2007; Komurov & White 2007; Ideker et al. 2002) or multiple instances of a single network (Tang et al. 2011) to express this dynamism. Using pathways as the primary units of biological activity and allowing genes to be included independently in multiple pathways results in a more concise, intuitive model. Results in yeast show that this is not a minor issue, with 44% of genes appearing in multiple pathways, and 83% of these genes having distinct functional profiles for each pathway they participate in (Section 4.4.8).

Molecular models must resort to using expression data (Ideker et al. 2002; Tang et al. 2011) or other molecular data (Costanzo et al. 2010; Ames et al. 2013) to capture these important contextual distinctions, while the knowledge recorded in pathways remains unused. Separate sets of expression data are required for each individual cellular context, such as multiple data points for each stage of the cell cycle (Tang et al. 2011). This makes generation of dynamic models expensive and time consuming. The use of pre-existing expression data in molecular models is impeded by the models' intrinsic detail, since if genes within the network are not included in the microarray, then the status of nodes is unknown, affecting edges and the local network topology. By utilising pathway data instead, the method developed in this thesis takes advantage of a wealth of information that is not used in the molecular models to generate more representative functional models.

7.2.2 Functional modules may not conform to molecular network clusters or pathways

Evidence suggests that of the three branches of the GO ontology, the cellular component ontology conforms more closely to molecular interaction networks than the Biological Process and Molecular Function ontologies (Dutkowski et al. 2013). They subjected an interaction map incorporating multiple types of interaction data to clustering analysis to reveal a hierarchical topology structure, referred to as NeXO. The NeXO clusters were then mapped to the three GO ontologies, with greatest similarity observed with the Cellular Component ontology. NeXO captures 58% of terms in the Cellular Component ontology, compared to only 25% of terms in the Biological Process and Molecular Function ontologies. While this supports the hypothesis that modules corresponding to biological processes are present, NeXO indicates cellular compartments have the highest influence on the topology of interaction networks (Figure 47). If traditional clustering approaches have a tendency to identify cellular compartments rather than biological processes, this may obscure attempts to extract functional insight from molecular models.

In contrast, the Biological Process ontology has been shown to be the most representative ontology for pathway data (Jain & Bader 2010; Guo et al. 2006). In these examples, the Biological Process ontology was used to predict PPI interactions.

Proteins from different cellular components interact within pathways, reducing the influence of the Cellular Component ontology. Pathways were not found to be significantly enriched for cellular component or molecular function GO terms, demonstrating the low influence that these factors have on pathway data (Guo et al. 2006). This supports the claim that pathways may be better representations of function than modules from molecular networks.

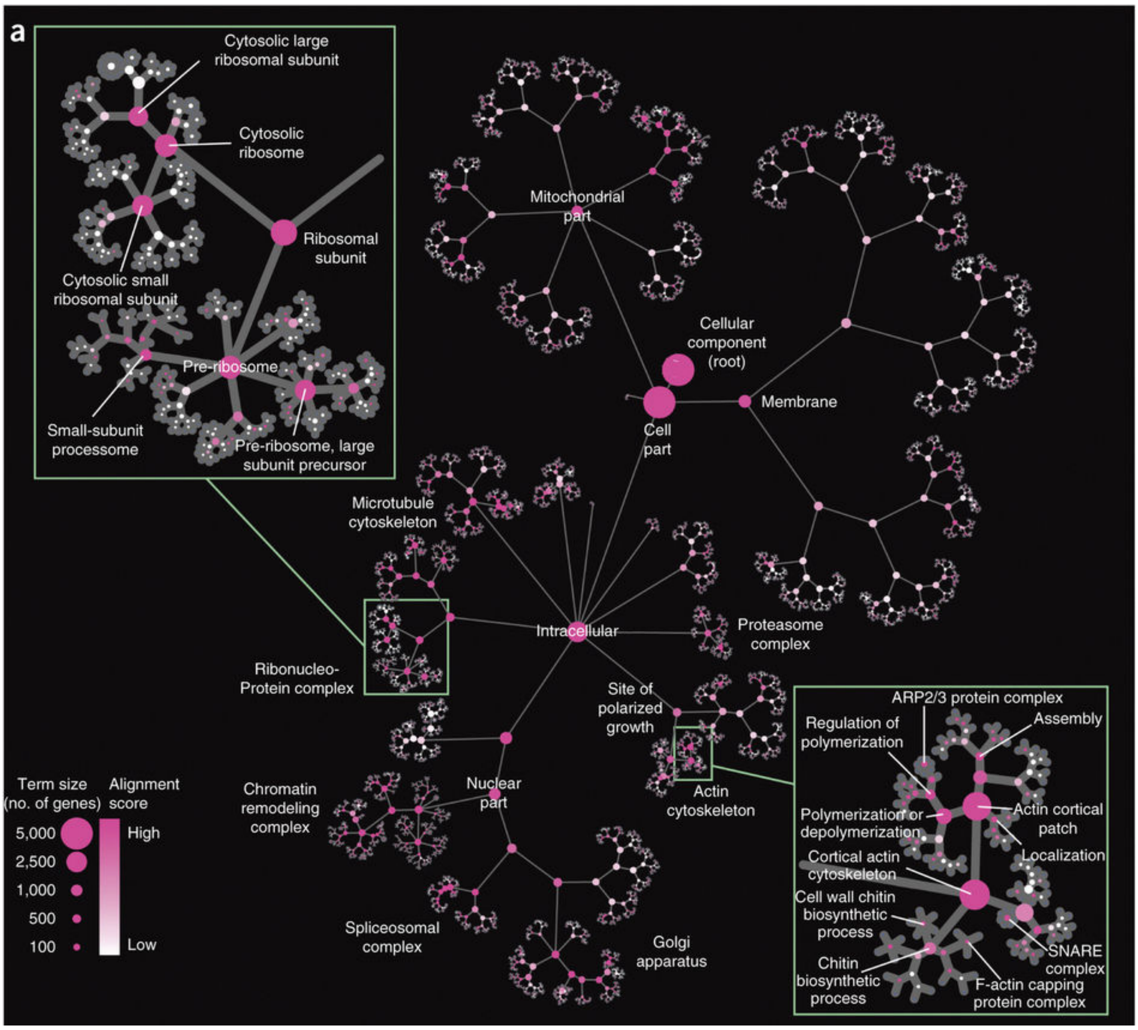


Figure 47: NeXo ontology. Nodes indicate terms and the node size indicates the number of genes allocated to a term. Node colours indicate agreement between the GO terms and the network cluster. Edges indicate hierarchical relationships between terms (Dutkowski et al. 2013).

The clustering methods used in molecular interaction networks identify specific relationships, in which each node shows a high density of interactions with all of the other nodes in the cluster (see Section 2.2.1). These methods are less suitable for identifying linear relationships, which are captured by pathways. It makes intuitive sense for cellular component clusters to form within the network. However, if the implementation of a function is generated through a linear set of interactions, then network clustering may not be the best approach (Figure 48). Pathways are sets of interactions, which were manually curated to adopt the most appropriate shape for the data, allowing them to identify linear implementations of function within the cell.

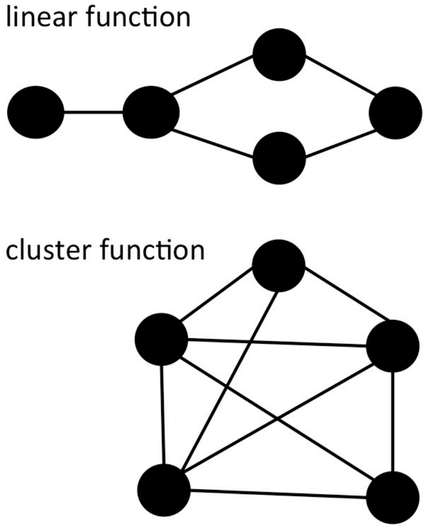


Figure 48: Network functional topologies

These issues are addressed by the simulated annealing algorithms, used to detect active sub-networks (Ideker et al. 2002; Bryant et al. 2013; Guo et al. 2007). This method scores nodes, often based on altered co-expression (see Section 2.2.5). Simulated annealing is capable of detecting linear modules, since it is not dependent on clustering coefficients. The method generates sets of interconnected high scoring nodes (Ideker et al. 2002) or edges (Guo et al. 2007) are extracted using probabilistic techniques. For example Ideker et al (2002) applied the approach to PPI and protein-DNA networks and was able to extract linear modules shown in Figure 3. However, this method is reliant on molecular data to generate the network topology and score the nodes. Linear chains of molecular interactions are particularly sensitive to incompleteness within interaction data, since a single missed link would break these chains. In addition, the requirement for data to score node is another source of error, since expression data can be undermined by post-transcriptional modification etc. (see Section 2.3). Pathways bring genes into close proximity clusters without the specifics of the molecular interactions being known. They also allow the allocation of function without expression data which is limited and is only applicable to the conditions tested.

Other studies have also approached the issue that functional modules may not be connected within molecular interaction networks. For example, proteins can be grouped based on their tendency to interact with other groups of proteins (Pinkert et al. 2010). Figure 49 shows four groups of proteins in which the defining characteristic of groups ‘b’ and ‘d’ is the nodes’ tendency to interact with groups ‘a’ and ‘c’. Application of this method to a human PPI network revealed grouping such as transmembrane proteins, which cannot be identified by searching for cohesive group of nodes (Pinkert et al. 2010).

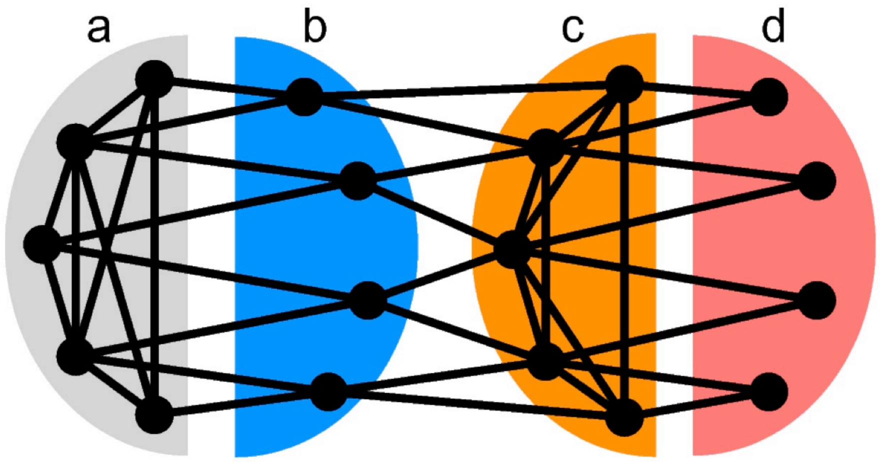


Figure 49: Node function identified by structural network position. Nodes in groups 'a' and 'c' form tight clusters, but groups 'b' and 'd' do not. Group 'b' is identified based on each nodes' shared tendancy to interact with nodes in groups 'a' and 'c' *(Pinkert et al. 2010)*

By applying enriched GO terms to pathways, functions can be captured even if they span multiple pathways or do not apply to all the genes within a pathway. Within the yeast network, many processes including cellular respiration, ribosome biogenesis, and cellular response to oxidative stress were covered by multiple pathways (see Section 4.4.6). Ames et al. (2013) showed that the functional modules generated using single molecular interaction networks, such as PPI networks, may not give comprehensive network clusters (see Section 2.2.6.1, Figure 4). This disjointed distribution of function within molecular data may make extraction of functional modules more difficult. Within the pathway network, knowledge of the physical interactions between functionally related gene sets is not required, as pathways are connected by functional links. As a result, rather than functional modules being defined by network clusters that may not provide an accurate representation of their structure, shared function defines the topology of the network.

The finding that functions span multiple pathway is supported by the Pathway Ontology, in which multiple pathways are gathered into functional groupings. The hierarchical nature of the ontology illustrates how multiple pathways collude to perform high-level tasks (see Section 2.6.1, Figure 6) Pathway suites are also generated by linking pathways by common concepts: for example the 'Glucose Homeostasis Pathway Suite Network' brings together pathways involved in glucose metabolism with related regulatory and signalling pathways (see Section 2.6.1).

Another example of the disparity between the borders of functional modules and pathways came from the multiple functions associated with each pathway. In the yeast network, 65% of pathways required multiple GO terms to represent their function (see Section 4.4.3), indicating that the GO term most significantly enriched to each pathway did not apply to every gene. Figure 23 in Section 4.4.4 shows that the functions covered within pathways may not be similar terms or parent/child nodes within the Gene Ontology, but reflect diverse functions. By allowing pathways to correspond to multiple functions, pathway enrichment is better able to represent this level of biological complexity.

Fuzzy clustering methods have been suggested to allow biological entities to represent multiple functions (Zhang et al. 2007) within a static network. Functional cluster are allowed to overlap and the likelihood that genes are positioned within a cluster is calculated. However, such methods still depend on functional clusters being generated from interaction networks, restricting detectable functions to those that form topological clusters.

7.2.3 The Wang best-match average best reflects the functional complexity of pathways

By studying the semantic links between pathway pairs, this study expands the current understanding of semantic distances between sets of GO terms (see Section 6.3.2.1). The method expands on previous approaches which assessed the ability of different semantic similarity measures to predict known PPIs (Guo et al. 2006; Jain & Bader 2010), co-expression patterns (Jain & Bader 2010) and sequence similarity data (Lord et al. 2003b). Our study was based on the premise that pathways should generally display some degree of functional homogeneity. This is intrinsic to the general definition of pathways and was confirmed by Guo et al (2006) and Mathur & Dinakarpandian (2012), with both papers showing that pathways have more similar GO terms than expected by chance. Comparing the ability of semantic similarity methods to distinguish semantic similarities between and within pathways allows the functional homogeneity of pathways to be applied to the validation of these methods. By measuring the distances between all pathways pairs, we generated a random distribution of semantic similarities against which the similarity of GO terms inside pathways could be compared (see Figure 39 in Section 6.4.2)

The distribution of semantic similarities within pathways offers insight into the degree of correlation between the structure of the Gene Ontology and the organisation of functions into pathways. For all methods tested, GO terms within pathways were more similar than GO terms between pathways. The tendency of the Wang method to outperform the Resnik method indicates that graph-based semantic similarity methods may present a more accurate depiction of cellular organisation. This supports the hypothesis that the current structure of the Gene Ontology reflects real relationships in the cell, information which is lost in information content-based methods (see Section 3.4.1).

The flexibility of the proposed methods to measure general patterns of similarity both between and within pathways is important, since the boundaries of pathways and functions show a general correlation, but do not directly conform to each other. The functional similarity of proteins within pathways is shown to decay as pathway length increases (Guo et al. 2006). This is intuitively explained as pathways being comprised as a series of different functional steps towards a discernable biological outcome. Directly interacting proteins show close functional relationships, possibly generating a gradient of function across the pathway. By generating pairwise distributions of semantic function between and within GO terms attributed to pathways, the frequency of divergent terms occurring within pathways can be observed (see Section 6.4.2).

The finding that the Wang method outperformed the Resnik method contradicted previous studies, which had measured similarity between gene annotations, as opposed to pathway annotations. Several previous studies found the Resnik measure to be the most suitable (Sevilla et al. 2005; Guo et al. 2006; Mistry & Pavlidis 2008; Guzzi et al. 2012; Jain & Bader 2010). This may be because the use of information content based methods is not suitable with pathway data, since these methods count the number of genes allocated to each GO term, allowing equal weighting to each gene. However, if a gene appears in multiple pathways, the frequency of the GO terms associated with the gene may increase. If the GO term does not apply to each pathway instance, a problem should not arise, since pathway enrichment analysis is unlikely to return the extraneous terms. However, if the GO term happened to be applicable in each pathway instance, then the term becomes more frequent in the network than accounted for by its information content.

It is worth noting that both Guo et al. (2006) and Mathur & Dinakarpandian (2012) found that Resnik was the most effective measure for generating semantic distances between genes within pathways. Guo et al (2006) used each method's ability to predict PPIs to measure efficacy. This method therefore only dealt with direct protein interactions without measuring more general, indirect relationships such as shared pathways, distinguishing the method used by Guo et al (2006) from the approach used in this thesis. Proteins that interact are more likely to have GO terms with a highly informative common ancestor. Mathur & Dinakarpandian (2012) tested each method's robustness to alterations in the lower levels of the Gene Ontology, therefore the information content of nodes close to the root were not affected. In contrast, using edges based methods, all measures that involved the perturbed terms would have been affected.

Studies have also shown the pairwise maximum method of comparing gene sets was more effective than the best match average and pairwise average (Jain & Bader 2010; Xu et al. 2008). The work presented in this thesis compared each method's ability to predict PPIs, concluding that the pairwise maximum method was effective because proteins only require a single basis of similarity to interact. If proteins have multiple pleiotropic functions, the existence of a second function should not make interactions connected to the first function less likely.

However, when connecting pathways based on shared functionality, we aimed to generate a modular network linking pathways with similar functions (see Section 6.3.2.1). If pathways performed random combinations of functions, then a modular network would not have been formed. The co-implementation of multiple functions within pathways is likely to represent some link between the functions, since pathways cannot be involved in independent functions in the same way genes can. This is because genes can be involved in independent functions by participating in different sets of interactions. In contrast, all the edges in the pathway network are assumed to be present whenever the incident pathway node is active. The co-occurrence of functions that the Gene Ontology regards as semantically diverse is likely to represent real biological connections, which are not represented with the Gene Ontology. This is supported by the finding that functionally diverse GO terms are often co-enriched in sub-graphs within molecular networks (Ames et al. 2013). This study found that around 40% of co-enriched GO terms within molecular networks were semantically diverse.

Based on the finding that functional diversity within pathways way common, network edges were generated using the best match average and pairwise average (see Section 4.4.8), since these methods allowed the entirety of each pathway's functions to be considered. The best match average measure was able to assign strong similarity to pairs of pathways with matching co-expressed GO terms, whereas the pairwise average method would reduce the score if the co-enriched terms were semantically diverse.

7.2.4 Exploring disease mechanisms through the functional pathway network

This chapter describes the opportunities provided by pathway networks to explore disease mechanisms.

7.2.4.1 Deciphering disease mechanisms and potential disease genes

The idea that polygenic diseases arise from combinations of mutations is well established (Goh et al. 2007). Sets of disease genes can be mapped onto molecular interaction networks, indicating the location of the perturbed function. Given this data the disease mechanism, along with a set of potential disease genes, should become apparent. However, gaining this insight is often non-trivial. Barabasi et al. (2011) noted that functional clusters and disease clusters overlap but are not equivalent. This could be due to difficulties in mapping functional and disease modules based on molecular network topology (see Section 7.2.2). In addition, disease modules incorporating multiple functional modules could add to this divergence. The ability of diseases to affect multiple functional modules is intuitive since biological processes do not work in isolation, but instead involve a high degree of crosstalk and co-regulation. Therefore, when generating disease modules it is not sufficient to isolate single functional modules, as any contributing processes in the functional periphery could also be involved in generating the disease phenotype.

Within molecular interaction networks, it is challenging to extract single functional modules (see Section 7.2.2), making identification of all functional modules within close proximity to a set of disease genes problematic. However, the pathway network proposed in this thesis reliably attributes functions to nodes, which are then directly linked to other functionally related pathways. As a result mapping diseases onto the pathway network reveals disease modules along with pathogenic mechanisms (see Section 6.3.3). Even if the disease pathways do not form network clusters, examination of enriched pathway functions and the full edge set (Section 6.3.2.2) will establish the closest functional links (see Section 7.3.3). Potential disease genes may then be identified within affected pathways and functions.

7.2.4.2 Pathway context affects the disease phenotypes of pleiotropic genes

Mapping diseases onto molecular interaction networks has contributed towards the elucidation of disease mechanisms (Janjić & Pržulj 2012c), identification of new disease-associated genes (Barabasi et al. 2011) and indication of potential drug targets (Yu et al. 2007). Molecular networks show the sum of each molecule's interactions without distinguishing between cellular contexts, therefore if a gene has multiple functions they may form a single cluster on the network. This becomes apparent when considering the ability of genes to contribute towards multiple disorders. If genes are limited to a single consolidated function, and diseases are considered to be phenotypic representations of a disrupted function, then a disrupted gene should result in a single disorder. If different mutations disrupt the function to varying degrees then this may produce a range of phenotypically similar disorders. However, gene mutations can be phenotypically diverse, such as *AKT1,* which is associated with schizophrenia, colorectal cancer, ovarian cancer and breast cancer (Chavali et al. 2010). Further evidence suggesting that diseases may act independently within different pathways comes from the finding that many disease pairs with shared genes do not show significant co-morbidity (Park et al. 2009). This indicates that a gene may be functionally disrupted in one cellular context, but perform normally within a different cellular context. In accordance with this hypothesis, comorbid associations are less likely if the causative mutations occur on different protein domains in each disease instance.

Molecular interaction data provides additional evidence that genes' ability to affect multiple diverse diseases may arise from their ability to act in multiple pathways. Genes associated with multiple disorders typically showed low clustering coefficients in molecular interaction networks, indicating that the nodes adjacent to them are less likely to interact than expected by chance (Chavali et al. 2010). This supports the hypothesis that these genes are acting in multiple separate pathways. Chavali et al. (2010) also showed that genes associated with multiple disorders have lower levels of co-expression with their interaction partners than genes associated with a single disorder. Co-expression decreases further if the gene is associated with functionally diverse diseases. Phenotypically diverse genes may also interact with different partners in different tissues: for example *AKT1* participates in a range of interactions dependent on tissue type, further supporting the hypothesis that the results arise from the gene acting in different pathways.

These results provided compelling evidence that pleiotropy allows genes to be involved in multiple disorders, as well as demonstrating that molecular networks are poorly suited to capture multiple functions of individual molecules. As a result, there is a need to develop methods, which are more suitable than molecular networks to study this occurrence. The pathway network provided in this thesis provides an established source of context depended interactions in which multiple functions can be assigned to genes depending on their pathway context (see Section 4.4.8). Therefore, the function of the perturbed pathways indicates the role that a disease gene may play in each disorder (see Sections 6.4.4 and 6.4.5).

7.2.4.3 Overlapping disease modules and functional cross talk

Co-morbidity analysis of the Human Phenotype Network showed that diseases that share genes tend to show increased levels of co-morbidity (see Section 2.5). This could be another effect of gene pleiotropy or it could represent crosstalk between functional modules. Mutations in one or more pleiotropic genes could simultaneously perturb the pathways corresponding to separate diseases. In this scenario, each disease pathway is independent in the cell, they just happen to share one or more pleiotropic genes with disease causing mutations.

In some cases, it may be more likely that pleiotropic disease genes represent overlap between disease pathways. Each disease module is a result of one or more functional modules and since functions are known to overlap in the cell, disease modules may also overlap. Highly related to this idea is the concept of cross talk between functional modules. Biological processes are highly interconnected, blurring the boundaries between functions. The pathway suites generated within the Pathway Ontology illustrate this interconnectivity (Petri, Jayaraman, et al. 2014). An example given in their recent paper is the 'Glucose Homeostasis Pathway Suite Network', which demonstrates the relationship between the glucose metabolic pathways and the signalling and regulatory pathways, which must collaborate to maintain suitable glucose levels.

Evidence for crosstalk between functions occurs from co-morbidities between seemingly unrelated diseases. Co-morbidity data linking seemingly unrelated disorders such as autism, cerebral palsy, schizophrenia and Parkinson's disease were predicted to resulting from genetic changes weakening the immune system (Rzhetsky et al. 2007). Of interest is the wide-ranging age of onset seen in these diseases, with autism typically manifesting before the age of three, schizophrenia developing during the teenage years and early twenties, and Parkinson's disease appearing in older adults. The diversity of these disease phenotypes as well as the diverse age of onset implies that these disorders have highly variable biological mechanisms involving a range of cellular functions. Female breast cancer was also found to be negatively correlated with bipolar disorder and schizophrenia, which was interpreted as corresponding to genes involved in the cell cycle and cell death, since female breast cancer is associated with abnormal cell proliferation, and schizophrenia and bipolar are associated with abnormal cell death in some tissues. This shows different proliferative effects that an affected gene can have on diverse biological processes.

In these instances it is clear that these disorders arise from different polygenic gene sets and have very different pathological mechanisms. To study the divergence of these disease phenotypes from their shared origins the relationship between cellular functions must be understood. Pathways with multiple functions provide platforms for crosstalk and co-regulation between the biological processes in the cell.

The pathway network provides a tool to examine the physical connectedness of divergent functions affected by shared gene mutations. If the functions are closer than expected within the network, the physical interactions linking them suggest cross talk and co-regulation through the shared disease mutations. If the network distance between the functions is great and the disease genes are present in multiple pathways, then pleiotropy may be more likely (see Section 6.3.3.2).

