

INFERRING CONDITION-SPECIFIC REGULATORY NETWORKS WITH SMALL SAMPLE SIZES: A
CASE STUDY IN *Bacillus subtilis* AND *Mus musculus* INFECTION BY THE PARASITE
Toxoplasma gondii.



Clare Elizabeth Pacini

DOWNING COLLEGE
UNIVERSITY OF CAMBRIDGE

THIS DISSERTATION IS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

September, 2016

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

As referenced in the text this work includes a testable experimental hypothesis identified by Orr Yarkoni of the Ajioka lab in the Pathology department, University of Cambridge. The RNA-seq data analysed in this thesis, as discussed in the text, was produced by Lalitha Sundaram also of the Ajioka lab.

Acknowledgements

First to my supervisor Gos Micklem for being someone I can always talk to and for trusting me enough to allow me to follow my own research directions. Thank you for being there when needed, giving your time, advice and support that meant the completion of this thesis. Jim Ajioka, for providing support and advice throughout my PhD and for continual kindness and enthusiasm. My advisor Simon Tavaré, for invaluable advice and having a calm and positive outlook that was always reassuring. Thank you to the Maths department, the EPSRC and Downing College for providing me with the opportunity and support to undertake this PhD. Many thanks also to members of the Pathology department, for sharing cakes, sweets, research advice and support including, Lalitha Sundaram, Orr Yarkoni, Bo Shiun, Paul Dear and Pete Davenport.

To my parents, for unconditional love and support in everything that I do and without whom this thesis wouldn't have been possible. Thank you for the many visits to Cambridge, to my Dad for lively DIY discussions and my Mum for expert gardening and trying to impart the difference between drinking tea in cups and mugs. My sister Adele, role model, best friend and plotter-in-chief, thank you, I couldn't ask for a better sister. Thanks to Ben for always welcoming me to the farm, providing cups of tea, gardening tips and unusual desserts. Special thanks to my niece Tilly for endless laughs and happiness and reminding me that there are other things in life than work, mainly mud and stickers.

My friend Caroline, the most genuine person I know, thank you for being someone I can always count on and for providing a much-needed Yorkshire perspective on life. To Cindy, for being there the last four years through every high, low and funny part in between. Thank you for your friendship, sharing day trips and unruly car stereos with me. I can't imagine having done this PhD without you there.

To Jackie, thank you for your unwavering support, faith and advice throughout. Thanks to Fran, for being an effusive Italian, loyal and caring, and introducing me to the wonders of a multi-course Italian wedding. Robin, thank you for your friendship, support and company over many years. Finally, to Kyle and Cindy for Bella and Rocket who are an endless source of entertainment.

Abstract

Modelling interactions between genes and their regulators is fundamental to understanding how, for example a disease progresses, or the impact of inserting a synthetic circuit into a cell. We use an existing method to infer regulatory networks under multiple conditions: the Joint Graphical Lasso (JGL), a shrinkage based Gaussian graphical model. We apply this method to two data sets: one, a publicly available set of microarray experiments perturbing the gram-positive bacteria *Bacillus subtilis* under multiple experimental conditions; the second, a set of RNA-seq samples of Mouse (*Mus musculus*) embryonic fibroblasts (MEFs) infected with different strains of the parasite *Toxoplasma gondii*. In both cases we infer a subset of the regulatory networks using relatively small sample sizes. For the *Bacillus subtilis* analysis we focused on the use of these regulatory networks in synthetic biology and found examples of transcriptional units active only under a subset of conditions, this information can be useful when designing circuits to have condition dependent behaviour. We developed methods for large network decomposition that made use of the condition information and showed a greater specificity of identifying single transcriptional units from the larger network using our method. Through annotating these results with known information we were able to identify novel connections and found supporting evidence for a selection of these from publicly available experimental results. Biological data collection is typically expensive and due to the relatively small sample sizes of our MEF data set we developed a novel empirical Bayes method for reducing the false discovery rate when estimating block diagonal covariance matrices. Using these methods we were able to infer regulatory networks for the host infected with either the ME49 or RH strain of the parasite. This enabled the identification of known and novel regulatory mechanisms. The *Toxoplasma gondii* parasite has shown to subvert host function using similar mechanisms as cancers and through our analysis we were able to identify genes, networks and ontologies associated with cancer, including connections that have not previously been associated with *T. gondii* infection. Finally a Shiny application was developed as an online resource giving access to the *Bacillus subtilis* inferred networks with interactive methods for exploring the networks including expansion of sub networks and large network decomposition.

Contents

1	<i>Network Biology</i>	1
1.1	<i>Transcriptional regulation</i>	3
1.2	<i>Measuring transcription</i>	5
1.3	<i>Regulatory networks</i>	6
1.4	<i>Mathematical analysis of regulatory networks</i>	8
1.5	<i>Regulatory models for a single condition</i>	10
1.5.1	<i>Boolean and logic models</i>	11
1.5.2	<i>Correlation, co-expression methods</i>	12
1.5.3	<i>Information-theoretic</i>	13
1.5.4	<i>Bayesian</i>	15
1.5.5	<i>Dynamic</i>	16
1.6	<i>Differential network methods</i>	16
1.7	<i>Modelling strengths and weaknesses</i>	19
2	<i>Synthetic Biology</i>	21
3	<i>Bacillus Subtilis and GGMs</i>	27
3.1	<i>Results and Discussion</i>	27
3.1.1	<i>Parameter selection</i>	36
3.1.2	<i>Analysis of effects of data inputs on JGL algorithm</i>	40
3.1.3	<i>Agglomerative clustering for JGL</i>	45
3.1.4	<i>Exploring subnetworks</i>	48

3.2	<i>Decomposing large networks</i>	51
3.2.1	<i>Affinity propagation clustering</i>	54
3.2.2	<i>Deterministic network separation</i>	56
3.2.3	<i>Simulation methods for splitting networks</i>	60
3.3	<i>Future analysis</i>	63
3.3.1	<i>Example networks</i>	63
3.3.2	<i>Experimental conditions</i>	67
3.4	<i>Conclusions</i>	69
3.5	<i>Methods</i>	78
3.5.1	<i>Mathematical preliminaries</i>	78
3.5.2	<i>Joint graphical lasso</i>	79
3.5.3	<i>Subnetwork analysis</i>	87
3.5.4	<i>Clustering for JGL</i>	88
3.5.5	<i>Network annotation and evaluation</i>	89
3.5.6	<i>Decomposing large networks</i>	89
4	<i>Empirical Bayes method for estimating covariances</i>	95
4.1	<i>Exploratory data analysis</i>	96
4.2	<i>Mathematical preliminaries</i>	99
4.3	<i>Empirical Bayes model</i>	100
4.3.1	<i>Calculating hyperparameters</i>	101
4.3.2	<i>Simulated data</i>	105
4.4	<i>Results and Discussion</i>	106
4.5	<i>Implementation</i>	107
5	<i>Toxoplasma gondii and GGMs</i>	109
5.1	<i>Introduction</i>	109
5.2	<i>The Hallmarks of Cancer</i>	111
5.3	<i>Cancer metabolism</i>	112
5.4	<i>Metabolism in Toxoplasma gondii</i>	115
5.5	<i>Results and Discussion</i>	118

5.5.1	<i>Experimental data</i>	118
5.5.2	<i>Aligning RNA-seq reads</i>	118
5.5.3	<i>Pre-processing the data</i>	120
5.5.4	<i>Initial Data analysis, parameter selection</i>	121
5.5.5	<i>Network annotation analysis</i>	121
5.5.6	<i>Empirical Bayes, aiding interpretability</i>	124
5.5.7	<i>Evaluating the network</i>	131
5.5.8	<i>Annotating the network</i>	134
5.5.9	<i>Functional and disease Networks</i>	142
5.6	<i>Conclusions</i>	153
5.7	<i>Methods</i>	163
5.7.1	<i>Aligning reads</i>	163
5.7.2	<i>Converting from reads to counts</i>	163
5.7.3	<i>Comparing the mouse genome to joint mouse and Toxoplasma genomes</i>	164
5.7.4	<i>Comparing the RH and GT1 strains</i>	165
5.7.5	<i>Comparing read counts to technical or biological factors</i>	167
5.7.6	<i>Analysis of block size</i>	173
5.7.7	<i>GO analysis of networks</i>	176
5.7.8	<i>P-value analysis</i>	176
6	<i>Web Application</i>	179
6.1	<i>Overview of BSN</i>	183
6.1.1	<i>Examples of network analysis</i>	186
6.1.2	<i>Differential expression network</i>	189
6.1.3	<i>Large network decomposition</i>	192
6.1.4	<i>Analyse new data</i>	193
6.1.5	<i>Exporting data</i>	195
6.2	<i>Code outline</i>	195
6.3	<i>Future developments</i>	196
	<i>Bibliography</i>	199

List of Figures

1.1	DNA double helix	1
1.2	Central Dogma of Molecular Biology	3
1.3	Regulatory network of multiple regulatory modules	7
1.4	Graphical relationship between parent and child nodes	10
3.1	Clustering of <i>B. subtilis</i> microarray samples	28
3.2	Heatmap of gene correlations in random and block diagonal order	30
3.3	Example subnetwork from JGL algorithm	33
3.4	Examples of subnetworks annotated with sigma factor data	35
3.5	Subnetwork under different shrinkage parameters	37
3.6	ECDF of node degrees for different shrinkage values	39
3.7	Runtime for JGL algorithm of five iterations of inverting covariance matrix	41
3.8	Maximum block size of covariance matrices for leave one out data sets	42
3.9	ECDF of block sizes for leave one out data input	43
3.10	Plot of maximum block size for different shrinkage parameters for different classes	45
3.11	Plot of number of blocks for different shrinkage parameters for different classes	46
3.12	Relationship between block size and shrinkage parameters	47
3.13	Example subnetwork after being expanded with 30 extra nodes.	49
3.14	The largest subnetwork from the JGL output.	53
3.15	Plots of networks decomposed using affinity propagation clustering with different penalty values	55
3.16	Plots of deterministic splitting of networks with different cutoff values.	57
3.17	Deterministic splitting of large networks under cutoff value 2 with different class conditions.	58
3.18	Assigning a node to a class for the decomposed network	60
3.19	Result of best fixing graph simulation from 50000 samples.	61
3.20	Example subnetwork found following network decomposition using edge conditions and a simulation method for separating the network.	62

3.21	Two examples of sub networks that have two known transcriptional units in them.	64
3.22	Subnetwork where all edges are in one class	66
3.23	An example of a network where the edges were found to be present in all the experimental classes.	67
3.24	Venn diagram of genes in three JGL models	69
3.25	Screening rules are used to identify the block diagonal structure. Dark blue squares represent correlations passing the shrinkage parameter thresholds. This stylistic representation shows the combining of block diagonal structure using the JGL screening for two classes. The block diagonal structure is determined for each class separately and then combined to form a single block diagonal structure, with significant correlations in dark blue.	84
3.26	Stylistic example of how the simulation based network decomposition is performed	93
4.1	ECDF for Pearson correlations	97
4.2	Ranks of Pearson correlation values	98
5.1	Warburg effect	114
5.2	Toxoplasma JGL network, edges coloured using miRNA target information	123
5.3	Toxoplasma JGL network using the EB covariance matrices, and miRNA edge information.	126
5.4	P-value plots for ME49 sample and EB correlation matrices	129
5.5	P-value plots for RH sample and EB correlation matrices	130
5.6	Selection frequencies for the ME49 and RH bootstrap samples for the edges in the EB network.	131
5.7	Selection frequencies of edges in a model with shrinkage parameter $\lambda_1 = 0.88$ for the ME49 and RH bootstrap samples	133
5.8	Glycolysis metabolism results	138
5.9	Zfp36 and Ier3 subnetwork	141
5.10	Ribosome Biogenesis subnetwork	143
5.11	Cell Motility networks	143
5.12	Overview of network with Cancer associated genes	146
5.13	Subnetwork containing a genes associated with cancer	147
5.14	Example subnetwork of genes connected to <i>Trib3</i> , <i>Akt3</i> or <i>Foxo3</i>	149
5.15	KEGG pathway for ERBB signalling	151
5.16	Subnetwork for genes connected to Gprc5a	153
5.17	Example of RNA sequence differences	163
5.18	Compare read counts for Mouse and Mouse and GT1 combined genomes	165
5.19	Figure of total read count for the samples aligned to RH or GT1 strain	167
5.20	Comparing mapping to Mouse genes with RH and GT1 strain	168
5.21	Read counts for RNA-seq data, by experimental factors	169

5.22	Histogram of the total read counts for the individual samples.	170
5.23	MA plots for ME49 strain before and after cqn	171
5.24	MA plots for RH strain before and after cqn.	172
5.25	MA plots for ME49 strain before and after cqn, coloured according to GC content.	172
5.26	MA plots for RH strain before and after cqn, coloured according to GC content.	173
5.27	Maximum block size for <i>T. gondii</i> infected MEF cells	174
5.28	Number of blocks for <i>T. gondii</i> infected MEF cells	174
5.29	Maximum block size for a subset of parameter values	175
5.30	Number of blocks for a subset of parameter values	176
6.1	Example of using the text based search to find a gene in the BSN network.	186
6.2	Selecting genes to expand the network around using BSN.	187
6.3	Output in BSN after performing PubMed search on selected genes.	189
6.4	Example sub network selected for expansion using BSN	190
6.5	Example sub network after expansion using BSN	190
6.6	Overlaying uploaded differential expression data onto the BSN network.	191
6.7	Example showing selection of network to decompose in BSN	192
6.8	The resulting network after using the simulation method to split a cluster in BSN	193
6.9	Analysing new data using BSN	194
6.10	Overview of the code for BSN	195
6.11	Graph structure of gene ontologies	197

List of Tables

3.1	Statistics of empirical p-values for transcriptional unit information in the network	31
3.2	Proportion of genes connected to their transcriptional unit	32
3.3	Percentage of the genes in a transcriptional unit included in the model	32
3.4	Statistics on model properties for different shrinkage values	38
3.5	Transcriptional unit information for networks with different shrinkage parameters	40
3.6	Maximum block sizes for each pair of data inputs.	46
3.7	Table of additional genes in expanded subnetwork, their known regulators and sigma factors.	50
3.8	Deterministic splitting networks scores for different parameter values	59
3.9	Comparison of JGL models	67
4.1	Simulation results for estimating block diagonal correlation matrices	107
5.1	Summary statistics of covariance eigenvalues	121
5.2	Table of over represented gene ontology terms within <i>Toxoplasma gondii</i> network.	135
5.3	Significant KEGG pathways in the network	137
5.4	Disease terms present in the <i>T. gondii</i> network	145
5.5	Comparison of regulatory network scope and number of input samples	159

Nomenclature

AIC	Akaikes information criterion
ATP	Adenosine Triphosphate
B. subtilis	Bacillus Subtilis
BIC	Bayesian information criterion
BINCO	Bootstrap Inference for Network COnstruction
BSN	<i>Bacillus subtilis</i> Networks
CDK	Cyclin dependent kinase
cDNA	complementary Deoxyribonucleic Acid
CH	Casein Hydrolysate
ChIP-chip	Chromatin immunoprecipitation chip
ChIP-seq	Chromatin immunoprecipitation sequencing
cpm	counts per million
cqn	conditional quantile normalisation
det	determinant
Dhfr	Dihydrofolate reductase
DNA	Deoxyribonucleic Acid
DO	Disease Ontology
DREAM	Dialogue on Reverse Engineering Assessment and Methods
EB	Empirical Bayes
ECDF	Empirical cumulative distribution function
FDR	False discovery rate
GEO	Gene Expression Omnibus

GGM Gaussian Graphical Model

GO Gene ontology

Gprc5a G Protein-Coupled Receptor Class C Group 5 Member a

GSEA Gene Set Enrichment Analysis

GTF Gene Transfer Format

GUI Graphical user interface

HIF1 α Hypoxia Inducible Factor 1, subunit α

HTML Hyper Text Markup Language

Ier3 Immediate Early Response 3

iff if and only if

JGL Joint Graphical Lasso

KEGG Kyoto Encyclopedia of Genes and Genomes

KKT Karush-Kuhn-Tucker

KO Knockout

LB Lysogeny broth

LDH Lactate Dehydrogenase

lncRNA long non-coding RNA

MEF Mouse Embryonic Fibroblast

MGI Mouse Genome Informatics

miRNA micro RNA

MLE Maximum Likelihood Estimator

MOI Multiplicity of Infection

mRNA messenger RNA

mtDNA Mitochondrial Deoxyribonucleic Acid

NAD Nicotinamide adenine dinucleotide

NC Normal Correlation

ODE Ordinary Differential Equations

ORFs Open Reading Frames

OXPHOS Oxidative phosphorylation
PCR Polymerase Chain Reaction
pdf probability distribution function
PDH Pyruvate Dehydrogenase
PV Parasitophorous Vacuole
RBS Ribosome Binding Site
Rhoc Ras Homolog Family Member C
RMA Robust Multi-Array Average
RNA Ribonucleic Acid
ROP Rhoptry
ROS Reactive oxygen species
RPKM Reads pre kilobase per million
sd standard deviation
shRNA Short hairpin Ribonucleic acid
SMM Spizizen minimal medium
SNP Single nucleotide polymorphism
TCA Tricarboxylic Acid
TPR True positive rate
Trib3 Tribbles Pseudokinase 3
VEGF Vascular Endothelial Growth Factor
WT wild-type
Zfp36 Zinc Finger Protein 36

Network Biology

OVER THE PAST FEW DECADES there have been huge advances in both the experimental methods for gathering data on cellular mechanisms and the computational tools used to analyse this information. These mechanisms include the processes by which a cell responds to stimuli and these stimuli can be either endogenous to or from outside the cell. Cells are the building blocks of organisms, providing physical structure and carrying out essential functions such as generating energy. Each cell within an organism is enclosed within a plasma membrane [Robertson, 1981]. Through the plasma membrane molecules are transported to and from the cell's environment. The plasma membrane forms a barrier between the external environment and the cell. This barrier is critical to the ability of the cell to respond to stimuli as a self-contained unit. These self-contained units can self-replicate; cells provide the hereditary information and material required to produce identical copies of themselves through cell division [Noireaux et al., 2011, Bell and Dutta, 2002, Sclafani and Holzen, 2007].

Living organisms can be classified into two basic types based on their cell types; prokaryotes and eukaryotes. Prokaryotes are simpler, lacking internal membrane bound structures such as the mitochondria and nuclei that are found in eukaryotic cells. The hereditary information, deoxyribonucleic acid (DNA) is found in different cellular locations for prokaryotes and eukaryotes. In prokaryotes, DNA is located within the cell cytoplasm. In contrast, eukaryotic DNA is primarily found within the nucleus. Some membrane-bound organelles carry their own genomes, for instance the mitochondria contain mitochondrial DNA (mtDNA). DNA has a double-stranded helical structure, first identified by Watson and Crick [Watson and Crick, 1953]. Each strand of DNA is a chain of nucleotides (adenosine A, guanine G, cytosine C, thymine T), including a sugar-phosphate backbone on the outside of the helical structure [Franklin and Gosling, 1953]. The two DNA strands are joined in complementary base pairing by hydrogen

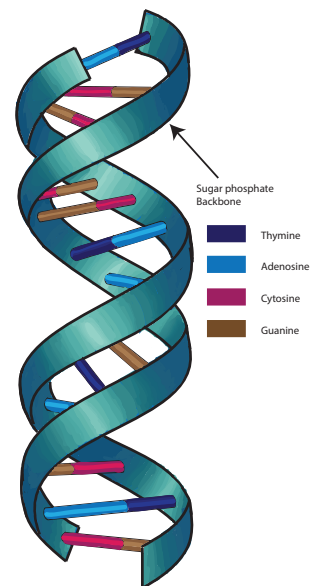


Figure 1.1: DNA has a double helical structure with the sugar phosphate backbone of the helix is on the outside of the structure. Nucleotides are joined in pairs by hydrogen bonds. These pairs are always guanine with cytosine and adenosine with thymine.

bonds. Complementary base pairing is the pairing of G with C bases and A with T bases across the two strands of DNA, Figure 1.1. The complete DNA sequence of an organism, its genome, is large and complex; the publication of the first full human DNA sequence contained over 3 billion bases [Lander et al., 2001].

The central dogma of molecular biology states that DNA is transcribed into ribonucleic acid (RNA) which can be translated into protein [Crick, 1970], Figure 1.2. In this case, the sequence order of the bases in DNA contains the coding information for a protein. In general, a gene is a DNA sequence that is transcribed to form a RNA transcript. RNAs can be either coding or non-coding; a coding RNA, or messenger RNA (mRNA) is translated to protein where a non-coding RNA is not. Coding RNAs are translated into three-dimensional protein structures. In eukaryotes, the enzyme, RNA Polymerase II transcribes DNA to mRNA in the 5' to 3' direction. In prokaryotes, sigma factors combine with RNA polymerase to form a holoenzyme that binds to a 'promoter' sequence upstream (5') of the gene and initiates transcription. RNA is a single stranded molecule like DNA but based on the ribose sugar rather than deoxyribose and with the thymine base replaced by uracil. Triplets of RNA bases, called codons, are translated into amino-acids by a large protein/RNA complex called the ribosome. The amino-acid chains created by translation fold into three dimensional structures. In this way, the heritable DNA sequence contains the information necessary to produce the proteins and RNAs required for cellular function and these in turn are responsible for mediating all other cellular functions.

A gene that has been transcribed into mRNA is said to be 'expressed'. The expression of a gene is required for activity of its corresponding RNA or protein. Typically, each cell contains the same DNA, however there is great diversity in the functions performed by different cell types [Gurdon, 1968]. Cellular processes are therefore coordinated through the activity of specific subsets of genes and different cells tend to express different combinations of cell type-specific genes. In addition, the activity or expression of genes is altered in response to the needs of the organism. These varying gene expression profiles arising from biological samples are often attributed to a phenotype. The regulation of biological processes occurs at many levels both transcriptional and post-transcriptional, however an important mode of regulation is the activation and inhibition of gene transcription by means of small sequence motifs known as regulatory elements.

Sigma factor proteins are required for the initiation of transcription in prokaryotes. Sigma factors direct the binding of the RNA polymerase to the DNA sequence. Sigma factors recognise different promoter sequences.

Phenotype: an observable trait in an organism, including disease state or physical traits.

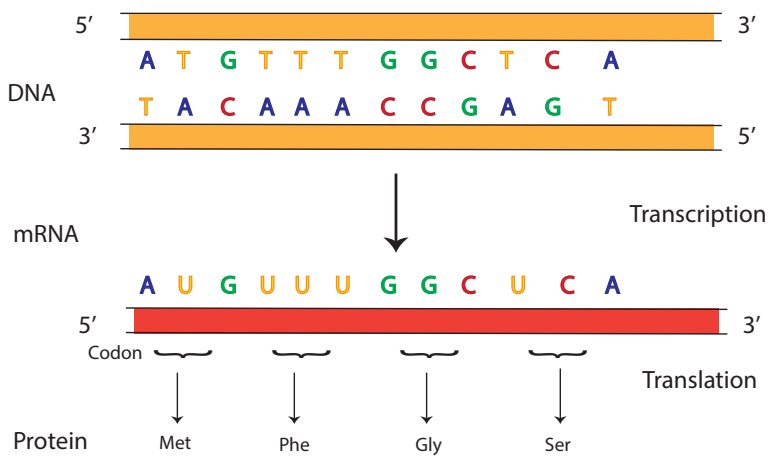


Figure 1.2: The Central Dogma of molecular biology. DNA is transcribed into mRNA. The three base codons of mRNA each are translated into one amino acid. In this example, the amino-acid sequence is Methionine (Met), Phenylalanine (Phe), Glycine (Gly) and Serine (Ser). Such amino-acid sequences fold into proteins.