The interesting ambiguity between pleiotropy and disease cross talk can also be observed within the diverse clinical manifestations associated with individual diseases (Zhou et al. 2014). Diseases with diverse ranges of symptoms were found to be associated with genes with high betweenness centrality in the PPI network. This indicates that these genes affect diverse cellular mechanisms. As discussed in Section

7.2.4.2 this may arise from pleiotropy if genes are in multiple pathways. However, if genes are restricted to a single pathway then functional crosstalk would allow local perturbations to extend their influence, giving rise to a range of diverse symptom phenotypes.

7.3 Limitations of the current research

The generation of the novel network structure presented in this thesis required the innovative re-evaluation of existing methods as well as the development of new approaches. Some shortcomings within the methods used are explored within this Section, along with their potential influence on the resulting networks.

7.3.1 Issues with heuristic methods

A heuristic set cover method was used in section Chapter 5 to successfully reduce redundancy between the human pathways. However, the set of pathways returned is dependent on the selection of early pathways. Within the first stage of this algorithm, all uncovered sets are assigned a score of one, plus a small modifier indicating how close each set is to a desired set size (such as the median size). The purpose of the modifier is to ensure that if two sets have the same proportion of uncovered elements, the set whose size conforms most closely to the median is selected. At the beginning of the algorithm, every median sized set will have exactly the same score, therefore a set will be selected at random from this subset. For example in Figure 50A, sets a and b both match the median size (4) and will therefore have the same score. Set selection will continue until all of the sets with the median number of elements have at least one element covered. The process is then repeated on the sets whose size differs from the median by one, and so on. As a result, substantial numbers of sets are selected randomly at the beginning of the algorithm.

This may result in the algorithm producing different results when run multiple times. For example, Figure 50B and Figure 50C show the different outputs generated depending on whether set ‘a’ or set ‘b’ is selected first. If set ‘a’ is selected first, the result shown in Figure 50B will be produced. The next set to be selected will be set ‘e’ since it is the only set remaining with all of its elements uncovered. Finally, set ‘d’ will be selected as all of the genes in set ‘c’ are covered. In contrast, if set ‘b’ is selected first, set ‘c’ will replace set ‘d’ in the final output (Figure 50C).

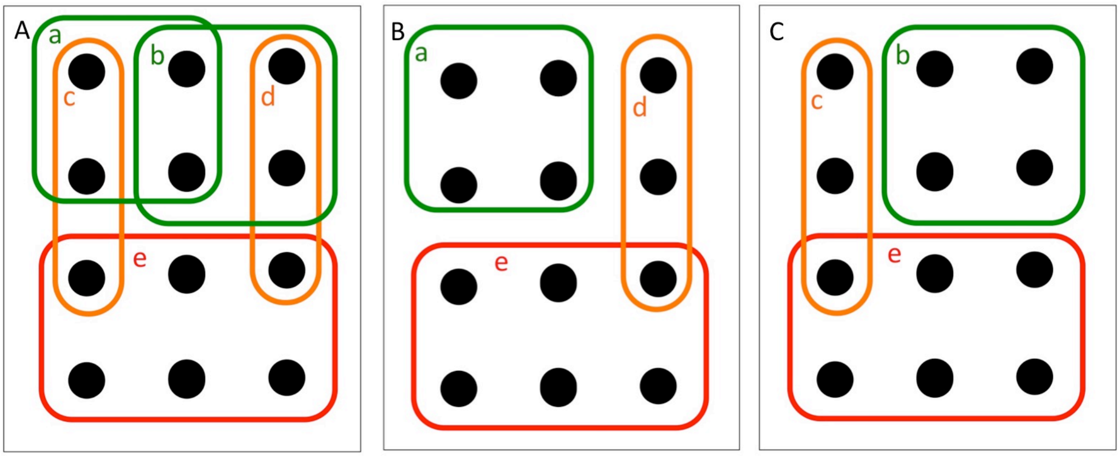


Figure 50: Heuristic effects of set cover output. A) Contains five sets with a median size of four. Pathways a and b are exactly median sized (green), sets ‘c’ and ‘d’ deviate from the median size by 1 (orange), and pathway ‘e’ deviates from the median size by 2 (red). Pathways are selected primarily on the proportion of uncovered genes they contain. If two sets have an equal score, a decision is based on the size, preferentially selecting pathways close to the median size of four. At the beginning of the algorithm, pathways ‘a’ and ‘b’ have the same proportion of uncovered elements and the size therefore a decision is made at random. B) and C) show the different outputs generated by this decision.

Each application of the algorithm to the pathway dataset results in a set of pathways with reduced redundancy and a similar size distribution, suitable for use within the network. Each pathway subset also covers exactly the same set of genes so the only difference is in the position of the pathway boundaries. There is a chance that some of the functions present within the pathways will change, since the results of enrichment are dependent on the gene sets subjected to analysis. The thesis presents a network based on a single output from the set cover algorithm. It represents a set of pathways that provide gene coverage of the entire dataset and extensive GO coverage. The network generated portrays the organisation of function across these pathways. Future work could explore the network variations provided by different pathway set covers. However, since all sets of pathways share have reduced redundancy and equal biological validity, there is little reason to select one set cover other another.

7.3.2 Conflicting current opinions over inclusion of ‘part-of’ GO edges

There is debate in the literature about whether both ‘is-a’ and ‘part-of’ GO links should be included in GO graphs when calculating semantic similarities (Sheehan et al. 2008; Jain & Bader 2010; Guo et al. 2006; Yu et al. 2007; Resnik 1999). The Wang method resolves the problem by including both but denoting a higher weighting to ‘is-a’ links than ‘part-of’ links (Wang et al. 2007). Since the Wang method required integration of both types of edges, both edge types were used throughout the project for consistency. The Resnik method, however, classically excludes ‘part-of’ links since they denote less specific information (Resnik 1999). Guo et al. (2006) included both types of edges, but stated that inclusion of ‘part-of’ links, which are particularly prevalent in the

Cellular Compartment ontology may have resulted in poorer predictive power between pairs of interacting proteins. It is possible that if 'part-of' edges had been excluded a different network may have been produced. This network may have had fewer spurious connections, in cases where 'part-of' relationships had low biological validity, however the loss of information may have resulted in a less complete portrayal of function.

7.3.3 Discarded functional edges may have biological significance

This project was ambitious with its goal to map the functional connectivity of all pathways and functions within humans and yeast. This is a complex task since within cells functions are highly interconnected and interdependent. This was reflected by the dense network edges representing pairwise semantic similarities (see Section 6.3.2.1.2). Unfortunately, the density of this network obscured the major functional groupings and made analysis extremely computationally slow. We therefore omitted some of the weaker links within the network, using a threshold optimised to maximise network cohesiveness while minimising density (see Section 6.3.2.2). The threshold used effectively satisfied both criteria, however many edges below the threshold may have been biologically relevant. We succeeded in producing networks representing comprehensive gene cover and extensive representation of global functions for each organism. Although the major functional connections have been mapped, due to the scale of the project some detail has been lost. Details of all functional connections are retrievable through examination of the full set of semantic distances between pathways. The semantic annotations of each pathway are also available, so that the functional links can also be retrieved for any pathway within the dataset (https://data.mendeley.com/datasets/3pbwkxjxg9/1).

A related issue is that the structure of the network produced reflects the structure of the GO. While this is largely advantageous since the GO represents a well-established model of biological function, there are instances where it may not closely represent the inner-workings of the cell. The clusters of signalling and regulatory pathways are examples of this, since their position within the network is dominated by their functional similarity to each other, rather than the processes they control. Their influence could be inferred by examining weaker links to other pathways; however it may be preferable if these relationships were apparent from the network topology. It could be possible to link signalling pathways more directly to the processes they control by excluding GO terms relating to signalling and information transfer when calculating the network edges.

7.3.4 Multiple testing within disease pathway detection

To detect disease pathways (see Section 6.3.3.1) we used the Fisher's exact test without correcting for multiple testing. This decision was made because implementation of the correction reduced the number of pathways per disease from a mean of 18.9 to 2.4, representing a large reduction in the number of pathways detected. This would make exploratory network analysis to identify regions associated with disease impossible, since it is not meaningful to search for disease modules with less than 3 nodes. Methods that could be used to resolve this issue are discussed in Section 3.3.2.

An alternative approach would have been to simply refer to all pathways containing a disease gene as disease pathways, however, this would be contradictory to the project's primary premise, which is to explore and distinguish the effects of pathway context on genes. Assigning a pathway disease status as a result of a single gene, which may not be affected by disease perturbations within the context of this pathway, would reintroduce the problems of molecular networks that this project seeks to avoid. Using enrichment analysis without correcting for multiple testing provides a compromise, since it detects pathways that contain proportionally high levels of disease genes. The p-values used to generate the disease pathways cannot be considered an accurate indication of probability; instead they should be used as an indication that the pathways selected using the Fisher's exact test are more likely to be associated with disease than pathways that were not selected. It is known that disease genes cluster into pathways rather than being randomly distributed, minimising the concern that high numbers of false positives will arise through the selection of random pathways.(Peng et al. 2010; Koutsogiannouli et al. 2013; Furlong 2013; Ghamlouch et al. 2017; Barabasi et al. 2011). All of the disease pathways examined within this project have strong literature support, affirming the validity of this approach.

7.4 Future research

Within this final section we suggest some future work that could follow from this thesis. We suggest applying controllability analysis to the network, to uncover pathways responsible for controlling the flow of information between cellular processes. This would further elucidate the physical mechanisms responsible for cellular organisation and disease phenotypes. In addition, future work could assess the suitability of the full range of available semantic similarity methods for forming network edges. Finally, the generation of disease modules could be enhanced through improved development of disease pathways. We suggest the application of genome-wide association studies to disease SNPs could be used to aid identification of disease pathways.

7.4.1 Creation of gene panels

Gene panels are commonly used in diagnostics for many disorders such as congenital muscular dystrophy (Valencia et al. 2013) and epilepsy (Lemke et al. 2012). In these situations genetic locus heterogeneity makes the use of gene panels preferable over single gene testing (Xue et al. 2015). Exome sequencing is also exceedingly common, particularly for disorders with high genetic heterogeneity; however, assessing the pathogenicity of the high numbers of sequencing variants detected can be problematic. Gene panels can provide a more targeted approach and offer four to five fold greater coverage than exome sequencing. Genetic panels are able to provide differential diagnosis between disorders with overlapping phenotypes. For example, Fanconi-Bickel syndrome and glycogen storage disease both present with fasting hypoglycaemia, therefore genetic markers for Fanconi-Bickel syndrome can be included in panels for glycogen storage disease.

There is, however, poor consensus between the sets of genes selected for different disorders, for example, the number of genes provided in an epilepsy gene panel can vary from 70 to 377 (Xue et al. 2015). Some laboratories may choose to include all genes remotely associated with a phenotypes, hoping to give a better diagnostic, however, genes selected based on association studies or single reports are frequently found to be non-causative. This may complicate interpretation of the panel results. By highlighting sets of pathways involved with disorders the pathway network could assist in the generation of effective gene panels. Firstly the method generates enriched pathway gene sets for diseases, a method known to indicate likely disease genes (Yu et al. 2007; Liu & Chance 2013). By linking pathways with similar functions, the network assembles pathways relevant to a particular disease aspect. This approach may identify the genes likely to contribute towards particular disease characteristics, enabling the generation of gene panels appropriate to patient symptoms.

In addition, the networks ability to represent a comprehensive range of diseases could contribute towards the development of differential gene panels. The screening of pathways associated with multiple disorders, as well as pathways capable of distinguishing between disorders could give insight into the relationship between these diseases. This could lead to better diagnostics as well as increased understanding of disease mechanics.

7.4.2 Controllability analysis

The pathway networks show the arrangement of functional modules within the cell, therefore it was of interest to identify the pathway nodes responsible for conveying information between these modules. To examine the possibility of controlling the flow of information within the cell network controllability analysis was applied to to the human pathway network. Controllability analysis aims to find minimum dominating sets capable of controlling all nodes within the network (Nacher & Akutsu 2012; Liu et al. 2011). The minimum dominating set (MDS) is the minimum subset of nodes capable of determining the state of any of the nodes in the network.

Liu et al. (2011) proposed a model in which 'driver nodes' within a directed network control all of the nodes downstream of them. In this model, it is possible for a driver node to control a chain of downstream nodes, without restriction on the length of the chain (

Figure 51a). However, driver nodes with multiple outgoing edges can only control a single down-stream node

Figure 51). The reasoning behind this rule is that it is impossible to generate the full range of network states if a driver node influences all of its outgoing edges collectively. Multiple MDS are often possible for a single network, meaning that different, equally sized sets of driver nodes can be used to control the network. Therefore, the edges are categorised based on the frequency of their presence within the minimum dominating sets. An edge is critical if it is used in every MDS to confer control between a driver node and its downstream counterpart. Redundant edges never have negative incidence to a driver node in a MDS and ordinary edges are required by some of the MDS.

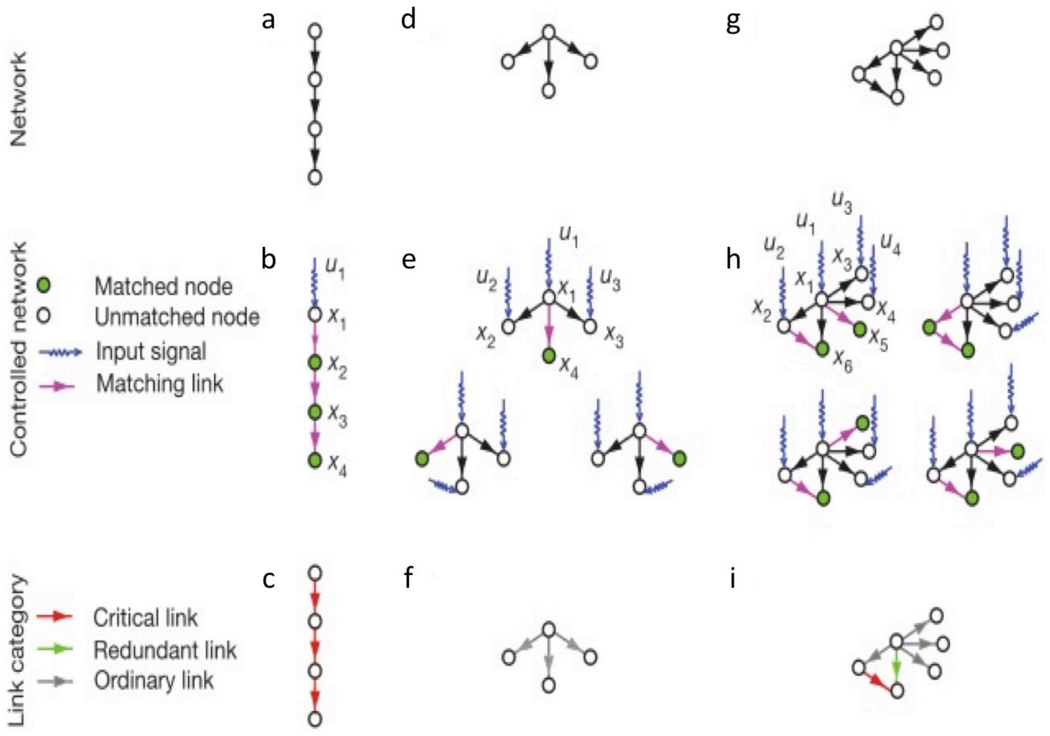


Figure 51: Model of controllability, taken from *Liu et al. (2011)* 'a', 'd' and 'g' show three directed networks to be controlled. 'b', 'e', 'h' show possible minimum dominating sets for each network. The white nodes are the controlled nodes, with the blue arrows indicating input. The multiple networks in 'e' and 'h' show all available control models. The purple arrows indicate edges being used by driver nodes to control down stream nodes (green). 'c', 'f' and 'i' show whether the edges in each model are critical, redundant or ordinary, based on whether they are in all, none or some of the above minimum dominating sets.

Nacher & Akutsu (2013) developed an alternative model in which driver nodes could control all of their outgoing edges independently, allowing driver nodes to control multiple nodes. Control could only be passed from driver nodes to nodes directly adjacent to them, disallowing chains controlled by a single driver node. By allowing nodes to utilise all of their edges independently, the focus was shifted from edges to nodes. Therefore, nodes were classified as critical if they were present in all MDS, redundant if they were absent from all MDS and intermittent if they were present in some MDS (Ishitsuka et al. 2016). The method was also expanded to incorporate undirected graphs. This approach generated very different results from the previous methods. In Liu et al. (2011) driver nodes tend to avoid hubs and the number of driver nodes required to control sparse networks is high. This is particularly true if the network has a scale free distribution, seen in most biological networks. In contrast, when the method described in Nacher and Akutsu (2013) was applied to PPI networks from many organisms including yeast and human, hub nodes were significantly more likely to be critical across all networks. The idea that hub nodes have special importance within biology is well established, since hubs are more likely to be essential genes (He & Zhang 2006b), have higher evolutionary conservation (Barabási & Oltvai 2004) and frequently act as global connectors between functional modules in the cell (Han et al. 2004). Nodes with high betweenness centrality are also enriched within minimum dominating sets (Wuchty 2014) supporting the hypothesis that these nodes are involved in global cellular connectiveness. For these reasons, the approach suggested by Nacher and Akutsu (2013) was considered more suitable for biological network analysis.

Controllability analysis was applied to the human pathway network, to establish if disease pathways, cancer pathways and pathways enriched with essential genes (essential pathways) were more or less likely to be critical nodes, however no significant relationships were observed using the Fishers Exact test. Preliminary results suggested that in the pathway network the node degree was less important than node betweenness centrality. This is most likely due to the pathway network not having a scale-free topology, reducing degree inequality. Therefore pathway nodes between connecting pathway clusters were most likely to be critical or intermittent pathway nodes. Pathways enriched with disease genes, cancer genes and essential genes did not significantly correspond to these pathways.

Future work should take a more targeted approach, identifying pathways involved in multiple types of cancer (or other diseases), which are expected to have enhanced control within disease sub-networks. Cancer genes are known to form interconnected clusters (Wu et al. 2010) and are therefore likely to localise to cancer pathway modules within the network. However, the prevalence of genes such as *P53* which are involved in a large number of cancers made us question whether some pathways could be central to the control of the cancer sub-network. Pathways associated with many types of cancer should be tested to see if they are control nodes in the cancer sub-network, since they should link modules associated with different cancer types.

7.4.3 Comprehensive exploration of semantic similarity measures

Through analysis of semantic similarity, insight into the relationship between pathways and biological processes has been gained. We used the Wang and Resnik methods since they have been shown to perform well in previous studies, and they cover both information content and topology based approaches, respectively (see Section 3.4). More semantic similarity methods are available and a more thorough examination of the options could deepen insight into the distribution of function within pathways.

Many papers reported that the Resnik method exceeded the Wang method, however, in our study the Resnik method performed poorly. Possible reasons include the Resnik method’s failure to consider the proximity of the test terms to the most informative ancestor, or the multiple instances of genes in pathways, or the inclusion of ‘is_a’ terms.

Lin (1998) proposed a method capable of calculating biological process similarities that show strong agreement with sequence similarity (Lord et al. 2003b; Guo et al. 2006). The Lin method is similar to the Resnik method in that it uses the doubled information content of the lowest common ancestor, divided by the summed information content of the two test terms. The Lin method is designed to overcome the failure of the Resnik method to consider the distance between the test terms and the most informative common ancestor. This would provide insight into the cause of the Resnik method’s poor performance in this study. The Lin method could potentially exceed both the Wang and Resnik approaches, if the Resnik method’s effectiveness is reduced by the distance between test terms and ancestors. Jiang & Conrath (1997) produced a similar method in which the information content of the test terms is subtracted from the doubled information content of the most informative ancestor, which could also produce promising results. Additionally the Resnik method could be reapplied with the ‘is_a’ links removed.

To join sets of GO terms within the links we used the best match average and pairwise average methods, since these approaches incorporated all of the functions of each pathway (see Section 3.4.2). However, some studies have shown that the maximum distance approach outperforms other methods in predicting protein interactions, making it an interesting option for generating network edges (Jain & Bader 2010; Xu et al. 2008). Testing all semantic similarity measures using the maximum semantic similarity would be ideal.

7.4.4 Using GWAS to generate a more complete set of disease pathways

In this thesis, established disease genes from the human phenotype database (HPD) were used. However, more detailed SNP data is available. Methods such as Genome-wide association studies (GWAS) are able to identify pathways significantly enriched for disease genes in situations where very low numbers of individual genes were statistically significant (Torkamani et al. 2008). This is thought to occur in many common diseases, which arise from large numbers of low risk genetic factors, referred to as polygenes (Peng et al. 2010; Torkamani et al. 2008). Polygenic disease mutations at multiple loci make small contributions to disease susceptibility. For example, a large study into seven common disorders (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes) demonstrated that pathways significantly enriched for disease SNPs were able to explain disease phenotypes (Torkamani et al. 2008). GWAS is also capable of identifying disease pathways in situations where very low numbers of disease genes are statistically significant. In another example, hypertension was associated with low numbers of individual genes, however a long list of interconnected enriched pathways were identified spanning processes such as dopamine signalling, calcium signalling, glucose metabolism, cell-cell interactions and cytoskeletal remodelling (Torkamani et al. 2008). Gene expression studies have also been used to identify KEGG pathways with varying expression in Alzheimer’s patients (Miller et al. 2013). Different pathways were enriched in different brain regions demonstrating that this method is capable of detecting cell type specific disease mechanisms.

GWAS has the potential to provide inclusive, accurate sets of disease pathway data, which could improve the mapping of disease modules. The application of GWAS to SNP data covering all diseases is a non-trivial undertaking that was beyond the scope of this thesis, however the development of pathway networks facilitates development in this direction.

Chapter 8

Conclusion

In this thesis, we have presented a new approach to modelling biological function, which addressess many of the issues affecting current used molecular functional networks. We showed that networks generated using functionally annotated nodes can generate an alternative model of cellular function and demonstrated the validity of this method using genetic interaction and disease data.

The use of pathway networks to organise cellular processes based on function, was first developed in yeast. A network model of functional pathways was generated and major functional modules were examined, taking advantage of the simplified single cell system. The general structure of the network made intuitive biological sense (Section 4.4.6) and genetic interaction data was used to confirm the biological validity of the network clusters generated (Section 4.4.7). Application of the pathway network to model function could be particularly applicable in situations where molecular interaction data is sparse or uncertain, or gene pleiotropy is suspected. In addition, the pathway network provides a solution in situations where clusters in molecular networks have shown a poor correlation with function.

Pathway multi-functionality was examined, revealing high variation in the number and variety of functions that pathways facilitate (Section 4.4.5). The ability of pathway nodes to represent multiple functions demonstrated the system's capability to handle biological complexity. The presence of pleiotropic genes was confirmed within the network, demonstrating the ability of the model to capture independent functional instances.

Expansion of the model into more complex human data required additional pre-processing to avoid excessive pathway redundancy. Combinatorial algorithms were developed to facilitate to the competing aims of maximising coverage, minimising redundancy and controlling size. Size homogeneity is essential to obtain comparable pathway nodes and maintain functional specificity. Three set theory algorithms are presented to allow varying compromises between reducing pathway redundancy, controlling pathway size and maximising coverage.

Using the new pathway set, a network covering global human function was generated. Multiple methods for measuring semantic similarity between pathways were explored in order to determine the optimal solution for generating edges in the pathway network. The major functional clusters were examined and validated using disease data. Diseases pathways were shown to be more closely connected within the network than random pathways (Figure 43). I addition, GO terms assigned to disease pathways reflected disease mechanisms and phenotypes found in the literature (Section 6.4.5). The generation of disease modules demonstrated the potential of the pathway network to study disease mechanisms. Mapping disease genes into disease modules provides new insights into disease mechanisms, forming an intermediate between molecular data and histological data.

An advantage of pathway networks compared to molecular networks, is that they do not depend on the graph topology of molecular interactions to attribute function, allowing them be used in cases where clustering methods are ineffective or molecular interactions are poorly understood. Additionally, pathway networks do not allow pleiotropic genes and temporal interactions to distort the distribution of function within the network. They could therefore clarify situations where molecular networks bring unlikely pairs of functions into close proximity. The disadvantage of the pathway network topology is that the Gene Ontology may not always reflect the intracellular organisation of function; for example, regulatory pathways cluster together rather than clustering with the processes that they control. Molecular networks also provide more biological detail, showing the physical interactions that comprise each function.

Pathway networks and molecular networks have different advantages and disadvantages, making the combined use of both methods promising. Although developing the methods to generate pathway models required intensive innovation, running analyses on an existing pathway network is relatively straightforward given the small size of the network. The incorporation of pathway network analysis with molecular network analysis is therefore highly promising for functional studies. The combined use of pathway and molecular networks is expected to be more efficient that generating co-expression or protein expression/interaction data for individual cellular conditions.

References

Akira, S. & Takeda, K., 2004. Toll-like receptor signalling. *Nature reviews. Immunology*, 4(July), pp.499–511.

Albert, R., 2005. Scale-free networks in cell biology. *Journal of cell science*, 118(Pt 21), pp.4947–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16254242 [Accessed October 19, 2013].

Alberts, B. et al., 2002. *Molecular Biology of the Cell. 4th edition* 4th ed., Garland Science.

Alexa, A., Rahnenführer, J. & Lengauer, T., 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13), pp.1600–1607.

Altelaar, A.F.M., Munoz, J. & Heck, A.J.R., 2012. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature reviews Genetics*, 14(1), pp.35–48. Available at: http://dx.doi.org/10.1038/nrg3356%5Cnpapers3://publication/doi/10.1038/nrg3356.

Amersi, F., Agustin, M. & Ko, C.Y., 2005. Colorectal cancer: Epidemiology, risk factors, and health services. *Clinics in Colon and Rectal Surgery*, 18(3), pp.133–140.

Ames, R.M. et al., 2013. Modular biological function is most effectively captured by combining molecular interaction data types. *PloS one*, 8(5), p.e62670. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3643936&tool=pmcentrez&rendertype=abstract [Accessed January 9, 2015].

Ao, S.I. et al., 2005. CLUSTAG: Hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, 21(8), pp.1735–1736.

Ashburner, M. et al., 2000. Gene Ontology : tool for the unification of biology. *nature genetics*, 25(1), pp.25–29.

Barabasi, A.-L., Gulbahce, N. & Loscalzo, J., 2011. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1), pp.56–68. Available at: http://dx.doi.org/10.1038/nrg2918.

Barabási, A.-L. & Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2), pp.101–113.

Belinky, F. et al., 2015. PathCards: multi-source consolidation of human biological pathways. *Database*, 2015(2), p.bav006-bav006. Available at: http://database.oxfordjournals.org/cgi/doi/10.1093/database/bav006.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1), pp.289–300. Available at: http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini and Y FDR.pdf%5Cnhttp://engr.case.edu/ray_soumya/mlrg/controlling_fdr_benjamini95.pdf.

Blake, J.A. et al., 2015. Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1), pp.D1049–D1056.

Blondel, V. & Guillaume, J., 2008. Fast unfolding of communities in large networks.

Journal of Statistical …, pp.1–12. Available at: http://iopscience.iop.org/1742-5468/2008/10/P10008 [Accessed July 13, 2014].

Borneman, J. et al., 2001. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics (Oxford, England)*, 17 Suppl 1, pp.S39-48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11472991.

Brückner, A. et al., 2009. Yeast Two-Hybrid , a Powerful Tool for Systems Biology. *International Journal of Molecular Sciences*, 10, pp.2763–2788.

Bryant, W.A., Sternberg, M.J. & Pinney, J.W., 2013. AMBIENT: Active Modules for Bipartite Networks - using high-throughput transcriptomic data to dissect metabolic response. *BMC Systems Biology*, 7(1), p.26. Available at: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-26.

Camon, E. et al., 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research*, 32, pp.D262–D266.

Carroll, S.Y. et al., 2012. Analysis of yeast endocytic site formation and maturation through a regulatory transition point. *Molecular Biology of the Cell*, 23, pp.657–668.

Cerami, E.G. et al., 2011. Pathway Commons , a web resource for biological pathway data. , 39(November 2010), pp.685–690.

Chatr-Aryamontri, A. et al., 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43(D1), pp.D470–D478.

Chavali, S. et al., 2010. Network properties of human disease genes with pleiotropic effects. *BMC systems biology*, 4(78), p.78. Available at:

http://www.biomedcentral.com/1752-0509/4/78.

Chen, B. et al., 2014. Identifying protein complexes and functional modules-from static PPI networks to dynamic PPI networks. *Briefings in Bioinformatics*, 15(2), pp.177–194.

Chen, J. & Yuan, B., 2006. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics (Oxford, England)*, 22(18), pp.2283–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16837529 [Accessed March 22, 2015].

Chen, X., Sun, R. & Yu, J., 2011. Approximating the double-cut-and-join distance between unsigned genomes. *BMC bioinformatics*, 12 Suppl 9(Suppl 9), p.S17. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3283313&tool=pmcentrez&rendertype=abstract.

Cherry, J.M. et al., 2012. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1), pp.700–705.

Chowbina, S.R. et al., 2009. HPD: an online integrated human pathway database enabling systems biology studies. *BMC bioinformatics*, 14(Suppl 11), p.S5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19811689%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3226194.

Copley, S., 2003. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*, 7(2), pp.265–272. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1367593103000322 [Accessed July 1, 2014].

Copley, S.D., 2012. Moonlighting is mainstream: paradigm adjustment required. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 34(7), pp.578–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22696112 [Accessed May 30, 2014].

Costanzo, M. et al., 2010. The genetic landscape of a cell. *Science (New York, N.Y.)*, 327(5964), pp.425–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20093466 [Accessed July 10, 2014].

Couto, F.M., Silva, M.J. & Coutinho, P.M., 2005. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. *CIKM*, pp.343–344. Available at: http://portal.acm.org/citation.cfm?id=1099554.1099658.

Croft, D. et al., 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(November 2013), pp.472–477.

Csardi, G. & Nepusz, T., 2006. The igraph software package for complex network research. *InterJournal*, Complex Sy, p.1695. Available at: http://igraph.org.

Cusick, M.E. et al., 2009. Literature-curated protein interaction datasets. *Nature Methods*, 6(1), pp.39–46. Available at: http://www.nature.com/doifinder/10.1038/nmeth.1284.

Denoyer, D. et al., 2015. Targeting copper in cancer therapy: “Copper That Cancer”. *Metallomics*, 7(11), pp.1459–1476. Available at: http://xlink.rsc.org/?DOI=C5MT00149H%5Cnpapers3://publication/doi/10.1039/c5mt00149h.

Din, O.S. & Woll, P.J., 2008. Treatment of Gastrointestinal Stromal Tumor: Focus on

Imatinib Mesylate. *Therapeutics and Clinical Risk Management*, 4(1), pp.149–162. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2503651&tool=pmcentrez&rendertype=abstract.

Doderer, M.S. et al., 2012. Pathway Distiller - multisource biological pathway consolidation. *BMC genomics*, 13 Suppl 6(Suppl 6), p.S18. Available at: http://www.biomedcentral.com/1471-2164/13/S6/S18.

Downward, J., 2003. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews*, 3(1), pp.11–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12509763.

Dudek, J., Rehling, P. & van der Laan, M., 2013. Mitochondrial protein import: Common principles and physiological networks. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1833(2), pp.274–285. Available at: http://dx.doi.org/10.1016/j.bbamcr.2012.05.028.

Dudley, A.M. et al., 2005. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular systems biology*, 1, p.2005.0001. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1681449&tool=pmcentrez&rendertype=abstract [Accessed June 26, 2014].

Dunham, W.H., Mullin, M. & Gingras, A., 2012. Affinity-purification coupled to mass spectrometry : Basic principles and strategies. *Proteomics*, 12, pp.1576–1590.

Dunn, S.L. et al., 2016. Gene expression changes in damaged osteoarthritic cartilage identify a signature of non-chondrogenic and mechanical responses. *Osteoarthritis and Cartilage*, 24(8), pp.1431–1440. Available at:

http://dx.doi.org/10.1016/j.joca.2016.03.007.

Dutkowski, J. et al., 2013. A gene ontology inferred from molecular networks. *Nature biotechnology*, 31(1), pp.38–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23242164 [Accessed November 8, 2013].

Fabregat, A. et al., 2016. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1), pp.D481–D487.

Firth, H. V. et al., 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4), pp.524–533. Available at: http://dx.doi.org/10.1016/j.ajhg.2009.03.010.

Furlong, L.I., 2013. Human diseases through the lens of network biology. *Trends in Genetics*, 29(3), pp.150–159. Available at: http://dx.doi.org/10.1016/j.tig.2012.11.004.

Gagiano, M., Bauer, F.F. & Pretorius, I.S., 2002. The sensing of nutritional status and the relationship to fillamentous growth in Saccharomyces cerevisiae. *FEMS Yeast Research*, 2.

Gavin, A.-C. et al., 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084), pp.631–636. Available at: http://www.nature.com/doifinder/10.1038/nature04532.

Geer, L.Y. et al., 2009. The NCBI BioSystems database. *Nucleic Acids Research*, 38(SUPPL.1), pp.492–496.

Gentleman, R., 2005. Visualizing and distances using GO. *URL http://www. bioconductor. org/docs/vignettes. …*, pp.1–5. Available at:

http://www.bioconductor.org/packages/release/bioc/vignettes/GOstats/inst/doc/GOvis.pdf.

Gerlach, J.Q., Sharma, S. & Leister, K.J., 2012. *A Tight-Knit Group : Protein Glycosylation , Endoplasmic Reticulum Stress and the Unfolded Protein Response* P. Agostinis & S. Afshin, eds., springer, Dordrecht.

Ghaemmaghami, S. et al., 2003. Global analysis of protein expression in yeast. *Nature*, 425(1997), pp.737–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14562106.

Ghamlouch, H., Nguyen-Khac, F. & Bernard, O.A., 2017. Chronic lymphocytic leukaemia genomics and the precision medicine era. *British Journal of Haematology*, pp.1–19. Available at: http://doi.wiley.com/10.1111/bjh.14719.

Gillis, J. & Pavlidis, P., 2011. The impact of multifunctional genes on "guilt by association" analysis. *PloS one*, 6(2), p.e17258. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041792&tool=pmcentrez&rendertype=abstract [Accessed June 3, 2014].

Glass, K. & Girvan, M., 2014. Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Scientific reports*, 4, p.4191. Available at: http://www.nature.com/srep/2014/140226/srep04191/full/srep04191.html.

Goh, K. et al., 2007. The human disease network. *PNAS*, 104(21), pp.8685–8690. Available at: http://www.pnas.org/content/104/21/8685.short [Accessed September 24, 2013].

Goudarzi, K.M. & Lindström, M.S., 2016. Role of ribosomal protein mutations in tumor development (Review). *International Journal of Oncology*, 48(4), pp.1313–1324.

Greenbaum, D. et al., 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, 4, p.117. Available at: http://www.biomedcentral.com/content/pdf/gb-2003-4-9-117.pdf.

Guan, Y. et al., 2008. A genomewide functional network for the laboratory mouse. *PLoS computational biology*, 4(9), p.e1000165. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2527685&tool=pmcentrez&rendertype=abstract [Accessed July 28, 2014].

Guo, X. et al., 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8), pp.967–973.

Guo, Z. et al., 2007. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, 23(16), pp.2121–2128.

Guzzi, P.H. et al., 2012. Semantic similarity analysis of protein data: Assessment with biological features and issues. *Briefings in Bioinformatics*, 13(5), pp.569–585.

Gygi, S.P. et al., 1999. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, 19(3), pp.1720–1730.

Hagberg, A.A., Schult, D.A. & Swart, P.J., 2008. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, (SciPy), pp.11–15.

Hakes, L. et al., 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology*, 8(10), p.R209. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-10-r209.

Hakes, L. et al., 2008. Protein-protein interaction networks and biology—what's the connection? *Nature Biotechnology*, 26(1), pp.69–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18183023%5Cnhttp://www.nature.com/doifinder/10.1038/nbt0108-69.

Han, J.J. et al., 2004. Evidence for dynamically organized modularity in the yeast protein – protein interaction network. , 430(July).

Hart, G.T., Ramani, A.K. & Marcotte, E.M., 2006. How complete are current yeast and human protein-interaction networks? *Genome biology*, 7(11), p.120. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794583&tool=pmcentrez&rendertype=abstract.

He, X. & Zhang, J., 2006a. Toward a molecular understanding of pleiotropy. *Genetics*, 173(4), pp.1885–91. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1569710&tool=pmcentrez&rendertype=abstract [Accessed June 27, 2014].

He, X. & Zhang, J., 2006b. Why do hubs tend to be essential in protein networks? *PLoS genetics*, 2(6), p.e88. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1473040&tool=pmcentrez&rendertype=abstract [Accessed December 16, 2013].

Hegde, S.R., Manimaran, P. & Mande, S.C., 2008. Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Computational Biology*, 4(11).

Hewett, M. et al., 2002. BiGCaT Bioinformatics\rPharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*, 30(1), pp.163–165. Available at:

http://www.bigcat.nl.

Howe, K., Bateman, a. & Durbin, R., 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, 18(11), pp.1546–1547. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/18.11.1546.

Hu, Z., Mellor, J., et al., 2007. Towards zoomable multidimensional maps of the cell. *Nature Biotechnology*, 25(5), pp.547–554. Available at: http://www.nature.com/doifinder/10.1038/nbt1304.

Hu, Z., Ng, D.M., et al., 2007. VisANT 3.0: New modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Research*, 35(SUPPL.2), pp.625–632.

Hu, Z. et al., 2009. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Research*, 37(May), pp.115–121.

Huang, C. et al., 2007. Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach. *Quality*, 4(1), pp.78–87.

Huang, D.W. et al., 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8(9), p.R183. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375021&tool=pmcentrez&rendertype=abstract.

Huang, D.W., Lempicki, R. a & Sherman, B.T., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), pp.44–57.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), pp.1–13.

Huttlin, E.L. et al., 2015. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2), pp.425–440.

Hyduke, D.R. & Palsson, B.Ø., 2010. Towards genome-scale signalling network reconstructions. *Nature reviews. Genetics*, 11(4), pp.297–307. Available at: http://dx.doi.org/10.1038/nrg2750.

Ideker, T. et al., 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18, pp.233–240.

Ideker, T. et al., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518), pp.929–934. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11340206.

Ideker, T. & Krogan, N.J., 2012. Differential network biology. *Molecular Systems Biology*, 8(565), pp.1–9. Available at: http://dx.doi.org/10.1038/msb.2011.99.

Ihmels, J. et al., 2002. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(august). Available at: http://www.nature.com/doifinder/10.1038/ng941.

Isaev, A., 2006. *Introduction to Mathematical Methods in Bioinformatics*, Springer-Verlag Berlin Heidelberg.

Ishitsuka, M., Akutsu, T. & Nacher, J.C., 2016. Critical controllability in proteome- wide protein interaction network integrating transcriptome. *Nature Publishing Group*, (April), pp.1–13. Available at: http://dx.doi.org/10.1038/srep23541.

Jain, S. & Bader, G.D., 2010. An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology. *BMC Bioinformatics*, 11(1), p.562. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-562.

Janeway, K.A. et al., 2011. Defects in succinate dehydrogenase in gastrointestinal stromal tumors lacking KIT and PDGFRA mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1), pp.314–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3017134&tool=pmcentrez&rendertype=abstract.

Janjić, V. & Pržulj, N., 2012a. Biological function through network topology: a survey of the human diseasome. *Briefings in functional genomics*, 11(6), pp.522–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22962330 [Accessed March 19, 2014].

Janjić, V. & Pržulj, N., 2012b. Biological function through network topology: a survey of the human diseasome. *Briefings in functional genomics*, 11(6), pp.522–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22962330 [Accessed March 19, 2014].

Janjić, V. & Pržulj, N., 2012c. The Core Diseasome. *Molecular bioSystems*, 8(10), pp.2614–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22820726 [Accessed March 19, 2014].

Janjić, V., Sharan, R. & Pržulj, N., 2014. Modelling the yeast interactome. *Scientific* *reports*, 4, p.4273. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3940977&tool=pmcentrez&rendertype=abstract [Accessed February 16, 2015].

Ji, J. et al., 2014. Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), pp.261–277.

Jiang, J.J. & Conrath, D.W., 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics*, (Rocling X), pp.19–33. Available at: http://arxiv.org/abs/cmp-lg/9709008.

Jiang, X. et al., 2015. Characterizing the Diverse Mutational Pathways Associated with R5-Tropic Maraviroc Resistance: HIV-1 That Uses the Drug-Bound CCR5 Coreceptor. *Journal of Virology*, 89(22), pp.11457–11472. Available at: http://jvi.asm.org/lookup/doi/10.1128/JVI.01384-15.

Johnsson, N. & Varshavsky, A., 1994. Split ubiquitin as a sensor of protein interactions in vivo. *PNAS*, 91(22), pp.10340–10344.

Jozefczak, M. et al., 2012. Glutathione Is a Key Player in Metal-Induced Oxidative Stress Defenses. *International journal of Molecular Sciences*, 13, pp.3145–3175.

Jules, M. et al., 2008. New insights into trehalose metabolism by Saccharomyces cerevisiae: NTH2 encodes a functional cytosolic trehalase, and deletion of TPS1 reveals Ath1p-dependent trehalose mobilization. *Applied and Environmental Microbiology*, 74(3), pp.605–614.

Kamburov, A. et al., 2009. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic acids research*, 37(Database issue), pp.D623-8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686562&tool=pmcentrez&rendertype=abstract [Accessed July 3, 2014].

Kamburov, A. et al., 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research*, 41(Database issue), pp.D793-800. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531102&tool=pmcentrez&rendertype=abstract [Accessed February 21, 2014].

Kanehisa, M. et al., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, 42(Database issue), pp.D199-205. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24214961 [Accessed March 22, 2014].

Kanehisa, M. et al., 2017. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), pp.D353–D361.

Kanehisa, M. et al., 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(SUPPL. 1), pp.480–484.

Karagoz, K. & Arga, K.Y., 2013. Assessment of high-confidence protein-protein interactome in yeast. *Computational Biology and Chemistry*, 45, pp.1–8. Available at: http://dx.doi.org/10.1016/j.compbiolchem.2013.03.002.

Kelder, T. et al., 2018. WikiPathways : building research communities on biological pathways. , 40(January), pp.1301–1307.

Kelley, R. & Ideker, T., 2005. Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology*, 23(5), pp.561–6. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2814446&tool=pmcentrez&rendertype=abstract [Accessed December 4, 2014].

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2), p.e1002375. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3285573&tool=pmcentrez&rendertype=abstract [Accessed May 23, 2014].

Klienberg, J. & Tardos, E., 2003. *Introduction To Algorithms*, Available at: http://books.google.com.br/books/about/Introduction_To_Algorithms.html?hl=pt-BR&id=NLngYyWFl_YC.

Köhler, S. et al., 2017. The human phenotype ontology in 2017. *Nucleic Acids Research*, 45(D1), pp.D865–D876.

Komurov, K. & White, M., 2007. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol*, 3(110), p.110. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17453049.

Kordalewski, D., 2013. New Greedy Heuristics For Set Cover and Set Packing. , (April), p.49. Available at: http://arxiv.org/abs/1305.3584.

Koutsogiannouli, E., Papavassiliou, A.G. & Papanikolaou, N.A., 2013. Complexity in cancer biology: is systems biology the answer? *Cancer Medicine*, 2(2), pp.164–177. Available at: http://doi.wiley.com/10.1002/cam4.62.

Kuchaiev, O. et al., 2009. Geometric de-noising of protein-protein interaction networks. *PLoS Computational Biology*, 5(8), pp.1–10.

Kunze, M. et al., 2006. A central role for the peroxisomal membrane in glyoxylate cycle function. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1763, pp.1441–1452.

Kutmon, M. et al., 2016. WikiPathways: Capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1), pp.D488–D494.

Lee, H.K. et al., 2004. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome research*, 14, pp.1085–1094.

Lee, I. et al., 2004. A Probabilistic Network of Yeast Genes. *Science*, 306(November), pp.1555–1558.

Lee, J. & Lee, J., 2013. Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction. *PloS one*, 8(4), p.e60372. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3618231&tool=pmcentrez&rendertype=abstract [Accessed March 19, 2014].

Lemke, J.R. et al., 2012. Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia*, 53(8), pp.1387–1398.

Li, M. et al., 2012. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. *BMC Bioinformatics*, 13(1), p.109.

Li, X.-L., Foo, C.-S. & See-Kiong, N., 2007. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. *Computational Systems Bioinformatics Conference*, 6, pp.157–168.

Liberzon, A. et al., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12), pp.1739–1740.

Lim, J. et al., 2006. A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell*, 125(4), pp.801–814.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. *Proceedings of ICML*, pp.296–304.

Liu, W. et al., 2015. Integrative analysis of human protein, function and disease networks. *Scientific Reports*, 5, p.14344. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4585831&tool=pmcentrez&rendertype=abstract.

Liu, Y.-Y., Slotine, J.-J. & Barabási, A.-L., 2011. Controllability of complex networks. *Nature*, 473(7346), pp.167–73. Available at: http://dx.doi.org/10.1038/nature10011.

Liu, Y. & Chance, M.R., 2013. Pathway analyses and understanding disease associations. *Current genetic medicine reports*, 1(4), pp.230–238. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24319650.

Lord, P.W. et al., 2003a. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics*, 19(10), pp.1275–1283.

Lord, P.W. et al., 2003b. Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 612, pp.601–612. Available at: http://www.cs.manchester.ac.uk/~stevensr/papers/psb-similarity-2003-paper.pdf%5Cnhttp://view.ncbi.nlm.nih.gov/pubmed/12603061.

Lu, H. et al., 2006. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, 345(1), pp.302–309.

Lu, S. & Lu, X., 2013. Using graph models to find transcription factor modules: the hitting set problem and an exact algorithm. *Algorithms for molecular biology : AMB*, 8(1), p.2. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3622577&tool=pmcentrez&rendertype=abstract.

Lutz, S., Lichter, J. & Liu, L., 2007. Exploiting temperature-dependent substrate promiscuity for nucleoside analogue activation by thymidine kinase from Thermotoga maritima. *Journal of the American Chemical Society*, 129(28), pp.8714–8715. Available at: http://pubs.acs.org/doi/abs/10.1021/ja0734391 [Accessed July 2, 2014].

Maier, T., Güell, M. & Serrano, L., 2009a. Correlation of mRNA and protein in complex biological samples. *FEBS Letters*, 583(24), pp.3966–3973. Available at: http://dx.doi.org/10.1016/j.febslet.2009.10.036.

Maier, T., Güell, M. & Serrano, L., 2009b. Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24), pp.3966–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19850042 [Accessed May 24, 2014].

Mathur, S. & Dinakarpandian, D., 2012. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, 45(2), pp.363–371. Available at: http://dx.doi.org/10.1016/j.jbi.2011.11.017.

McKusic, V.A., 2009. Online Mendelian Inheritance in Man, OMIM. *McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and*

National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

Milewski, S., Gabriel, I. & Olchowy, J., 2006. Enzymes of UDP-GlcNAc biosynthesis in yeast. , pp.1–14.

Miller, J. a et al., 2013. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome medicine*, 5(5), p.48. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3706780&tool=pmcentrez&rendertype=abstract.

Mistry, M. & Pavlidis, P., 2008. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9(1), p.327. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-327.

Mitra, K. et al., 2013. Integrative approaches for finding modular structure in biological networks. *Nature reviews. Genetics*, 14(10), pp.719–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24045689 [Accessed November 7, 2013].

Nacher, J.C. & Akutsu, T., 2012. Dominating scale-free networks with variable scaling exponent: Heterogeneous networks are not difficult to control. *New Journal of Physics*, 14.

Nacher, J.C. & Akutsu, T., 2013. Structural controllability of unidirectional bipartite networks. *Scientific reports*, 3, p.1647. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3622082&tool=pmcentrez&rendertype=abstract.

Newman, M.E.J., 2001. Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality. *Phys. Rev. E*, 64(1), p.16132. Available at:

http://link.aps.org/doi/10.1103/PhysRevE.64.016132.

Nishimura, D., 2001. A View From the Web. *Biotech Software & Internet Report*, 2(3), pp.117–120.

Niwa, T., 2007. Protein glutathionylation and oxidative stress. *Journal of chromatography B*, 855, pp.59–65.

Nobeli, I., Favia, A.D. & Thornton, J.M., 2009. Protein promiscuity and its implications for biotechnology. *Nature biotechnology*, 27(2), pp.157–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19204698 [Accessed March 19, 2014].

Ohmuro-Matsuyama, Y., Chung, C.-I. & Ueda, H., 2013. Demonstration of protein-fragment complementation assay using purified firefly luciferase fragments. *BMC Biotechnology*, 13(1), p.31. Available at: http://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-13-31.