1.1 Transcriptional regulation

In eukaryotes, regulatory elements are included within regions called enhancers and promoters. These regions are both *cis*-acting elements, DNA sequence regions that combinatorially regulate the transcription of nearby genes [Wittkopp and Kalay, 2012]. *Cis* motifs are recognised by proteins called transcription factors that bind to DNA and influence the transcription of the gene. *Cis*-regulatory elements act as binding sites for *trans*-acting factors, this includes the class of transcription factors. In contrast, the DNA sequence encoding these *trans*-acting factors is typically far from the location of the *cis* elements that they bind to [Gilad et al., 2008]. *Trans*-acting factors bind to the constituent factors within the transcription machinery and direct the assembly of this machinery at the target promoter sequence. Transcription factors are a subset of *trans* factors that can bind to both DNA and protein. Transcription factors typically have a specific set of sequence motifs that are found in the DNA sequence of all genes that they regulate. Transcription factors are required to direct RNA Polymerase to DNA to initiate transcription [Butler and Kadonaga, 2002]. In the simpler prokaryotic system, a single transcription factor, called a sigma factor, is required to initiate transcription whereas eukaryotic transcription often requires multiple *trans*-acting factors that bind both to the promoter and to enhancer regions that can be

far away [Paget, 2015]. As transcription factors for eukaryotes have been identified, databases of information on them and specifically their target sequences have been curated. This sequence information can be integrated with gene expression profiles to further refine the functional groups of genes and identify new targets of a transcription factor.

It was hypothesised by Britten and Davidson in 1969 that RNAs are more likely than proteins to evolve or form new regulatory elements. Regulatory RNAs do not need to undergo an additional step to be translated into protein to have function [Britten and Davidson, 1969]. It has now been shown that RNAs can have function as regulatory molecules without being translated to protein. Indeed, the ENCODE project showed that around 80% of the genome is pervasively transcribed; pervasive transcription is the transcription of DNA that does not apparently, code for protein [Encode and Consortium, 2007, The ENCODE Project Consortium, 2012]. However, the functionality of these low-level transcripts is a current area of debate [van Bakel et al., 2010]; transcription is an imperfect mechanism [Raj and van Oudenaarden, 2008], transcription can occur at low-levels with no functional implications [Graur et al., 2013]. Though some non-coding RNAs have been shown to have function, the extent of functional RNAs in the genome is still not well understood [Bakel et al., 2011, Clark et al., 2011]. Recent studies of the human genome identified 19,175 potential lncRNAs, with potential regulatory function [Hon et al., 2017]. This is still a relatively small proportion of the pervasively transcribed RNAs [Palazzo and Lee, 2015]. However, multiple classes of functional non-coding RNAs have now been identified.

MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) are two examples of classes of non-coding RNAs that can regulate the transcription of genes. LncRNAs are defined as those non-coding RNAs over 200 nucleotides in length. The RNA structure of these regulatory elements means that they can in principle interact with both proteins and DNA, resulting in a wide range of mechanisms through which they can influence regulation [Fatica and Bozzoni, 2013]. The identification of miRNAs and lncRNAs confirmed the hypothesis of RNA based regulatory elements first proposed by Britten and Davidson. For example, the lncRNA HOTAIR represses the expression of the HOXD10 transcription factor through recruitment of a transcription repression complex [Rinn et al., 2007].

In 1993, the first miRNAs were identified by Lee *et. al* [Lee et al., 1993] in the worm, *Caenorhabditis elegans*. miRNAs are short nucleotide sequences that have multiple mRNA targets and affect gene expression through cleavage or destabilisation of mRNA. They can also inhibit translation of mRNA to mature protein [Bartel, 2009]. The two short miRNAs of *lin-4*, 22 and 61 nucleotides in length, were

found to be unlikely to encode for protein whilst nevertheless regulating the expression of the gene *lin-14*. Expression of *lin-4* resulted in a decrease of *lin-14* gene expression and a concordant decrease in LIN-14 protein expression. These examples highlight the role of non-coding RNAs in the regulation of gene expression. Identifying miRNAs, lncRNAs and transcription factors and their targets can help to understand and infer regulatory mechanism. Therefore, measuring the transcription of coding genes and regulatory RNAs is important for understanding cellular function.

1.2 Measuring transcription

Transcriptomics is the simultaneous measurement and study of all transcripts in a biological sample. The ability to study genome-wide expression has advanced our understanding of biological function [Deng et al., 2015]. Advances in experimental methods have enabled the simultaneous genome-wide measurement of transcript expression. Microarrays are a hybridisation based methodology. The microarray contains probes of known gene DNA sequences that hybridise to the fluorescently labelled sample ‘cDNA’, DNA that has been reverse-transcribed from the sample RNA [Hegde et al., 2000]. Microarray protocols fluorescently tag the sample cDNA. Then, through imaging, they measure the amount of cDNA that hybridises to each of the probe sequences on the array is measured. The relative strength of the different fluorescence probe signals determines the gene expression profile of the sample. Using a pre-defined set of probes, microarrays can identify all known genes but not novel genes. They also produce noisy datasets because of the hybridisation of mismatched sequences and probes.

RNA-seq is an approach that is becoming dominant in which cDNA sequences of length between 30-500bp, these fragments are then sequenced [Head et al., 2014]. As the fragments are fractions of the full RNA sequence the sequenced fragments are aligned to a reference genome to establish the read counts per gene. RNA-seq can find novel genes or transcripts and has a larger dynamic range than microarrays. [Conesa et al., 2016] Microarrays are limited at the lower end by background noise and the maximum is capped by signal saturation. By contrast, RNA-seq is not in principle restricted. In practice however, the sequencing depth does place constraints on the range of gene expression values. The greater the number of sequencing reads, the greater the number of genes that can be measured. This is because the expression values vary for different genes. For a small number of total reads the genes identified can be dominated by a few highly-expressed genes. This has consequences for genes expressed

cDNA complementary DNA is a single strand of DNA, obtained from the reverse transcription of a strand of RNA such as mRNA or a miRNA. Reverse transcription is the opposite of transcription as shown in Figure 1.2

Dynamic range defined as the ratio between the smallest and largest value a variable can take. For Microarrays, the ratio of fluorescence probe signals. For sequencing, the ratio of read counts per gene.

Sequencing depth refers to the total number of reads sequenced.

at relatively low levels, which require higher sequencing depth to be identified by RNA-seq [Wang et al., 2009b].

Chromatin immunoprecipitation (ChIP) methods, ChIP-chip or ChIP-seq, are used to measure *in vivo* interactions between proteins and DNA [Buck and Lieb, 2004, Park, 2009]. In chromatin immunoprecipitation experiments the protein of interest is cross-linked to its target DNA. After the protein is cross-linked to the DNA, the DNA is sheared into small fragments of several hundred base pairs. Using an antibody that recognises the protein, DNA sequences where the protein has bound are purified. These DNA sequences are then either hybridised to a microarray (ChIP-chip) or sequenced using high-throughput sequencing (ChIP-seq). These methods are a useful method for identifying the target DNA sequences of a transcription factor (protein) [Valouev et al., 2008]. Information on transcription factors and their targets is often used to validate or infer regulatory networks.

1.3 Regulatory networks

Microarray and RNA-seq provide a rich source of measurements for the activity of molecules which are responsible for biological function [Malone and Oliver, 2011]. Understanding the interactions between genes, proteins and regulatory elements such as transcription factors and lncRNAs is essential to elucidate the mechanisms of biological functions [Zhou et al., 2015]. Recently, with the improvement in proteomic methods, it has become experimentally tractable to measure the activity of proteins including transcription factors. Developing methods to analyse proteomic data is an active area of research. However, it is currently still easier and more cost effective to produce genomic data, and large repositories of publicly available gene expression datasets such as ArrayExpress [Rustici et al., 2013] and Gene Expression Omnibus (GEO) [Barrett et al., 2013a] have been curated. These databases contain expression sets from microarray and RNA-seq technologies. Further, it has been shown that the activity of many proteins can be approximated using the expression of the genes that encode them. This is despite the post-transcriptional regulation of some mRNAs through factors such as miRNAs [Pe'er and Hacothen, 2011]. Therefore, the expression levels of transcription factor proteins can be estimated using transcriptomic expression data. Using transcriptomic data, the regulation of gene expression in the cell can be represented as a network.

Proteomics is the study of all proteins within an organism

Regulatory networks can include interactions between proteins, genes and regulatory elements. At an 'omics' level, the term regulatory network is used to refer to the interaction of regulatory elements and

their targets. These regulatory networks, and their corresponding gene expression profiles, are specific to different phenotypes [Vidal et al., 2011]. This includes the mis-regulation of gene expression that can result in a disease phenotype. In these large-scale regulatory networks, multiple biological functions are represented [Hartwell et al., 1999]. A regulatory module is a set of genes in the networks under the control of a group of *cis* and *trans* acting elements that are sufficient to control the transcription of the gene set in a coordinated fashion, Figure 1.3. We view inference of regulatory networks in this thesis as inference of one or more interacting regulatory modules for a given phenotype. Using transcriptomic data from microarrays or RNA-seq the resulting regulatory network specifically refers to a transcriptional regulatory network [Blais and Dynlacht, 2005]. Given the large number of transcripts measured, computational/statistical analysis is required to understand these data. This relies on the use of robust statistics to analyse large data sets. Transcriptomic data can be used to infer interactions between multiple genes. To identify regulatory networks, mathematical methods can be used [Karlebach and Shamir, 2008].

We are interested in two different areas where regulatory networks can be inferred from experimental genomic data generated using microarrays or RNA-seq. The first area is synthetic biology where information on regulatory networks can aid the design of new constructs by providing insight into functional regulatory modules [Lim et al., 2013]. In synthetic biology, not only are the interactions between promoters and DNA sequences important, it is also important to understand how these functional units can vary under different conditions. A recent paper produced and analysed data on *Bacillus subtilis* under 104 different conditions [Nicolas et al., 2012]. The analysis included identification of sigma factors, which control gene expression for some genes and the hierarchical clustering of genes and conditions. Recently network analysis has been done on this data set combining all samples and conditions. However, no network analysis has been performed to allow for regulatory networks over different conditions to be inferred. By comparing the regulatory networks over different conditions, we identified commonalities and differences between them. By decomposing subnetworks according to which condition they are present in we identified smaller, regulatory modules from transcriptomic data.

The second area is in understanding how the *Toxoplasma gondii* parasite subverts its host cells. *Toxoplasma gondii* is a protozoan from the Apicomplexan group that infects nearly a quarter of the adults world-wide causing birth defects and perinatal deaths. Additionally, it is an opportunistic pathogen and is responsible for $\sim 15\%$ of deaths in the AIDS epidemic. In general, identifying potential drug targets

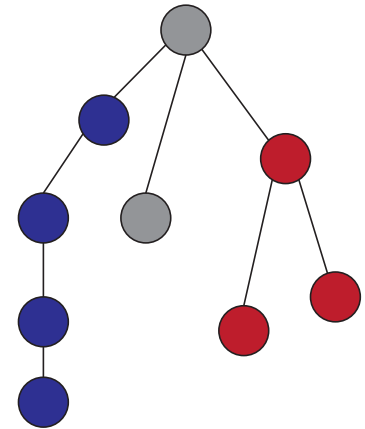


Figure 1.3: Representation of a regulatory network that contains multiple regulatory modules. Two regulatory modules, shown in blue and red that are under the control of a single regulatory element. A regulatory network that results in a phenotype may contain one or more regulatory modules.

involves finding those signalling proteins and pathways that are affected by the invading parasite. Understanding how the parasite can co-opt the host cells function to proliferate and survive should suggest potential approaches to protect against it. By considering two different applications, one the regulatory network associated with a disease phenotype and the second the identification of regulatory modules in synthetic biology we demonstrate how, transcriptomic data analysis can be used to infer these networks.

The interactions between regulatory elements and the expression of their target RNA transcripts can be written as a mathematical network. A mathematical network is defined by a set of nodes and edges. In a regulatory network a node represents a gene. An edge between two nodes exists when an interaction between the nodes exists. The area of graph theory is the analysis and inference of networks. Graph theory was first introduced by Euler in 1736 [Shields, 2012]. Network analysis can identify subnetworks that could relate to regulatory modules. To infer regulatory networks from gene expression data there are different mathematical frameworks that can be used [Pe'er and Hacoen, 2011].

In this chapter, first we consider the role of network analysis in molecular biology, and how graphical models are used to model regulatory interactions. We then discuss methods for inferring networks for a single biological condition. Finally, we discuss differential network analysis for analysing networks under different biological conditions and how the methods used to infer these networks compare to each other.

1.4 *Mathematical analysis of regulatory networks*

IN RECENT YEARS, THE APPLICATION OF MATHEMATICAL NETWORKS to the analysis of many real world situations has increased. Applications include modelling of social and communication networks, transport networks or the file systems for computers. These networks use a node and edge structure to represent at the simplest level undirected topological links between nodes. More advanced network models can also be used to represent hierarchical causal relationships or directed connections between elements. This gives the flow or direction of the network as well as its connectivity structure.

Many networks that have been modelled in real world situations follow random network structures. In these cases, the degrees of the nodes follow a Poisson distribution. As the Poisson distribution tends to the Normal distribution we expect a network whereby the spread of degree values is close to symmetrical around a mean value.

Degree of a node is defined as the number of edges connected to it.

Thereby giving a network of nodes with similar degree values, for all nodes. However, it has been shown that biological networks are scale-free networks [Jeong and Albert, 2000]. This means that the degree distribution of the nodes follows a power law. The power law distribution has a heavy right tail and is not symmetric. This gives networks whereby there are relatively few nodes with high degree value that are connected to many nodes and many of the nodes have low connectivity, or low degree value. Together this results in sparse networks or several hubs, where groups of nodes are highly interconnected within a hub but not between hubs.

Results on biological networks have shown that there are often several sub networks that are highly connected and that these networks centre around functionally important genes. From a biological perspective, these central hubs can be categorised in different ways, either functional, disease or co-expressed/topological [Barabási et al., 2011]. This is due to the highly-interconnected nature of these hubs; within them there are likely to be multiple functional units as well as subnetworks that are specific to a disease type, while the overall hub demonstrates a co-expression or topological network. This means that it is possible to decompose and categorise these hubs via different metrics depending on the aim of the analysis. Networks can be used in a wide range of contexts to model interactions between molecular components including genes, proteins and metabolites. Networks can be inferred using quantitative data to identify functional or causal interactions. Qualitative information such as known interactions found from experimental results can also be represented as networks. With graphical models the regulatory networks are connections between genes. Each gene is represented by a node and the edges of the network represent probabilistic relationships between genes [Jordan, 2004].

There exist many methods for inferring regulatory networks based on different data assumptions [Lefebvre et al., 2012]. Broadly these can be grouped into either, correlation, Bayesian, dynamic, or information-theoretic models [Marbach et al., 2010]. These can be used to infer biological networks between molecular components. Some of these networks are between one type of molecular component, such as protein-protein interaction (PPI) networks. Alternatively, the relationships between two different molecular components, for example transcription factor proteins and the genes they regulate can be inferred. The combination of data from different experimental methods facilitates this analysis. Binding data of transcription factors from ChIP-chip combined with gene expression data of target genes provides a more comprehensive view of the networks than either data set in isolation.

Poisson distribution a discrete probability distribution that is used to model the number of events over a fixed interval

Normal distribution, model for a continuous random variable that has a symmetric range around a mean value.

Recently models have also been developed for integrating proteomic and transcriptomic data into one framework [Rogers et al., 2008, Kholodenko et al., 2012]. In addition, it has been shown that transcriptomic data can be used to infer signalling networks. The nested effects model (NEM) uses the subset effects on transcriptional response to infer the signalling hierarchy [Markowitz et al., 2007]. This model uses the observation that, for a given signalling network, knocking out a child node in the network will result in a subset of changes in the gene expression that is caused by the parent node, Figure 1.4. This model allows inference of signalling networks from genome-wide transcriptomic data. To use this model, the perturbations must be specific knockouts. Therefore, some prior information on elements of the signalling network would be useful to ensure the overlap in the expression profiles that is needed to infer the signalling networks. These examples highlight the use of networks in modelling biological processes. These processes can be activated or modified in response to disease. Therefore, understanding these networks is critical to understand disease progression.

In the analysis of disease progression, it can be particularly useful to take a genome-wide view because we do not necessarily have prior knowledge on all the genes or cellular functions being targeted by the disease. While other methods may give greater detail at a smaller level of tens as opposed to thousands of genes, these are not high level enough to capture all the transcriptional effects and consequently limits the ability to capture the full development of disease phenotype. In synthetic biology, optimising constructs can benefit from the detailed low-level view, however, an equally important aspect is to understand how changing a part of a regulatory network will impact on the rest of the cell, as many off-target or unintended effects can be seen when altering part of a system. To answer these questions, bioinformatics, can provide a useful tool by giving a genome-wide context specific view of the cell that can help to design circuits to minimise these off-target effects. This should help to design more specific constructs and reduce the laboratory costs by providing some of the testing without the need for experimental validation.

1.5 Regulatory models for a single condition

We first consider the inference of regulatory networks, the interactions between genes, under a single experimental condition. Later we discuss advances in inferring and comparing networks over multiple experimental conditions and the emergence of differential network biology. In both cases one common constraint is to view these networks at a single time point. It is argued that at this level, the models represent

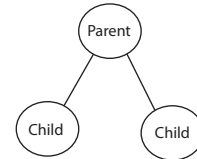


Figure 1.4: In a graphical sense a parent node is above a child node in a hierarchy. For example, in gene ontology Regulation of gene expression is a parent node of negative regulation of gene expression. A parent node can also indicate the flow of a network. For example, a Transcription factor (parent) regulating its target genes (children).

a steady-state of the network. It has been shown that this assumption can still result in useful models that can accurately identify novel interactions between elements that have been experimentally validated. The static as opposed to dynamic models focus on genetic interactions at a single time point, ignoring feed-back loops or post-translation modifications. We broadly categorise and give examples of different modelling frameworks and data types that have been used to infer regulatory mechanisms in both prokaryotic and eukaryotic organisms.

Steady-state A stable condition of the system. Gene or protein expression would be constant.

1.5.1 Boolean and logic models

Two closely related sets of models are boolean and logic models. A boolean model is a two-state discrete model where variables are given the value 0 or 1. In reference to regulatory networks, gene expression values would be discretised to 0 and 1 and would be interpreted as not expressed or expressed respectively. This is a simplification of the model that makes it easier to infer the network, computationally it is easier to work with binary and integer values than real numbers and gives a simple and finite set of states that two genes can be in together, either both not expressed (0,0), one expressed (0,1 or 1,0) or both expressed (1,1). Even with this simplification boolean networks have been successfully used to infer regulatory networks. One natural partner of boolean networks are logic models. These extend the boolean network to include logic functions such as, AND, OR, NOT, which are used to model the relationships between elements. For example, a gene may be activated by one of two transcription factors in which case the model would have two edges between the gene and each of the transcription factors with an OR gate governing them. Alternatively, it could require both transcription factors to activate transcription, in the case it would be an AND operator. Another advantage of logic models is that they can be used to describe and combine several different data sources. For example, using this framework it has been possible to both discretise experimental data and describe known prior information from the literature in logic form and therefore combine the two to improve network inference [Saez-Rodriguez et al., 2009].

Davidson *et al.* used logic models to model the *cis*-regulatory interactions for sea urchins [Davidson et al., 2002]. *Cis*-regulatory elements such as promoters direct the binding of transcription factors to switch on transcription. Viewed as an on-off switch, it is sensible to model the activation of these *cis*-regulatory elements using a logic framework. Logic models can be used to represent whether an element is active and combinations between elements. The authors refer to this model as a ‘first-stage regulatory model’ and this information enabled them

to identify genes and interactions that are involved in the developmental process of the organism. While these modelling frameworks are unable to capture the kinetic reactions between transcription factors and their targets, they are able at a lower level, to model the combinatorial interactions between elements. The analysis was based on perturbation data of gene expression and this data was subsequently combined with information on the *cis*-regulatory motifs for genes that enabled the authors to determine which interactions were direct or indirect. Using known interactions can improve model inference by reducing model search space or resolving two equally fitting models mathematically where only one is consistent with the prior knowledge.

1.5.2 Correlation, co-expression methods

Correlation or co-expression methods model linear relationships between genes. These methods often use gene expression data from biological replicates or under different conditions. Intuitively these models assume that genes under the similar regulation will have similar expression profiles. These models can be applied to genome wide transcriptomic data given the relative simplicity in calculating these metrics. There are several different correlation metrics including the parametric Pearson's correlation and the non-parametric Kendall's tau. Parametric metrics assume that the data follow a known distribution, whereas non-parametric measures do not.

Correlation based metrics do not however, determine causal relationships. Therefore, a significant correlation cannot be used to definitively say that two genes are connected in the same regulatory units or that one gene is controlled by a transcription factor as this correlation could be due to an indirect regulatory relationship. However, by comparing DNA microarray expression profiles over multiple organisms, functional and conserved genes have been identified due to their co-expression [Stuart, 2003]. In this case, the conservation across different organisms including human, mouse and yeast, provides further evidence for positive selection of these genes. Consequently, they are more likely to be functionally important. Further, those with similar expression profiles may also share similar function. The value of co-expression was measured by the Pearson's correlation coefficient. Statistical significance of these correlations was assessed by comparing them to values that would be observed by chance.

A further step in correlation analysis is to calculate the partial correlations, as opposed to correlation. Although computationally more demanding, the partial correlation does give causal information and allows a hierarchical network to be inferred. The partial correlation is defined as the correlation between two variables conditional on all

Pearson's correlation ρ , for two variables X, Y , where the mean and standard deviation of X are μ_x, σ_x and the mean and standard deviation of Y are μ_y, σ_y is $\rho = \frac{E(X-\mu_x)(Y-\mu_y)}{\sigma_x\sigma_y}$

other variables. It is this conditioning that allows the hierarchy to be inferred. By conditioning on the expression of all other genes, a significant correlation between two genes then implies a causal relationship, intuitively conditioning on all other variables means removing any other direct or indirect relationships and consequently if a significant relationship still exists after removing all other possible controlling factors the interaction is said to be direct or causal. Where the Pearson's correlation matrix is used, the network defined by the partial correlations is called a Gaussian Graphical Model (GGM).

Shrinkage methods have been used successfully with partial correlation methods, these shrinkage methods aid interpretability of the model by shrinking low value correlations to zero and consequently removing the respective edge. These shrinkage methods therefore result in sparse graphs which are consistent with the power law properties observed in biological networks. One such method is the Graphical Least Absolute Squares Shrinkage Operator (glasso) that shrinks the parameter estimates based on a penalised maximum likelihood [Friedman et al., 2008]. The maximum likelihood estimate Θ , is the parameter set that maximises the likelihood $L(X|\Theta)$ which gives the likelihood of observing the data X given the set of parameter values Θ :

$$\operatorname{argmax}_{\Theta} L(X|\Theta)$$

The general form of a penalised maximum likelihood, with penalty term $P(\Theta)$ is then:

$$\operatorname{argmax}_{\Theta} \{L(X|\Theta) - P(\Theta)\}$$

The penalty used in glasso is the L1 norm. For a graph defined by a partial correlation matrix Θ , the L1 norm penalises the sum of the absolute values of the parameters: $P(\Theta) = \sum_{i,j} |\theta_{i,j}|$

1.5.3 Information-theoretic

Given the assumption of multivariate normality, the Gaussian graphical models can infer conditional dependencies for linear relationships. However, regulatory networks may include non-linear relationships. For example, time series data could capture feed-back mechanisms or switches that repress or activate gene expression. These non-linearities can be captured by different measures such as dcor [Székely et al., 2007] or MIC statistic [Reshef et al., 2011]. However, these are single pairwise distance measures, that have not, so far, been used to create graphical models. Another measure capable of capturing non-linear relationships, mutual information, is an information theoretic measure that has been used to infer networks. Using information theoretic

Gaussian Graphical Model a model of continuous multivariate normal observations that are represented in a network structure. In this graphical model, probabilistic connections between nodes are determined by the significance of the partial correlations.

measures clearly gives an advantage over correlation statistics when the relationship is non-linear. However, if the relationship is linear, the information-theoretic methods are less powerful in detecting these relationships in comparison to Pearson's correlation measure. Power in this context is a statistical measure that quantifies the ability of a test to accurately find a positive association where one exists. Information-theoretic models similarly cannot infer causal relationships, they provide a single measure of similarity between two genes, which does not distinguish direct from indirect links.