Oldfield, C.J. et al., 2008. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC genomics*, 9 Suppl 1, p.S1. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2386051&tool=pmcentrez&rendertype=abstract [Accessed June 16, 2014].

Oliver, S., 2000. Guilt-by-association goes global. *Nature*, 403(6770), pp.601–603.

Park, J. et al., 2009. The impact of cellular networks on disease comorbidity. *Molecular systems biology*, 5(262), p.262. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683720&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2013].

Peña-castillo, L. et al., 2008. Open Access A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, 9, pp.1–

19.

Peng, G. et al., 2010. Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics*, 18(1), pp.111–117. Available at: http://www.nature.com/doifinder/10.1038/ejhg.2009.115.

Peng, W. et al., 2017. Predicting Protein Functions by Using Unbalanced Random Walk Algorithm on Three Biological Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(2), pp.360–369.

Pesquita, C. et al., 2009. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), p.e1000443.

Pesquita, C., Faria, D. & Bastos, H., 2007. Evaluating gobased semantic similarity measures. *Proc. 10th Annual Bio-*, 2007, pp.37–40.

Petri, V., Hayman, G.T., et al., 2014. Disease pathways at the Rat Genome Database Pathway Portal: genes in context-a network approach to understanding the molecular mechanisms of disease. *Human genomics*, 8(1), p.17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25265995%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4191248.

Petri, V., Jayaraman, P., et al., 2014. The pathway ontology - updates and applications. *Journal of biomedical semantics*, 5(1), p.7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3922094&tool=pmcentrez&rendertype=abstract.

Philpott, C.C., Klausner, R.D. & Rouault, T. a, 1994. The bifunctional iron-responsive element binding protein/cytosolic aconitase: the role of active-site residues in ligand binding and regulation. *Proceedings of the National Academy of Sciences of* *the United States of America*, 91(15), pp.7321–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=44391&tool=pmcentrez&rendertype=abstract.

Piatigorsky, J. & Wistow, G., 1989. Enzyme/Crystallins: Gene Sharing as an Evolutionary Strategy. *Cell*, 57, pp.197–199.

Pinkert, S., Schultz, J. & Reichardt, J., 2010. Protein interaction networks--more than mere modules. *PLoS computational biology*, 6(1), p.e1000659. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-76749084811&partnerID=tZOtx3y1.

du Plessis, L., Škunca, N. & Dessimoz, C., 2011. The what, where, how and why of gene ontology-A primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6), pp.723–735.

Promislow, D.E.L., 2004. Protein networks, pleiotropy and the evolution of senescence. *Proceedings. Biological sciences / The Royal Society*, 271(1545), pp.1225–34. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1691725&tool=pmcentrez&rendertype=abstract [Accessed June 3, 2014].

Przulj, N., Wigle, D. a & Jurisica, I., 2004. Functional topology in a network of protein interactions. *Bioinformatics (Oxford, England)*, 20(3), pp.340–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14960460 [Accessed May 25, 2014].

Przytycka, T.M., Singh, M. & Slonim, D.K., 2010. Toward the dynamic interactome: It’s about time. *Briefings in Bioinformatics*, 11(1), pp.15–29.

Raghav, K. & Overman, M.J., 2013. Small bowel adenocarcinomas--existing evidence and evolving paradigms. *Nature reviews. Clinical oncology*, 10(9), pp.534–44. Available at: http://dx.doi.org/10.1038/nrclinonc.2013.132.

Rath, A. et al., 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5), pp.803–808.

Ravasz, E. et al., 2002. Networks Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(August), pp.1551–1555.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Articial Intelligence Research*, 11(3398), pp.95–130.

Revelle, W., 2017. psych: Procedures for Psychological, Psychometric, and Personality Research.

Robinson, P.N. et al., 2008. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, 83(5), pp.610–615. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0002929708005351.

Rodrıguez, P. et al., 2005. Redefining the Facilitated Transport of Mannose in Human Cells : Absence of a Glucose-Insensitive , High-Affinity Facilitated Mannose Transport System †. *Biochemistry*, 44(1), pp.313–320.

Roling, W.F.M. & Head, I.M., 2005. *Molecular Microbial Ecology* A. M. Osborn & C. J. Smith, eds., Taylor and Francis.

Rowe, J.M., 2016. AML in 2016: Where we are now? *Best Practice and Research: Clinical Haematology*, 29(4), pp.315–319.

Ryan, C.J. et al., 2012. Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell*, 46(5), pp.691–704. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3380636&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2013].

Rzhetsky, A. et al., 2007. Probing genetic overlap among complex human phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(28), pp.11694–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17609372.

Saitou, N. & Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4), pp.406–425.

Schaefer, C.F. et al., 2009. PID: The pathway interaction database. *Nucleic Acids Research*, 37(SUPPL. 1), pp.674–679.

Scheerer, U. et al., 2010. Sulphur flux through the sulphate assimilation pathway is differently controlled by adenosine 5' -phosphosulphate reductase under stress and in transgenic poplar plants overexpressing g - ECS , SO , or APR. *Journal of experimental botany*, 61(2), pp.609–622.

Schwanhausser, B., 2011. Global quantification of mammalian gene expression control. *Nature*, 473, pp.337–342. Available at: http://dx.doi.org/10.1038/nature10098.

Schwikowski, B., Uetz, P. & Fields, S., 2000. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12), pp.1257–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11101803.

Segal, E. et al., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2), pp.166–176. Available at: http://www.nature.com/doifinder/10.1038/ng1165.

Sethi, N. et al., 2013. Juvenile myelomonocytic leukemia. *Indian J Hematol Blood Transfus*, 29(3), pp.164–166. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24426365.

Sevilla, J.L. et al., 2005. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4), pp.330–337.

Shannon, P. et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), pp.2498–2504.

Shapiro, S. & Wilk, B., 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4), pp.591–611.

Sharan, R., Ulitsky, I. & Shamir, R., 2007. Network-based prediction of protein function. *Molecular systems biology*, 3(88), p.88. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1847944&tool=pmcentrez&rendertype=abstract [Accessed May 28, 2014].

Sheehan, B. et al., 2008. A relation based measure of semantic similarity for Gene Ontology annotations. *BMC Bioinformatics*, 9(1), p.468. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-468.

Skiena, S., 2008. *The Algorithm Design Manual*, Springer-Verlag London Limited.

Slenter, D.N. et al., 2018. WikiPathways : a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46(January), pp.661–667.

Snider, J. et al., 2015. Fundamentals of protein interaction network mapping.

Molecular systems biology, 11(12), p.848. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26681426%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4704491.

Sokal, R.R. & Michener, C.D., 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28, pp.1409–1438.

del Sol, A. et al., 2010. Diseases as network perturbations. *Current Opinion in Biotechnology*, 21(4), pp.566–571. Available at: http://dx.doi.org/10.1016/j.copbio.2010.07.010.

Song, J. & Singh, M., 2009. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics (Oxford, England)*, 25(23), pp.3143–50. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3167697&tool=pmcentrez&rendertype=abstract [Accessed June 26, 2014].

Song, J.J. et al., 2002. Role of Glutaredoxin in Metabolic Oxidative Stress. *Jpn J Pharmacol*, 277(48), pp.46566–46575.

Song, L. & Florea, L., 2013. CLASS: constrained transcript assembly of RNA-seq reads. *BMC bioinformatics*, 14 Suppl 5(Suppl 5), p.S14. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3622639&tool=pmcentrez&rendertype=abstract.

Spirin, V. & Mirny, L. a, 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21), pp.12123–12128.

Sprinzak, E., Sattath, S. & Margalit, H., 2003. How reliable are experimental protein-

protein interaction data? *Journal of Molecular Biology*, 327(5), pp.919–923.

Srihari, S. & Leong, H.W., 2012. Temporal dynamics of protein complexes in PPI Networks: a case study using yeast cell cycle dynamics. *BMC Bioinformatics*, 13(Suppl 17), p.S16. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3521212&tool=pmcentrez&rendertype=abstract.

Stanley, P., 2011. Golgi Glycosylation. *Cold Spring Harbour Perspective Biology*, 3, pp.1–13.

Stark, C. et al., 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue), pp.D535–D539.

Stoney, R.A. et al., 2015. Disentangling the multigenic and pleiotropic nature of molecular function. *BMC Systems Biology*, 9(Suppl 6), p.S3. Available at: http://www.biomedcentral.com/1752-0509/9/S6/S3/abstract.

Stuart, J.M. et al., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643), pp.249–255. Available at: http://www.sciencemag.org/cgi/content/abstract/302/5643/249%5Cnpapers://d2952c50-9509-4ba2-9a03-22fbc04267d4/Paper/p908.

Suthram, S. et al., 2010. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6(2), pp.1–10.

Takabayashi, H. et al., 2014. Anti-inflammatory activity of bone morphogenetic protein signaling pathways in stomachs of mice. *Gastroenterology*, 147(2), pp.396–406.

Tang, X. et al., 2011. A comparison of the functional modules identified from time course and static PPI network data. *BMC Bioinformatics*, 12, p.339. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21849017&retmode=ref&cmd=prlinks%5Cnpapers3://publication/doi/10.1186/1471-2105-12-339.

Tarassov, K. et al., 2008. An in Vivo Map of the Yeast Protein Interactome. *Science*, 320(5882), pp.1465–1470. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1153878.

Tarca, A.L., Bhatti, G. & Romero, R., 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE*, 8(11).

Taylor, I.W. et al., 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2), pp.199–204. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19182785.

Tordjman, R. et al., 2002. A neuronal receptor, neuropilin-1, is essential for the initiation of the primary immune response. *Nature Immunology*, 3(5), pp.477–482. Available at: http://www.nature.com/doifinder/10.1038/ni789.

Tsai, C.J., Ma, B. & Nussinov, R., 2009. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends in Biochemical Sciences*, 34(12), pp.594–600.

Vader, P., Breakefield, X.O. & Wood, M.J.., 2014. Extracellular vesicles: Emerging targets for cancer therapy. *Trends in Molecular Medicine*, 20(7), pp.385–393.

Valencia, C.A. et al., 2013. Comprehensive Mutation Analysis for Congenital Muscular Dystrophy: A Clinical PCR-Based Enrichment and Next-Generation Sequencing Panel. *PLoS ONE*, 8(1), pp.1–11.

Velásquez, R. & Melo, M.T., 2005. A Set Packing Approach for Scheduling Elective Surgical Procedures. *Operations Research Proceedings 2005*, pp.425–430.

Vidal, M., Cusick, M.E. & Barabási, A.-L., 2011. Interactome networks and human disease. *Cell*, 144(6), pp.986–98. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102045&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2013].

Vivar, J.C. et al., 2013. Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. *Omics*, 17(8), pp.414–422. Available at: http://online.liebertpub.com/doi/pdfplus/10.1089/omi.2012.0083.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4), pp.227–232. Available at: http://dx.doi.org/10.1038/nrg3185.

Wagner, G.P., Pavlicev, M. & Cheverud, J.M., 2007. The road to modularity. *Nature reviews. Genetics*, 8(12), pp.921–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18007649 [Accessed May 26, 2014].

Wagner, G.P. & Zhang, J., 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature reviews. Genetics*, 12(3), pp.204–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21331091 [Accessed May 28, 2014].

Wang, J. et al., 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10), pp.1274–1281.

Wang, J. et al., 2013. An effective method for refining predicted protein complexes based on protein activity and the mechanism of protein complex formation. *BMC Systems Biology*, 7, p.12.

Wang, J. et al., 2013. Construction and application of dynamic protein interaction network based on time course gene expression data. *Proteomics*, 13(2), pp.301–312.

Wang, J. et al., 2017. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *The American Society for Biochemistry and Molecular Biology*, 16(1), pp.121–134.

Wang, J. et al., 2010. Recent advances in clustering methods for protein interaction networks. *BMC genomics*, 11 Suppl 3(Suppl 3), p.S10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2999340&tool=pmcentrez&rendertype=abstract [Accessed October 26, 2014].

Wang, J.-T. et al., 2005. Detection of Epstein-Barr virus BGLF4 protein kinase in virus replication compartments and virus particles. *The Journal of general virology*, 86(Pt 12), pp.3215–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16298966 [Accessed July 2, 2014].

Wang, X., Gulbahce, N. & Yu, H., 2011. Network-based methods for human disease gene prediction. *Briefings in functional genomics*, 10(5), pp.280–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21764832 [Accessed November 6, 2013].

Warsow, G. et al., 2010. ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Systems Biology*, 4(1), p.164. Available at: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-4-164.

Winterbach, W. et al., 2013. Topology of molecular interaction networks. *BMC systems biology*, 7, p.90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24041013.

Wu, G., Feng, X. & Stein, L., 2010. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(53).

Wuchty, S., 2014. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19), pp.7156–7160.

Xu, T., Du, L. & Zhou, Y., 2008. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9(1), p.472. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-472.

Xue, Y. et al., 2015. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: Single-gene, gene panel, or exome/genome sequencing. *Genetics in Medicine*, 17(6), pp.444–451.

Yon Rhee, S. et al., 2008. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7), pp.509–515. Available at: http://www.nature.com/doifinder/10.1038/nrg2363.

Yook, S.-H., Oltvai, Z.N. & Barabási, A.-L., 2004. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4), pp.928–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15048975 [Accessed December 12, 2013].

Young, M.D. et al., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2), p.R14. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872874&tool=pmcentrez&rendertype=abstract.

Yu, G. et al., 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5), pp.284–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3339379&tool=pmcentrez&rendertype=abstract.

Yu, G. et al., 2010. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7), pp.976–978.

Yu, J.X. et al., 2007. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC cancer*, 7(1), p.182. Available at: http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-7-182.

Yu, N. et al., 2012. hiPathDB: A human-integrated pathway database with facile visualization. *Nucleic Acids Research*, 40(D1), pp.797–802.

Yurimoto, H., Kato, N. & Sakai, Y., 2005. Assimilation, dissimilation, and detoxification of formaldehyde, a central metabolic intermediate of methylotrophic metabolism. *The Chemical Record*, 5(6), pp.367–375.

Zhang, S., Wang, R.-S. & Zhang, X.-S., 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1), pp.483–490.

Zhou, X. et al., 2014. Human symptoms–disease network. *Nature communications*, 5(4212).

Zhu, C. et al., 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research*, 19(4), pp.556–566.

Zhu, J. et al., 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 40(7), pp.854–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2573859&tool=pmcentrez&rendertype=abstract [Accessed May 29, 2014].

Appendix A

Supplementary materials for Disentangling the multigenic and pleiotropic nature of function

Supplementary Table 1 GO annotations considered too frequent to be informative (>50% of annotations) and removed from the data set.

GO term	**% Genes**
Biological process	100
cellular process	93
metabolic process	80
cellular metabolic process	77
primary metabolic process	77
organic substance metabolic process	76

single organism process	73
single-organism cellular process	66

Supplementary Data 1 The GO terms and associated genes within each pathway. The pathways are separated into those with pleiotropic terms, listed first, and those without pleiotropic terms, listed second. Pleiotropic GO terms are listed following the header "PLEIOTROPIC FUNCTIONS", other GO terms are listed after the header "ORIGINAL FUNCTIONS". If there are any genes within a pathway that are not covered by GO terms those are also listed following the header "UNCOVERED GENES"

pleiotropic pathways

acetaldehyde biosynthesis ORIGINAL FUNCTIONSamino acid catabolic process to alcohol via Ehrlich pathway:ADH3,ADH2,ADH1,ADH5,ADH4,PDC1,PDC5,PDC6 PLIEOTROPIC FUNCTIONS fermentation:ADH3,ADH1,ADH5,ADH4,PDC1,PDC5

dolichyl glucosyl phosphate biosynthesis ORIGINAL FUNCTIONSUDP-glucose metabolic process:UGP1,PGM2,PGM1 PLIEOTROPIC FUNCTIONS nucleotide-sugar metabolic process:UGP1,PGM2,PGM1

ethanol degradation IV ORIGINAL FUNCTIONSacetate metabolic process:ALD6,ALD4,ACS1 thioester biosynthetic process:ACS1,ACS2 response to reactive oxygen species:CTT1,CTA1 PLIEOTROPIC FUNCTIONS acetyl-CoA biosynthetic process:ACS1,ACS2 internal protein amino acid acetylation:ACS1,ACS2 UNCOVERED GENES ALD3

fatty AcylCoA Biosynthesis ORIGINAL FUNCTIONSfatty acid metabolic process:ACC1,FAA3,HFA1,TSC13 organic acid transport:CTP1,FAA4,FAA1 PLIEOTROPIC FUNCTIONS long-chain fatty-acyl-CoA metabolic process:FAA4,FAA1

fatty acid betaoxidation I ORIGINAL FUNCTIONSfatty acid metabolic process:FAT1,POT1,ECI1,FOX2,FAA2,FAA3 long-chain fatty acid transport:FAA4,FAT1,FAA1 PLIEOTROPIC FUNCTIONS fatty-acyl-CoA metabolic process:FAA4,FAA1

glucose transport ORIGINAL FUNCTIONSglucose import:HXK1,GLK1,HXK2 single-organism transport:YBR241C,HXK1,GLK1,VPS73,HXK2 PLIEOTROPIC FUNCTIONS mannose metabolic process:HXK1,GLK1,HXK2 UNCOVERED GENES EMI2

glutamate degradation III ORIGINAL FUNCTIONSsuccinyl-CoA metabolic process:LSC2,LSC1 PLIEOTROPIC FUNCTIONS thioester metabolic process:LSC2,LSC1 UNCOVERED GENES KGD2,GDH2

glutathioneglutaredoxin system ORIGINAL FUNCTIONScellular response to oxidative stress:GRX3,GRX2,GRX1,GRX5,GRX4,GLR1 PLIEOTROPIC FUNCTIONS protein glutathionylation:GRX1,GLR1transition metal ion homeostasis:GRX3,GRX4 negative regulation of binding:GRX3,GRX4

glycogen breakdown glycogenolysis ORIGINAL FUNCTIONSsingle-organism catabolic process:GPH1,PGM2,PGM1,GDB1,PGM3 energy reserve metabolic process:GPH1,GDB1,PGM1,PGM2,GLG1,GLG2 PLIEOTROPIC FUNCTIONS glucose 1-phosphate metabolic process:PGM1,PGM2

lactose degradation ORIGINAL FUNCTIONSgalactose catabolic process:GAL10,GAL7,GAL1,PGM2,PGM1 UDP-glucose metabolic process:PGM1,PGM2,UGP1 PLIEOTROPIC FUNCTIONS nucleotide-sugar metabolic process:PGM1,PGM2,UGP1

mannose degradation ORIGINAL FUNCTIONSfructose import:HXK1,HXK2 PLIEOTROPIC

FUNCTIONS fructose metabolic process:HXK1,HXK2 UNCOVERED GENES PMI40

methylglyoxal catabolism ORIGINAL FUNCTIONSlactate metabolic process:GLO1,GLO2,DLD3,GLO4single-organism metabolic process:GLO1,GLO2,DLD3,GLO4,DLD1 PLIEOTROPIC FUNCTIONS methylglyoxal metabolic process:GLO1,GLO2,GLO4 cellular aldehyde metabolic process:GLO1,GLO2,GLO4

other types of Oglycan biosynthesis ORIGINAL FUNCTIONSprotein O-linked glycosylation:KRE2,PMT6,MNN1,KTR3,KTR1,PMT5,PMT4,MNT3,MNT2,PMT1,PMT3,PMT2 PLIEOTROPIC FUNCTIONS regulation of response to stress:PMT4,PMT1,PMT2 protein exit from endoplasmic reticulum:PMT1,PMT2

phase 1 Functionalization of compounds ORIGINAL FUNCTIONSacetate metabolic process:ALD6,ALD5,ALD4,ACS1 thioester biosynthetic process:ACS1,ACS2 PLIEOTROPIC FUNCTIONS acetyl-CoA biosynthetic process:ACS1,ACS2 internal protein amino acid acetylation:ACS1,ACS2 UNCOVERED GENES ERG11

pyruvate dehydrogenase complex ORIGINAL FUNCTIONSacetyl-CoA biosynthetic process from pyruvate:PDA1,LAT1,PDB1 pyruvate metabolic process:PDA1,LAT1,LPD1,PDB1 PLIEOTROPIC FUNCTIONS thioester metabolic process:PDA1,LAT1,PDB1

sucrose degradation ORIGINAL FUNCTIONScellular carbohydrate catabolic process:SUC2,HXK1,HXK2 PLIEOTROPIC FUNCTIONS fructose import:HXK1,HXK2

sulfur relay system ORIGINAL FUNCTIONStRNA thio-modification:SLM3,NCS6,URM1,NCS2,NFS1,UBA4,TUM1 cellular response to oxidative stress:UBA4,URM1,AHP1 PLIEOTROPIC FUNCTIONS adenosine metabolic process:NFS1,SLM3

superoxide radicals degradation ORIGINAL FUNCTIONSresponse to reactive oxygen species:CTT1,SOD1,SOD2,CTA1 PLIEOTROPIC FUNCTIONS age-dependent general metabolic decline:SOD1,SOD2,CTA1

synthesis of UDPNacetylglucosamine ORIGINAL FUNCTIONSamino sugar biosynthetic process:PCM1,QRI1,GNA1,GFA1 PLIEOTROPIC FUNCTIONS nucleotide-sugar biosynthetic process:QRI1,GNA1 fungal-type cell wall biogenesis:PCM1,GFA1

thioredoxin system ORIGINAL FUNCTIONScellular response to oxidative stress:TRX2,TRX3,TRX1,TRR1,TRR2 PLIEOTROPIC FUNCTIONS protein deglutathionylation:TRX2,TRX1

transport to the Golgi and subsequent modification ORIGINAL FUNCTIONSCOPII-coated vesicle budding:SEC13,SFB2,SEC24,SFB3,SEC31 regulation of vesicle targeting, to, from or within Golgi:SAR1,SEC23 PLIEOTROPIC FUNCTIONS purine ribonucleoside triphosphate catabolic process:SAR1,SEC23

trehalose Anabolism ORIGINAL FUNCTIONSoligosaccharide biosynthetic process:PGM1,PGM2,TPS1,TPS3,TPS2,TSL1,UGP1 PLIEOTROPIC FUNCTIONS nucleotide-sugar metabolic process:PGM1,PGM2,UGP1

trehalose degradation II trehalase ORIGINAL FUNCTIONScellular carbohydrate catabolic process:NTH2,NTH1,ATH1,HXK1,GLK1,HXK2 PLIEOTROPIC FUNCTIONS fructose import:HXK1,HXK2 UNCOVERED GENES EMI2

unwinding of DNA ORIGINAL FUNCTIONSdouble-strand break repair via break-induced replication:SLD5,CDC45,MCM7,MCM6,MCM5,MCM4,MCM3,MCM2,PSF1,PSF2 PLIEOTROPIC FUNCTIONS DNA duplex unwinding:MCM7,MCM6,MCM4

xylulosemonophosphate cycle ORIGINAL FUNCTIONSsingle-organism carbohydrate metabolic process:DAK1,DAK2,FBP1,TKL1,FBA1 single-organism carbohydrate catabolic process:DAK1,DAK2,FBA1,TKL1 glycerol catabolic process:DAK1,DAK2 PLIEOTROPIC FUNCTIONS response to toxin:DAK1,DAK2

all other pathways

2ketoglutarate dehydrogenase complex ORIGINAL FUNCTIONS GO:0000949:PDC1,PDC5,PDC6 GO:0006103:KGD2,LPD1 GO:1901606:PDC1,PDC5,LPD1,PDC6

5aminoimidazole ribonucleotide biosynthesis I ORIGINAL FUNCTIONS GO:0006164:ADE4,ADE6,ADE8

aBC transporters ORIGINAL FUNCTIONSGO:0065008:STE6,ATM1 GO:0015718:PXA2,PXA1 GO:0071702:STE6,PXA2,PXA1 UNCOVERED GENES PDR5

aBCfamily proteins mediated transport ORIGINAL FUNCTIONS GO:0015849:YCF1,PXA2,BPT1 GO:0055085:ADP1,PXA2,YCF1

UNCOVERED GENES ATM1

adaptive Immune System ORIGINAL FUNCTIONS

GO:0006468:PKC1,SKM1,STE20,CLA4,PKH3,PKH2,PKH1

GO:0023052:PKC1,STE20,PKH3,PKH2,PKH1,CDC42

adenosine deoxyribonucleotides ide novoi biosynthesis I ORIGINAL FUNCTIONS

GO:0080058:TRX2,TRX1 GO:0009263:RNR2,RNR1,RNR4

GO:0009165:RNR2,YNK1,RNR1,RNR4GO:0034599:TRX2,TRX3,TRX1

adenosine ribonucleotides ide novoi biosynthesis ORIGINAL FUNCTIONS

GO:0033615:ATP14,ATP11,ATP12,ATP19,ATP10

GO:0015992:TIM11,ATP8,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,VPH1,ATP20,ATP14,ATP19,ATP17,ATP16,ATP15,ATP4 GO:0044260:VMA9,TFP1,ADK1

GO:0009206:TIM11,ATP8,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,ATP4,ATP20,ATP19,ATP17,ATP16,ATP15,ATP14

GO:0009127:TIM11,ATP8,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,ADK1,ATP20,ATP14,ADE13,ATP19,ATP17,ATP16,ATP15,ATP4

GO:0044765:TIM11,ATP8,VMA6,ATP4,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,VPH1,ATP20,VMA2,ATP19,VMA9,ATP17,ATP16,ATP15,ATP14,CUP5

GO:0009117:TIM11,ATP8,ATP4,ATP3,ATP2,ATP1,OLI1,ADK2,ATP6,ATP5,ADK1,ATP7,ATP20,ADE12,ADE13,ATP19,ATP17,ATP16,ATP15,ATP14

GO:0006164:TIM11,ATP8,ATP4,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,ADK1,ATP20,ADE12,ADE13,ATP19,ATP17,ATP16,ATP15,ATP14

GO:0055085:TIM11,ATP8,VMA2,ATP3,ATP2,ATP1,OLI1,ATP7,ATP6,ATP5,ATP4,ATP20,ATP19,VMA9,ATP17,ATP16,ATP15,ATP14 aerobic respiration linear view ORIGINAL FUNCTIONS

GO:0033108:COX20,COX23,COX12,COX16,COX19

GO:0044710:COX8,COX9,YJL045W,COX5A,COX5B,COX20,SDH1,COX3,SDH3,SDH2,COX6,SDH4,COX4,COX7,COX2,COX15,COX10,COX11,COX13,NDI1,COX1

GO:0045333:COX8,COX9,YJL045W,COX5A,COX5B,COX20,SDH1,COX3,SDH3,SDH2,COX6,SDH4,COX4,COX7,COX2,COX11,COX13,NDI1,COX1

GO:0007005:COX19,SDH3,COX18,COX20,COX23,COX12,COX13,COX16

alphaLinolenic acid metabolism ORIGINAL FUNCTIONS

GO:0016042:POX1,POT1,TGL4

amino acid and oligopeptide SLC transporters ORIGINAL FUNCTIONS

GO:0006810:ESBP6,PTR2,DIC1,AVT1,MCH2,MCH4,MCH5

amino acid synthesis and interconversion transamination ORIGINAL FUNCTIONS

GO:0046394:AAT2,SER2,SER1,PRO3,PRO2,GLN1,BNA3,ALT1

GO:1901605:AAT2,SER2,SER1,CAR2,PRO3,PRO2,GLN1,ALT1

apoptosis ORIGINAL FUNCTIONSGO:0048308:STE20,CLA4,DNM1

GO:0006122:CYC7,CYC1 GO:0035376:STE20,CLA4,SKM1

arachidonic acid metabolism ORIGINAL FUNCTIONS

GO:0042221:ECM38,HYR1,GRE3,YCF1,GPX1,GPX2 GO:0015723:YCF1,BPT1

UNCOVERED GENES LAP2,SBA1,TGL4

arginine biosynthesis ORIGINAL FUNCTIONS

GO:1901605:CPA1,URA2,CPA2,ARG1,ARG2,ARG3,ARG4,ARG7

GO:0006526:CPA2,ARG1,CPA1,ARG4,ARG3

GO:0009064:CPA1,URA2,CPA2,ARG1,ARG3,ARG4 UNCOVERED GENES

ARG5,6

arginine degradation ORIGINAL FUNCTIONSGO:0006591:CAR2,CAR1

GO:0009064:PUT1,CAR1,PRO3

arginine degradation I arginase pathway ORIGINAL FUNCTIONS

GO:0006591:CAR2,CAR1 GO:0009065:CAR1,PUT2

asparagine biosynthesis ORIGINAL FUNCTIONSGO:0006529:ASN2,ASN1

GO:0009067:AAT2,AAT1,ASN2,ASN1

asparagine degradation ORIGINAL FUNCTIONS

GO:0009066:AAT2,AAT1,ASP1,ASP3-4

aspartate Biosynthesis ORIGINAL FUNCTIONSGO:0019319:PYC1,PYC2

GO:0006532:AAT2,AAT1

aspartate degradation II ORIGINAL FUNCTIONSGO:0006531:AAT2,AAT1

GO:0001300:AAT1,MDH1 UNCOVERED GENES MDH3,MDH2

assembly of the ORC complex at the origin of replication ORIGINAL FUNCTIONS

GO:0030466:ORC4,ORC5,ORC2,ORC1

basal transcription factors ORIGINAL FUNCTIONS

GO:0016070:CCL1,FAP7,TAF13,TAF12,TAF11,TAF10,RAD3,TAF14,TFA1,TFA2,TAF7,SPT15,TFG2,TFG1,KIN28,SSL2,TAF6,TAF5,SSL1,TAF3,TAF2,TAF1,TAF9,TAF8,TFB1,TFB2,TFB3,TFB4,TFB5,SUA7,TOA1,TOA2

GO:0006366:CCL1,TAF13,TAF12,TAF11,TAF10,RAD3,TAF14,TFA1,TFA2,TAF7,SPT15,TFG2,TFG1,KIN28,SSL2,TAF6,TAF5,SSL1,TAF3,TAF2,TAF1,TAF9,TAF8,TFB1,TFB2,TFB3,TFB4,TFB5,SUA7,TOA1,TOA2

baseExcision Repair AP Site Formation ORIGINAL FUNCTIONS

GO:0006285:NTG2,NTG1,OGG1

betaAlanine metabolism ORIGINAL FUNCTIONS

GO:0015939:SPE4,FMS1,PAN6,SPE3

GO:0019752:EHD3,PAN6,ALD6,ALD5,ALD4,ALD3,SPE3,UGA1,ALD2,GAD1,SPE4,FMS1 GO:0006595:SPE4,FMS1,ALD3,SPE3,ALD2

bile acid and bile salt metabolism ORIGINAL FUNCTIONS

GO:0044712:FOX2,TES1,GRE3 GO:0006631:FAT1,FOX2,TES1

GO:0044255:FAT1,FOX2,SUR2,TES1

bile salt and organic anion SLC transporters ORIGINAL FUNCTIONS

GO:0006810:ESBP6,MCH5,YHR032W,MCH2,MCH4,YDR338C

biosynthesis of unsaturated fatty acids ORIGINAL FUNCTIONS

GO:0006631:TSC13,ELO1,IFA38,SUR4,FEN1,POX1,TES1,OLE1

GO:0044255:PHS1,ELO1,OLE1,IFA38,SUR4,FEN1,TSC13,POX1,TES1

biotin metabolism ORIGINAL FUNCTIONSGO:0006768:BIO2,BIO3,BIO4

GO:0072330:BIO2,BIO3,CEM1,BIO4 UNCOVERED GENES BPL1

biotin transport and metabolism ORIGINAL FUNCTIONSGO:0006633:ACC1,HFA1

GO:0019319:PYC1,PYC2 UNCOVERED GENES BPL1

carnitine shuttle ORIGINAL FUNCTIONSGO:0006577:YAT2,CAT2,YAT1

cell Cycle CheckpointsORIGINAL FUNCTIONS

GO:0065007:TEL1,CDC28,CHK1,CLB3,CLB2,CLB1,CLB6,CLB5,CLB4

GO:0000079:CLB3,CLB2,CLB1,CLB6,CLB5,CLB4

GO:0022402:CDC28,CHK1,CLB3,CLB2,CLB1,CLB6,CLB5,CLB4

GO:0010696:CDC28,CLB3,CLB2,CLB1,CLB5,CLB4 chaperoninmediated protein folding ORIGINAL FUNCTIONS

GO:0006457:CCT8,CCT2,CCT3,CCT6,CCT7,CCT4,CCT5,TCP1

choline biosynthesis ORIGINAL FUNCTIONSGO:0006644:CPT1,PCT1,SPO14,EPT1

GO:0006656:CPT1,PCT1,EPT1

citric acid cycle TCA cycle ORIGINAL FUNCTIONSGO:0006104:LSC2,LSC1

GO:0045333:YJL045W,SDH1,SDH3,SDH2,ACO1,FUM1

GO:0006103:KGD2,LPD1

citrulline biosynthesis ORIGINAL FUNCTIONS

GO:1901605:ARG3,PUT1,CAR2,CAR1,PRO2,PRO1

GO:0009064:CAR1,ARG3,PUT1,PRO2,PRO1

coenzyme A biosynthesis ORIGINAL FUNCTIONS

GO:0015936:VHS3,CAB2,CAB3,SIS2,CAB4,CAB5 UNCOVERED GENES LEU5

colanic Acid Building Blocks Biosynthesis ORIGINAL FUNCTIONS

GO:0044723:GAL7,PMI40,GAL1,GAL10,UGP1

GO:0033499:GAL7,GAL1,GAL10 UNCOVERED GENES SEC53

cyanoamino acid metabolism ORIGINAL FUNCTIONSGO:0006730:SHM2,SHM1

GO:0006530:ASP1,ASP3-4 GO:0043603:ECM38,ASP3-4,ASP1

cysteine biosynthesis IV fungi ORIGINAL FUNCTIONS

GO:0000096:MET2,MET17,STR3,CYS4,IRC7,CYS3

cytoplasmic Ribosomal Proteins ORIGINAL FUNCTIONS

GO:0002181:RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL41A,RPL2A,RPL6A,RPL6B,RPL31A,RPL33B,RPL33A

GO:0042254:RPS13,RPS21A,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,RPL40A,RPS9A,RPL5,RPL3,RPL35B,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL8A,RPL11B,RPS9B,RPS6A,RPS0A,RPS0B,RPL8B,RPS14B,RPS27A,RPS16B,RPS20,RPS26A,RLP7,RPL34A,RPS26B,RPS24B,RPP0,RPL10,RPS10A,RPS8B,RPL6A,RPL6B,RPS28B,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B,RPS14A

GO:0006412:RPP1A,RPP1B,RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPP2B,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP2A,RPP0,RPL10,RPL41A,RPL2A,RPL6A,RPL6B,RPL31A,RPL33B,RPL15A,RPL33A

cytoplasmic tRNA Synthetases ORIGINAL FUNCTIONS

GO:0043038:DPS1,TYS1,CDC60,GRS2,YDR341C,GUS1,MES1,KRS1,FRS1,SES1,FRS2,ALA1,YNL247W,ILS1,THS1,DED81,GLN4

cytosolic tRNA aminoacylation ORIGINAL FUNCTIONS

GO:0043038:SES1,DED81,GRS2,GRS1,FRS1,FRS2,ALA1,THS1,VAS1,YNL247W

UNCOVERED GENES PPA2

dAP12 signaling ORIGINAL FUNCTIONS

GO:0023052:PKC1,FUS3,TPK3,TPK1,KSS1,TPK2,PKH3,PKH2,PKH1

GO:0050794:PKC1,FUS3,TPK3,BCY1,TPK1,KSS1,TPK2,PKH3,PKH2,PKH1

dTMP ide novoi biosynthesis mitochondrial ORIGINAL FUNCTIONS

GO:0006730:SHM2,SHM1 UNCOVERED GENES DFR1,CDC21

de Novo Biosyn of Pyrimidine Deoxyribonucleotides ORIGINAL FUNCTIONS

GO:0009165:CDC21,CDC8,YNK1,RNR2,RNR3,RNR1,RNR4

GO:0009262:DUT1,CDC21,CDC8,RNR2,RNR3,RNR1,RNR4 de Novo NAD Biosynthesis ORIGINAL FUNCTIONS

GO:0009435:NMA1,NMA2,BNA6,QNS1,BNA5,BNA4,BNA2,BNA1

GO:0019438:NMA1,NMA2,BNA6,QNS1,BNA5,BNA4,BNA3,BNA2,BNA1

deadenylationdependent mRNA decay ORIGINAL FUNCTIONS

GO:0016071:LSM5,LSM6,LSM7,DCS2,LSM1,LSM2,DCS1,KEM1,LSM3

dehydroDarabinono14lactone biosynthesis ORIGINAL FUNCTIONS

GO:0070485:ARA2,ALO1 UNCOVERED GENES ARA1

dimyoiinositol 134trisphosphate biosynthesis ORIGINAL FUNCTIONS

GO:0046488:INP54,ARG82,INP53,INP52

dimyoiinositol 145trisphosphate biosynthesis ORIGINAL FUNCTIONS

GO:0006644:LSB6,STT4,MSS4,PLC1,PIK1,PIS1

GO:0046488:PIS1,STT4,MSS4,PIK1,LSB6 GO:0046854:STT4,MSS4,PIK1,LSB6

diphthamide biosynthesis ORIGINAL FUNCTIONS

GO:0017183:YLR143W,DPH2,DPH1,DPH5,KTI11

dolichyl phosphate Dmannose biosynthesis ORIGINAL FUNCTIONS

GO:0043413:DPM1,PMI40 UNCOVERED GENES SEC53

dolichyldiphosphooligosaccharide biosynthesis ORIGINAL FUNCTIONS

GO:1901137:ALG8,ALG9,ALG2,ALG3,ALG1,ALG6,KRE2,ALG7,KTR3,KTR2,KTR1,DIE2,KTR7,KTR6,KTR4,ALG11,ALG12,ALG13,ALG14,YUR1,DPM1

GO:0070085:ALG8,ALG9,DIE2,ALG3,ALG1,ALG6,KRE2,ALG7,KTR3,KTR2,KTR1,KTR7,KTR6,KTR4,ALG11,ALG12,YUR1,DPM1

dual incision reaction in GGNER ORIGINAL FUNCTIONS

GO:0090304:SSL2,SSL1,TFB4,RFA1,RAD14,RSE1,KIN28,TFB3,RAD10,TFB1,TFB2,R

AD4,CCL1,RAD2,RAD1,RAD3,RFA2

GO:0036211:SSL2,SSL1,RAD23,RFA1,KIN28,TFB1,TFB2,TFB3,CCL1,TFB4,RAD3,RFA2 GO:0070816:SSL2,RAD3,SSL1,KIN28,TFB1,TFB2,TFB3,CCL1,TFB4

GO:0006289:SSL2,SSL1,RFA2,RFA1,RAD14,TFB3,RAD10,TFB1,TFB2,RAD4,RAD3,RAD2,RAD1,TFB4

endocytosis ORIGINAL FUNCTIONS

GO:0016197:VPS20,YPT32,VPS27,VPS24,VTA1,HSE1,RSP5,DID4,STP22,DID2,YPT52,SNF7,YPT7,VPS60,VPS4 GO:0044710:YPT7,MSS4,SPO14,VPS4

GO:0016192:VPS20,YPT32,VPS24,GCS1,DOA4,RSP5,AGE2,DID4,ENT2,CHC1,SPO14,YPT52,SNF7,GLO3,YPT7,VPS4

GO:0007034:VPS20,VPS27,VPS24,VPS25,HSE1,VPS36,RSP5,DID4,MVB12,STP22,SNF8,DID2,YPT52,SNF7,VTA1,VPS28,VPS60,VPS4

GO:0000122:RME1,VPS36,SNF8,VPS25

GO:0043170:SSB2,SSA1,SSA3,SSA2,RSP5,SSA4,DOA4,RIM20

GO:0044087:MVB12,RHO1,VTA1,RSP5,CDC42

GO:0051649:YPT32,GCS1,SSA1,SSA3,SSA2,RSP5,SSA4,STP22,YPT52,VPS4,VPS20,VPS27,VPS24,VPS25,DOA4,VPS28,MVB12,GLO3,YPT7,VPS60,SSB2,DID4,SNF8,DID2,HSE1,SNF7,VPS36,VTA1,AGE2,SPO14

endosomal Sorting Complex Required For Transport ESCRT ORIGINAL FUNCTIONS

GO:0007034:VPS20,VPS24,VPS36,VPS25,DID4,SNF8,SNF7 UNCOVERED GENES

RPS31

ergosterol biosynthesis ORIGINAL FUNCTIONS

GO:0016126:ERG4,ERG5,ERG6,ERG7,ERG1,ERG2,ERG3,ERG24,ERG8,ERG9,MVD

1,ERG20,ERG12,ERG13,ERG10,ERG11,HMG1,HMG2

GO:0008610:ERG4,ERG5,ERG6,ERG7,ERG1,ERG2,ERG3,IDI1,ERG24,ERG8,ERG9,MVD1,ERG20,ERG12,ERG13,ERG10,ERG11,HMG1,HMG2

ethanol degradation I ORIGINAL FUNCTIONS

GO:0000947:ADH3,ADH2,ADH1,ADH5,ADH4

GO:0046496:ADH3,ADH2,ADH1,ADH5,ALD6,ALD4

ether lipid metabolism ORIGINAL FUNCTIONS

GO:0006644:AYR1,TGL4,CPT1,ALE1,EPT1,SPO14 UNCOVERED GENES PAC1

ethylene biosynthesis ORIGINAL FUNCTIONSGO:0001320:SOD2,SOD1

GO:0015891:FRE3,FRE4 GO:0035434:FRE2,FRE1,FRE7

GO:0055076:FRE3,FRE4,SOD1 UNCOVERED GENES ARO9

eukaryotic Translation Elongation ORIGINAL FUNCTIONS

GO:0002181:RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL2A,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A

GO:0042254:RPS13,RPS21A,RPS21B,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,RPS9B,RPS9A,RPS27A,RPL3,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL11B,RPL11A,RPS6A,RPS0A,RPS0B,RPL8B,RPL8A,RPL5,RPS16B,RPS20,RPS26A,RPL34A,RPS26B,RPS24B,RPP0,RPL10,RPS10A,RPS8B,RPL6A,RPL6B,RPS28B,RPS28A,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B

GO:0006412:RPL9A,RPL9B,EFT2,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL10,RPL2A,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A fatty Acid Biosynthesis Initial Steps ORIGINAL FUNCTIONS

GO:0006631:HFA1,FAS2,FAS1,CEM1,OAR1,ACC1,MCT1

fatty acid elongation ORIGINAL FUNCTIONS

GO:0006631:ELO1,FAS2,FAS1,CEM1,SUR4,FEN1,IFA38,OAR1,TSC13,ETR1,TES1

GO:0044255:PHS1,TSC13,FAS2,FAS1,CEM1,SUR4,ELO1,FEN1,IFA38,OAR1,ETR1, TES1

fc epsilon receptor FCERI signaling ORIGINAL FUNCTIONS

GO:0023052:PKC1,FUS3,HOG1,PBS2,KSS1,PKH3,PKH2,PKH1

fcgamma receptor FCGR dependent phagocytosis ORIGINAL FUNCTIONS

GO:0035382:STE20,CLA4,SKM1 GO:0007010:CLA4,PKC1,LAS17,CDC42

GO:0019236:STE20,CLA4,SPO14,CDC42 UNCOVERED GENES DPP1

folate biosynthesis ORIGINAL FUNCTIONSGO:0006732:PHO8,DFR1,FOL1,FOL3,FOL2

GO:0009396:DFR1,FOL1,FOL3,FOL2

GO:0044281:FOL1,MET7,FOL3,FOL2,PHO8,DFR1,ABZ1

GO:0008652:DFR1,FOL1,ABZ1,FOL3,FOL2

folate polyglutamylation ORIGINAL FUNCTIONSGO:0006730:SHM2,MET7,SHM1

GO:1901605:MET13,FOL3,ADE3,SHM1,DFR1,MIS1

GO:0042558:DFR1,MIS1,FOL3,ADE3

folate transformations ORIGINAL FUNCTIONS

GO:0006730:MTD1,GCV1,GCV2,GCV3,SHM2,SHM1

GO:1901605:MTD1,MET13,ADE3,GCV1,GCV2,GCV3,SHM1,FAU1,LPD1,MIS1

folic acid biosynthesis ORIGINAL FUNCTIONS

GO:0008652:FOL1,FOL3,FOL2,SHM1,MIS1,ABZ2,ABZ1

GO:0006730:SHM2,MET7,SHM1

GO:0042558:FOL1,FOL3,ADE3,FOL2,ABZ2,MIS1

formation of the ternary complex and subsequently the 43S complex ORIGINAL FUNCTIONS

GO:0022613:RPS13,RPS21A,RPS21B,RPS17A,RPS8B,RPS15,RPS17B,RPS27B,RPS9B,RPS9A,RPS27A,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS19B,RPS19A,RPS6A,HCR1,RPS0A,RPS0B,RPS16B,RPS20,RPS26A,RPS26B,SUI3,SUI2,RPS10A,RPS10B,GCD11,RPS28B,RPS28A,RPS24B,RPS18B,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B

GO:0006413:PRT1,SUI3,SUI2,TIF35,TIF34,GCD11,NIP1

formation of transcriptioncoupled NER TCNER repair complex ORIGINAL FUNCTIONS

GO:0090304:RAD10,RPB10,SYF1,CCL1,RAD3,RAD2,RAD1,RPB11,RPB4,RPB5,RPB7,RPB2,RPB3,RPB8,RPB9,DST1,RPO21,RPO26,RAD26,KIN28,RAD28,SSL2,SSL1,TFB1,TFB2,TFB3,TFB4

GO:0006289:RAD26,SSL2,RAD10,TFB3,RPB9,TFB1,TFB2,SSL1,RAD3,RAD2,RAD1,TFB4

GO:0006366:CCL1,RAD3,RAD2,RPB10,RPB11,RPB4,RPB5,RPB7,RPB2,RPB3,RPB8,RPB9,DST1,RPO21,RPO26,KIN28,SSL2,SSL1,TFB1,TFB2,TFB3,TFB4

GO:0006281:RAD26,SSL2,RAD10,TFB3,RPB9,RAD28,TFB1,TFB2,SSL1,RAD3,RAD2,RAD1,TFB4,RPO21

gDPmannose biosynthesis ORIGINAL FUNCTIONSGO:0019673:PSA1,PMI40

GO:0009117:PGI1,PSA1,PMI40 UNCOVERED GENES SEC53

gPCR downstream signaling ORIGINAL FUNCTIONS

GO:0035556:RHO1,PKC1,ROM2,PDE2,ROM1

GO:0051128:RHO1,PKC1,ROM2,CDC24 UNCOVERED GENES YJU3

gTP hydrolysis and joining of the 60S ribosomal subunit ORIGINAL FUNCTIONS

GO:0002181:RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL2A,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A