ARACNe was the first method to use information-theoretic metrics to infer biological networks. This model calculated pairwise values of mutual information between gene expression from microarray profiles of human B cells then used the data processing inequality to remove potentially indirect links [Basso et al., 2005]. This simplified the resulting network, aided interpretability and allowed greater focus on direct links or causal links. From this network, the transcription factor MYC was found within a hub and novel connections to other gene targets were experimentally validated.

[Carro et al., 2010] used ARACNe on a set of meta-data of gene expression of grade III and IV glioblastoma brain tumours. This showed how, as well as cell specific models, disease or context specific models can be inferred and can be effective in identifying master regulators of transcription. Master regulators are those that are at the top of the regulatory hierarchy and are therefore not under the control of any other transcriptional regulators [Kin Chan, 2013]. The analysis of brain tumour gene expression data identified two novel master regulators of the disease network, which the authors then experimentally validated.

MINDy is a method that is used to find post-translational modifiers of a transcription factor. This demonstrates the utility of correlation or information theoretic metrics to identify relationships in different data types. It calculates the conditional mutual information between the gene expression level of the transcription factor and its targets given the value of the potential modulators. By conditioning on the modulators, the statistic can find causal links as opposed to only numerical relationships. It is cell context specific, and requires many gene expression profiles as well as input from the user on the list of transcription factor targets and modulators to be tested [Wang et al., 2009a]. Wang *et. al* could identify and experimentally validate four post-translational modifiers of the MYC transcription factor within Human B-cells.

1.5.4 Bayesian

The correlation or information-theoretic methods are both frequentist methods. These methods estimate the parameters of interest, θ , from the observed data. In contrast, Bayesian methods allow for uncertainty in the data in the form of a prior distribution of the parameters. Prior distributions can be used to incorporate knowledge of the parameters into the inference. For example, Bayesian inference of regulatory networks can use information of known transcription factors and their regulators. This information is encoded in the prior distribution. Bayes theorem gives the following relationship between the likelihood $L(X|\theta)$, posterior $p(\theta|X)$ and prior distribution $p(\theta)$:

$$p(\theta|X) \propto L(X|\theta).p(\theta)$$

From a statistical perspective, a common method for selecting models is based on minimising the Bayesian Information Criterion (BIC) which is defined as:

$$BIC = -2\ln L + 2k\ln(n)$$

where L is the likelihood and k the number of parameters estimated from a dataset with n observations.

The prior distribution on the parameter space may be either informative or uninformative. The informative prior is taken from known information on the data. For example, if the network is based on correlations between genes, the prior can support the inclusion of edges between a transcription factor and its known targets. Alternatively, an uninformative prior is one which has no basis in the knowledge of the system for example, a uniform prior which gives equal weights to all parameter possibilities so we do not bias the inference to any interactions. In this instance, the purpose of the prior is to allow for uncertainty or incomplete information in the data. That is, data sets are samples of the population and do not give complete information on the system due to the constraints of data collection (not all scenarios can be covered with maximal numbers of replicates). It is this uncertainty in the data the prior is used to model. Informative priors are taken from experimentally validated interactions. Analysis of this type has been made increasingly possible through numerous on line databases. These databases contain information on transcription factors and their targets, for example TRANSFAC [Matys et al., 2006] and DBTBS [Sierro et al., 2008] are databases of transcription factors in human or mouse and *Bacillus subtilis* respectively.

A Bayesian framework was first proposed for analysing gene expression data by Friedman *et al* [Friedman et al., 2000]. The authors used a Bayesian graphical approach to analyse time series data for

cell-cycle processes based on measurements of mRNA levels of *S. cerevisiae* ORFs. The Bayesian framework covers a wide range of model formulations that vary in their definition of the likelihood functions and prior distributions. This means that Bayesian models have been used in many different applications in biological sciences, for example in inferring protein signalling networks using single cell data [Sachs et al., 2005] as well as regulatory networks [Pe'er et al., 2001]. Other examples include using a Bayesian approach to infer functional networks that combined multiple data inputs including mRNA, protein levels and literature evidence to generate a probability network of the connections between genes in *S. cerevisiae* [Lee, 2004].

1.5.5 *Dynamic*

To model the dynamics of the cell-cycle and identify certain control mechanisms, time series data are required. Time series data can infer network dynamics such as feed forward or feed-back loops in networks that are not identifiable by static data. As well as specifically looking at gene sets or processes that may vary over time, time series data can also be used as perturbation data that can be used with many network inference algorithms.

Ordinary differential equations (ODE) models have been used to model a number of regulatory networks. These models require a large amount of prior information that give the network interactions. It is necessary to know which elements of the regulatory network interact with each other and to represent these interactions using differential equations. Solving these ODEs gives kinetic parameters that can recreate the regulatory networks and the quantitative mRNA and protein levels. Chen *et al.* calculated parameters of an ODE model for the cell cycle of budding yeast, this included around 30 interacting elements in the network [Chen et al., 2004].

1.6 *Differential network methods*

IT IS KNOWN THAT different regulatory networks are active under different conditions and that these networks may be mis-regulated in the presence of disease. Therefore, understanding the differences in these networks between cellular conditions is an active area of research [Pe'er and Hacohen, 2011]. The methods discussed above have focused on providing a network under a single experimental condition or a single cell type. However, it is also possible to consider the similarities and differences between networks for different experimental conditions, organisms or cell types. Moving to differential network analysis from

differential expression of groups of genes or proteins has several advantages; the network formulation gives a more complete view of the cell as genes do not act independently as differential expression analysis assumes.

Gene set enrichment analysis was one of the first methods to address the fact that genes act together. GSEA identifies enrichment of biological processes whose gene sets are ranked highly together. However, gene set enrichment does not give a network view, in that group of related genes can be identified but not the hierarchy or structure of how these genes interact. Moreover, differential expression is limited in detecting regulatory interactions as often transcription factors do not show differential expression but are constitutively expressed when active [Hudson et al., 2012]. Within synthetic biology, the ability to design new circuits crucially depends on available information on how the components interact with each other as well as how function may change under different conditions [Kwok, 2010]. In parasite host interactions, mechanisms of action of a cell or parasite can be modelled; understanding how these networks are affected can help identify therapeutic targets [Pe'er and Hacohen, 2011, Iorio et al., 2012]. Differential expression analysis is severely limited in answering these questions, a network view is essential to be able to understand and modify these networks. Therefore, we were interested in finding differences in networks between conditions.

Within differential network biology methods have been developed to analyse differences between networks under different conditions [Ideker and Krogan, 2012]. These methods infer both the underlying network as well as identifying changes to the network following perturbation. Examples include qualitative [Miller et al., 2009a] and quantitative measures [Bisson et al., 2011] that have been used to detect differences in protein-protein interaction networks [Ideker and Krogan, 2012]. From a qualitative perspective, manual curation of the literature has been used to combine known information into a comprehensive metabolic network for *Homo sapiens*. This network provided a basis for comparison of functional units under different conditions [Duarte and Becker, 2007].

Differential networks have been used to identify regulators and functional units of disease. In one example a protein protein interaction (PPI) network based on two-hybrid yeast data was mined using the gene-expression profiles taken from two studies of patients with or without breast tumour metastasis. The protein-protein network was searched using genome-wide profiles of disease effects. The genomic expression data was used to identify subnetworks within the PPI that were statistically discriminative of metastasis. These subnetworks of interest were identified according to their activity, defined as the

Gene set enrichment analysis uses a non-parametric test to compare the ranked list of genes (usually according to their differential expression), to a list of genes involved in the same biological or molecular process

Two-hybrid yeast: this system uses two separately encoded protein domains which, when physically nearby, activate the transcription of a reporter gene. Each of these two protein domains are attached to a different protein of interest. If the two target proteins interact, the separately encoded protein domains will bind together and switch on the reporter gene. [Brückner et al., 2009]

average normalised expression values of the genes in each network. By comparing the activity values of the networks across the two groups, those with and without metastasis, they could determine their discriminatory capability. This was done by calculating mutual information between gene expression levels and the disease status within the PPI network and comparing the results to those of random networks [Chuang et al., 2007]. The authors found that viewing expression data in a network context and by incorporating information on the interactions between genes, the statistically significant sub networks, according to disease status, were better predictors than the individual component genes.

Gambardella *et al* curated gene sets from pathway information in the KEGG database and combined this prior information with a large set of gene expression profiles obtained from ArrayExpress. They generated 30 tissue or condition specific correlation based networks from the expression profiles. Using a score based on the number of edges in each of the 30 networks between the genes in each set obtained from KEGG they could determine the differential activity of pathways between these 30 networks. This approach relied heavily on both existing knowledge on pathways, from KEGG, and on the expression data obtained from ArrayExpress [Gambardella et al., 2013].

Ergun *et al.* developed an algorithm, the Mode of Network Identification (MNI), and used this to identify the AR gene and its associated pathway as involved in prostate cancer metastasis [Ergün et al., 2007]. The analysis showed that the network approach could find this result where differential expression and gene set enrichment was not. The MNI is a two-stage process. First several expression data sets were combined for 7 different cancers. The requirement being that these profiles cause a wide range of perturbations to the cell that will consequently allow for the modelling of the network. Second, these networks were mined using the specific prostate cancer profiles. Networks were scored according to their inconsistency with the disease profiles, assuming the disease causes disruption to the normal network.

Recent advances in Gaussian graphical models allow for inference of networks across experimental conditions [Danaher et al., 2014]. This Joint Graphical Lasso (JGL) model is applicable for microarray gene expression data as it is log normally distributed. For RNA-seq data, the count data can also be approximated by a normal distribution [Hansen et al., 2012] so Gaussian graphical models can also be applied. The authors demonstrated the utility of the JGL model on a meta study of microarray gene expression data for patients' biopsies of normal and cancerous cells. The JGL model extends the glasso model [Friedman et al., 2008] to add a secondary term that penalises

differences across the different experimental conditions.

1.7 Modelling strengths and weaknesses

THERE ARE MANY METHODS for inferring regulatory networks, there is also interest in comparing and assessing these using a consistent standardised set of tests. The Dialogue on Reverse Engineering Assessment and Methods (DREAM) consortium was established to provide a systematic comparison of the available methods. They found that all methods have their strengths and weaknesses and that error rates do not tend to be consistent across all methods [Marbach et al., 2012]. Therefore, we can select different models depending on the question to be answered and the priorities of the analysis. The constraints of the available data determine which models can be used as the data must match the assumptions of the model. The analysis from the DREAM consortium also showed increased accuracy when combining results from different methods.

The multiple methods of inferring regulatory networks that have been developed are applicable depending on the available data, such as genome wide studies or subset analysis and often the type of perturbations. Examples of perturbations include gene knockouts or treatment with compounds, see [Markowitz, 2010, Pe'er and Hachohen, 2011] for reviews. In addition, there is a computational and data trade-off between the detail and scope of the model. Correlation or co-expression studies do not provide causal relations. In contrast this information is available from ordinary differential equation (ODE) models as well as giving quantitative information on the interactions [Lefebvre et al., 2012]. However, whilst ODE models provide greater detail than correlation or co-expression networks they usually contain smaller numbers of genes or proteins, whereas correlation or information-theoretic studies can be performed at a genome-wide level.

Arguably the Gaussian graphical models lie between correlation and ODE models in terms of complexity and scalability. As these models use partial correlations to infer conditional independence relations they consequently provide evidence for causal relations as well as correlation. For genome-wide data sets, the number of observations n tends to be much less than number of parameters p (genes), which we write $n \ll p$. Estimates will therefore exhibit high variance and lack of identifiability. Moreover, standard numerical estimators, such as maximum likelihood will not give estimates that are exactly zero. Consequently, shrinkage methods have become increasingly popular to estimate networks from high dimensional data [Chun et al., 2014, Jung et al., 2015, Xia et al., 2015, Kling et al., 2015].

The Joint Graphical Lasso approach is a purely data driven model. In comparison to literature based methods this offers the ability to find new links. These two methods are not mutually exclusive however, and can synergistically behave together to produce a better model. As an example, the weighted glasso model was developed to combine the benefits of the shrinkage methodology with prior information by allowing different penalty terms for different edges based on prior information about gene interactions. The authors showed improved inference over glasso on simulated and biological data from *Arabidopsis thaliana* when comparing different error rates [Li and Jackson, 2015]. In this work, we used the JGL model to analyse microarray and RNA-seq data given that it can take input at a genome-wide level and provide output containing networks that are smaller and consequently more interpretable.

The rest of the thesis is outlined as follows, the second chapter discusses the bacterium *Bacillus subtilis* and its role in synthetic biology. The third chapter uses the JGL model to analyse the microarray expression data from Nicolas *et al* to infer regulatory networks of *Bacillus subtilis* under different experimental conditions. In the fourth chapter, we provide an empirical Bayes method for estimating correlation matrices. The fifth chapter contains analysis of RNA-seq data for Mouse embryonic fibroblast cells that have been infected with *Toxoplasma gondii*. In the final chapter we develop the results of this analysis into a publicly available application that can be hosted on-line as an interactive web resource. Throughout we discuss annotation and interpretation of these networks from both a biological and mathematical perspective, as well as developing methods for controlling error rates and finally identifying novel hypotheses that could be tested experimentally, with the aim of improving our understanding of *Bacillus subtilis* and the parasite *Toxoplasma gondii*.

Synthetic Biology

We define synthetic biology as the design and engineering of new genetic circuits that are used to re-wire or expand the existing cellular mechanisms within an organism. The aim is to modify the organism to produce a desired phenotype or product. These methods can be used in industrial applications including drug discovery and the production of renewable biofuels and other chemical products [Purnick and Weiss, 2009, Nandagopal and Elowitz, 2011]. In designing these circuits recent research has focused on characterising and defining elements that can be used to create genetic circuits. These circuits are designed to affect different parts of the cell cycle processes using, for example, transcriptional, translational or post-translational modifications [Liang et al., 2011]. We first consider the general mechanisms of cellular regulation, with a focus on prokaryotes/bacteria before discussing examples of specific synthetic control that have been implemented within *Bacillus subtilis*.

At a translational level, early synthetic controls used protein based methods. For example, identifying proteins that facilitate binding of the ribosome to an mRNA sequence and using these to activate or repress the target's translation in to protein. However, there has been a shift to generating synthetic RNA based modules instead of proteins to control gene expression. As the understanding of the RNAs functions increased so did the interest in generating synthetic versions to be used in genetic circuit design. There are multiple ways in which RNA can control gene expression [Liang et al., 2011]. From a structural perspective secondary structures of mRNA, for example loops and hairpins, can be changed to restrict access to its Ribosome binding site (RBS) [de Smit and van Duin, 1990]. The ribosome binding to mRNA is necessary to allow translation of the mRNA into protein. Consequently, inhibiting this binding activity is one method for controlling the rate of gene expression into mature protein. One class of RNAs that control gene expression are catalytic RNAs or ribozymes. In bacteria examples of cleaving ribozymes are the

Genetic circuit a system of biological parts designed to provide logic control of cellular functions. For example, activating or repressing gene activity as an on/off model.

Ribosome Molecular machinery for the translation of messenger RNA to protein. The ribosome recruits tRNAs to translate the RNA into the correct amino acid sequence

glmS ribozyme and RNase P. The glmS gene produces the coenzyme GlcN6P [Winkler et al., 2004]. When GlcN6P is present in the cell it combines with the glmS ribozyme to cleave the mRNA of glmS this in turn deactivates its expression, therefore providing a mechanism of self-regulation. RNase P is a ubiquitous cleaving ribozyme that cleaves sequences of pre-tRNAs to produce tRNAs [Guerrier-Takada et al., 1983]. Another important class of RNA control mechanisms are RNA switches. This refers to mRNA sequences that function as riboswitches which respond to cellular metabolites and co factors by altering their secondary structures to either allow or restrict access to their RBS and consequently control their translation [Serganov and Patel, 2007]. This provides another method for controlling the genes that are responsible for the breakdown of widely available organic compounds such as glucose [Henkin, 2008].

Finally, within the prokaryotes there exist transcripts that form small RNAs that are used in conjunction with RNA binding molecules to inhibit or promote the translation of their target mRNAs. This is done by controlling the mechanism that binds the ribosome to the mRNA and so controlling protein synthesis. Using RNA to control the cellular response to organic compounds is perhaps most easily combined with our results as we are categorising the changes and activation of regulatory networks (or targets) under different conditions including compounds, or co factors, such as glucose and malate [Cochrane and Strobel, 2008]. Regulatory networks can provide information on the genes responsible for the processing of these compounds. This may provide synthetic biologists with genes to target to influence an organism's response to these compounds. Further, by taking a wider network view it may also help researchers to understand additional or knock on effects of perturbing one part of the system. This will help in designing systems with greater efficacy and fewer unwanted responses.

In prokaryotes, sigma factors regulate transcription. The sequences of the sigma factors vary and therefore bind to different genes in the organism to initiate transcription. From a transcriptional perspective, the gene expression mechanism in bacteria often involves activation of a promoter by a transcription factor to start transcription. The transcription factor alters the three-dimensional structure of the DNA sequence to allow access to the promoter for the complex of sigma factor and RNA polymerase to bind. Therefore, the transcriptional process provides synthetic biology with multiple mechanisms to control the level of mRNA production by altering promoters, transcription factors, termination sequences or a combination of these. For promoters, progress has been made to identify different promoters for prokaryotes.

In *E. coli* for example, promoters in four different categories are

tRNA recognise specific mRNA sequences and translate them into amino acid sequences

known, these are either constitutive, activator, repressor or combinatorial promoters. As their names suggest, constitutive promoters are always active, that is they do not require additional molecular factors (either sigma or transcription factors) to allow the RNA polymerase to bind and bring transcription of the genes under their control. In *E. coli* the $\hat{\sigma}70$ factor, that is always present in the cell combines with RNA polymerase and transcribes genes that contain constitutive promoters so that these genes are always being transcribed. Activator and repressor promoters control gene expression by either activating or repressing gene transcription following binding by its associated transcription factor [Singh, 2014]. Combinatorial promoters are under the control of more than one transcription factor and are therefore able to create a greater dynamic range of transcripts. The identification of transcription factors and their targets is obviously critical to the success of understanding and subsequently controlling gene regulatory networks. Combining transcription factors and their associated promoters allows for the design of synthetic constructs that can control gene expression.

Once the DNA sequence has been transcribed into mRNA there are also numerous controls that can affect the rate of translating mRNA into protein. However, as we are inferring regulatory networks from transcriptomic data, our results are more useful where synthetic circuits are using a transcriptional level control which assumes a linear progression from transcription to translation and protein production. In addition, being able to see how the regulatory mechanisms change under different experimental conditions will also have value for the design of synthetic controls as it has been found that in some bacteria, the translation of RNA can be controlled by external temperature [Liang et al., 2011].

To date, synthetic systems have been designed and implemented that can create numerous effects on the cell. These include, circuits with logic controls or switches and those that create oscillations in cell process or simply activate a network. These circuits can help in the understanding of the cell mechanisms as well as re-programming them for the optimisation of the yield of new or existing cell products. However, these synthetic circuits, are designed on a small scale, that is they usually include only a few genes. In practice, the prokaryotic cells have more complicated regulatory networks that interact with each other depending on different external conditions or signals passed to the cell. Therefore, we view the inference of the regulatory networks at a genome-wide scale and under multiple conditions to be complementary to the design of synthetic circuits or regulatory elements.

Bacillus Subtilis (*B. subtilis*) is a gram positive bacterium capable of secreting multiple enzymes and proteins, many of which have

diverse uses among pharmaceutical and therapeutic industries. It is used as a system in many synthetic biology approaches due to its relatively simple structure, ability to produce enzymes with industrial applications and its lack of toxic byproducts [van Dijl and Hecker, 2013, Jeong et al., 2015]. Consequently there has been an effort to understand *Bacillus subtilis*, in order to reverse engineer its system to produce higher yields with higher specificity.

The *B. subtilis* genome has been well annotated, and advances have been made to understand the organism from multiple perspectives including transcriptomics, proteomics and metabolomics. Integrating this information in the future will likely benefit the design of synthetic biology circuits as well, due to the interactions between these different cellular mechanisms. To design and implement more complex circuits it is necessary to consider the various levels of control within the cell and how these interact with each other. So far, the synthetic designs focus on one area, for example post-translational or transcriptional but does not combine these concepts. As more information becomes available from different ‘omics’ resources it would be beneficial to combine these into a more complete view of the cell and how this can be harnessed in synthetic biology [Khalil and Collins, 2010].

For *Bacillus subtilis*, the genome is categorised into groups of genes called operons. Initially it was thought that all genes within an operon were under the control of the same regulatory system, leading to the same levels of gene expression for each gene within the operon [Ermolaeva et al., 2001]. In this traditional model, each operon is regulated by a single sigma factor which binds to the promoters to initiate transcription. However, recently it has been shown that the system of regulation is more complex. There are also interactions between termination signals within genes, shRNA and riboswitches, causing different levels of expression for genes within the same operon. Additionally, there are combinatorial effects of multiple sigma factors on the same operon [Güell et al., 2011]. This means that under different experimental conditions in our model it is possible we may see different elements of an operon active and interaction between different regulatory networks.

There currently exist a number of resources that contain information on the combined knowledge of regulatory elements within *B. subtilis*. These include the database of transcription factors DBTBS [Sierro et al., 2008], and SubtiWiki [Michna et al., 2013]. BsubCyc [Caspi et al., 2014] and SynBioMine [Micklem group, unpublished] are online databases that include information on *B. subtilis*. The BsubCyc website database includes known transcriptional units for each gene. A transcriptional unit is defined as a gene or set of genes under the control of the same promoter. The transcriptional unit entry for a

Metabolomics is the study of the metabolic elements and relationships between the metabolic

transcription factor can have multiple transcriptional units associated with it. Although the BsubCyc databases contains information on known transcriptional units and the local context of genes within the *B. subtilis* genome, it does not have condition specific or hierarchical information, on the interactions between them. These are gaps in our understanding of the regulatory networks that we wanted to address.

By taking a wider view of the networks we may identify combinatorial regulatory patterns under specific conditions in addition to off-target effects of introducing a synthetic system into the cell. That is, regulatory networks and their biological processes that may also be induced or repressed due to the synthetic circuits but whose activation was not part of the initial design specification. Recently Kobayashi *et al* took a systematic approach to decouple the functional modules, to design toggle switch circuits [Kobayashi et al., 2004]. This highlighted the benefit of being able to categorise functional or regulatory modules that, ideally, can be independently targeted for design modification. They made use of two known signalling pathways, including the SOS signalling pathway, and used an ODE model of the designed circuit to simulate the parameter values that would give the best efficacy of the system. In this case, the mathematical modelling was used to optimise the synthetic circuit. This design required the prior knowledge of the native pathways that the synthetic circuit interfaced with. In our case, we use mathematical methods to increase this prior knowledge, categorising pathways and functional modules in *B. subtilis* that could be used in synthetic biology.