GO:0042254:RPS13,RPS21A,RPS21B,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,RPS9B,RPS9A,RPS27A,RPL3,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL11B,RPL11A,RPS6A,RPS0A,RPS0B,RPL8B,RPL8A,RPL5,RPS16B,RPS20,RPS26A,RPL34A,RPS26B,RPS24B,RPP0,RPL10,RPS10A,RPS8B,FUN12,RPL6A,RPL6B,RPS28B,RPS28A,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B

galactose catabolism ORIGINAL FUNCTIONS

GO:0044712:PGM2,GAL7,PGM1,GAL10,PGM3

GO:0019388:PGM2,GAL7,PGM1,GAL10

gammaglutamyl cycle ORIGINAL FUNCTIONS

GO:0006749:GSH2,DUG1,YKL215C,GSH1,ECM38

gastrinCREB signalling pathway via PKC and MAPK ORIGINAL FUNCTIONS

GO:0044700:KSS1,PKC1,FUS3GO:0007166:KSS1,FUS3 UNCOVERED GENES YJU3

genes of Meiotic Recombination ORIGINAL FUNCTIONS

GO:0007126:XRS2,RED1,RAD51,MRE11,HOP1,RAD55,RAD57,MEK1,SPO11,REC104,REC114,RAD50,REC107,MEI4,REC102,RAD52

gluconeogenesis ORIGINAL FUNCTIONS

GO:0044712:FBA1,ENO1,ENO2,PGI1,TPI1,GPM1,MDH3,TDH1,TDH2,TDH3,PGK1

GO:0044710:AAT2,PCK1,MDH1,FBA1,ENO1,ENO2,MAE1,PGI1,TPI1,GPM1,MDH3,MDH2,TDH2,TDH3,PGK1,TDH1,PYC1,PYC2,FBP1

GO:0044711:AAT2,PCK1,FBA1,ENO1,PGI1,ENO2,GPM1,MDH2,TDH2,TDH3,PGK1,TDH1,PYC1,PYC2,FBP1 GO:0015849:CTP1,DIC1,AGC1

GO:0006091:TDH2,FBA1,ENO1,ENO2,PGI1,TPI1,GPM1,TDH1,MDH1,TDH3,PGK1

GO:0006096:FBA1,ENO1,ENO2,PGI1,TPI1,GPM1,TDH1,TDH2,TDH3,PGK1

GO:0006094:PCK1,FBA1,ENO1,FBP1,PGI1,ENO2,GPM1,MDH2,TDH2,TDH3,PGK1,PYC1,PYC2,TDH1 UNCOVERED GENES DET1,YOR283W

glutamate degradation I ORIGINAL FUNCTIONSGO:0043649:GAD1,UGA2

GO:0009448:UGA2,UGA1 UNCOVERED GENES GDH2

glutamate degradation VII ORIGINAL FUNCTIONSGO:0006104:LSC2,LSC1

GO:0045333:SDH1,SDH3,SDH2,SDH4,FUM1 GO:0043648:AAT2,AAT1,FUM1

glutathioneGlutaredoxin Redox Reaction ORIGINAL FUNCTIONS

GO:0034599:GPX1,GPX2,GLR1,HYR1 GO:0006749:GTT1,GTT2

glycerol biosynthesis ORIGINAL FUNCTIONSGO:0006116:GPD1,GPD2

GO:0006071:HOR2,RHR2,GPD2

glycerol3phosphate shuttle ORIGINAL FUNCTIONSGO:0006116:GUT2,GPD1,GPD2

glycine biosynthesis ORIGINAL FUNCTIONSGO:0006545:AGX1,GLY1

GO:0006730:SHM2,SHM1 GO:0009070:AGX1,GLY1,SHM1

glycogen biosynthesis ORIGINAL FUNCTIONS

GO:0005978:GLC3,GLG1,GLG2,GSY1,GSY2,UGP1

glycolysis ORIGINAL FUNCTIONSGO:0016311:DET1,YOR283W

GO:0044710:PYK2,FBA1,PGI1,CDC19,PFK2,PFK1,ENO1,TPI1,GPM1,TDH1,TDH2,TDH3,ENO2,PGK1

GO:0006096:FBA1,PGI1,CDC19,PFK2,PFK1,ENO1,TPI1,GPM1,TDH1,TDH2,TDH3,ENO2,PGK1

glycosphingolipid metabolism ORIGINAL FUNCTIONSGO:0009141:NPP2,NPP1

UNCOVERED GENES ISC1

glyoxylate and dicarboxylate metabolism ORIGINAL FUNCTIONS

GO:0044710:CIT3,MDH1,AGX1,FDH1,SHM2,SHM1,ACO1,MLS1,MDH3,MDH2,ERG10,CTA1,BNA7,GCV3,GLN1,ICL2 GO:0006730:GCV3,SHM2,SHM1

GO:0032787:FDH1,ICL2,CIT3,MDH3,MLS1

GO:0044248:CIT3,CIT1,FDH1,DAL7,MDH3,CTA1,GCV3,ICL2

GO:0006099:ACO1,CIT3,MDH1

GO:0044281:CIT3,GLN1,AGX1,FDH1,SHM2,SHM1,MLS1,MDH3,ERG10,BNA7,GCV3,ICL2 GO:0072329:FDH1,ICL2,CIT3,MDH3 GO:0000302:CTT1,CTA1

glyoxylate cycle ORIGINAL FUNCTIONSGO:0044248:CIT1,DAL7,CIT3,MDH3

GO:0006099:ACO1,CIT3,MDH1 GO:0055114:ACO1,CIT3,MDH3,MDH1

GO:0032787:MDH3,MLS1,CIT3 UNCOVERED GENES MDH2

golgi to ER Retrograde Transport ORIGINAL FUNCTIONS

GO:0016236:ARF2,ARF1,GEA2,GEA1

guanosine ribonucleotides ide novoi biosynthesis ORIGINAL FUNCTIONS

GO:0009165:YNK1,GUA1

hexaprenyl Diphosphate Biosynthesis ORIGINAL FUNCTIONS

GO:0045337:IDI1,ERG20 UNCOVERED GENES COQ1,ARG5,6 histidine Biosynthesis ORIGINAL FUNCTIONS

GO:0000105:HIS4,HIS5,HIS6,HIS7,HIS1,HIS2,HIS3

homocysteine and Cysteine Interconversion ORIGINAL FUNCTIONS

GO:0006534:STR3,STR2,IRC7,CYS3,CYS4

homologous recombination ORIGINAL FUNCTIONS

GO:0006310:XRS2,RFA2,RFA1,MRE11,POL32,TOP3,SGS1,RAD57,MUS81,RAD55,RAD54,RAD52,RAD51,RAD50,RDH54,RAD59

GO:0065007:XRS2,RFA2,RFA1,MRE11,TOP3,SGS1,RAD57,RAD54,RAD52,RAD51,RAD50,RAD59,SEM1

GO:0006259:XRS2,POL3,RFA2,RFA1,MRE11,POL32,TOP3,SGS1,RAD57,MUS81,RAD55,RAD54,POL31,RAD52,RAD51,RAD50,RDH54,RAD59

homologous recombination repair of replicationindependent doublestrand breaks ORIGINAL FUNCTIONS

GO:0006281:TEL1,MRE11,CDC9,RAD52,RAD51,RAD50,HTA2,HTA1

hypusine synthesis from eIF5Alysine ORIGINAL FUNCTIONSGO:0045905:ANB1,HYP2

GO:0008612:DYS1,LIA1

import of palmitoylCoA into the mitochondrial matrix ORIGINAL FUNCTIONS

GO:0006631:HFA1,CRC1,ACC1

inosine5phosphate biosynthesis II ORIGINAL FUNCTIONS

GO:0072521:ADE16,ADE17,ADE1,ADE2,ADE13

GO:0006189:ADE16,ADE17,ADE1,ADE13

inositol phosphate biosynthesis ORIGINAL FUNCTIONS

GO:0043647:VIP1,KCS1,IPK1,DDP1,PLC1,ARG82 integration of energy metabolism ORIGINAL FUNCTIONS

GO:0006796:TPK1,TPK2,TPK3,PKC1,TAL1 GO:0010737:TPK1,TPK3,TPK2

GO:0035556:TPK1,TPK3,PKC1,TPK2 GO:0050794:TPK1,TPK3,PKC1,TPK2,BCY1

UNCOVERED GENES YBR241C,VPS73

interconversion of 2oxoglutarate and 2hydroxyglutarate ORIGINAL FUNCTIONS

GO:1901615:DLD3,ADH4

ion transport by Ptype ATPases ORIGINAL FUNCTIONS

GO:0006811:PMR1,DRS2,DNF2,DNF1 GO:0045332:DNF1,DNF2,DRS2

iron uptake and transport ORIGINAL FUNCTIONS

GO:0051453:VMA2,TFP1,STV1,VPH1 GO:0030003:VMA2,TFP1,CUP5,STV1,VPH1

GO:0007034:CUP5,VMA6

isoleucine biosynthesis ORIGINAL FUNCTIONS

GO:0006520:ILV5,ILV6,ILV1,ILV3,ILV2,CHA1,BAT2,BAT1

GO:0009082:ILV5,ILV6,ILV1,ILV3,ILV2,BAT2,BAT1

kinesins ORIGINAL FUNCTIONSGO:0007067:KAR3,KIP1,KIP2,KIP3

GO:0090307:KIP1,KIP2,KIP3 UNCOVERED GENES SMY1

leucine Degradation ORIGINAL FUNCTIONS

GO:0009063:ADH3,ADH2,ADH1,ADH5,ADH4,BAT2,BAT1,ARO10

GO:0000947:ADH3,ADH2,ADH1,ADH5,ADH4,ARO10 UNCOVERED GENES

THI3

leucine biosynthesis ORIGINAL FUNCTIONS

GO:0009082:LEU4,LEU2,BAT2,BAT1,LEU1,LEU9

linoleic acid metabolism ORIGINAL FUNCTIONS

GO:0044710:GCY1,TGL4,AAD10,AAD3,AAD6,AAD4,AAD14,YPR1

GO:0019568:YPR1,GCY1 GO:0006081:AAD10,AAD3,AAD6,AAD14,AAD4

lipases biosynthesis ORIGINAL FUNCTIONSGO:0044242:PLC1,ISC1 UNCOVERED GENES SPO22,SPO14

lipid digestion mobilization and transport ORIGINAL FUNCTIONS

GO:0044255:YJU3,NCR1 UNCOVERED GENES ADP1

lipidLinked Oligosaccharide Biosynthesis ORIGINAL FUNCTIONS

GO:0006488:ALG2,ALG11,ALG8,ALG6,ALG12

GO:0006486:ALG8,ALG9,DIE2,ALG6,ALG11,ALG12

GO:1901576:ALG8,ALG9,ALG2,ALG6,ARG7,DIE2,ALG11,ALG12

lipoic acid metabolism ORIGINAL FUNCTIONSGO:0009249:LIP2,LIP5,AIM22

lysine Biosynthesis ORIGINAL FUNCTIONS

GO:0006520:ARO8,HOM6,HOM2,HOM3,LYS9,LYS21,LYS1,LYS2,LYS20,LYS4

GO:0009067:HOM6,HOM2,HOM3,LYS9,LYS21,LYS1,LYS2,LYS20,LYS4

GO:0006553:LYS21,LYS20,LYS9,LYS1,LYS2,LYS4 UNCOVERED GENES LYS5

lysine catabolism ORIGINAL FUNCTIONSGO:0006839:ODC1,ODC2

GO:0006103:KGD2,LPD1 UNCOVERED GENES LYS9

lysine degradation ORIGINAL FUNCTIONSGO:0019413:ALD6,ALD5,ALD4

GO:0016571:DOT1,SET1,SET2 GO:0006553:LYS9,LYS2,LYS1 UNCOVERED GENES KGD2,ERG10

mRNA Splicing ORIGINAL FUNCTIONS

GO:0016070:PRP6,RPB10,RPB11,PRP8,RPB4,RPB5,RPB7,RPB2,RPB3,RPB8,RPB9, SMX2,RPO21,RPO26,SME1,CUS1,SNU114,TFG2,LSM2,HSH49,DIB1,SNU13,CBC2,S

MX3,STO1,SMB1,HSH155,TFG1,BRR2,SMD1,SMD2,SMD3

GO:0000377:CUS1,CBC2,SNU114,BRR2,SMD2,STO1,LSM2,SMB1,PRP6,HSH49,DIB1,SMD3,PRP8,SMX2,SMX3,HSH155,SMD1,SME1,SNU13

GO:0016071:RPB4,CUS1,RPB7,SNU114,CBC2,BRR2,SMD2,STO1,LSM2,SMB1,PRP6,HSH49,DIB1,SMD3,PRP8,SMX2,SMX3,HSH155,SMD1,SME1,SNU13

mcresol degradation ORIGINAL FUNCTIONS

GO:0006081:AAD10,AAD3,AAD14,AAD15,AAD4,AAD6

metabolism of folate and pterines ORIGINAL FUNCTIONS

GO:1901605:DFR1,MET13,MIS1,ADE3 GO:0006760:DFR1,MIS1,ADE3

UNCOVERED GENES FLX1,MET7

metabolism of polyamines ORIGINAL FUNCTIONS

GO:0019509:MDE1,MRI1,MEU1,ADI1,UTR4

GO:0009067:AAT2,ADI1,UTR4,MEU1,MDE1,MRI1

GO:0008652:AAT2,ADI1,SPE2,MRI1,MEU1,SPE3,SPE1,MDE1,UTR4

methionine Biosynthesis ORIGINAL FUNCTIONS

GO:0000096:MET2,MET6,MET17,STR3,HOM6,HOM2,HOM3,IRC7,MHT1,SAM4

GO:0009066:MET2,MET6,MET17,STR3,HOM6,HOM2,HOM3,SAM4,THR1

methionine Degradation ORIGINAL FUNCTIONSGO:0006556:SAM2,SAM1

GO:0006575:CYS4,SAM1,SAH1,SAM2

methionine salvage pathway ORIGINAL FUNCTIONS

GO:0019509:MDE1,MRI1,MEU1,ADI1,UTR4

GO:0006520:AAT2,ADI1,UTR4,MEU1,MDE1,MRI1,ARO9

GO:0009067:AAT2,ADI1,UTR4,MEU1,MDE1,MRI1 mevalonate Pathway ORIGINAL FUNCTIONS

GO:0016126:ERG8,MVD1,ERG12,ERG13,ERG10,HMG1,HMG2

GO:0008610:IDI1,ERG8,MVD1,ERG12,ERG13,ERG10,HMG1,HMG2 UNCOVERED GENES ARG5,6

mismatch repair ORIGINAL FUNCTIONS

GO:0006259:SRS2,MSH6,MSH3,POL3,RFA2,RFA1,RFC5,RFC4,MSH2,EXO1,RFC1,MLH1,RFC2,CDC9,POL31,RFC3,POL32,PMS1,MLH3,POL30

mitochondrial IronSulfur Cluster Biogenesis ORIGINAL FUNCTIONS

GO:0055072:NFS1,ISU2,ISU1,ARH1,YFH1 GO:0048250:MRS3,MRS4

GO:0016226:NFS1,ISU2,ISU1,YAH1,YFH1,ISD11

mitochondrial Protein ImportORIGINAL FUNCTIONS

GO:1901576:ATP2,ATP1,OLI1,TAZ1,AAC1,COQ2

GO:0016043:SSC1,TIM8,TIM9,TOM5,TOM6,TOM7,TOM22,TOM20,TOM40,JAC1,PAM16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,YFH1,POR1,TOM70,MAS2,MAS1,ACO1,MGE1,TIM10,BCS1,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,MDM35,HSP60,COX19,TAZ1,ERV1,TIM54,SAM50,TIM50,COX17

GO:0044765:SSC1,TIM8,TIM9,TOM5,TOM6,TOM7,TOM22,TOM20,TOM40,PAM16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,ATP2,ATP1,OLI1,POR1,CCS1,TOM70,MAS2,MAS1,MGE1,TIM10,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,HSP60,COX19,ERV1,TIM54,AAC1,SAM50,TIM50,MIR1,COX17

GO:0045333:COX5A,MIC17,COX4,JAC1,TAZ1,CYT1,ACO1,DLD1,AAC1,MIC14

GO:0044281:CYB2,ATP2,ATP1,OLI1,YFH1,COQ2

GO:0072655:SSC1,TIM8,TIM9,TOM5,TOM6,TOM7,TOM22,TOM20,TOM40,PAM

16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,TOM70,MAS2,MAS1,HSP60,TIM10,BCS1,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,MGE1,ERV1,TIM54,SAM50,TIM50

GO:0006839:SSC1,TIM8,TIM9,TOM5,TOM6,TOM7,TOM22,TOM20,TOM40,PAM16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,TOM70,MAS2,MAS1,HSP60,TIM10,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,MGE1,ERV1,TIM54,AAC1,SAM50,TIM50

GO:0055085:SSC1,TIM8,TIM9,TOM6,TOM7,TOM22,TOM20,TOM40,PAM16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,ATP2,ATP1,OLI1,TOM70,MGE1,TIM10,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,HSP60,ERV1,TIM54,SAM50,TIM50,MIR1

GO:0007005:SSC1,TIM8,TIM9,TOM5,TOM6,TOM7,TOM22,TOM20,TOM40,PAM16,PAM17,TIM44,TIM21,TIM23,TIM22,PAM18,MIA40,POR1,TOM70,MAS2,MAS1,ACO1,MGE1,TIM10,BCS1,TIM12,TIM13,SAM35,SAM37,TIM17,TIM18,MDM35,HSP60,COX19,TAZ1,ERV1,TIM54,SAM50,TIM50,COX17 UNCOVERED GENES CIT1

mitochondrial tRNA Synthetases ORIGINAL FUNCTIONS

GO:0043038:MSK1,VAS1,MSM1,SLM5,ISM1,MST1,HTS1,NAM2,MSW1,MSY1,MSF1,MSD1,MSE1,MSR1

GO:0032543:MSK1,SLM5,ISM1,MST1,HTS1,NAM2,MSW1,MSY1,MSF1,MSD1,MSE1,MSR1

mitochondrial tRNA aminoacylation ORIGINAL FUNCTIONS

GO:0043038:VAS1,GRS2,YDR341C,MSM1,SLM5,MSR1,ISM1,KRS1,DIA4,ALA1,NAM2,MSW1,GRS1,THS1,MSY1,MSF1,MSD1,MSE1,GLN4 UNCOVERED GENES

PPA2

mitotic Prophase ORIGINAL FUNCTIONS

GO:0000079:CLB3,CLB2,CLB1,CLB6,CLB5,CLB4

GO:0010696:CDC28,CLB3,CLB2,CLB1,CLB5,CLB4

GO:0051726:CDC28,CLB3,CLB2,CLB1,RIM15,CLB6,CLB5,CLB4

nAD biosynthesis II from tryptophan ORIGINAL FUNCTIONS

GO:0009435:NMA1,NMA2,BNA6,BNA7,QNS1,BNA5,BNA4,BNA2,BNA1

nAD salvage pathway ORIGINAL FUNCTIONSGO:0000183:SIR2,NPT1,PNC1

GO:0019363:NMA1,NMA2,NPT1,PNC1,QNS1

GO:0019362:NMA1,NMA2,NPY1,QNS1,NPT1,PNC1

nGlycan biosynthesis ORIGINAL FUNCTIONS

GO:1901137:ALG8,ALG9,ALG2,ALG3,ALG1,ALG6,ALG7,STT3,DIE2,CAX4,OST6,OST4,OST5,OST2,OST3,DPM1,OST1,SEC59,ALG11,ALG12,ALG13,ALG14,CWH41,WBP1,SWP1 GO:0009272:CWH41,ROT2

GO:0009100:ALG8,ALG9,ALG3,ALG1,ALG6,ALG7,STT3,DIE2,CAX4,OST6,OST4,OST5,OST2,OST3,DPM1,OST1,SEC59,MNS1,ALG11,ALG12,CWH41,WBP1,SWP1

GO:0044267:ALG8,ALG9,ALG3,ALG1,ALG6,ALG7,STT3,DIE2,CAX4,OST6,OST4,YLR057W,OST2,OST3,DPM1,OST1,SEC59,MNS1,ALG11,ALG12,OST5,CWH41,WBP1,SWP1 UNCOVERED GENES RFT1

nglycan trimming in the ER and CalnexinCalreticulin cycle ORIGINAL FUNCTIONS

GO:0009100:CWH41,MNS1 GO:0030433:MNL1,MNS1

nicotinamide riboside salvage pathway I ORIGINAL FUNCTIONS

GO:0009435:NMA1,NMA2,NRK1

nitrogen metabolism ORIGINAL FUNCTIONSGO:0006541:GLT1,GLN1,GDH1,GDH3

UNCOVERED GENES GDH2

nonOxidative Branch of the Pentose Pathway ORIGINAL FUNCTIONS

GO:0006098:TAL1,GND2,GND1,TKL2,TKL1,RKI1,RPE1,ZWF1

nonhomologous endjoining ORIGINAL FUNCTIONS

GO:0006303:NEJ1,XRS2,POL4,RAD27,MRE11,YKU70,DNL4,LIF1,RAD50

GO:0006302:NEJ1,XRS2,POL4,YKU80,RAD27,MRE11,YKU70,DNL4,LIF1,RAD50

nonsense Mediated Decay Independent of the Exon Junction Complex ORIGINAL FUNCTIONS

GO:0002181:RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL2A,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A

GO:0010467:RPS13,RPS21A,RPS21B,RPS23B,RPS27B,RPS9B,RPS9A,RPS27A,RPL9A,RPL9B,CBC2,RPS11B,RPS18B,NAM7,RPS31,RPL7B,RPL7A,RPL29,RPL17B,SUP45,RPL15B,RPL22A,RPL22B,RPS6A,RPS0A,RPS0B,RPL36A,RPL8B,RPL8A,RPL36B,RPS16B,RPL17A,RPS20,RPL26A,RPL26B,RPS24B,RPP0,RPL10,RPS8B,RPL2A,SUP35,RPL6A,RPL6B,STO1,RPL30,RPS1B,RPS1A,RPL31A,RPL31B,RPL33B,RPL15A,RPL33A

GO:0042254:RPS13,RPS21A,RPS21B,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,RPS9B,RPS9A,RPS27A,RPL3,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL11B,RPL11A,RPS6A,RPS0A,RPS0B,RPL8B,RPL8A,RPL5,RPS16B,RPS20,RPS26A,RPL34A,RPS26B,RPS24B,RPP0,RPL10,RPS10A,RPS8B,RPL6A,RPL6B,RPS28B,RPS28A,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B

GO:0006412:RPL9A,RPL9B,NAM7,RPL7A,RPL29,SUP45,RPL15B,RPL22A,RPL22B, RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL10,RPL2A,SUP35,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A

nucleotidebinding domain leucine rich repeat containing receptor NLR signaling pathways ORIGINAL FUNCTIONSGO:0000491:HSC82,HSP82

GO:0071822:PBS2,HSC82,HSP82,SGT1

GO:0009628:PBS2,HOG1,HSC82,HSP82

p75 NTR receptormediated signalling ORIGINAL FUNCTIONS

GO:0044700:RHO1,HOG1,ROM2,ROM1 GO:0007264:RHO1,ROM2,ROM1

GO:0006970:HOG1,ISC1

pI3KAKT activation ORIGINAL FUNCTIONSGO:0035556:RHO1,PKH3,PKH2,PKH1

GO:0000196:PKH3,PKH2,PKH1

pKBmediated events ORIGINAL FUNCTIONSGO:0035556:TOR2,TOR1,PDE2,LST8

GO:0001558:KOG1,TOR1,LST8 GO:0031929:TOR2,TOR1,LST8

UNCOVERED GENES RPS6A,RHB1

pRPP biosynthesis ORIGINAL FUNCTIONSGO:0046391:PRS3,PRS2,PRS1,PRS5,PRS4

pentose Phosphate Pathway 2 ORIGINAL FUNCTIONSGO:0006409:SOL2,SOL1

GO:0006098:GND2,ZWF1,TKL2,SOL3

pentose and glucuronate interconversions ORIGINAL FUNCTIONS

GO:0005997:XYL2,XKS1 GO:0019321:XKS1,XYL2,GRE3

GO:0005975:PGU1,GRE3,XYL2,RPE1,XKS1,UGP1

GO:0019413:ALD6,ALD5,ALD4

pentose phosphate pathway hexose monophosphate shunt ORIGINAL FUNCTIONS GO:0006409:SOL2,SOL1

GO:0006098:TAL1,GND2,GND1,RKI1,RPE1,ZWF1,SOL3

peroxisomal lipid metabolismORIGINAL FUNCTIONS

GO:0034440:FOX2,ANT1,TES1,POT1 GO:0006631:FAT1,FOX2,ANT1,TES1,POT1

GO:0015711:FAT1,ANT1,PXA2

GO:0032787:ANT1,FAT1,POT1,CAT2,FOX2,TES1

phase II conjugation ORIGINAL FUNCTIONS

GO:0006575:ECM38,DUG1,SAH1,GSH2,SAM2,SAM1 UNCOVERED GENES

UGP1

phenylalanine and Tyrosine Biosynthesis ORIGINAL FUNCTIONS

GO:0009072:ARO8,ARO9,PHA2,ARO7

phenylalanine metabolism ORIGINAL FUNCTIONS

GO:0006520:AAT2,AAT1,ARO8,ARO9,ALD3,ALD2,HIS5,ARO10

phosphatidate biosynthesis I the dihydroxyacetone pathway ORIGINAL FUNCTIONS GO:0008654:SLC1,SCT1,AYR1,GPT2

phosphatidate biosynthesis II the glycerol3phosphate pathway ORIGINAL FUNCTIONS GO:0019637:SLC1,SCT1,GPD1,GPD2,GPT2

GO:0008654:SLC1,SCT1,GPT2

phosphatidylglycerol biosynthesis ORIGINAL FUNCTIONS

GO:0008654:SLC1,CDS1,SCT1,GPT2,PGS1

phosphatidylinositol signaling system ORIGINAL FUNCTIONS

GO:0019637:INM2,INM1,FAB1,VPS34,STT4,CMD1,MSS4,PLC1,PIK1,CDS1,INP54, PIS1,INP53,INP52

GO:0006644:FAB1,VPS34,STT4,CMD1,MSS4,PLC1,PIK1,CDS1,INP54,PIS1,INP53,I

NP52

GO:0046488:FAB1,VPS34,STT4,CMD1,MSS4,PIK1,CDS1,INP54,PIS1,INP53,INP52

GO:0006796:INM2,PKC1,FAB1,VPS34,STT4,CMD1,MSS4,PLC1,INM1,PIK1,CDS1,INP54,PIS1,INP53,INP52

phospholipid Biosynthesis ORIGINAL FUNCTIONS

GO:0046474:PSD2,CDS1,PSD1,OPI3,CHO1,CHO2,PGS1

GO:0008610:PSD2,CDS1,CRD1,OPI3,CHO2,CHO1,PSD1,PGS1

phospholipid biosynthesis Kennedy pathway ORIGINAL FUNCTIONS

GO:0006656:CKI1,CPT1,PCT1,EPT1

phosphopantothenate biosynthesis I ORIGINAL FUNCTIONS

GO:0015940:ECM31,PAN6,PAN5 GO:0009108:CAB1,ECM31,PAN6,PAN5

polyamine Biosynthesis ORIGINAL FUNCTIONSGO:0006596:SPE4,SPE2,SPE3,SPE1

polymerase switching on the Cstrand of the telomere ORIGINAL FUNCTIONS

GO:0006260:POL1,RFC5,RFC4,RFC1,RFC3,RFC2,POL12,POL30,PRI1

postElongation Processing of the Transcript ORIGINAL FUNCTIONS

GO:0043631:CFT2,CFT1,YSH1 GO:0006397:CBC2,YSH1,STO1,CFT1,CFT2

proline biosynthesis ORIGINAL FUNCTIONSGO:0006561:PRO3,PRO2,PRO1

proteasome ORIGINAL FUNCTIONS

GO:0044267:RPN2,RPN5,RPN6,PUP1,RPN9,PUP3,PUP2,PRE9,PRE8,PRE5,PRE4,PRE7,PRE6,PRE1,PRE3,PRE2,UMP1,RPT6,RPT4,RPT5,RPT2,RPT3,RPT1,SCL1,SEM1,RPN12,RPN13,RPN10,RPN11,PRE10

GO:0019941:RPN2,RPN6,PUP1,RPN9,PUP3,PUP2,PRE9,PRE8,PRE5,PRE4,PRE7,PRE6,PRE1,PRE3,PRE2,UMP1,RPT6,RPT4,RPT5,RPT2,RPT3,RPT1,SCL1,SEM1,RPN12, RPN13,RPN10,RPN11,PRE10

GO:0043248:RPT6,RPN2,RPT5,RPT2,RPT3,RPN6,UMP1,RPN9,RPT1,RPT4,BLM10,PRE9,PRE4,SEM1,PRE2

protein Modifications ORIGINAL FUNCTIONS

GO:0035268:PMT5,PMT4,PMT6,PMT1,PMT3,PMT2

GO:0018342:BET2,RAM2,RAM1,CDC43

GO:0006464:BET2,BPL1,RAM1,RAM2,CDC43,PMT5,PMT4,PMT6,PMT1,PMT3,PMT2

protein export ORIGINAL FUNCTIONS

GO:0072599:SPC1,KAR2,SRP54,SEC11,SEC65,SBH1,SSH1,SEC61,SEC62,SEC63,SSS1,SRP101,SRP102,SRP14,SRP72,SRP68,SBH2,SPC2,SPC3

GO:0006605:SEC65,SEC61,SEC62,SEC63,SRP101,SRP102,SPC1,SBH1,SPC3,SRP54,OXA1,SSS1,SPC2,SRP14,SRP72,IMP2,IMP1,SEC11,SBH2,SSH1,KAR2,SRP68

protein processing in endoplasmic reticulum ORIGINAL FUNCTIONS

GO:0051603:YDJ1,UFD2,PNG1,CUE1,DFM1,SEC61,UBC6,NPL4,SCJ1,EPS1,JEM1,SKP1,SSM4,DOA1,DSK2,CNE1,UFD1,HRD3,YLR057W,SHP1,CDC53,YOS9,RAD23,SEC13,MNS1,HLJ1,UBX2,KAR2,HRD1,HRT1,UBC4,DER1,UBC7,CDC48,USA1

GO:0070972:YDJ1,KAR2,SIL1,SSA1,SSA3,SSA2,SSA4,SEC62,SEC63,LHS1,SSS1,SEC61,SSH1,SBH2,SBH1,EPS1

GO:0006457:YDJ1,JEM1,SSB2,SSE1,CNE1,SSE2,SSA1,SSA3,SSA2,SSA4,EUG1,HSC82,HSP26,MPD1,SNL1,HSP82,SCJ1,PDI1

GO:0044267:YDJ1,UFD2,PNG1,SSE1,CUE1,SSE2,SSA1,PDI1,SSA2,DFM1,SSA4,STT3,HSC82,UBC6,NPL4,YLR057W,EPS1,JEM1,UFD1,SSM4,DOA1,EUG1,IRE1,DSK2,HSP

82,OST6,HRD3,OST4,OST5,OST2,CDC53,OST1,YOS9,SSB2,RAD23,OST3,SEC13,MNS1,SNL1,SEC61,HSP26,SUI2,GCN2,HRD1,MPD1,HLJ1,SCJ1,OTU1,WBP1,KAR2,UBX2,SSA3,CNE1,SHP1,HRT1,UBC4,UBC5,DER1,UBC7,SKP1,CWH41,CDC48,USA1,SWP1

GO:0016043:YDJ1,CUE1,SSA1,SSA3,SSA2,SSA4,SEC62,SEC63,NPL4,SBH2,SBH1,JEM1,SKP1,HSC82,IRE1,SSS1,DSK2,SEC31,OST6,UBX2,SHP1,OST3,HSP42,HSP82,SEC13,SNL1,SEC61,SUI2,SSH1,KAR2,SIL1,HRD1,UBC7,LHS1,SEC24,CDC48

GO:0061024:KAR2,SIL1,CUE1,SSA1,SEC13,SSA2,SSA4,SEC62,SEC63,SSA3,SSS1,SEC24,SEC61,SEC31,SSH1,LHS1,SHP1,SBH2,SBH1,CDC48

GO:0006807:SSB2,DOA1,HSC82,GCN2,SEC23,CDC48,NPL4,UFD1,USA1,SAR1

GO:0030433:YDJ1,UFD2,CUE1,DFM1,SEC61,UBC6,NPL4,CNE1,JEM1,SSM4,DSK2,UFD1,HRD3,YLR057W,SHP1,HRD1,YOS9,RAD23,SEC13,MNS1,HLJ1,SCJ1,KAR2,UBX2,DER1,UBC7,EPS1,CDC48,USA1GO:0071852:UBC7,HRD1,CWH41,ROT2,IRE1

purine Fermentation ORIGINAL FUNCTIONSGO:0009113:AAH1,ADE3

GO:0006760:MIS1,ADE3 UNCOVERED GENES FDH1

purine catabolism ORIGINAL FUNCTIONSGO:0072521:GUD1,PNP1

GO:0000302:CTT1,CTA1

purine ribonucleoside monophosphate biosynthesis ORIGINAL FUNCTIONS

GO:0006164:ADE4,ADE6,ADE16,ADE17,ADE12,ADE13,GUA1

purine ribonucleosides degradation to ribose1phosphate ORIGINAL FUNCTIONS

GO:0006400:TAD1,TAD2 UNCOVERED GENES PNP1

purine salvage ORIGINAL FUNCTIONSGO:0072521:APT1,PNP1,ADO1

pyrimidine biosynthesis ORIGINAL FUNCTIONS

GO:0072527:URA2,URA3,URA1,DUT1,DCD1,CDC21 rNA Polymerase I Transcription Initiation	ORIGINAL FUNCTIONS

GO:0006360:RPC40,RPB8,RRN3,RPA135,RPA190,RPC19

rNA Polymerase II Promoter Escape	ORIGINAL FUNCTIONS

GO:0006366:CCL1,RAD3,RPB10,RPB11,RPB4,RPB5,RPB7,RPB2,RPB3,RPB8,RPB9,RPO21,RPO26,TFG2,TFG1,KIN28,SSL2,SSL1,TFB1,TFB2,TFB3,TFB4

rNA Polymerase III Transcription Termination	ORIGINAL FUNCTIONS

GO:0009304:RPC40,RPB5,RPC53,RPB10,RPB8,RPC11,RPC34,RET1,RPC10,RPC31,RPC82,RPC25,RPC17,RPC19,RPO31,RPO26

rNA polymerase	ORIGINAL FUNCTIONS

GO:0006351:RPC53,RPA14,RPA34,RPC34,RPC37,RPC10,RPC31,RPB10,RPB11,RPC19,RPB4,RPB5,RPB7,RPB2,RPB3,RPB8,RPB9,RPA43,RPC82,RPA49,RPO26,RPC40,RPA135,RET1,RPC25,RPA12,RPO21,RPC11,RPA190,RPO31

recycling of eIF2GDP	ORIGINAL FUNCTIONS

GO:0006417:GCD6,GCD7,GCD2,GCD1,GCN3,GCD11

GO:0006446:GCD6,GCD7,GCN3,GCD2,GCD1 GO:0001731:SUI3,SUI2,GCD11

regulation of APCC activators between G1S and early anaphase	ORIGINAL FUNCTIONS	GO:0000079:CLB3,CLB2,CLB1,CLB6,CLB5,CLB4

GO:0010695:CDC28,CLB3,CLB2,CLB1,CLB5,CLB4,CDH1

regulation of Water Balance by Renal Aquaporins	ORIGINAL FUNCTIONS

GO:0007265:TPK1,TPK3,TPK2GO:0050794:TPK1,TPK3,TPK2,BCY1	UNCOVERED GENES	YFL054C

regulation of autophagy	ORIGINAL FUNCTIONS

GO:0032258:VPS30,VAC8,ATG4,ATG3,ATG1,ATG10,ATG11,ATG12,ATG13,ATG1

4,ATG7,ATG16,ATG5,ATG8

GO:0016236:VPS30,VAC8,ATG16,ATG13,VPS15,ATG5,ATG10,ATG3,ATG12,ATG1,ATG14,ATG7,ATG4,ATG17,ATG8

GO:0006914:VPS30,VAC8,ATG16,ATG13,ATG11,VPS15,ATG5,ATG10,ATG3,ATG12,ATG1,ATG14,ATG7,ATG4,ATG17,ATG8,VPS34

GO:0034727:VPS30,VAC8,ATG16,ATG13,ATG11,ATG5,ATG10,ATG3,ATG12,ATG1,ATG14,ATG7,ATG4,ATG17,ATG8

regulatory RNA pathways ORIGINAL FUNCTIONSGO:0006997:GSP2,GSP1

UNCOVERED GENES MSN5

respiratory electron transport ATP synthesis by chemiosmotic coupling and heat production by uncoupling proteins ORIGINAL FUNCTIONS

GO:0045333:SDH1,SDH3,SDH2,YJL045W

response to elevated platelet cytosolic Ca2 ORIGINAL FUNCTIONS

GO:0007015:COF1,PKC1,AIP1GO:0044703:KAR2,SSO2,SSO1

GO:0016192:SEC1,COF1,SSO1

GO:1902589:PKC1,SSO1,COF1,NTO1,STM1,AIP1 GO:0030042:COF1,AIP1

riboflavin FMN and FAD BiosynthesisORIGINAL FUNCTIONS

GO:0042727:FAD1,RIB1,RIB5,RIB3,RIB2,FMN1,RIB4,RIB7

ribose and Deoxyribose Phosphate Metabolism ORIGINAL FUNCTIONS

GO:0005996:RBK1,TKL2,TKL1,RKI1 GO:0006098:TKL2,TKL1,RKI1

GO:0055086:CDD1,TKL2,TKL1,RKI1

ribosome biogenesis in eukaryotes ORIGINAL FUNCTIONS

GO:0016072:MDN1,UTP6,UTP5,UTP4,RRP7,NOP1,NOG1,UTP9,UTP8,NOP10,NH

P2,RMP1,GSP1,POP7,NUG1,RIO2,UTP13,RIO1,NOP58,DIP2,NOP56,RNH70,FAP7,NOB1,UTP10,UTP15,UTP14,TIF6,UTP18,REX2,RNT1,CBF5,POP5,POP4,RCL1,NOP4,RPP1,SNU13,KEM1,NAN1,PWP2,SNM1,BMS1,RAT1,FCF1,POP3,POP1,UTP22,POP6,EMG1,UTP21,MPP10

GO:0034470:MDN1,UTP6,UTP5,UTP4,RRP7,NOP1,NOG1,UTP9,UTP8,NOP10,NHP2,RMP1,GSP1,POP7,NUG1,RIO2,UTP13,RIO1,NOP58,DIP2,NOP56,RNH70,FAP7,NOB1,UTP10,UTP15,UTP14,TIF6,UTP18,REX2,RNT1,CBF5,POP5,MPP10,POP4,RCL1,NOP4,RPP1,SNU13,NAN1,PWP2,HRR25,SNM1,BMS1,RAT1,FCF1,POP3,POP1,UTP22,POP6,EMG1,UTP21,POP8

GO:0042254:MDN1,RIA1,UTP5,UTP4,RRP7,NOP1,NOG1,UTP9,UTP8,NOP10,KRE33,NHP2,RMP1,MTR2,GSP1,POP7,NUG1,RIX7,RIO2,UTP13,RIO1,NMD3,NOP58,DIP2,NOP56,RNH70,FAP7,NOB1,UTP10,UTP15,UTP14,TIF6,UTP18,REX2,UTP6,RNT1,CBF5,POP5,POP4,RCL1,NOP4,RPP1,NOG2,MEX67,NAN1,PWP2,HRR25,SDO1,CRM1,SNM1,BMS1,RAT1,FCF1,AFG2,POP3,POP1,LSG1,UTP22,POP6,EMG1,UTP21,MPP10,SNU13

GO:0006356:CKA1,UTP5,CKA2,UTP9,UTP8,UTP4,UTP15,CKB1,NAN1,UTP10,CKB2

GO:0034660:MDN1,UTP6,UTP5,UTP4,RRP7,NOP1,NOG1,UTP9,UTP8,NOP10,NHP2,RMP1,GSP1,POP7,NUG1,RIO2,UTP13,RIO1,NOP58,DIP2,NOP56,RNH70,FAP7,NOB1,UTP10,UTP15,UTP14,TIF6,UTP18,GAR1,RNT1,CBF5,POP5,MPP10,POP4,RCL1,NOP4,RPP1,SNU13,KEM1,NAN1,PWP2,HRR25,SNM1,BMS1,RAT1,REX2,FCF1,POP3,POP1,UTP22,POP6,EMG1,UTP21,POP8

GO:0071840:MDN1,RIA1,UTP5,UTP4,RRP7,NOP1,NOG1,UTP9,UTP8,NOP10,KRE

33,NHP2,RMP1,MTR2,GSP1,POP7,NUG1,RIX7,RIO2,GSP2,RIO1,NMD3,NOP58,DIP2,NOP56,RNH70,FAP7,NOB1,UTP10,UTP15,UTP14,TIF6,UTP18,UTP13,UTP6,RNT1,CBF5,POP5,POP4,RCL1,NOP4,RPP1,NOG2,MEX67,NAN1,PWP2,HRR25,SDO1,CRM1,SNM1,BMS1,RAT1,REX2,FCF1,AFG2,POP3,POP1,LSG1,UTP22,POP6,EMG1,UTP21,MPP10,SNU13

ruMP cycle and formaldehyde assimilation ORIGINAL FUNCTIONS

GO:0044712:PGI1,GND1,FDH1,GND2 GO:0006098:PGI1,GND1,GND2

sNARE interactions in vesicular transport ORIGINAL FUNCTIONS

GO:0006906:SED5,BET1,SNC2,SSO1,SNC1,GOS1,SFT1,BOS1,TLG1,SEC22,TLG2,VTI1,YKT6,SEC9,VAM3,VAM7

GO:0016192:SED5,BET1,PEP12,SNC2,SEC20,SSO1,SNC1,GOS1,BOS1,SFT1,TLG1,SEC22,TLG2,VTI1,YKT6,USE1,VAM7,SEC9,VAM3,UFE1

GO:0006810:SED5,BET1,PEP12,SNC2,SEC20,SSO1,SNC1,GOS1,USE1,BOS1,SFT1,TLG1,SEC22,TLG2,VTI1,YKT6,SYN8,VAM7,SEC9,VAM3,UFE1

GO:0016043:SED5,BET1,PEP12,SNC2,SSO2,SSO1,SNC1,VAM3,GOS1,BOS1,SFT1,TLG1,SEC22,TLG2,VTI1,YKT6,VAM7,SEC9,UFE1,SPO20

GO:0044801:SED5,BET1,SNC2,SSO1,SNC1,GOS1,BOS1,SFT1,TLG1,SEC22,TLG2,VTI1,YKT6,VAM7,SEC9,VAM3,UFE1

sRPdependent cotranslational protein targeting to membrane ORIGINAL FUNCTIONS

GO:0002181:RPL9A,RPL9B,RPL7A,RPL29,RPL15B,RPL22A,RPL22B,RPL15A,RPS0A,RPS0B,RPL36A,RPL36B,RPL17A,RPL17B,RPL26A,RPL26B,RPP0,RPL2A,RPL6A,RPL6B,RPL31A,RPL31B,RPL33B,RPL33A

GO:0033036:RPS28A,SRP54,RPS10B,RPS15,RPS26A,SEC65,RPS19A,RPS26B,SRP68,RPS18B,SRP101,RPS5,RPS3,RPS2,SRP102,RPS19B,RPS28B,SRP72,RPS0A,RPS0B,RPS10A

GO:0042254:RPS13,RPS21A,RPS21B,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,RPS9B,RPS9A,RPS27A,RPL3,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL11B,RPL11A,RPS6A,RPS0A,RPS0B,RPL8B,RPL8A,RPL5,RPS16B,RPS20,RPS26A,RPL34A,RPS26B,RPS24B,RPP0,RPL10,RPS10A,RPS8B,RPL6A,RPL6B,RPS28B,RPS28A,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B

GO:0044085:RPS13,RPS21A,RPS21B,RPS17A,RPS10B,RPS15,RPS17B,RPS27B,SEC65,RPS9A,RPS27A,RPL3,RPS11B,RPS5,RPS3,RPS2,RPS31,RPS18B,RPL7B,RPL7A,RPL25,RPS19B,RPS19A,RPL11B,RPS9B,RPS6A,RPS0A,RPS0B,RPL8B,RPL8A,RPL5,RPS16B,RPS20,RPS26A,RPL34A,RPL11A,RPS26B,RPS24B,RPP0,SRP54,RPL10,RPS10A,RPS8B,RPL6A,RPL6B,RPS28B,RPS28A,RPL30,RPS1B,RPS1A,RPS7A,RPS23B,RPS7B,RPL12B

sUMOylation ORIGINAL FUNCTIONSGO:0016925:AOS1,UBC9,UBA2,SMT3

GO:0070647:AOS1,UBC9,UBA2,ULP1,SMT3

salvage pathways of adenine hypoxanthine and their nucleosides ORIGINAL FUNCTIONS GO:0043094:XPT1,PNP1,HPT1,APT1,AMD1,AAH1

GO:0043101:AAH1,XPT1,APT1,AMD1,HPT1

GO:0006144:XPT1,APT1,ADO1,AMD1,AAH1

salvage pathways of pyrimidine ribonucleotides ORIGINAL FUNCTIONS

GO:0034654:FUR1,FCY1,URK1,URH1,YNK1,CDD1

GO:0008655:URH1,URK1,FCY1,CDD1,FUR1

selenocompound metabolism ORIGINAL FUNCTIONS

GO:0006520:MET3,MET6,MSM1,STR3,STR2,MES1,IRC7,CYS3

GO:0006790:MET3,MET6,STR3,STR2,YLL058W,YML082W,IRC7,CYS3

UNCOVERED GENES TRR1,TRR2

serineisocitrate lyase pathway ORIGINAL FUNCTIONS

GO:0044712:ENO1,ENO2,CIT3,MDH3,GPM1

GO:0006091:CIT3,ENO1,ENO2,GPM1,MDH1,ACO1 GO:0006730:SHM2,SHM1

GO:0006094:ENO1,ENO2,GPM1,MDH2

sphingolipid biosynthesis ORIGINAL FUNCTIONS

GO:0006665:LIP1,TSC10,AUR1,IPT1,CSG2,LAG1,LCB1,LAC1,LCB2,SUR2,SUR1,SCS7

sphingolipid de novo biosynthesis ORIGINAL FUNCTIONS

GO:0006665:LAG1,LCB1,LAC1,LCB2,YSR3,DPL1 GO:0019722:LCB3,DPL1

sphingolipid recycling and degradation ORIGINAL FUNCTIONS

GO:0006665:YPC1,LCB5,LCB4,YSR3,DPL1,YDC1

GO:0019722:LCB3,LCB5,LCB4,DPL1

spliceosome ORIGINAL FUNCTIONS

GO:0016071:SNU66,SYF2,PRP22,RSE1,ECM2,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,DBP2,PRP38,CWC15,CBC2,PRP6,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

GO:0044260:SNU66,SYF2,PRP22,SSA1,SSA3,RSE1,SSA4,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,SSB2,SSA2,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,ECM2,HSH49,DIB1,CEF1,SNU13,DBP2,LSM3,PRP38,CWC15,CBC2,PRP6,FAL1,YRA1,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

GO:0090304:SNU66,SYF2,PRP22,RSE1,ECM2,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,SSB2,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,DBP2,PRP38,CWC15,CBC2,PRP6,FAL1,YRA1,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

GO:0010467:SNU66,SYF2,PRP22,SSA1,RSE1,ECM2,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,SSB2,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,DBP2,PRP38,CWC15,CBC2,PRP6,FAL1,YRA1,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

GO:0000377:SNU66,SYF2,PRP22,RSE1,ECM2,SLU7,PRP6,PRP4,LEA1,PRP2,PRP3,PRP18,YSF3,PRP8,PRP9,CUS1,LSM8,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,SMX2,SMX3,PRP16,SME1,SNU114,PRP19,PRP31,LSM4,LSM5,LSM6,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,PRP38,CWC15,SNU23,STO1,SMB1,CBC2,PRP46,PRP45,PRP43,ISY1,PRP40,BUD31,BRR2,CDC40,SMD2,SMD3

GO:0006396:SNU66,SYF2,PRP22,RSE1,ECM2,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,SSB2,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,DBP2,PRP38,CWC15,CBC2,PRP6,FAL1,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

GO:0006397:SNU66,SYF2,PRP22,RSE1,ECM2,THO2,SNU23,PRP4,LEA1,PRP2,PRP3,LSM8,SLU7,YSF3,PRP8,PRP9,CUS1,SUB2,SNP1,YHC1,SYF1,HSH155,CLF1,SAD1,SMD1,RDS3,BUD31,SMX2,SMX3,LSM6,SME1,PRP18,PRP19,PRP31,LSM4,LSM5,PRP16,LSM7,PRP11,LSM2,LSM3,HSH49,DIB1,CEF1,SNU13,PRP38,CWC15,CBC2,PRP6,STO1,SMB1,PRP46,PRP45,PRP43,ISY1,PRP40,SNU114,BRR2,CDC40,SMD2,SMD3

sulfate assimilation pathway II ORIGINAL FUNCTIONS

GO:0000103:MET10,MET3,MET14,MET16,MET5

GO:0006790:MET10,MET3,MET14,MET16,MET17,MET5

sulfate reduction I assimilatory ORIGINAL FUNCTIONSGO:0080058:TRX2,TRX1

GO:0034599:TRX2,TRX3,TRX1

GO:0000103:TRX2,MET3,MET14,MET16,MET5,MET10

sulfur degradation ORIGINAL FUNCTIONSGO:0000096:STR3,MET3,MET6

superpathway of Glutamate Biosynthesis ORIGINAL FUNCTIONS

GO:0009084:GLT1,IDP1,GDH1,GLN1,GDH3

GO:0006537:GLT1,IDP1,GDH1,GDH3

GO:0019752:GDH1,GDH3,GLT1,IDP3,IDP2,IDP1,GLN1

superpathway of acetoin and butanediol biosynthesis ORIGINAL FUNCTIONS

GO:0000949:PDC1,PDC5,PDC6 GO:0046165:PDC1,BDH1,PDC5,PDC6

GO:0009082:ILV6,ILV2

superpathway of allantoin degradation in ORIGINAL FUNCTIONS

GO:0043605:DUR1,2,DAL1,DAL2

superpathway of chorismate metabolism ORIGINAL FUNCTIONS

GO:0042181:COQ5,COQ6,CAT5,COQ3,COQ2

GO:0044283:ARO2,CAT5,COQ5,COQ6,COQ3,COQ2

superpathway of geranylgeranyldiphosphate biosynthesis I via mevalonate

ORIGINAL FUNCTIONS

GO:0006694:ERG8,MVD1,ERG20,ERG12,ERG13,ERG10,HMG1,HMG2

GO:0008610:IDI1,ERG8,MVD1,ERG20,ERG12,ERG13,ERG10,BTS1,HMG1,HMG2

superpathway of heme biosynthesis ORIGINAL FUNCTIONS

GO:0006778:HEM12,HEM13,HEM14,HEM15,HEM2,HEM3,HEM1,HEM4

superpathway of serine and glycine biosynthesis I ORIGINAL FUNCTIONS

GO:0006730:SHM2,SHM1 GO:0009070:SER33,SHM1,SER2,SER3,SER1

superpathway of tetrahydrofolate biosynthesis and salvage ORIGINAL FUNCTIONS GO:0008652:FOL1,FOL3,FOL2,DFR1,MIS1,ABZ2,ABZ1