Recent advances in understanding the regulatory networks of *B. subtilis* have included both genome-wide inference of regulatory networks and the integration of multiple data sources to understand network dynamics. Buescher *et al* combined information from transcriptomics, proteomics, metabolite levels as well as measuring promoter activity and transcription factor binding through ChIP-chip data [Buescher et al., 2012]. By combining a wide range of data types, they could identify novel coding sequences from the mRNA sequencing, binding regions from the ChIP-chip data and establish potential functional classification of genes through clustering of expression profiles. They could identify post-transcriptional regulation through comparison of mRNA and protein levels and identify transcription factors with differential activity under changing conditions (either changes in glucose or malate levels in the growth medium). For the transcriptomic data, the authors generated three replicates per condition. As this sample size does not provide enough information alone to infer networks or dynamic models, prior knowledge of the transcription factors and their targets was required to infer the changes in their activity.

We were interested in using the publicly available data from Nicolas

et al that gives three replicates per condition. Despite the small number of replicates, they have 104 different conditions thereby providing a rich source of information for network analysis. Arrieta-Ortiz *et al* generate an additional 38 microarray experiments and combined them with the Nicolas *et al* data set to infer regulatory networks for *B. subtilis* [Arrieta-Ortiz et al., 2015]. Again, to infer the networks the authors combined the transcriptomic data with prior information from the transcription factor databases. The authors focused on combining data networks between two different strains and the two different data sets into one regulatory network model. From here they could identify and experimentally validate several novel interactions.

Our focus, however, will be on identifying differences between these networks under different conditions. This will involve the assessment of the viability of inferring these networks with fewer samples per condition. We split the samples according to their experimental conditions into multiple groups as opposed to combining them into one group as with meta studies of the organism. In the next chapter, we select both the data inputs and parameters of the model and assess the network output for accuracy and identify potentially novel connections in and between transcriptional units.

Bacillus Subtilis and GGMs

3.1 Results and Discussion

We used an existing microarray data set on *Bacillus subtilis* that contains cells under 104 different conditions with three biological replicates for each condition. The conditions include cells grown in different mediums or those treated with different carbon sources such as malate or glucose, as well as drug compounds. Three replicates are not sufficient for correlation analysis therefore, the hybridisations were clustered to combine information over conditions. This gave meta-conditions where each meta-condition contains a set of related experimental conditions that have similar expression profiles. We assumed a commonality of regulatory networks active under different experimental conditions and consequently a similarity of gene correlation patterns between them. This assumption is reasonable as many of the 104 conditions share co factors, such as cells grown with glucose but harvested at different time points.

Using Euclidean distance and affinity propagation clustering [Frey and Dueck, 2007], the meta-conditions containing groups of similar expression profiles were identified. The results of the affinity propagation clustering are shown in Figure 3.1. The results show clustering of replicates from the same condition and hybridisations under similar conditions. A subset of these clusters (initially by trial and error) were then used as input to JGL, an automated method for selecting data to input in the model introduced in Section 3.5.4 To do this we iteratively investigated the clusters with the largest number of observations in the clusters to give largest potential statistical power in model inference and used these as input into the JGL algorithm.

From this, three clusters were used as input to the JGL algorithm. As a pre-processing step the data is standardised to zero mean and unit variance. This standardisation does not affect the inference of the JGL model as it is a Gaussian graphical and the Pearson's correlation for Gaussian data is scale and location invariant. The JGL model was

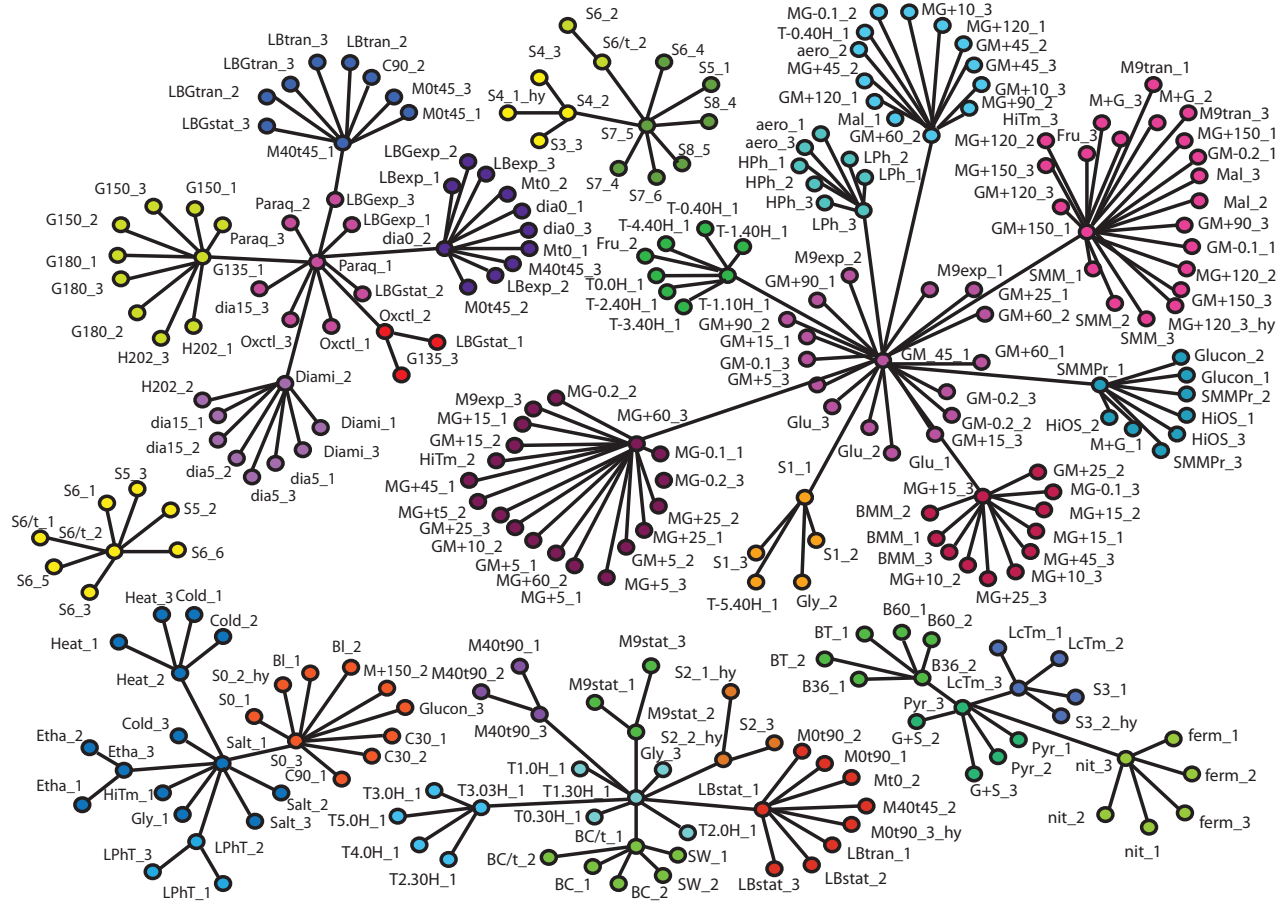


Figure 3.1: Microarray data are clustered using affinity propagation clustering. Each of the clusters is based on the Euclidean distance between all hybridisations from the *Bacillus subtilis* data set. The labels for each node are the condition of the hybridisation and the underscore denotes the number of the biological replicate. Hybridisations cluster by replicate as well as similar conditions such as GM+15 and GM+25 representing malate treatment for 15 and 25 minutes respectively.

run with shrinkage parameters $\lambda_1 = 0.925$ and $\lambda_2 = 0.005$, throughout the rest of this section this model will be referred to as the inferred network. Varying the shrinkage parameters are discussed in Section 3.1.1.

We viewed the correlations between genes in a random order and after the genes have been ordered and separated into their blocks using the JGL algorithm. This shows the validity of the block diagonal assumption and gives a visual overview of the correlation structure. The Figures 3.2 a) and 3.2 b) are the heatmaps for the 944 genes that were connected in the output of the JGL algorithm that had input shrinkage parameter values $\lambda_1 = 0.925$ and $\lambda_2 = 0.005$. Figure 3.2 a) shows the randomly ordered genes and as expected there is no obvious pattern to the correlations in the heatmap. In contrast, Figure 3.2 b) shows the genes ordered according to the block they are assigned to by the JGL algorithm ordered left to right from largest to smallest block. Here we see there are rectangles or blocks of blue indicating sets of genes with strong correlations to each other.

Block diagonal matrices can be written as a set of square matrices on the diagonal. Off-diagonal elements are zero.

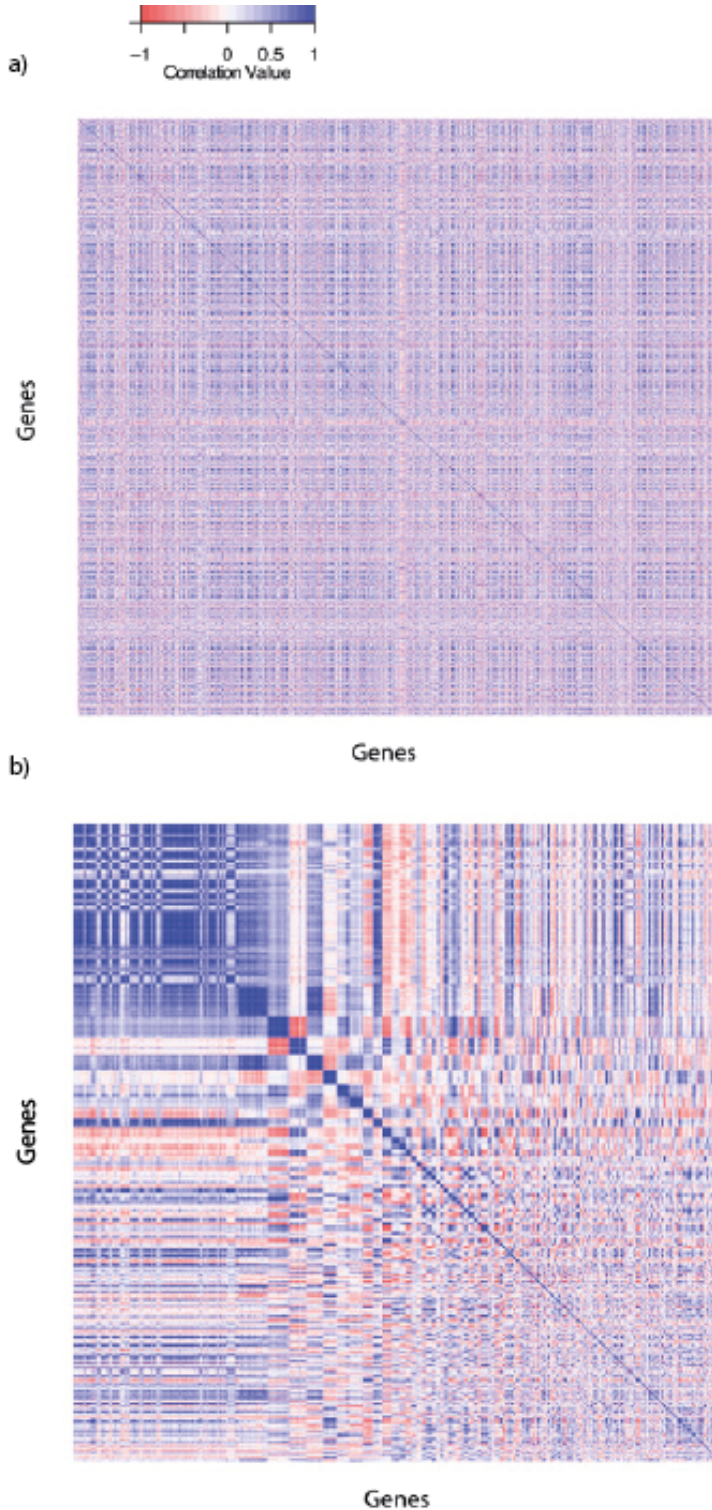


Figure 3.2: a) These are the correlations between all the genes in the network. These are the genes randomly ordered and show no obvious pattern. b) These are the correlations between the genes where genes are ordered according to the block they have been assigned by the JGL algorithm, this shows blocks of colour representing groups of correlated genes. The scale for the heatmaps go from blue to red with strong positive to negative $[1,-1]$ correlations. Uncorrelated (zero value) correlated genes in white. There are patterns of four blocks which have one diagonal of two blue or red blocks, the other two are white blocks. It is this pattern that enables the separation of the genes in the two blue or red rectangles as the off-diagonal, low correlation, white blocks indicate that the genes in one blue or red block are not strongly correlated with those in the second blue or red block.

The data input included three different meta-conditions as selected by the affinity propagation clustering. For each of the three conditions the resulting network had 1295, 2219, 772, edges that were contained within 131, 175, 102 subnetworks or blocks for each meta condition respectively. There were 637 edges that were shared between all three conditions. We annotated the inferred network to both evaluate the accuracy of the results and generate hypotheses of novel interactions. To do this, we used publicly available resources including the ontology terms of the genes in the network and information on sigma factors associated with the genes as well as known transcriptional units within the *Bacillus subtilis* genome. We created a list of the transcriptional unit information parsed from BsubCyc.org. This gives the transcriptional unit information for each of the 5873 genes contained in the BsubCyc online database. As each transcriptional unit may contain multiple genes, we derived the unique set of 522 transcriptional units from BsubCyc, the empirical p-values included are for those transcriptional units that have at least one of their genes in the inferred network, this gave 113 transcriptional units.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Empirical P-values	0.00	0.00	0.00	3.487e-05	0.00	6.000e-04

Using the transcriptional unit information, we globally compared the connected components to known transcription information. The results in Table 3.1 show for each known transcriptional unit the empirical p-value of the number of known connections found by the model occurring by chance, see section 3.5 for details. This showed the number of connections found is highly unlikely to have occurred by chance giving greater confidence in the model output.

We then looked at the specificity of the inferred network. First, we compared the proportion of the genes in a transcriptional unit that are connected in the network result, where the number of genes from a transcriptional unit is greater than one. We were then able to establish the proportion of genes in the transcriptional unit that are in the network result. Second, we calculated how many genes within a transcriptional unit are connected to a gene not in the same transcriptional unit. For the first Table 3.2 we have the proportion of genes in a transcriptional unit in the network result that are connected to at least one other gene in the transcriptional unit. Again, for the 113 transcriptional units present in the network, 100 percent of the genes in those transcriptional units are connected to at least one other member of their transcriptional unit.

For the same 113 transcriptional units, we calculated the proportion

Transcriptional unit is a set of genes that share a common regulation mechanism. Usually this includes shared promoter sequence and proteins that initiate transcription.

Table 3.1: Summary statistics for the multiple hypothesis corrected (using Benjamini-Hochberg) empirical p-values of the transitional unit information. The empirical p-values were generated by randomly perturbing the network node labels, retaining the graph degree structure. The number of connected components in each of the transcriptional units was calculated under each permutation and compared to those found in the network.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1	1	1	1	1

of the transcriptional unit in the inferred network, Table 3.3. Although there is a low minimum value, 75% of the transcriptional units have over 60% of their genes connected and the median value shows that 50% of the transcriptional units have at least 80% of their genes connected

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1818	0.6667	0.8000	0.7760	1.0000	1.0000

By assigning colours to the annotations of different transcriptional units we were also able to identify some potentially interesting results. For example, Figure 3.3 shows two known transcriptional units, denoted by green edges, linked by a previously unknown (blue) edge. This example is useful because the network terminates in a gene involved in sporulation (*dppA*). This makes it easier to test experimentally given the explicit phenotype involved in the hypothesis, as we predict that a difference in sporulation would be observable in the cells if the upper network (*ykfA-D*) is perturbed. As well as annotating the edges in the inferred network we used the Gene Ontology to annotate the nodes (genes) in the network [Blake et al., 2015]. There are shared Gene Ontology terms between genes in the same transcriptional unit and between the two transcriptional units. Genes tend to have more than one ontology term associated with them, therefore the visualisation does not always capture the shared terms between all nodes. However, we can still annotate and colour the nodes in the networks according to the full ontology terms. From a visual perspective, the transcriptional unit edge information was more useful as each edge can only be one of three possible values; both genes are in the same transcriptional unit, the genes are in two different transcriptional units or one or both genes have no known transcriptional unit information.

As well as the Gene Ontology annotation used in Figure 3.3 we also have sigma factor information for the genes, Figure 3.4 shows some examples of the network where nodes are coloured according to their sigma factors. The information on sigma factors is fairly sparse, however Figure 3.4 contains a subset of the genes where there is relatively dense coverage of sigma factor information and suggests that genes under the same sigma factor have been connected in the network. Nodes coloured white did not have an associated sigma factor in the database. Figure 3.4 also demonstrates the utility of annotating the network with sigma factor information. Where genes without a known sigma factor are present in a network that contains genes annotated with sigma factor information it is possible to generate

Table 3.2: Proportion of the genes in the network that are connected to at least one other gene in its transcriptional unit. This shows clearly that each gene is connected to at least one other member of its transcriptional unit.

Table 3.3: Percentage of the genes in a transcriptional unit included in the model. This shows overall a high level of connectivity between genes in the same transcriptional unit in the inferred network. The value of the first quartile shows that 75% of the transcriptional units have 67% of their genes connected.

The Gene Ontology is a database of annotations for genes classified into three groups, biological processes, molecular functions and cellular location. The Gene Ontology provides a common set of annotation terms for the research community to use that allows results from multiple experimental sources to be collated and compared.

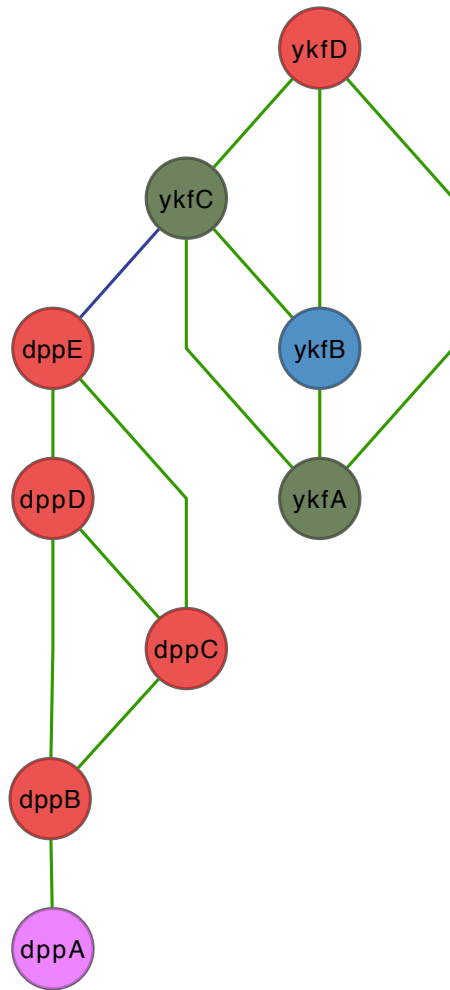


Figure 3.3: One sub network of the output that shows two previously unconnected transcriptional units linked together (as shown by the blue edge). The green edges denote connections between genes known to be in the same transcriptional unit. The nodes are coloured according to their GO terms. The red nodes have GO term, ‘Transport’, blue refers to ‘Metabolic process’, green to ‘Cell wall organisation’ and purple to ‘sporulation resulting in formation of cellular spore’.

hypotheses of the sigma factors for genes that are not annotated. This gives a potential mechanism for gene control that can be tested experimentally. Additionally, transcriptional units can be controlled by more than one transcription factor. By combining the experimental condition information and comparing any connected transcriptional units, sigma factors can provide information on the likely sigma factor controlling gene expression under an experimental condition.

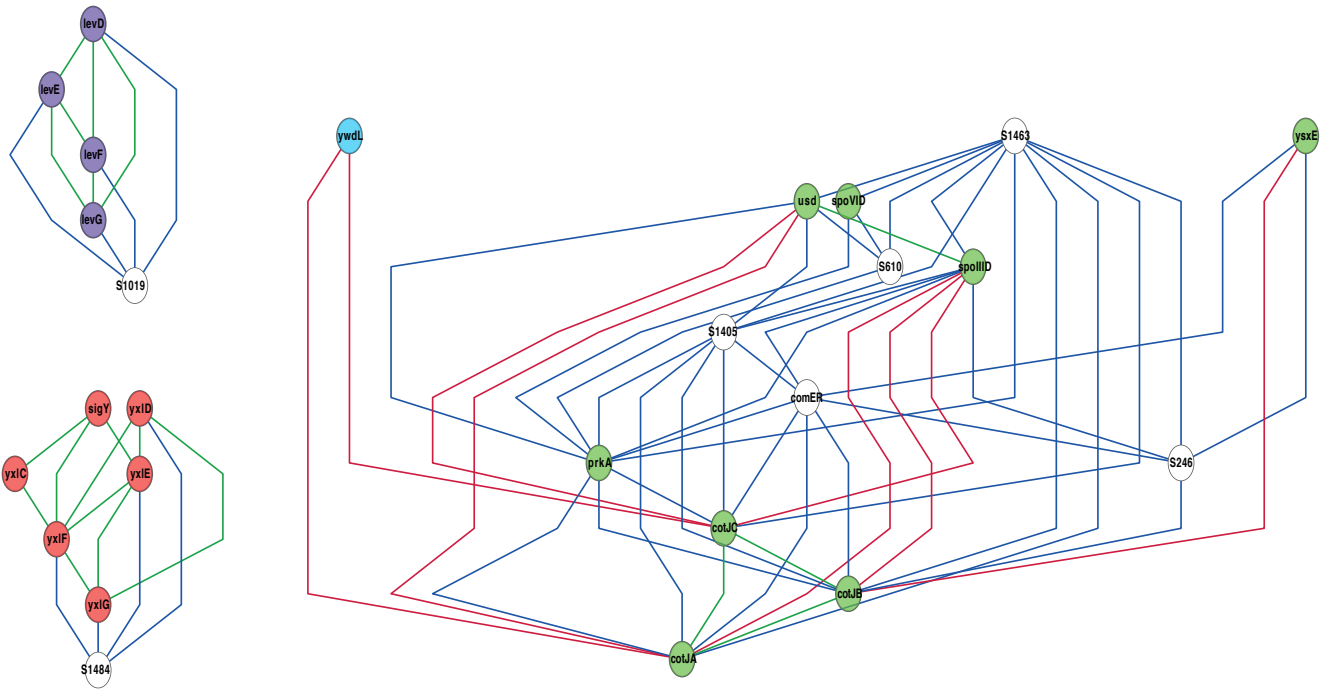


Figure 3.4: For the *Bacillus subtilis* data, these are examples of sub networks where the genes have known sigma factors. Nodes are coloured according to their controlling sigma factors. A white node indicates that the sigma factor is currently unknown. Edges are coloured according to known transcriptional units; green edges for genes in the same transcriptional unit, blue for nodes with no known transcriptional units and red edges between nodes known to be in two different transcriptional units. Annotating the network using the known sigma factor information can be used to generate testable hypotheses on the sigma factors of genes which are currently unknown. Knowledge of sigma factors controlling genes and the conditions they are active in will help in the design of synthetic biology circuits.