GO:0042559:FOL1,FOL3,FOL2,DFR1,ABZ2,MIS1

GO:0019438:FOL1,FOL3,FOL2,ADE8,DFR1,MIS1,ABZ2,ABZ1

switching of origins to a postreplicative state ORIGINAL FUNCTIONS

GO:0036388:ORC1,MCM7,MCM6,MCM5,MCM4,MCM3,MCM2,CDC6

GO:0022607:RPS31,CDC28,ORC1,MCM7,MCM6,MCM5,MCM4,MCM3,MCM2,CDC6 GO:0006302:CDC28,MCM7,MCM6,MCM5,MCM4,MCM3,MCM2 synthesis and interconversion of nucleotide di and triphosphates ORIGINAL FUNCTIONS GO:0009148:CDC8,URA7,URA8

GO:0072528:URA6,URA7,CDC8,URA8 GO:0009133:CDC8,ADK1

UNCOVERED GENES GLR1

synthesis of IP2 IP and Ins in the cytosol ORIGINAL FUNCTIONS

GO:0016311:INM2,INM1,INP51,INP53,INP52

GO:0019751:INM2,INM1,INO1

synthesis of PC ORIGINAL FUNCTIONSGO:0045017:PCT1,PAH1,OPI3

GO:1901615:PCT1,CAT2,OPI3

synthesis of PE ORIGINAL FUNCTIONSGO:0045017:MUQ1,PSD1,PAH1

synthesis of PIPs at the early endosome membrane ORIGINAL FUNCTIONS

GO:0046488:VAC14,FIG4,FAB1,VPS34,LSB6 GO:0034243:VPS15,VPS34

synthesis of PIPs at the late endosome membrane ORIGINAL FUNCTIONS

GO:0046488:VAC14,FIG4,FAB1,VPS34,YMR1GO:0034243:VPS15,VPS34

synthesis of PIPs at the plasma membrane ORIGINAL FUNCTIONS

GO:0046488:LSB6,INP51,YMR1,INP53,INP52

GO:0046856:INP51,YMR1,INP53,INP52

synthesis of glycosylphosphatidylinositol GPI ORIGINAL FUNCTIONS

GO:0006506:MCD4,GPI18,GWT1,GPI10,GPI12

tCA Cycle biocyc ORIGINAL FUNCTIONSGO:0035383:CIT1,LSC2,LSC1

GO:0000002:ACO1,KGD2 GO:0019319:MDH2,PYC2,PYC1

GO:0045333:CIT3,SDH1,SDH3,SDH2,SDH4,MDH1,ACO1,FUM1

GO:0055114:CIT3,SDH1,SDH3,SDH2,SDH4,MDH3,MDH1,ACO1,FUM1 tRNA splicing ORIGINAL FUNCTIONSGO:0008033:SEN54,SEN1,TPT1

GO:0006388:SEN54,TPT1 UNCOVERED GENES CPD1

taurine and hypotaurine metabolismORIGINAL FUNCTIONSGO:0009063:ECM38,GAD1

UNCOVERED GENES GDH2

thiamin diphosphate biosynthesis IV eukaryotes ORIGINAL FUNCTIONS

GO:0072527:THI80,PHO3,THI6 UNCOVERED GENES PHO5,DIA3

thiamine metabolism ORIGINAL FUNCTIONSGO:0072528:THI80,THI20,THI21,THI6

UNCOVERED GENES NFS1

threonine and Methionine biosynthesis ORIGINAL FUNCTIONS

GO:0009066:MET2,MET6,MET17,HOM6,HOM2,HOM3,THR4,THR1

threonine degradation ORIGINAL FUNCTIONSGO:0006567:CHA1,ILV1

GO:0009069:STR2,CHA1,CYS3 GO:0006790:YML082W,STR2,CYS3

threonine degradation IV ORIGINAL FUNCTIONSGO:0019413:ALD6,ALD4

GO:0046394:ALD6,ALD4,GLY1

tollLike Receptors Cascades ORIGINAL FUNCTIONS

GO:0023052:PBS2,KSS1,HOG1,FUS3 GO:0006888:ERP2,ERP4

GO:0016192:ERP3,ERP2,ERP4

transmission across Chemical Synapses ORIGINAL FUNCTIONS

GO:0019752:ALD6,ALD5,ALD4,UGA2,UGA1,CAT2,GLN1

GO:0032787:ALD6,ALD5,ALD4,UGA2,UGA1,CAT2

transport of vitamins nucleosides and related molecules ORIGINAL FUNCTIONS

GO:0015711:YEA4,FAT1 GO:1901264:YEA4,FUN26

trehalose Degradation Low Osmolarity ORIGINAL FUNCTIONS

GO:0044275:NTH2,ATH1,NTH1,GLK1 GO:0005993:NTH2,ATH1,NTH1

GO:0005991:NTH2,ATH1,TPS2,NTH1

tryptophan biosynthesis ORIGINAL FUNCTIONS

GO:0000162:TRP2,TRP3,TRP1,TRP4,TRP5

tyrosine metabolism ORIGINAL FUNCTIONS

GO:0006520:ADH3,ADH2,ADH1,AAT1,ADH5,ADH4,ARO8,ARO9,ALD3,ALD2,UGA2,HIS5,SFA1,AAT2 GO:0000947:ADH3,ADH2,ADH1,ADH5,SFA1,ADH4

GO:1901565:ADH3,ADH2,ADH1,ADH5,ADH4,ALD3,ALD2,UGA2,SFA1

UNCOVERED GENES RMT2

uDPGlucose Conversion ORIGINAL FUNCTIONSGO:0015980:UGP1,PPA2

UNCOVERED GENES GAL10

uMP biosynthesis ORIGINAL FUNCTIONS

GO:0006207:URA10,URA4,URA5,URA2,URA3

GO:1901566:URA4,URA5,URA2,URA3,CPA2,CPA1,URA10

ubiquinone Biosynthesis ORIGINAL FUNCTIONS

GO:0044711:ERG20,CAT5,COQ5,BTS1,COQ6,COQ1,COQ3,COQ2

GO:1901663:CAT5,COQ5,COQ6,COQ1,COQ3,COQ2

GO:0044283:ERG20,CAT5,COQ5,COQ6,COQ1,COQ3,COQ2

ubiquinone and other terpenoidquinone biosynthesis ORIGINAL FUNCTIONS

GO:0042181:COQ5,COQ6,CAT5,COQ3,COQ2

GO:0044281:ARO8,CAT5,COQ5,COQ6,COQ3,COQ2

urea cycle ORIGINAL FUNCTIONSGO:0006591:ARG2,CAR1,ARG4

GO:0006525:CPA2,ARG1,CAR1,ARG4,ARG3

valine leucine and isoleucine degradation ORIGINAL FUNCTIONS

GO:0046395:EHD3,UGA1,POT1,BAT1,LPD1,BAT2

GO:0032787:EHD3,ALD6,ALD5,ALD4,UGA1,POT1,LPD1

GO:0044283:ERG10,ALD6,ALD5,ALD4,BAT2,BAT1,LPD1,ERG13 UNCOVERED GENES IRC15

various types of Nglycan biosynthesis ORIGINAL FUNCTIONS

GO:1901137:ALG9,ALG2,ALG3,ALG1,STT3,ANP1,KTR6,VAN1,OCH1,OST6,OST4,OST5,OST2,OST3,OST1,ALG11,ALG12,ALG13,ALG14,MNN9,MNN2,MNN1,MNN5,WBP1,SWP1,MNN11,MNN10

GO:0009100:ALG9,ALG3,ALG1,STT3,ANP1,KTR6,VAN1,OCH1,OST6,OST4,OST5,OST2,OST3,OST1,MNS1,ALG11,ALG12,MNN9,MNN2,MNN1,MNN5,WBP1,SWP1,MNN11,MNN10

GO:0044267:ALG9,ALG3,ALG1,STT3,ANP1,KTR6,VAN1,OCH1,OST6,OST4,YLR057W,OST2,OST3,OST1,MNS1,ALG11,ALG12,MNN9,OST5,MNN2,MNN1,MNN5,WBP1,SWP1,MNN11,MNN10

vitamin B2 riboflavin metabolism ORIGINAL FUNCTIONSGO:0009141:NPP2,NPP1

GO:0006753:FMN1,NPP2,NPP1

vitamin B6 metabolism ORIGINAL FUNCTIONS

GO:0008614:SNO2,SNO3,SNZ3,SNO1,SNZ1

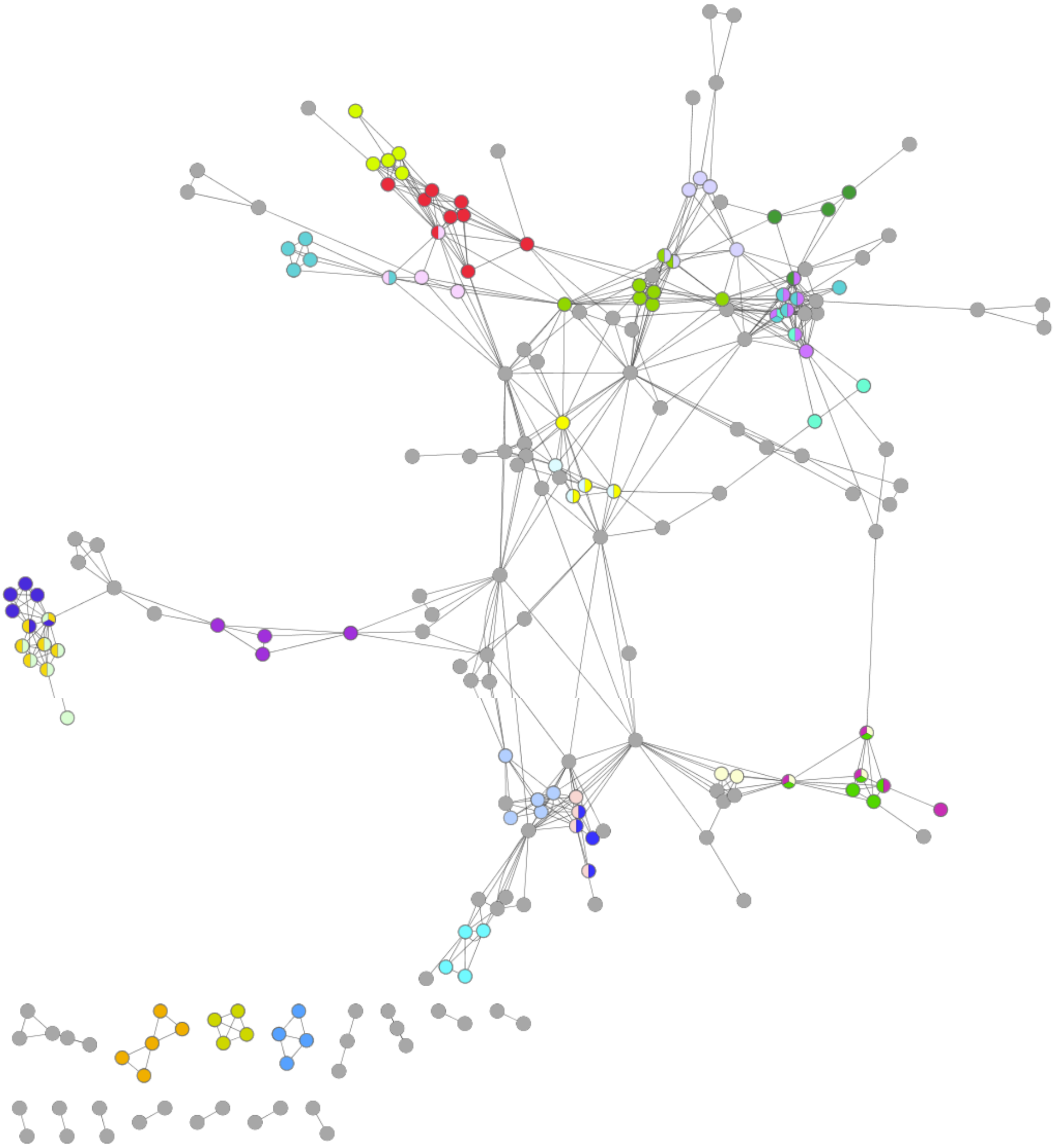
GO:0072524:BUD16,SNZ1,SNO3,SNZ3,SNO1,SNO2

GO:1901566:BUD16,SER1,SNZ1,SNO3,SNZ3,SNO1,SNO2,THR4

GO:0007105:BUD17,BUD16

zymosterol biosynthesis ORIGINAL FUNCTIONS

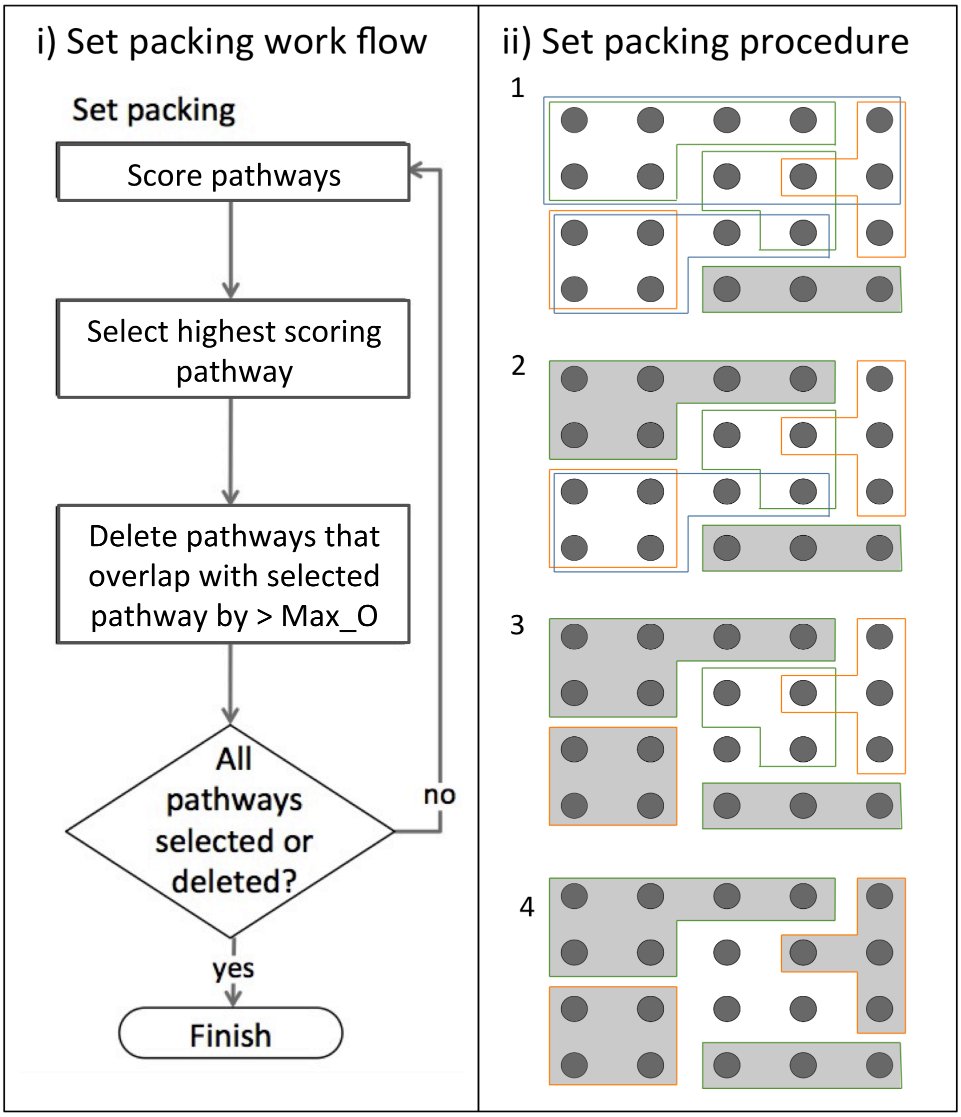
GO:0006696:ERG24,ERG27,ERG26,ERG25,ERG11



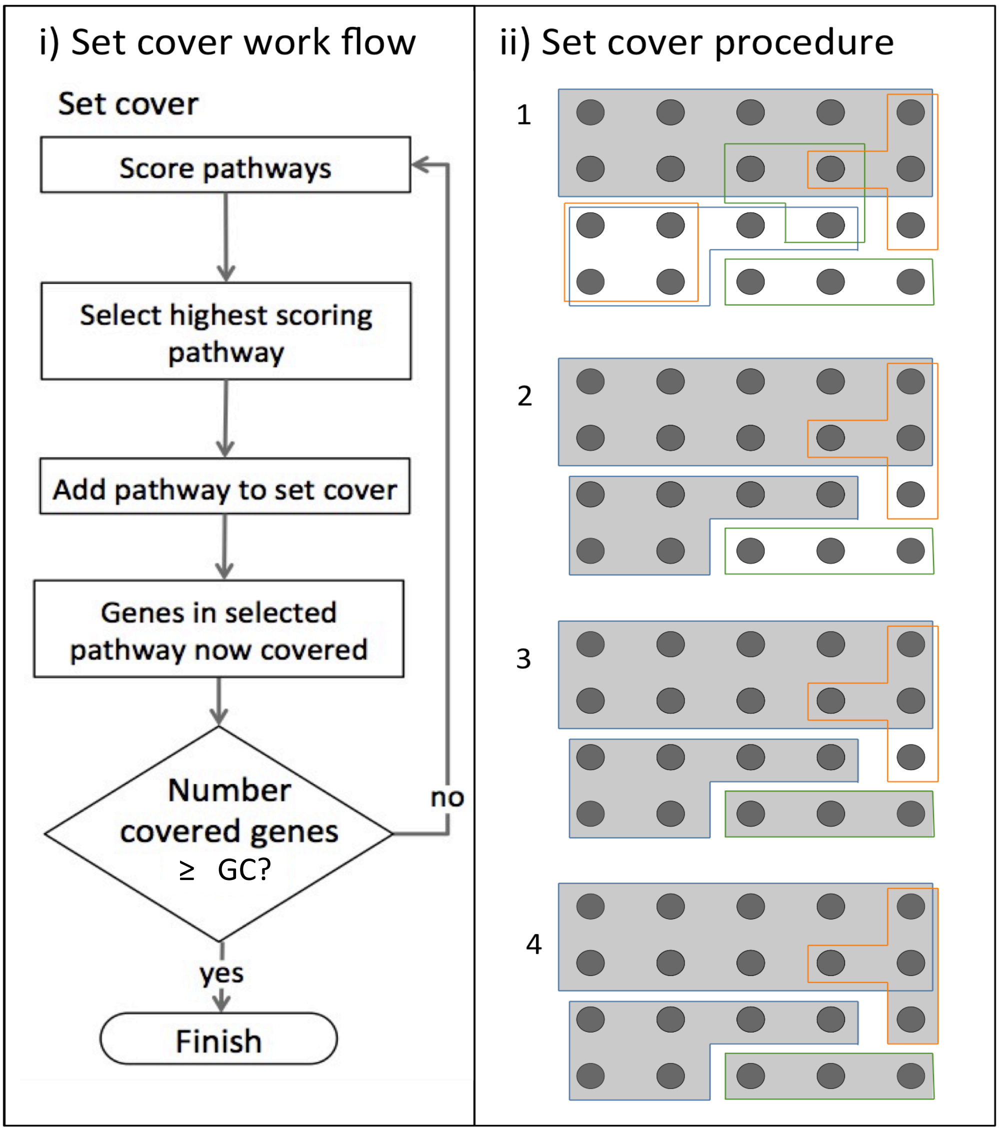
Supplementary Figure 1 Network clusters created using ONECLUST.

Appendix B

Supplementary materials for Reducing pathway redundancy using set theory algorithms


i) Set packing work flow
Set packing
Score pathways
Select highest scoring pathway
Delete pathways that overlap with selected pathway by > Max_O
All pathways selected or deleted?
no
yes
Finish
ii) Set packing procedure
1
2
3
4

Supplementary Figure 2 A Set packing visualisation



Supplementary Figure 3 Set cover visualisation

Supplementary Table 2: Enriched pathways from the osteoarthritis dataset (p-value<0.05). The set cover column indicated the 23 pathways that were included in the set cover.

p-value	Set cover	pathway
0.0000	1	extracellular matrix organization
0.0000	0	collagen biosynthesis and modifying enzymes
0.0000	0	collagen formation
0.0000	1	gpcr signaling g alpha q
0.0000	1	signal transduction
0.0000	1	protein digestion and absorption homo sapiens human
0.0000	1	pathways in cancer homo sapiens human
0.0000	0	gpcr signaling cholera toxin
0.0000	0	gpcr signaling pertussis toxin
0.0000	0	class b 2 secretin family receptors
0.0000	0	gpcr ligand binding
0.0001	0	gpcr signaling g alpha s epac and erk
0.0001	0	gpcr signaling g alpha s pka and erk
0.0003	0	integrin cell surface interactions
0.0003	1	vitamin d receptor pathway
0.0004	0	signaling by gpcr
0.0006	1	integrin
0.0006	0	basal cell carcinoma homo sapiens human
0.0010	0	ecm proteoglycans
0.0015	1	wnt signaling network
0.0016	1	o linked glycosylation
0.0022	1	ecm receptor interaction homo sapiens human
0.0027	1	small cell lung cancer homo sapiens human
0.0067	1	wnt signaling pathway
0.0082	0	degradation of the extracellular matrix
0.0115	1	signaling pathways regulating pluripotency of stem cells homo sapiens
0.0136	1	beta1 integrin cell surface interactions
0.0142	1	complement and coagulation cascades homo sapiens human
0.0146	1	cell adhesion molecules cams homo sapiens human
0.0262	1	pi3k akt signaling pathway homo sapiens human
0.0354	0	collagen degradation
0.0381	0	wnt5a dependent internalization of fzd2 fzd5 and ror2
0.0389	1	hippo signaling pathway homo sapiens human
0.0415	0	gpcr downstream signaling
0.0415	1	benzo a pyrene metabolism
0.0430	0	o linked glycosylation of mucins
0.0430	1	axon guidance
0.0430	1	prostaglandin synthesis and regulation
0.0436	0	activation of trka receptors
0.0436	1	neuroactive ligand receptor interaction homo sapiens human
0.0445	0	small ligand gpcrs
0.0453	1	wnt signaling pathway and pluripotency

Appendix C

Supplementary materials for Mapping biological process relationships and disease perturbations within a pathway network

Supplementary Table 3: disease terms used to remove disease pathways from the dataset.

- action pathway
- addiction
- aflatoxin
- alcohol
- allograft rejection
- amphetamine
- amyloids
- anaesthetic
- anthrax
- anxiety
- argininemia
- arsenate
- arthritis
- asthma
- bacterial
- biotin
- bipolar
- blastoma
- botulinum
- cancer
- carcinoma
- cardiomyopathy
- carnosinemia
- cholerae
- clostridium
- constitutive
- cystic fibrosis
- cytoma
- deficiency
- depression
- diabete
- diphtheria
- disease
- disorder
- disulfiduria
- drug
- epsteinbarr
- escherichia
- ethanol
- fanconi anemia
- glioma
- hepatitis
- hereditary
- heroin
- herpes
- hiv
- htlvi
- hypertrophy
- hypophosphatasia
- iasis
- immunodeficiency
- infection
- infertility
- influenza
- intolerance
- legionellosis
- leishmaniasis
- leukemia
- longterm depression
- lupus
- malaria
- measles
- melanoma
- morphine cocaine
- mutant
- nicotine
- obesity
- obsess
- pathogenic
- pathological
- pertussis
- pharmacodynamics
- pharmacokinetics
- protection
- resistance
- salmonella
- schizophrenia
- sclerosis
- shigellosis
- spinal cord injury
- staphylococcus
- substance abuse
- susceptibility
- syndrome
- tetanus
- toma
- toxin
- toxoplasmosis
- trypanosomiasis
- tuberculosis
- tumor
- tumour
- uria
- vibrio
- viral
- virion
- virus
- west nile