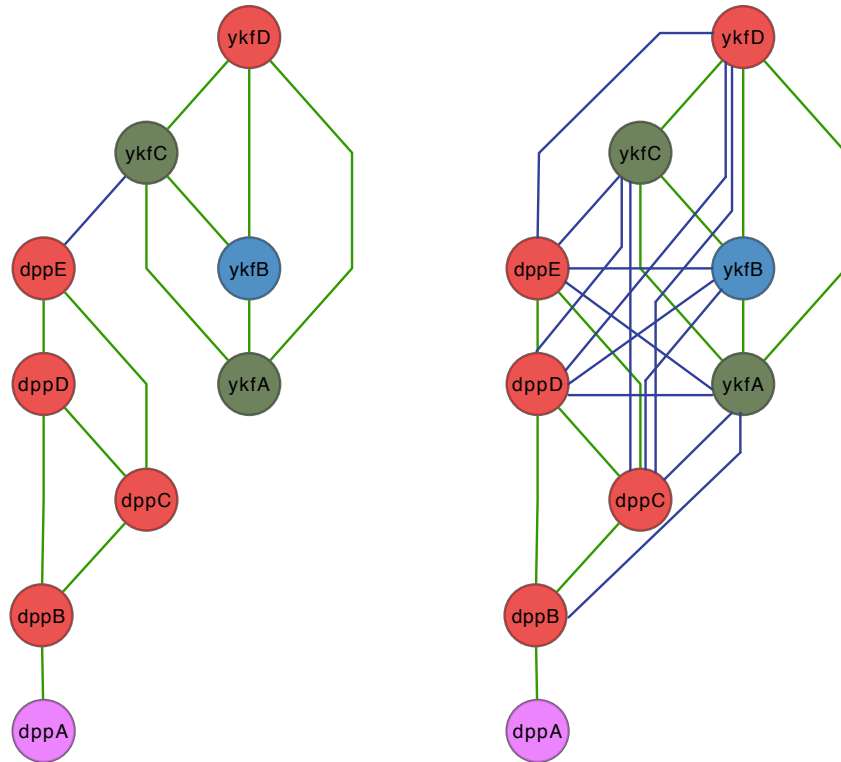
3.1.1 Parameter selection

The JGL model has two shrinkage parameters, the first penalises the size of the model to include only those with highest correlation values, the second penalises discrepancies between the models for the different conditions. The results from the previous Sections 3.1 are based on two shrinkage parameters that were chosen to give a parsimonious model, in terms of the trade-off between the size of the model and the time it would take to run the inference. This was a practical consideration, to allow the evaluation of the potential for using the JGL model with the selected data set. From a biological perspective, we may look to find the model that gives the best fit in terms of sensitivity and specificity to known interactions, or to minimise the false discovery rate of the interactions. This is particularly relevant for a biologist who would prefer fewer possible hypotheses to test as these are both expensive in time and cost.

Statistically, parameter selection would optimise an information criterion such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). An analytical form of AIC does exist for the JGL model that in theory can be used to select the two λ shrinkage parameters. For the JGL model where we have a solution Θ_k , which is the inverse covariance matrix for class k , the AIC is defined as:

$$AIC = \sum_k \{2\log(L(X|\Theta)) - 2p_k\}$$

Where p_k is the number of non-zero edges in Θ_k . Using the AIC criterion, the shrinkage parameters would be selected to maximise AIC value. The initial analysis was run with $\lambda_1 = 0.925$, and to begin to investigate the effect of this parameter on the output this was reduced to $\lambda_1 = 0.91$. The increase in computation time given this shift in the parameter was substantial. The second computation taking over 4 hours compared to 21 minutes with $\lambda_1 = 0.925$. Although a search of the parameter space could be run using a computer cluster, in addition to the computational expense the AIC only accounts for the statistical fit of the model, excluding the biological interpretability in parameter selection. This is indicated in Figure 3.5 where the change of the parameter used on a subnetwork of interest did not increase the number of nodes in the network. The effect of changing the parameter was to increase the number of significant edges between the nodes instead. Additionally, Figure 3.5 also shows how the parameter reduction resulted in more edges that are blue connections, that is more edges between genes not in the same transcriptional unit. We now had 10 blue edges to the single blue edge in the original sub network, Figure (a). This observation was also shown globally when comparing overall the number of different edge values based on the transcriptional unit



(a) One sub network that shows two previously unconnected transcriptional units linked together as shown by the blue edge. Green edges denote connections between genes known to be in the same transcriptional unit. The nodes are coloured according to their GO terms. The red nodes have GO term *Transport*, blue refers to *Metabolic Process*, green to *Cell Wall Organisation* and purple to *Sporulation resulting in formation of cellular spore*.

(b) Sub network of the output with parameter $\lambda_1 = 0.91$. We see that we largely retain the hierarchical structure at the lower shrinkage level but with more connections, including full connectivity between *ykfDBA* and *dppCDE*. Again genes in the same transcriptional unit are connected by green edges, those in different transcriptional units by blue edges.

Figure 3.5:

information.

Globally comparing this result with the previous output, with lambda parameter $\lambda_1 = 0.925$ the network contained 944 nodes and 8,821 edges. Moving the parameter to 0.91 we had 2,329 nodes and 32,370 edges, in both cases the second parameter remained the same, $\lambda_2 = 0.005$. The λ_2 selection was relatively small to allow the JGL algorithm to find differences between meta-conditions. A zero value for λ_2 is equivalent to calculating the networks for each meta-condition independently. Conversely, a large λ_2 value would force the networks to be the same or very similar for each class.

The L1 norm of off-diagonal elements of classes' Thetas for this model: 103, 143 and 65, in comparison to 17, 31 and 11 for the parameter values 0.925 and 0.005, this showed how there is a noticeable increase in the number of non-zero off diagonal values (edges) due to changing the parameters. Table 3.4 shows that as well as having more subnetworks reducing the shrinkage parameter increases the density of the subnetworks with a greater than two-fold increase in the average number of edges per subnetwork with a two-fold increase in the number of genes. This means that in addition to an expected increase in the number of genes included in the model the networks sparsity is reduced for both the initial and additional genes in the network. We also saw this in the plot of the empirical cumulative distribution function (ECDF) of the degrees in Figure 3.6. At the lower shrinkage level, there were a larger range of degree values with a maximum degree of genes up to 400 connections compared to around 100 for the higher shrinkage value. The ECDF for $\lambda_1 = 0.91$ was also consistently to the right of the first curve. This indicated a higher proportion of genes with larger degree values.

	$\lambda_1 = 0.925$	$\lambda_1 = 0.91$
Connected Genes	944	1811
Subnetworks 1	131	216
Subnetworks 2	175	247
Subnetworks 3	102	169
Avg. Edges 1	9.88	24.88
Avg. Edges 2	12.68	27.24
Avg. Edges 3	7.57	20.02

To evaluate the information content in the network results we summarised the number of different edge types according to the transcriptional unit information. Red is an edge between two genes that are within two different known transcriptional unit. Green are edges between genes in the same transcriptional unit. Blue edges connect to a gene that currently has no known transcriptional unit. Table 3.5 shows there are more edges between genes in the same as opposed

ECDF For a set of i.i.d random variables x_1, \dots, x_n with cumulative distribution function $F(m)$, the ECDF is an empirical approximation of $F(m)$ defined as:

$$\hat{F}(m) = \frac{1}{n} \sum_i \mathbb{1}_{x_i \leq m}$$

Table 3.4: Summary statistics of the networks for two different shrinkage values. For the inferred network three different meta-conditions were input into the JGL model. Overall the total number of connected genes shows an increase from 944 to 1811 due to the change in shrinkage parameters. We compared the different network structures for each of the meta-conditions, which gave a more detailed view of the network structure than the combined network. The values show the number of subnetworks, for example 131 subnetworks for condition 1 with $\lambda_1 = 0.925$ and 216 when $\lambda_1 = 0.91$ and the average number of edges per subnetwork which shows consistently more edges per subnetwork for lower shrinkage parameter.

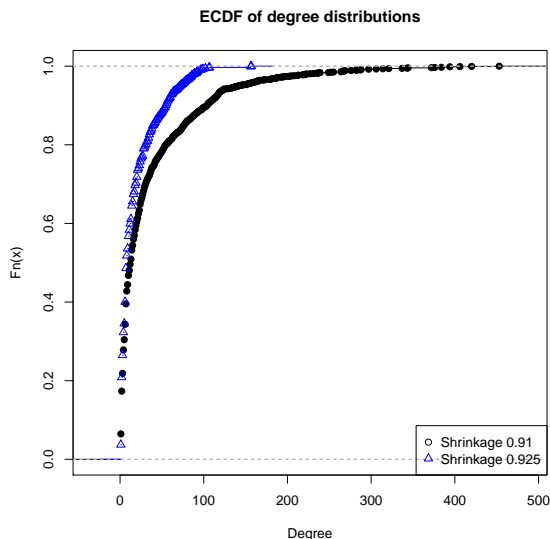


Figure 3.6: Empirical cumulative distribution functions for the degree of each of the nodes in the two networks with different shrinkage parameter values. The ECDF curve for the lower shrinkage value is to the right of the curve for $\lambda = 0.925$ indicating a consistently higher degree structure, greater connectivity of nodes, at the lower shrinkage value.

to different transcriptional units and that there are a large number of genes with still unknown transcriptional units. This gives many potentially novel interactions to investigate from the results. In terms of parameter selection Table 3.5 shows the number and percent of edges in each of the three transcriptional unit classes for two different λ_1 values. A reduction of the shrinkage parameter will either make no difference to the network, result in more genes being included in the network or more edges in the network, or both. The example subnetwork in Figure 3.5 shows an increase in edges but not genes for this subnetwork. However, this is not necessarily true for all subnetworks. Hence, we globally compared the number of genes, edges and the proportion of edge types between the two models. The results indicate a reduction in specificity of the model with the reduction in the shrinkage parameters. This is due to the fall in the percentage of the green (known) edges between genes with a simultaneous increase in red edges that indicate a connection between genes in different transcriptional units. Ideally, we might expect the network to be highly connective between genes in the same transcriptional unit but with minimal connectivity between different transcriptional units. By minimal it is important to note that transcriptional units do not always act independently of all other transcriptional units. Therefore, it is interesting to see connections between different units particularly under different conditions. However, once two genes in different units are connected, additional connections between genes in the two different units do not provide more information on units acting in concert and lead to a model which is more difficult to interpret. Here we implicitly separate those edges that connect two unconnected genes from

two different transcriptional units to those edges that connect two genes from different transcriptional units that are already connected to another member of the second transcriptional unit. For example, if we had two transcriptional units the first containing three genes a, b, c and the second f, g, h . If the two genes a, g were not connected to any other member of the second transcriptional unit, an edge between them would be informative in connecting the two transcriptional units. In contrast, if a were connected to g , b connected to f then a further edge connecting a to f would not be informative, though it could provide stronger evidence for the two transcriptional units being functionally related.

Model	Red	Green	Blue
$\lambda_1 = 0.925$	227	525	1753
$\lambda_1 = 0.91$	1102	949	7557
$\lambda_1 = 0.925$ Percent	9	21	70
$\lambda_1 = 0.91$ Percent	11	10	79

3.1.2 Analysis of effects of data inputs on JGL algorithm

We noticed large differences in running time of different data inputs within the same cluster of observations that are close in Euclidean distance (according to affinity propagation clustering). Using the same shrinkage parameters (λ 's), we wanted to identify the main reasons for the increases in computation time. This could provide a method for parameter selection, selecting data inputs, or give an *a priori* indication of how long the analysis could take to run before computing it. We found that using all 9 conditions in the cluster the algorithm had not converged within several hours. This cluster contains 9 different groups, of these 9 the results showed in previous sections include group numbers 1, 2 and 7. We used leave-one-out analysis on this data set of 9 conditions to see if there was one cluster that was causing the variation in computation times.

The JGL algorithm can be broadly separated into two stages. The first is the screening of the covariance matrix to determine its block diagonal structure. The second is to find the inverse of each of the blocks identified. The system times for the first stage, identifying the block structure, was 36 seconds for the subset and 219 seconds for all nine groups in the cluster. This implies the difference in running time is due to the second stage of the JGL algorithm.

We restricted the number of iterations of the convergence algorithm (the second stage of the JGL algorithm) to 5 to see if we can identify the data that is causing the difference in results by comparing computation time. Figure 3.7 shows the computation time after leave-one-out

Table 3.5: Summary statistics of the different transcriptional units for the two different network results under different shrinkage parameters. The red edges are those where two genes are connected that are not in the same transcriptional unit. Green edges are for connections between genes in the same transcriptional unit and blue edges denote an edge between connecting genes where at least one currently has no known transcriptional unit listed in the BsubCyc database. The absolute number of edges show that in both cases there are the most edges with no known transcriptional unit information, though with the smaller shrinkage parameter there are now more red than green edges. This, alongside the percentage values of the three different edge values shows that when reducing the parameter to 0.91 we have a relative increase in connections between genes not in the same transcriptional unit.

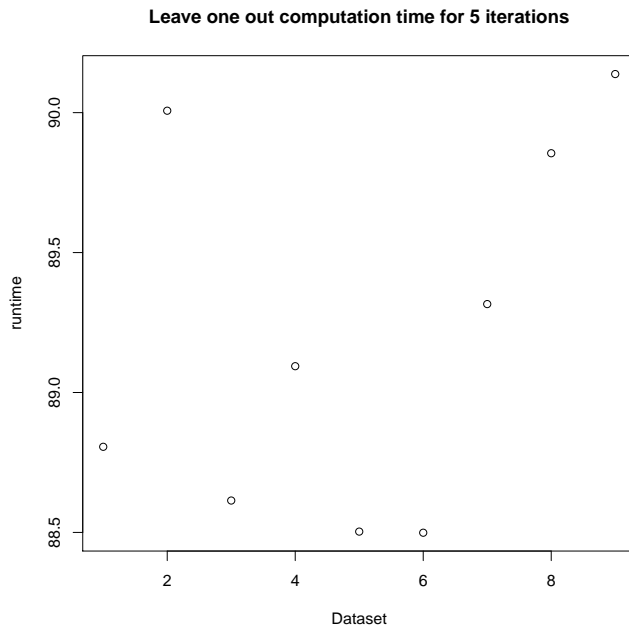


Figure 3.7: The plot shows the run time (in seconds) for 5 iterations of the algorithm to find the inverse covariance matrix. The x-axis denotes which one out of the 9 data sets was excluded from the input. There is very little difference between the iteration times after removing one of the classes indicating that this isn't the bottleneck in the algorithm

analysis for each data set. There are no obvious differences between these timings, implying that the difference in input groups will be shown by the different rates of convergence rather than the time of a single iteration. That is, the data inputs are more likely to impact the number of iterations to convergence rather than the time it takes to complete one iteration of the algorithm to find the inverse of the covariance matrix. Consequently, we looked instead at the blocks generated for the different data inputs to try and identify what would cause different convergence rates.

Figure 3.8 shows the maximum block size after removing each data input (class) in turn. This shows a large difference in the maximum block size after removing class 8, whilst leaving data sets 1, 2 and 7 that were used to infer the network from Section 3.1. The two classes that make a difference to the maximum block size are 2 and 8. When both are included in the model the maximum block size is close to 400. The removal of class 8 reduces the maximum blocks size to 100. This makes the computation of the inverse covariance matrix easily tractable. This shows that the maximum block size may be a good indicator of how long the JGL algorithm will take to run. This is understandable as the size of a covariance matrix determines the number of parameters to be estimated. Figure 3.9 supports this observation as there is little difference between the overall distribution of the block sizes with the exception of the maximum block size for

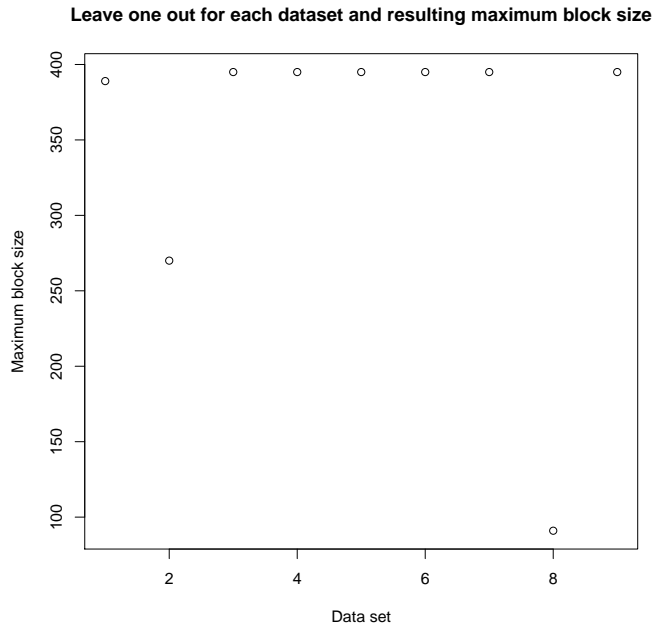


Figure 3.8: The maximum block size, after finding the block structure of the covariance matrix is shown after removing each data set from the input in turn. The maximum block size is all those genes for which the covariance is larger than the selected shrinkage parameter in any one of the data sets. There is a large drop to under 100 for the maximum block size after removing data set 8.

each of the leave-one-out class models.

We may therefore be able to use the maximum block size to select shrinkage parameter values as well as selecting experimental conditions to input into the JGL algorithm. It can also be used as a metric to combine data sets, that is, as an alternative to the euclidean distance metric that we used initially. We hypothesised that using the block statistics we may be able to select reasonable shrinkage parameters, or parameter ranges, prior to running the model. This has the advantage of being computationally more efficient as well as taking into account our knowledge of biological networks as sparse networks, in contrast to the statistical measures AIC and Bayesian information criterion (BIC) that do not take biological sparsity into account. Although the FDR methods are a more intuitive and relevant metric they do still require the estimation of error probabilities which, for regularisation models, require recalculating the model with multiple shrinkage parameter values in addition to bootstrap methods for the data. This means that these measures will be computationally demanding.

From a biological perspective, we assumed that the shrinkage parameters will result in a sparse network given a high enough shrinkage value. A more interesting perspective for parameter selection would be to consider the alternative situation whereby, following the reduction of the shrinkage parameter the sparsity of the network is reduced and the signal to noise ratio is decreased. To investigate the signal to

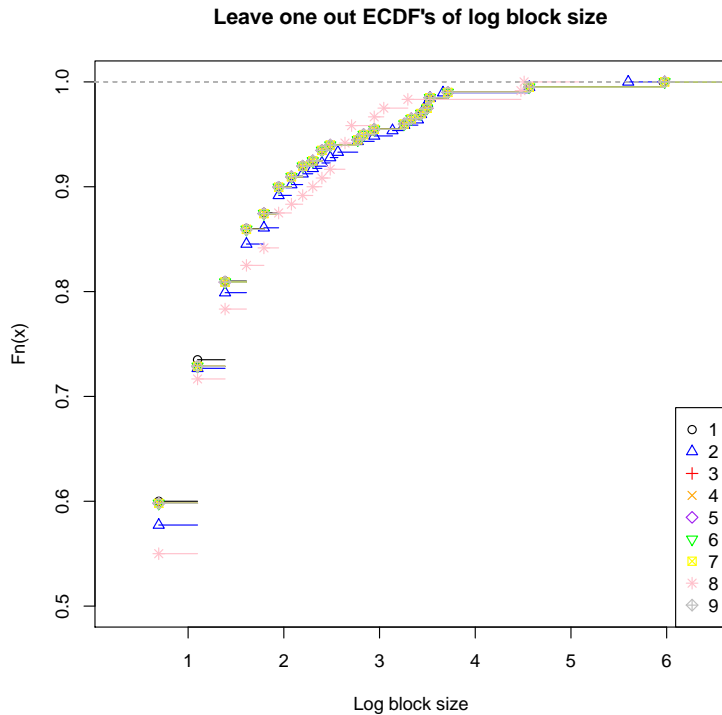


Figure 3.9: The empirical cumulative distribution functions for the block sizes of the data set with leave-one-out on each of the classes. This shows little difference in the distributions of the block size except for the maximum block size for each. The symbol for 8 where we have removed sample 8, is shown to have a lower maximum block size than after leaving out each of the other samples, as would be expected.

noise ratio in the networks we looked at the behaviour of the block structure for various shrinkage parameters. Figures 3.10, 3.11, 3.12 show the maximum block size, the number of blocks and the change in the maximum block size for each of the classes individually as well as combined. These metrics were chosen because we have already shown that the size of the blocks has the largest impact on the computation time for the JGL algorithm. Moreover, intuitively we know that the resulting network should be sparse due to the sparse nature of biological networks. Therefore, a network containing a single large hub would be expected to contain more noise or false positive results. This situation can arise since two genes may not be functionally connected, yet their correlation value is likely not to be exactly zero. The shrinkage parameter (λ_1) is therefore useful to shrink those correlations below the threshold (λ_1) to zero. We assumed there is a baseline level of correlation in the data that will be observed due to random noise as opposed to biological signal. By observing the behaviour of the network with varying shrinkage parameters we aimed to identify this baseline level of correlation to maximise the signal to noise ratio within the model.

By definition the maximum block size must be strictly decreasing with increasing shrinkage values, therefore we also looked at the change in the maximum block size to see if there were significant

changes when varying the parameter value that could indicate that the signal/noise barrier has been crossed, Figure 3.12. Arguably considering only the maximum block size is more of a computational constraint. On the Figures 3.10,3.11, 3.12 we have added lines for shrinkage values 0.925 and 0.91. From our previous analysis, we know that this change in parameters results in a significant increase in computation time from 21 minutes to over four hours. This is highlighted by the number of genes in the largest block doubling in size, Figure 3.10. However, from a biological perspective it may be more useful to consider the summary of the full network as shown by Figure 3.11. Here we see a bell shape form for the change in the number of blocks as the shrinkage parameter varies. The number of genes or edges included in the model must be strictly increasing as the parameter value decreases. As the λ_1 falls below a particular value, here around 0.9 the sparsity in the network is lost and previous sub networks are combined resulting in fewer blocks overall. Therefore, we argue that the λ_1 value corresponding to the maximum number of blocks is a useful greatest lower bound for λ_1 selection. In our case, we have used a shrinkage value above this bound. We have also considered the computational time that would be required to analyse more of the network. However, this analysis suggests we have more biological signal than noise in the inferred network and stringency in our analysis enables us to focus on smaller networks in greater detail as opposed to doing global analysis on larger networks that is not the focus of this study.

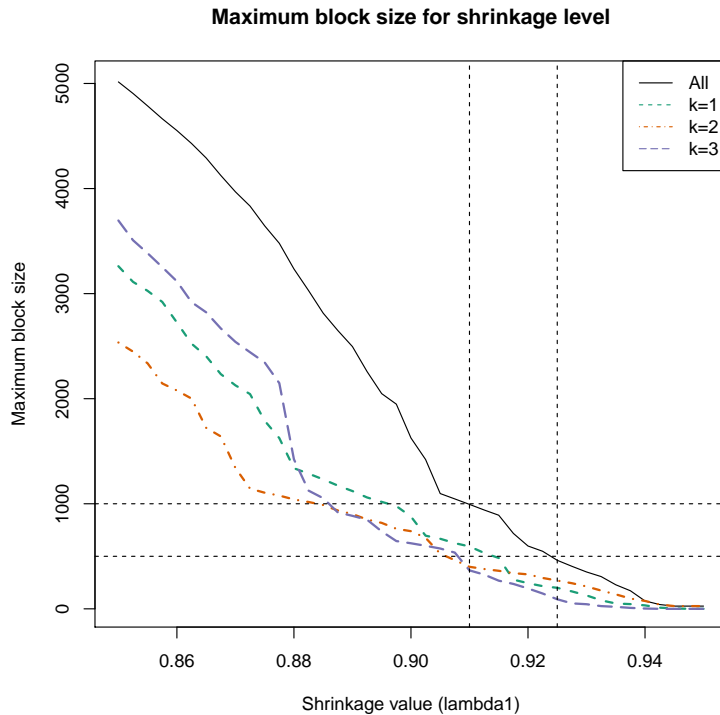


Figure 3.10: The plot shows the maximum block size, number of genes, for each of the three conditions individually as well as combined. These are shown for different levels of the shrinkage parameter λ_1 that controls the significance level of the correlations. We have also marked lines for the two λ_1 values used in the previous analysis at 0.925 and 0.91 as well as maximum block sizes at 500 and 1000. The movement from 0.925 to 0.91 results in over twice the genes in one block and this results in a large increase in computation time.

3.1.3 Agglomerative clustering for JGL

Following the analysis of the block structure for the JGL method, we needed a method that allows us to combine data sets which could further be used as input into the JGL model and is also able to identify data sets that will have the greatest impact on the maximum block size as outliers. We used the data from the previous analysis which had 9 different experimental groups. Our previous analysis showed that group 8 gave the largest change, in the maximum block size and consequently we wanted our clustering method to identify group 8 as an outlier. If we take the median expression value across the clusters to combine into one gene-expression profile for each cluster we can then use affinity propagation clustering to see how this method clusters the data sets.

Clusters:

Cluster 1, exemplar 2:

1 2 3 6 7 8

Cluster 2, exemplar 4:

4

Cluster 3, exemplar 5:

5

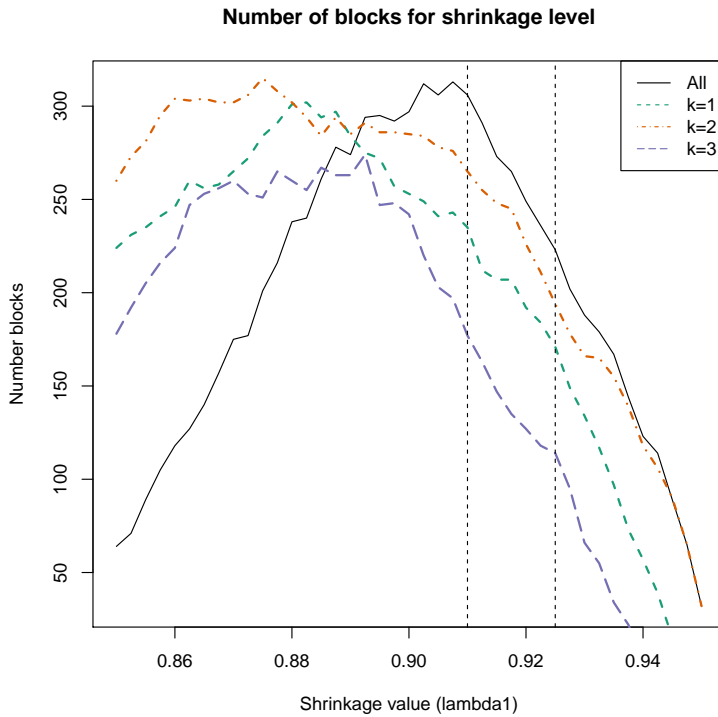


Figure 3.11: The plot shows the number of blocks for each of the conditions individually as well as combined at different shrinkage, λ_1 values. The number of blocks gives us an overview of the level of sparsity we can potentially expect for a given shrinkage value. For example, with low numbers of blocks we have two possibilities, few blocks with few genes (large shrinkage parameter, here approximately 0.94) and a few blocks with many genes (small shrinkage parameter, in the plot at around 0.85).

Cluster 4, exemplar 9:
9

We can see that by using this Euclidean distance-based clustering between the profiles we are unable to identify group 8 as an outlier.

We developed an agglomerative clustering method for combining data sets. This used the maximum block size as a measure of distance between different groups.

	1	2	3	4	5	6	7	8	9
1	1	91	16	16	16	16	16	270	16
2	91	1	91	91	91	91	91	389	91
3	16	91	1	1	1	1	1	269	1
4	16	91	1	1	1	1	1	269	1
5	16	91	1	1	1	1	1	269	1
6	16	91	1	1	1	1	1	269	1
7	16	91	1	1	1	1	1	269	1
8	270	389	269	269	269	269	269	1	269
9	16	91	1	1	1	1	1	269	1

Table 3.6 shows the pairwise maximum block sizes for each of the groups. This shows an increase in the size of each pairwise match with

Table 3.6: Maximum block sizes for each pair of data inputs. These are used for input into the agglomerative clustering algorithm.

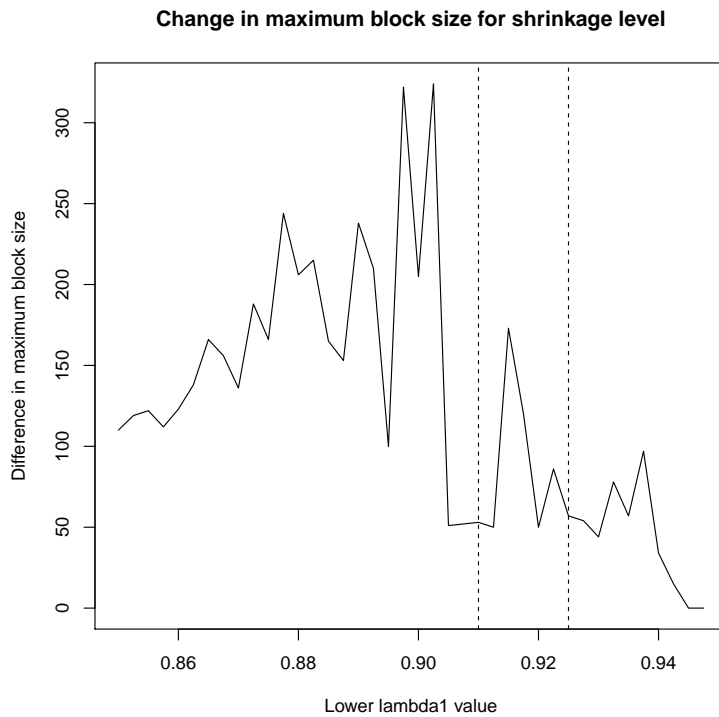


Figure 3.12: We plot the change in maximum block size with respect to the shrinkage parameter λ_1 . The difference in maximum block size for the model between two different λ_1 values is shown on the y-axis. λ_1 values are varied at increments of 0.0025. The λ_1 value on the x-axis relates to the lower value. For example, the line vertical line at 0.925 indicates the difference in maximum block size between two models, one where $\lambda_1 = 0.9275$ and the second when $\lambda_1 = 0.925$.

group 8 as was expected. With the number of clusters set to 2 we have the following result. The method uses agglomerative clustering and recalculates the scores for each group based on the maximum block size of all conditions in one group after each iteration. To determine the maximum block size, the two shrinkage parameters of the JGL need to be selected. In this example we use the same values as our initial model: $\lambda_1 = 0.925$, $\lambda_2 = 0.005$.

Cluster 1:

8

Cluster 2:

2 1 3 4 5 6 7 9

The time in seconds for this was 222. The algorithm was also run with a maximum number of 3 clusters, in this case the computation time increased to 260 seconds. The result we have when 3 clusters are allowed is as follows:

Cluster 1:

8

Cluster 2:

2

Cluster 3:

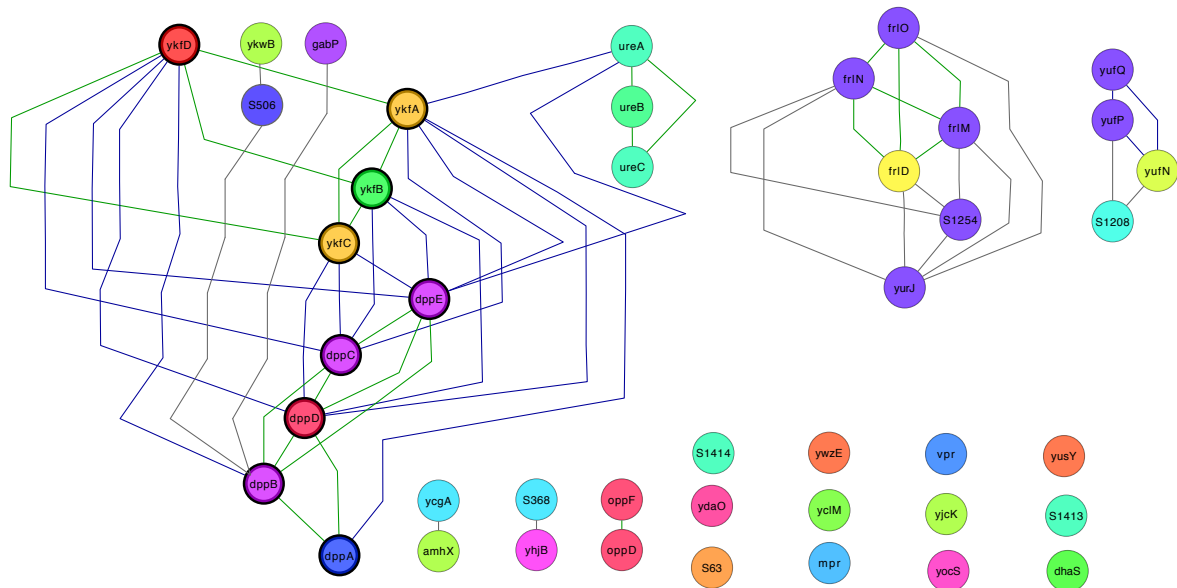
3 1 4 5 6 7 9

We used the agglomerative clustering method with maximum block size as a measure of distance. Setting the number of clusters to three we found group 8 to be an outlier as it has separated from all other conditions, as shown above. The algorithm combines into groups the data sets closest to each other first. Therefore, the order of data sets within a group are according to the distances between them. For example, with 2 clusters the second cluster has the ordering, 2 1 3 4 5 6 7 9. Table 3.6 showed that after data set 8, data set 2 has the largest distances, this is reflected in its first position in the second cluster.

3.1.4 Exploring subnetworks

In the two previous sections, we looked at the effect of changing the shrinkage parameters on the network. Changing the parameter from 0.925 to 0.91 resulted in an increase in computation time from 21 minutes to around 4 hours. Moreover, this change in parameters and increase in computation time did not provide additional genes connected to our example subnetwork. Ideally, we would have liked to expand the subnetwork to find additional genes or transcriptional units that are interacting with our network. From our analysis of the JGL algorithm we also know that the block structure is the main determinant of computation time as well as determining the possible genes which could be connected to each other at the completion of the algorithm; two genes not in the same block cannot be connected in the resulting network. Consequently, we used the block structure of the covariance matrix to analyse individual sub networks we were interested in. Given the example, as shown in Figure 3.3, we can rank the remaining genes according to their covariance with the nodes in the sub network and use this to select genes to include in the block of our subnetwork.

The results in Figure 3.13 show the inference from JGL with the selected nodes and additionally up to 30 of those closest to the sub network, giving a maximum of 39 nodes. The shrinkage parameter input into the model is the greatest lower bound of the lambda parameters needed to find a connection between the additional genes and those in the sub network. In this case, the λ_1 parameter used is 0.825. The inference for these 38 nodes took 2 seconds. This shows how, for a sub network of interest, it is computationally better to analyse the sub network on its own rather than with the rest of the network. This is



because having shrinkage parameter 0.825 for all sub networks would result in a network that would take days to infer.

Figure 3.13 displays the expanded subnetwork rendered in Cytoscape [Shannon, 2003]. Of the additional genes selected by the subnetwork algorithm only a fraction of them are connected to the original sub network. Showing that not all the genes in Figure 3.13 are connected to each other. This shows that although some genes have high correlation to those in our subnetwork they have lower partial correlation, indicating that there may be an indirect as opposed to causal link between them. According to the BsubCyc database, these transcriptional units do share a common sigma factor A, and the presence or levels of the sigma factor may explain why these are the strongest 30 in terms of correlation. However, relative to our original network the shrinkage level is quite low, and we see that the partial correlation is not high enough to be significant in this network. Together this evidence suggests that these transcriptional units do not necessarily interact directly with each other but share some common regulators leading to the correlation of their expression levels. We would expect to observe these types of relationships, as there are only 10 sigma factors in *Bacillus subtilis*, that are used to control all transcription in the genome. The regulatory information is summarised in Table 3.7 The genes that are connected to our initial subnetwork of interest are *ykwB*, *gabP*, *S506* (all of which are connected to *dppB*).

Figure 3.13: Output rendered using Cytoscape showing the inference of up to an additional 30 nodes to a previous network containing 9 nodes. The original 9 nodes are contained within the largest subnetwork, highlighted with a black border around their nodes. Within the expanded subnetwork there are also nodes with GO terms for *Transport* included the network. GO terms are used to colour the nodes, with white nodes having no known GO terms. Transcriptional unit information is used to colour the edges. Green edges between genes in the same transcriptional unit. Orange edge between genes in different transcriptional units and blue for genes with no transcriptional unit annotation.

UreA, *ureB* and *ureC* which are connected to each other and *ykfA* from the original network. Both the known transcriptional units of *ureABC* and *ykfA* can be inhibited by the regulatory factor *codY* and their gene products are located within the cytoplasm. *UreABC* can be activated by either sigma factor A or sigma factor H. *YkfA* is involved in proteolysis, the breakdown of proteins. One method for the removal of excess amino acids is ureagenesis. *UreABC* are all primarily involved in urease activity therefore both *ykfA* and *ureABC* are involved in related processes. Of those genes connected to *dppB* the most annotated is *gabP*, both genes are inhibited by *codY*. *GabP* is activated by *tnrA* and both *gabP* and *dppB* are activated by Sigma A. *GabP* and *dppB* are both located in the plasma membrane and have Gene Ontology terms involved in transport processes. The transcriptional unit containing *frlMNOD* is involved in fructosamine catabolism and carbohydrate metabolism, that is, the breakdown of sugars for carbohydrate-based energy.

Genes	Activator	Inhibitor	Sigma Factors
ykfABCD		CodY	
dppABCDE		AbrB, CodY	A
ureABC 2	PucR	GlnR, CodY	A, H
gabP	TnrA	CodY	A
frlMNOD		FrIR, CodY	A
yufPQN		CodY	

Table 3.7: Given a sub network of interest containing *ykfABCD* and *dppABCDE*, we found the 30 most strongly correlated genes to those in the network. For these genes, using the JGL model, we found the partial correlation network. This inference found further transcriptional units *ureABC* and *gabP* and two additional subnetworks *frlMNOD* and *yufPQN*. The table summarises the transcriptional regulators of all genes in the sub networks for those where there are some known transcriptional regulators and the network contains at least 3 genes. This shows some commonality between the transcription promoters which could explain the correlation between the gene expression. However, no known shared transcription activators are expected because there is no significant partial correlation over all transcriptional units.

3.2 *Decomposing large networks*

CELLULAR RESPONSES TO ENDOGENOUS or exogenous stimuli are complicated and often include the interaction between multiple transcriptional units. It is therefore common that analysis of gene expression networks results in highly interconnected networks which comprise hundreds of genes with thousands of edges between them. However, transcriptional units usually contain fewer genes, in the order of tens as opposed to hundreds of genes. We wanted to identify transcriptional units from the larger networks. This would give information on sets of genes that we expect to act together. These gene sets may combine with other transcriptional or functional units depending on the state of the cell. This is relevant for synthetic biology where design of circuits is often at a low level, concentrating on the modification of single transcriptional units or regulatory modules. In this way, we aimed to identify smaller transcriptional units for design purposes and how they interact with the rest of the regulatory network under different conditions. This would then provide information on the network to be altered and how these modifications would affect the overall state of the cell.

In addition, genome wide networks are hard to visually explore, consequently these large networks are often analysed using global methods such as over representation of ontology terms or known interactions between transcription factors and their regulons. In this way information on, for example, master regulators or genes that have the widest influence on the cell under a given condition, are identified. This is useful for identifying critical transcription factors. This is perhaps more important for identifying drug targets where influencing factors central to the disease phenotype is advantageous. In contrast, to design synthetic circuits a detailed view of the transcriptional mechanisms in the cell are required. From an engineering perspective, detailed information enables the design of circuits constructed from cellular processes that have well defined inputs and outputs.

Using the JGL model the inferred network is not genome-wide, however, it still contains close to a thousand genes. The largest sub-network is shown in Figure 3.14, the edges are coloured according to the conditions in which the edges are present. This Figure shows that there are parts of the network that are predominantly one edge colour indicating that our method for decomposing the network using the edge information may be effective. From Figure 3.14 we can see that we may expect one sub network with yellow, one with green and one connected by red edges. This would be three out of the seven possible edge values in the inferred network. The input into the JGL for the

inferred network was three of the meta classes found from the affinity propagation clustering. The three meta classes are given below along with the different experimental conditions included within them.

- Class a: Malate, Glucose, M9 growth medium with glucose at exponential phase, cells at high temperature.
- Class b: M9 growth medium with glucose at trans phase, M9 medium with LB culture, with different carbon sources, cells at high temperature, cells grown in SMM, cells before and after addition of malate, cells after addition of glucose.
- Class c: M9 growth with glucose at exponential phase, cells before and after treatment with glucose with LB and/or M9 medium.

There are seven edge values that that can arise from these three meta classes, these are:

- Edge value 1: Class a
- Edge value 2: Class b
- Edge value 3: Class a and b
- Edge value 4: Class c
- Edge value 5: Class a and c
- Edge value 6: Class b and c
- Edge value 7: Class a, b and c.

The seven edge values are comprised of three values (1,2,4) for edges present in only one of the three classes, three edge values (3,5,6) that are combinations of edges present in two of the three classes and the final edge value (7) for edges that are present in all three classes. From a mathematical perspective, standard graphical algorithms exist to find certain elements of the network. These include, the minimal paths between two nodes and node(s) that are central hubs in the network [Langfelder et al., 2013, Managbanag et al., 2008]. These may be useful in other areas, for example, finding quickest routes in a transport network but minimal paths which are not as intuitively useful from a biological perspective. In the context of regulatory networks, we wanted to know which transcriptional units are active. The genes contained within these regulatory networks and ideally the main regulators of these networks. Causal effects and hierarchical information are relevant to regulatory network analysis but the distance the signals in the cells must pass is not. We investigated several methods

for decomposing networks from the JGL output based on the edge values. We hypothesised that we might find smaller regulatory modules, within a larger network. These smaller subnetworks were identified as nodes connected by edges with the same value. To decompose networks based on edges values we made two assumptions 1) that genes in the same regulatory network are active under the same experimental conditions and will have the same edge values 2) not all regulatory networks will be active under all conditions and therefore not all edges in the regulatory network will have the same edge values.

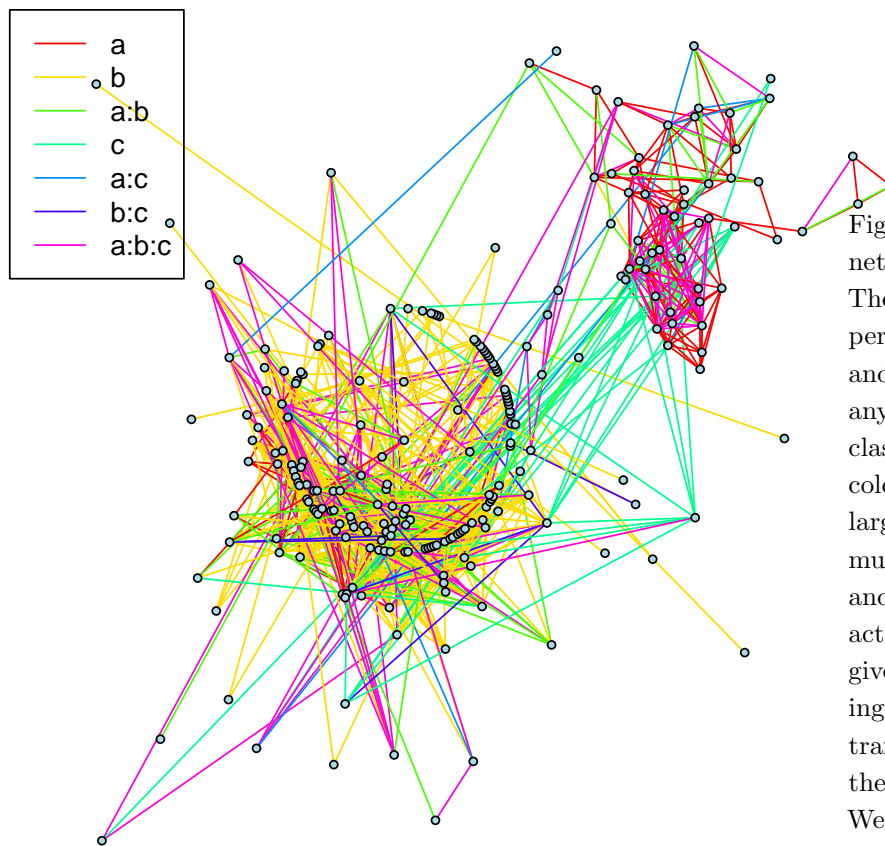


Figure 3.14: The largest sub-network from the JGL output. There are three different experimental classes (a, b and c) and each edge may appear in any combination of these three classes, as shown by the edge colours. We assume that this large network is composed of multiple transcriptional units and that not all units may be active under all conditions. This gives the potential for decomposing and identifying individual transcriptional units by using the edge condition information. We can see potential separate parts of this network that would have just yellow, green or red edges. The value of the conditions each edge corresponds to are shown in the legend.

3.2.1 Affinity propagation clustering

This first method uses an existing clustering method to decompose the network. The affinity propagation method takes as input a similarity matrix between all the nodes and uses this to determine the number of clusters and the elements of each cluster. Therefore in order to decompose a given subnetwork we needed to represent the network as a similarity matrix, see Section 3.5. Given the similarity based on the current network we wanted to penalise the connection of two hubs (transcriptional units) that are present under different combinations of edge classes. To do this we first identified those nodes that share common parents but where the edges between the nodes and the parent are different class combinations. Where the class membership differs a penalty between the two nodes was added to the similarity matrix, this penalty is a negative value that can be used as the affinity propagation clustering method will accept negative similarity scores. For different penalty values, we looked at the number of different classes in a cluster. We analysed the results of different size penalty values on the resulting clusters. The clustering varies according to overall number of clusters as well as the number of conditions within a cluster. If a cluster contains one condition, all the edges in the cluster are the same edge type. In our example, with three classes, this edge type can be any single combination of the classes. For example, a cluster containing all edges as $a:b:c$, which is present in all three classes, would have one condition as would a cluster where all edges are in class a only. Ideally, we would like as few conditions as possible in one cluster. As the penalty value is designed to penalise multiple conditions in one cluster, we would expect to see more clusters with one condition as the penalty increases and fewer with multiple conditions. However, although there are some falls in the number of clusters for three and five conditions this is not a linear relationship.

Figure 3.15 doesn't show a clear pattern in the number of conditions in each cluster as the penalty is varied. Ideally, we hoped to see a decrease in the larger number of conditions and more clusters with only one or two conditions as the penalty value increases. The lack of pattern with the increasing penalty value suggests that the affinity propagation clustering exemplar selection is having a larger effect than the penalty metric in determining the clusters.

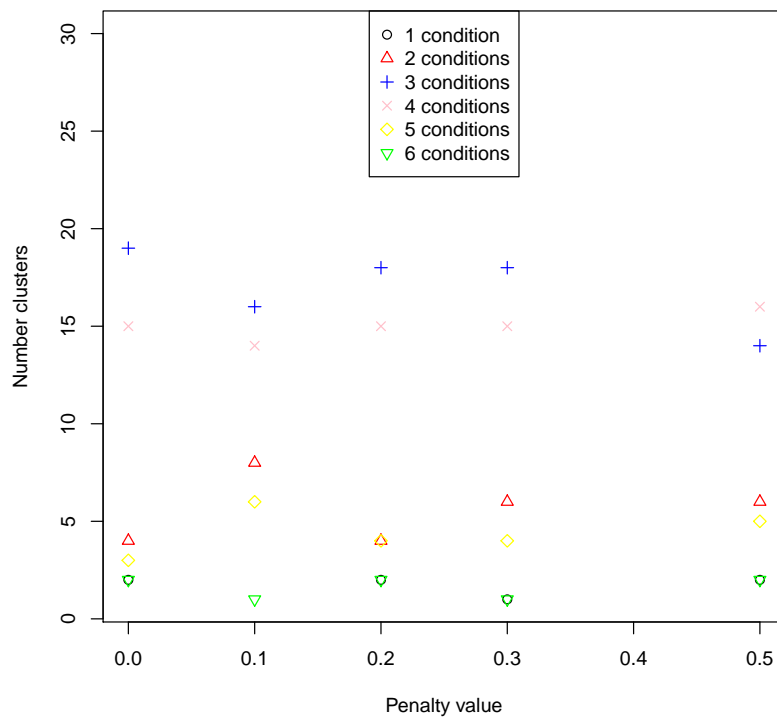


Figure 3.15: Plots of number of clusters that have between 1 and 6 different conditions in them, for different penalty values. These clusters were calculated using affinity propagation clustering of the original network. Penalty values were added to the similarity matrix when edges were present in different conditions. One condition can contain more than one class. That is, all edges could have edge type a:b:c, this would be one condition. The plot shows there is no obvious trend between the the penalty value and the number of clusters, for each of the conditions.

3.2.2 *Deterministic network separation*

By using a deterministic algorithm to separate large networks we know that a constant part of the graph will always separate in the same way irrespective of whether other parts of the network change. This in contrast to the affinity propagation method that can alter in number and categorisation of clusters for all nodes given any one change in the data input. The first deterministic method we used starts by identifying, for a pair of nodes, all those nodes they are both connected to but that are connected under different experimental classes. Where the number of these nodes exceeds the cutoff selected by the user all connections between the pair of nodes are removed. By using a cutoff parameter we allow for noise in the data as it is possible that we have false positive or negatives in the edges. The results for decomposing the network method using different cutoff values is shown in Figure 3.16. The first image with Cutoff 240 is the original network. As the cutoff value is decreased, the algorithm becomes stricter on the edge connections, consequently there is also a decrease in the number of edges in the network.

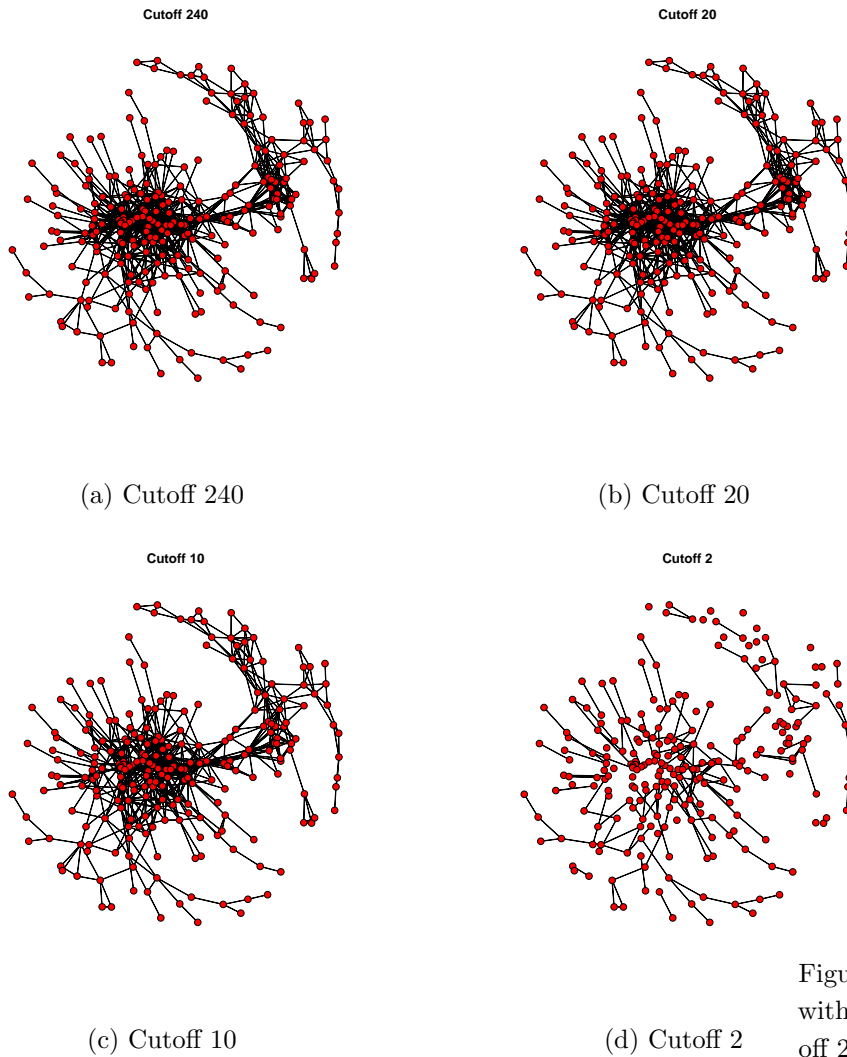
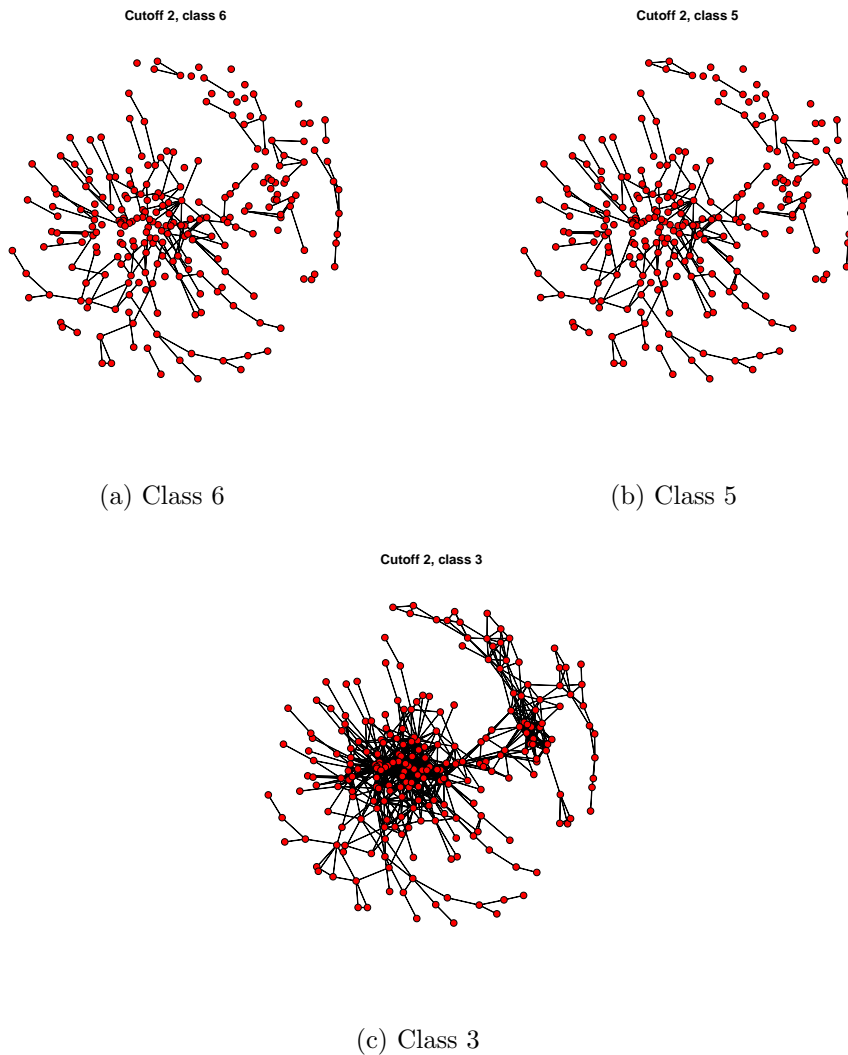


Figure 3.16: The graph split with different cutoff values. Cutoff 240 indicates no splitting of the network; the original network. The deterministic algorithm for splitting the network removes all edges between two nodes whose number of edges to neighbouring nodes with different class values exceeds the cutoff value. Reducing the cutoff value means more edges are removed, this is seen in the Figure as sparser networks are shown as the Cutoff value decreases.

A second deterministic algorithm considered the values of the class edges. That is, when counting differences in edges between a pair of nodes and their shared nodes, only edges in classes selected by the user were counted towards the total number of different edges. For each of the examples below, 2 differing edges are required to separate nodes. The varying parameter was those nodes that are counted towards the cutoff.



To help select which cutoff parameters to use we generated summary statistics over the different graphs. This statistic is a weighted average of the number of genes in a cluster and number of classes in the cluster, we added a penalty term to penalise sparseness by adding the number of single nodes to the score. We could then choose the parameter values that minimise this statistic. The values of this statistic for different parameters are given in Table 3.8. In this example, the

Figure 3.17: Deterministic splitting of large networks under cutoff value 2 with different class conditions. The lower the class value the less stringent the cutoff criteria; Class 3 has significantly more edges remaining after decomposition than either Class 5 or 6.

lowest value is for cutoff 2 with the sum of the different class values being less than 5.

\sum Class <	3	4	5	6	7
Cutoff 5	1680	1680	543	322	322
Cutoff 4	1680	1680	186	134	134
Cutoff 3	1680	834	91	77	77
Cutoff 2	1680	319	73	77	77
Cutoff 1	1668	155	204	228	228

For cutoff value 2 we also plot the graphs under different classes, Figure 3.17. Again, this shows how the class effects the splitting of the network. When few edges are defined as different, as shown in class 3, many of the edges remain in the network. In contrast, when the threshold is increased to 5 or 6, more edges meet the splitting criteria and this results in more nodes being separated from each other.

Table 3.8: The network scores for different parameter values. The cutoff parameter gives the number of different valued edges that need to be present for two nodes to be separated. For example, cutoff 1 means two nodes connected to a single third node under two different class values would be separated. With cutoff 5, two nodes would each have to be connected to five additional nodes under different class edges to be separated. Those edges that are counted as different are given by the class parameter. The class values for the edges can take values 1 to 7. The definition we have used is that the sum of the different edge values be less than a selected threshold. In this way, a smaller value is less stringent as there are fewer combinations of edges that would be under this threshold and thus less splitting. For example, if the threshold (as shown in the first column of the table) is 3, the only combination of edge classes to meet this criteria are for one edge to be class 1 and the other class 2, then the sum of these differing edges is 3.

3.2.3 Simulation methods for splitting networks

In the deterministic algorithm, once nodes are separated all shared nodes are disconnected from both nodes. Arguably we would have liked these nodes to be separated from only one of the nodes. Therefore, we developed a simulation-based method for separating large networks. Under the constraint that each node can only be assigned to one class, Figure 3.18.

We simulated class assignments for each node and take the graph that is closest to the original network, see Section 3.5 for details. The decomposed graph of the largest subnetwork following 50,000 simulations after nodes with differing class edges have been separated is shown in Figure 3.19.

From the network decomposition, we can visually inspect the resulting smaller subnetworks. One example of these, identified by Orr Yarkoni of the Ajioka lab in the Department of Pathology, University of Cambridge, is Figure 3.20 that shows a potentially novel connection between *spoVIF* and the known transcriptional unit between *cotVWX*. The literature shows *gerE* as the regulator of *cotVWX* [Driks, 1999]. Therefore *spoVIF* may be required together with *gerE* to control *cotVWX*, or alternatively only one of them is needed to regulate it. This is something that could be validated experimentally.

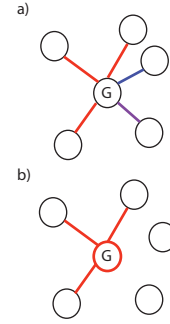


Figure 3.18: Each node in the decomposition method is assigned to one class. a) In the original network, node G is connected to 5 other genes, 3 through orange edges, one blue and one purple. The difference in edge colours indicates a different set of meta-conditions. b) Using the simulation method, node G is assigned to the orange group. Therefore, all edges apart from the orange edges are removed in the decomposed network.

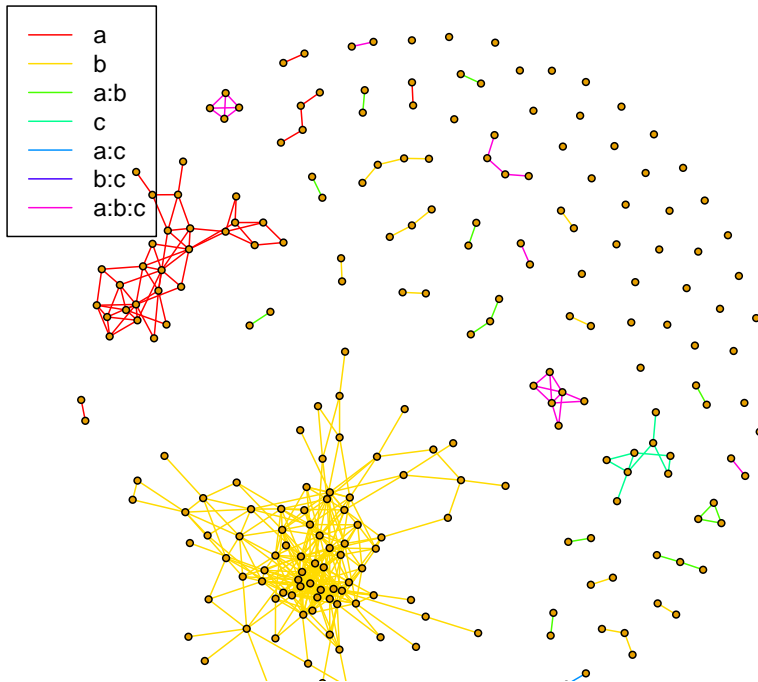
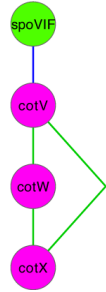


Figure 3.19: Result of best fitting graph simulation from 50000 samples. The best fit graph is further split by removing all edges between nodes that have been given different class assignments in the simulation. Each class is a combination of conditions as shown in the legend. This shows how the decomposition has resulted in one large network being separated into several subnetworks of one class as indicated by the single edge colours within them.



Our preferred method is to use the *Monte Carlo* methods. Although this has the drawback of being simulation based rather than deterministic, given the complexity of the network structure, we have found this method to be a good trade-off between computational demands and variation in the predictions. We also considered the parsimony of the model, both the affinity propagation method and the deterministic method require the user to select parameter values that effect the output. In the case of affinity propagation varying the penalty value leads to changes in the output that are not always intuitive. This is due to the interactions between the penalty terms and the clustering metrics that are optimising two different constraints. In the deterministic model, the parameters are consistently minimising the same constraint, namely that we minimise the number of nodes connected with differing class edges. However, this also requires both parameter selection as well as resulting in parent nodes being disconnected from both child nodes. Ideally, we wanted the parent to be separated from only one of the children but to do this deterministically would require propagation of the edge selection throughout the entire network with each possible combination of assigning each disconnected node to each of its neighbours separately. Clearly this is computationally very demanding and increasing in complexity with the size of the network, particularly given that degree structures of networks follow the power law.

We considered the global impact of the *Monte Carlo* method on the network by using this network decomposition on all sub networks with more than 20 genes. We compared the results of these networks to the original networks from the JGL output. We used the known transcriptional unit information from BsubCyc.org to compare the results before and after decomposition. The network decomposition resulted overall in a reduction of the number of edges from 750 to 353. Note here that we have used a subset of genes in the output that have known transcriptional information. Including the genes without transcriptional unit information would mean that we would be unable to categorise them as true or false positives or negatives. The true positive rate falls after the network decomposition as may be expected

Figure 3.20: A subnetwork found following network decomposition using edge conditions and a simulation method for separating the network. Analysis of the decomposed network showed an interesting network between *spoVIF* and *cotVWX*. Connections between *spoVIF* and the *cotVWX* genes and their transcriptional unit information are shown, green are known and blue edges are for those genes that currently have no transcriptional unit information. Hence while the interaction between *cotVWX* is already known, a relationship between these genes and *spoVIF* has not yet been shown. Annotations of the transcriptional unit information are generated automatically and have been collected from the BsubCyc website. The genes are automatically annotated with gene ontology terms that have been used to colour the nodes in this Figure. The green node of *spoVIF* has three Biological Process GO terms associated with it these are, *Transcription*, *DNA-templated*, *Regulation of transcription*, *DNA-templated* and *Sporulation resulting in formation of cellular spore*. Purple nodes, *cotVWX* all have one associated GO Biological Process term that is, *sporulation resulting in formation of cellular spore*.

since some correctly connected genes in the same transcriptional unit may be connected under different combinations of experimental classes due to the noise in the underlying data. However, the precision rose from 0.7 to 0.78 for the decomposed networks indicating that although some true positives are removed relatively more false positives are removed leading to an overall higher precision, or positive predictive value. Therefore, although there is a trade off in losing some information to aid identification and visualisation of networks by decomposing them, we can see that this decomposition is not arbitrary. That is, the removed edges overall correspond to connections between genes in different transcriptional units. Consequently, while we would not use the decomposed networks as an alternative inference method, their increased precision and interpretability makes them useful for interrogating and understanding the networks as a tool to understand the output from the JGL model. In this way, the decomposition methods are complementary to rather than competing with the JGL algorithm.

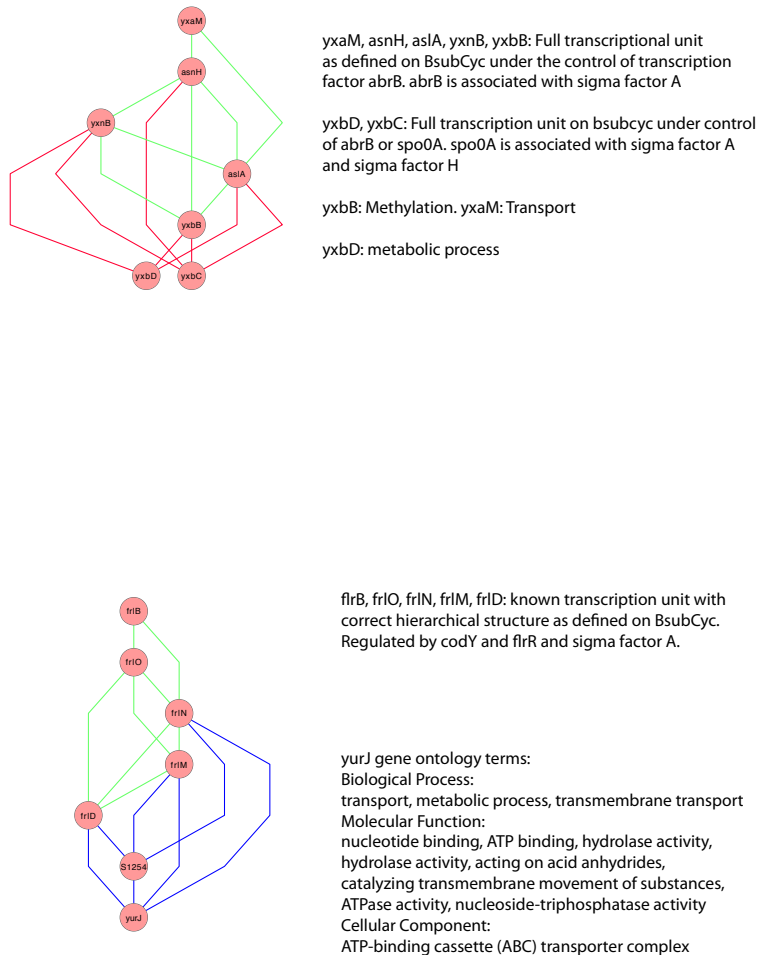
3.3 *Future analysis*

From the analysis, we have done so far, we can see that there are multiple networks that could be analysed further. A next step following identification of interesting networks would be to look at experimentally validating novel connections. Additionally, we have so far only used a sub section of the data. There may also be scope to run the algorithm on more of the experimental conditions. This depends on the similarity of the replicates to be combined into one class to give the required level of similarity and statistical power to infer the networks.

3.3.1 *Example networks*

In finding potentially interesting results we looked for edges connected to genes without a transcriptional unit, as well as potentially novel regulatory mechanisms identified by the model. This may include transcriptional regulators connected to a known transcriptional unit. As the model calculated the partial correlations we view edges as causal interactions. We may also identify interactions between more than one transcriptional unit and the different conditions that a regulatory process is active under. We have identified examples of these two situations. In Figure 3.21 we show two example subnetworks where the edges are coloured according to their transcriptional units. Green edges indicate that both genes are in the same transcriptional unit. Red edges are between in two genes in different transcriptional units and blue edges for genes that have no known transcriptional unit in-

formation. Using the experimental validation matrix from SubtiWiki created in the research by [Arrieta-Ortiz et al., 2015] between regulators and targets we also found that *yxB*, *yxA* and *yxD* are all known targets of *abrB*, *codY*. Given the connection to the other transcriptional units that are also under control of either *abrB* or *spo0A*, we would hypothesise from these results that under these experimental conditions these genes are under the regulation of *abrB*. Similarly using the experimental validation dataset, we found that *codY* is a regulator of *yurJ*.



four different growth phases. These are, in sequential order, the lag, exponential, transitional and stationary phases. The stationary phase is followed by the decline or death of the bacteria. In comparison to classes a and c, class b includes cells harvested in the transitional phase as opposed to the earlier exponential phase for classes a and c. Included in this subnetwork for class b is *cotJC* that is involved in the elimination of superoxide radicals. This could make sense for the transitional phase that occurs after the exponential phase. During the exponential phase the cell uses cofactors and responds to environmental cues, and in doing so creates byproducts including superoxide radicals. These radicals should be removed before the cell reaches its final stationary phase. In terms of regulation we also search the experimentally validated connections from [Arrieta-Ortiz et al., 2015] *et al.* Using this information source, we found that *ysxE*, *ywdl*, *cotJC*, *cotJA*, *cotJB*, *spoVID*, *usd*, *prkA*, and *spoIIID* (also a regulator of *cotJC*) are all regulated by *sigE*. *ComER* was not found in the database.

In contrast, the transport of sugars, that we would expect to be a more ubiquitous process is found in all three of our experimental classes as shown in Figure 3.23. The transcriptional unit including genes *levD*, *levE*, *levF* and *levG* is involved in reactions for the transport of sugars.

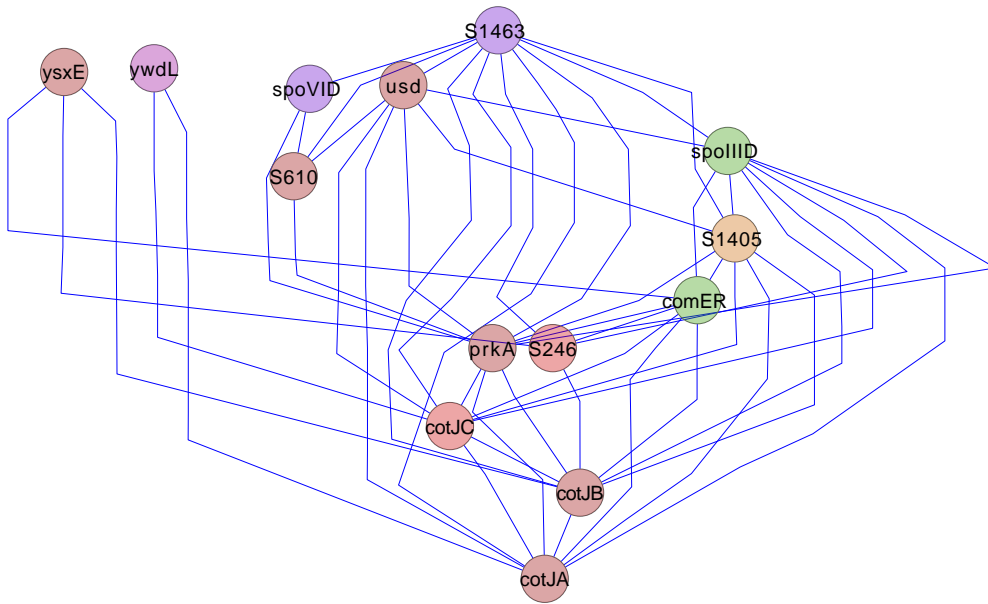


Figure 3.22: This is an example of a network where the edges were found to be present in only one of the experimental classes, class b. The transcriptional unit involving *cotJA*, *cotJB* and *cotJC*. *CotJC* is involved in the pathway for superoxide radicals' degradation. The experimental conditions included in class b involved cells harvested in the transitional phase. The nodes are coloured according to their gene ontologies.

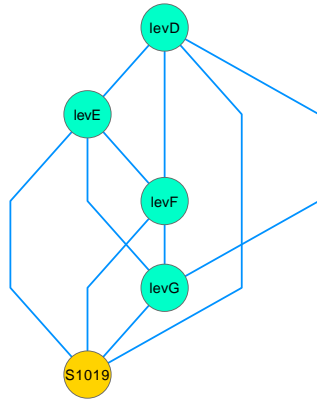


Figure 3.23: This is an example of a network where the edges were found to be present in all the experimental classes. The transcriptional unit involving *levDEFG* is involved in the transport of sugars.

3.3.2 Experimental conditions

We have used a subset of the original data as input into the network inference. We then also considered the remaining data clusters of expression profiles that could be used as input into the JGL model. To use the JGL algorithm we require a minimum of two classes, in this case these will be meta classes that cover multiple experimental conditions. Table 3.9 shows the 58 samples in the previous analysis as well as two additional data inputs each containing 21 samples in total. The summary results indicate that the network could be informative, in contrast to a network containing few genes or edges.

	Number classes	Number samples	Number genes	Number edges
JGL1	3	58	944	3649
JGL2	2	21	1534	4818
JGL3	2	21	943	4338

Table 3.9: Numbers of genes and edges in networks using different input data and the JGL model.

We looked at the original experimental conditions for the data for commonalities and differences between replicates for the different classes. The initial clustering based on euclidean distance was entirely data driven and therefore did not consider any prior information on the similarity of the experimental conditions. Being able to make potential similarities and differences to the edges or genes included in the network may aid our understanding of the conditions under which regulatory networks are active. The experimental conditions included in each of the three groups are outlined below. JGL1 refers

to the output that we have analysed in the previous sections. We can see a commonality in the growth medium and experimental cofactors, malate and glucose. In JGL2 the perturbations both include drug treatments to the cells. In summary, the experimental factors are for JGL3 are the same, though there may be differences in the time points that could contribute to differences in the networks.

1. JGL1:

- Class a: Malate, Glucose, M9 growth medium with glucose at exponential phase, cells at high temperature.
- Class b: M9 growth medium with glucose at trans phase, M9 medium with LB culture, with different carbon sources, cells at high temperature, cells grown in SMM, cells before and after addition of malate, cells after addition of glucose.
- Class c: M9 growth with glucose at exponential phase, cells before and after treatment with glucose with LB and/or M9 medium.

2. JGL2:

- Class a: Purified spores, cells LB medium with the herbicide paraquat or drug H202.
- Class b: Cells with and without glucose in exponential phase, exponentially growing cultures as controls with or without mitomycin

3. JGL3:

- Class a: Cells in CH medium, induced sporulation, cells harvested at different time points.
- Class b: Cells in CH medium, induced sporulation, cells harvested at different time points.

The overlap of genes in the three different JGL outputs is shown in the Venn diagram, Figure 3.24. The overlap of genes between the models is not surprising given that there are commonalities of experimental conditions between the different datasets. The overall number of genes covered by the three models is 2,450 or approximately half the genes in the full data set. This is for an initial set of parameters. It is also possible to reduce the shrinkage parameters and increase the number of genes covered by the model output.

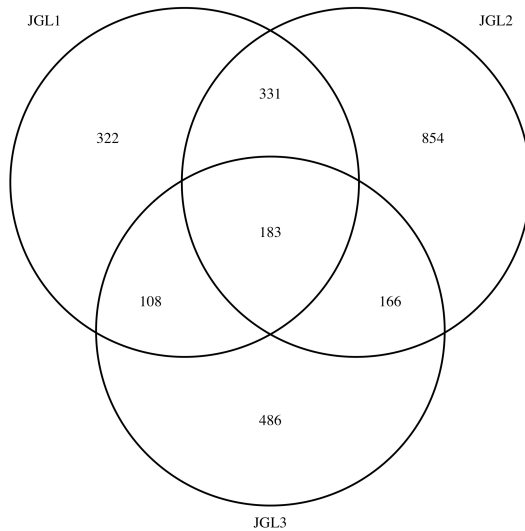


Figure 3.24: Venn diagram of the genes included in the three separate JGL models run for different subsets of the data provided by Nicolas *et al.* Each model has a subset of genes that are only included in them, indicating that all three models may be informative.

3.4 Conclusions

We have shown that we are able to infer regulatory networks within the bacterium *Bacillus subtilis* using relatively small sample sizes. Unlike methods that aim to infer the full genome-wide network through the use meta-analysis of data under multiple conditions, we restricted our view to fewer genes and can infer regulatory networks for these genes using smaller sample sizes. Although we only inferred regulatory networks for a portion of the genes the JGL model takes as input genome-wide expression and this consequently removes the need for any prior knowledge on the regulatory networks that are active under the condition of interest as we do not have to specify the genes to measure in advance. Comparing our results to those of a similar study by Arrieta-Ortiz *et al* [Arrieta-Ortiz et al., 2015], we see that we are similarly able to identify known transcriptional unit links. This is not surprising as we are using a subset of the data included in their inference. The main difference is that we can infer validated networks from a small sample size and that these are split according to experimental conditions. We also used the gold standard information from this paper to validate connections in our example networks that were not found on the BsubCyc database.

As the data set we used contained only three replicates per con-

dition we combined perturbations to create three different classes containing 58 samples in total. Using Euclidean distance and affinity propagation resulted in clusters containing mainly replicates of the same conditions as would be expected. In addition, related experimental conditions, for example the addition of glucose (MG) or malate (GM) at different concentrations and time points were also found in the same clusters. We have shown that combining perturbations to be treated as replicates of the same group increases sample size so that they can be used in the JGL method and that this produces biologically meaningful results. Future work could run this analysis on the remaining data set, where we have shown there may be an additional two models with two classes each that could result in an informative model. An alternative approach would be to use the biological prior knowledge on the experimental conditions to select groups as opposed to using data driven clustering. This may help to improve the biological interpretation of the results for example, by combining all MG concentration perturbations together into one group and all GM conditions into another.

Using known transcriptional unit information, we could confirm that these networks performed better than random networks, a possibility at low sample sizes according to previous simulation-based analysis [Li et al., 2013a]. There was also a high proportion of the genes in each of the transcriptional units contained within the same subnetwork. Using annotations from publicly available resources means we could identify both known connections and potentially novel edges in our network.

To control for the possibility of including a large amount of noise in the network, we considered the behaviour of the structure of the correlation matrices with varying values of the shrinkage parameters. We assumed that there is a baseline level of correlation that can occur between genes that represents noise as opposed to signal between them. By observing how the correlation structure behaves we can identify points at which the shrinkage parameter falls below the signal threshold and results in a proportionally large inclusion of significant edges. We interpret this change as being a result of including noisy connections into our model and consequently choose a shrinkage parameter that is more stringent than this.

While traditional methods for network selection maximise a statistical measure of goodness of fit of the model, these statistical metrics do not consider the biological interpretability of the model. In terms of parameter selection, we may be interested in minimising an error rate, such as the false discovery rate in the network. The false discovery rate is the expected number of type I errors. Type I errors occur when the null hypothesis is rejected when it is true. That is, a result

is shown as statistically significant when it is not. The false discovery rate differs from multiple hypothesis corrections such as the Bonferroni correction as these adjust individual p-values to control the probability of at least one false positive rather than the expected number of false positives [Armstrong, 2014]. Methods for estimating the false discovery rate (FDR) for graphical models include perturbing the data sets, using bootstrap samples, to estimate the selection probabilities for each edge [Li et al., 2013a]. The FDR is estimated by fitting a mixture model to the selection probabilities. This gives estimates of the proportions of the null and alternative models as well as their densities. Given the estimation of the FDR, parameters can also be optimised using the value of the FDR as a constraint. For methodologies that use a regularization or shrinkage parameter, controlling FDR and maximising power, allows for the selection of both the regularization parameter and selection threshold. A main constraint to this analysis is its computational demands. Not only can a single inference of the network at a lower shrinkage level take several hours, the bootstrap methodology means the algorithm would need to be run many times for a single shrinkage parameter. Further, to use error rates to select shrinkage parameters these bootstrap estimates will need to be calculated for a range of shrinkage parameters.

Controlling error rates such as the false discovery rate or family wise error rate are used for multiple hypothesis corrections. That is, when there are a many hypotheses, in our case edges between pairs of genes, to evaluate at once we know that the chance of edges being found as significant increases with the repeated testing by definition. This means that these methods are particularly relevant for genome-wide models where the number of individual hypotheses tested at once are in the thousands. With smaller models for tens of genes, these metrics are not usually employed for model selection as the size of the model is notably smaller. Because of the shrinkage method used by the JGL model, the output does result in smaller subnetworks of comparable size. This is because the shrinkage methods inherently control the false discovery rates through the shrinkage and selection criteria that result in these sparse networks with subnetwork of small size. We have taken an alternative approach to selecting parameter values by heuristically identifying a lower bound on the signal to noise ratio in the data. By using a shrinkage value above this bound we aim to control the signal to noise in the network.

It is advantageous to be able to dynamically explore the parameter and network space rather than using a fixed shrinkage value in the analysis. While desirable to control the global amount of noise in the network using these shrinkage parameters, small changes in these parameter values can result in large changes to the overall network.

By allowing for changes to single subnetworks not only is this method more computationally tractable but it also allows for the possibility that the same signal to noise ratios may not be present in all transcriptional units under the same experimental conditions. The method for expanding the subnetworks uses the known network structure to screen other potential members of the network. The results were more encouraging for genes that were chosen to have high covariance with all nodes in the network instead of any single node. This may be explained as an additional node must be consistently correlated with the graph structure. This means that we would expect the results to be further improved by using the logical structure of the network for selecting nodes to expand the network. We also see that moving the shrinkage parameters does not lead to a more informative model as no additional components of the network are found. In addition, the hierarchy and potential causal links are further obscured by the additional links added to the output due to the reduction of shrinkage in the model.

Given our network result, we developed methods for the interrogation of this network. We assumed that not all regulatory networks are active under all experimental conditions and used this to decompose large networks. Our methods for decomposing or expanding subnetworks are used for network exploration as opposed to parameter selection. The sensitivity and precision comparison of the networks before and after decomposition indicated that while some true positive edges are removed, overall the precision is increased meaning there is a relatively larger increase in the false positive edges removed. Therefore, while these decomposed networks are not taken as the full or final inferred network, the level of precision in the edges that remain means that we can still use the decomposed network to explore interactions between genes and generate hypotheses for experimental validation.

This approach differs from previous methods that identify hubs within networks structures as we focus on the edge classifications as opposed to the number of edges or degree structures within the network [Managbanag et al., 2008, Langfelder et al., 2013]. Particularly when the network has been inferred from literature or using a meta study of combined data sources, the resulting hubs can frequently be broader classifications and lack the scale and specificity of the smaller transcriptional units identified by our method.

The results supported our hypothesis that different combinations of transcriptional units will be present in different experimental conditions, and this information is made informative by the JGL for multiple classes. Therefore, it is not only possible to infer networks under different conditions but use this information to identify transcriptional units within the larger networks. This is a useful feature

for the researcher as a network containing hundreds to thousands of genes cannot be easily interpreted. Usually in these cases, the approach is to calculate global statistics to identify, for instance, a master regulator within the network [Kin Chan, 2013, Fujita and Losick, 2003]. However, we wanted to be able to identify the structure of the transcriptional units and how they connect to each other as from a synthetic biology perspective this low-level detail is useful. When designing constructs for synthetic biology circuits, the scope of the design is usually within a single operon or transcriptional unit. Therefore, global regulator identification is less relevant for synthetic biology, in contrast to, for example, finding regulators that control a disease response or phenotype. Although we took a lower level view of the regulatory network than global genome-wide methods, it is still a broader view than standard synthetic biology models. By using experimental data taken at a genome-wide level the network has the potential to capture the full effect of the circuit on the cell as opposed to those on the circuit alone.

Our results are comparable to the databases of transcriptional units that exist for *B. subtilis*, such as those on BsubCyc and DBTBS. The additional information our model provides is the hierarchical information on the direction or flow of the network as well as the condition information on which experimental conditions edges are present. For instance, the BsubCyc website provides all known transcriptional unit information, that we have used to annotate our network, but this information does not include the hierarchy of the transcriptional unit or provide information on the conditions in which these transcriptional units are active. The BsubCyc and DBTBS databases contain experimentally validated connections, while in contrast the JGL model is also able to identify novel connections inferred from the input data sets. These connections could be genes that are previously unknown members of a transcriptional unit or connections between multiple transcriptional units that work in concert under the given experimental conditions.

Specific synthetic biology resources exist that have been designed to create a framework for consistent and modular representation of the available parts and constructs for synthetic circuits. One main resource for this is BioBricks, an online database that contains information on, for example promoters, repressors and plasmid backbones for *Bacillus subtilis* and other organisms. BioBricks contains some functional annotation of these parts and indicates those constructs that have been used to create a response such as cell death or motility in the cell. For *Bacillus subtilis* there is also a category of parts that are designed for use with sigma factor A. In the future, we may expect these lists to increase to include other sigma factors, and this

also relates to our model output that is annotated with the sigma factor information. To the designer this gives a link to identify those transcriptional units under a sigma factor that could be controlled by the associated parts listed in Biobricks [Knight, 2003]. Currently, the Biobricks database contains information, in some cases, on the genes that a part is designed to affect [Nandagopal and Elowitz, 2011]. Parsing and annotating the network with this information could also provide a useful bridge between the information on the transcriptional or functional units from the experimental data and the availability of constructs to manipulate them. Although this engineering approach to synthetic biology has several intuitive ideas, such as the identification of parts that can be combined in multiple ways to produce alternative circuits, biological systems have a number of additional factors that make this more difficult than for example, an electrical system [Purnick and Weiss, 2009]. Amongst these are the different behaviours of circuits placed in different cell types or organisms, as well as the noise in the cell; stochastic variability in levels of gene or protein expression whilst initially may be considered as noise in data generation or natural fluctuations, have been shown to be necessary to maintain the correct balances within the cell. In one example, a synthetic circuit containing two transcription factors, one activator and one repressor was shown to function independently in a manner that produced stable oscillations. Measurement of these oscillations allowed the selection and optimisation of parameters, however, when the circuit was then introduced to the cell it was found that these oscillations were tighter, less noisy, than those observed *in vitro*. The authors identified a delay in the cells response mechanism that naturally shortens the time for these circuits to move from activated to repressed. This time differential meant that the predictions from the synthetic circuit alone were not accurate enough to engineer a response from the cell [Nandagopal and Elowitz, 2011].

Computational models within synthetic biology have focused on the optimisation of parameters for small networks to aid the design of synthetic circuits. That is, finding ideal concentrations of transcription factors or gene connections for a functional unit. The computational models at this low level require the use of differential equation models to accurately capture the system dynamics. Therefore, our modelling approach would be a precursor to this type of analysis. Its aim is to investigate from experimental data the larger impact of a circuit on the cell. Current differential equation models are usually applied to a few (tens) as opposed to thousands of genes and when searching the parameter space the probability of making a change to the kinetic parameters of genes currently in the model is chosen to be greater than the alternative probability of either adding or removing genes from the

models [Rodrigo et al., 2007]. Another area in which computational methods have been used in engineering circuits is to quantify the effects of regulators on their target gene expression. This is aimed at the design and optimisation of the systems immediate response. One example categorised over a thousand ‘parts’ or synthetic constructs in *E. Coli* according to their ability to specifically effect their target gene as well as the ‘quality’ of the part which they defined as the variation in its activity under different conditions [Mutalik et al., 2013]. This does not however, address the selection of the gene as the mechanism for producing the desired phenotype, or the optimal experimental conditions, such as growth media or the time at which to harvest cells.

From a computational perspective, we identified the size of the blocks in the block diagonal structures as an influential factor in determining computational demands of the running the JGL inference for different data inputs. Therefore, we were able to demonstrate its utility in finding similarities in expression correlation profiles that can be used to cluster data sets. The metric based on the maximum block sizes performed better at finding clusters or data sets for input into the JGL algorithm, in comparison to an alternative method based on Euclidean distances. This is not surprising as our method uses the block structure as the specific distance metric, thus tailoring it to the JGL method and the assumption that the data have correlation matrices that form block diagonal structures. When the aim is to infer correlation or partial correlation matrices, we would expect our method to be particularly useful in clustering data. Computationally we have seen that for the use of the JGL on a personal computer, the level of shrinkage needs to be high to get small enough blocks of genes. It is possible to combine profiles based on their expression similarity into groups that have similar enough correlation structures to result in useful models from JGL.

We have identified several networks of interest that contain both known and unknown connections between genes. These networks make sense from a biological perspective, in that we can find sensible annotations from ontological information that could explain the connection between these genes. From a synthetic biology perspective, identifying hierarchical structures within and between transcriptional units is useful to the researcher for understanding where a synthetic unit could be constructed to give greatest efficacy or reduce off-target effects. There has been a shift in recent years to addressing the difficulties of engineering circuits within biological systems. This includes the often-observed unintended effects of a circuit on the cell. For example, understanding that an upstream regulator of a phenotype of interest also controls a secondary phenotype that would ideally be left unperturbed can help to identify a secondary regulatory that controls only

the transcriptional unit target intended. One of our examples shows a possible control mechanism of *spoVIF* regulating the sporulation genes *cotVWX*. This may be a more targeted control of these genes when compared to the known regulator *gerE* that also controls several other transcriptional units e.g. *cgeAB* and *cgeCDE* that control formation of the outer spore layer, and *gerP-ABCDE* spore germination genes. *CotVWX* itself controls spore coat protein genes. Additionally, we have shown that the condition information can be useful from a synthetic biology perspective by identifying those networks present in different conditions. In one example the networks that are active differ for a subset of the conditions and these conditions contain samples of cells harvested in different growth phases. This shows how these methods could be used as an initial step in engineering genetic circuits as we know that the standardised parts will not behave the same in all cell types and under all conditions. In the example shown we have a subnetwork involved in the degradation of superoxide radicals that is present in the data set containing cells at the transitional phase of the cell cycle. This is consistent with superoxide radicals being produced in the proceeding exponential phase [Cabiscol et al., 2010].

Our results show commonality of included genes in the networks in comparison to those inferred using partial correlation methodology by Arrieta-Ortiz *et al.* The authors used the full data set from Nicolas *et al* with an additional data set of 403 microarray samples on a separate strain of *Bacillus subtilis* under 38 different experimental conditions. By combining the samples the authors therefore inferred the network on 671 samples within one model [Arrieta-Ortiz et al., 2015]. Clearly this is substantially more than we have used for each of our meta-conditions, however, given that these samples are replicates across a wide range of conditions these do not necessarily provide increased correlation signal for genes in the different transcriptional units, as we would not expect all transcriptional units to be active under all conditions. Our results support this idea as we could decompose larger networks using a model with multiple conditions and the fact that not all transcriptional units are active under all conditions. Therefore, the standard approach to combine multiple perturbations of a cell type or organism, while effective, could be made more efficient by selecting the perturbations more carefully thereby reducing the overall number of samples required. Previously, models have been developed that use the inference of regulatory networks to aid experimental design through selection of the highest value targets to perturb [Barrett and Palsson, 2006]. We have shown that even with lower sample sizes, our network contains high specificity to the known transcriptional unit information as well as identifying some potentially novel results.

Other methods for inferring signalling networks have used single

gene knockouts to infer hierarchy in the network [Markowitz et al., 2005], or to establish the impact of knocking out genes within a signalling network on cellular phenotypes [Wang et al., 2007]. In our analysis, the partial correlation methodology enables inference of the hierarchical structure, and the perturbations are experimental conditions as opposed to gene knockouts. Using gene knockouts arguably provides more specific perturbations than using experimental conditions such as growth factors or drug compounds. Our method for selecting data inputs is a balance between the large data sets of multiple, though not necessarily related experimental conditions, and sets of gene knockouts on groups of functionally or phenotypically related genes that are usually selected using prior information on the organism. The specific nature of the experiments means that this data set can more accurately infer the network of the perturbed genes than a data set based on general perturbations of the organism. There is clearly a trade-off between the amount of experimental data and the scope of the model. Using the JGL model on smaller sample sizes it can infer hierarchy and causal relations through the use of partial correlations for a subset of the genome. This may be particularly relevant for the synthetic biologist that is interested in a specific condition response. By selecting relevant perturbations, smaller data sets can be used to infer the hierarchical network of the active transcriptional units.

3.5 Methods

3.5.1 Mathematical preliminaries

The trace of a matrix A , $\text{tr}(A)$, is the sum of the diagonal elements, $\sum_i A_{ii}$ and has the following properties:

$$\text{tr}(AB) = \text{tr}(BA)$$

where AB and BA exist, and

$$\frac{d}{dX} \text{tr}(AX) = X^T$$

Properties of the determinant (\det) of a matrix:

$$\det(AB) = \det(A) \det(B) = \det(A) \det(B)$$

$$\det(I_n) = 1$$

where I_n is the $n \times n$ identity matrix

$$\frac{d}{dX} \log \det(X) = X^{-1}$$

for a matrix $X \in R^n$

Result 1 The (i,j) th element of a $p \times p$ precision matrix Ω is zero if and only if x_i, x_j are conditionally independent given all other variables z , where $z = 1, \dots, p, z \neq i, j$.

Proof We consider the standard linear regression model with normal errors. Given the vector of observations $x = x_1, \dots, x_p$ we can regress one variable (x_p) on the rest. From standard regression notation, we denote x_p as the regression variable Y . Let $z = x_1, \dots, x_{p-1}$. Then the covariance matrix can similarly be decomposed to give

$$\Sigma = \begin{bmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma_{ZY}^T & \sigma_{YY} \end{bmatrix}$$

The conditional distribution of $Y|Z$ is well known from multivariate normal theory and is given by:

$$Y|Z = z \sim N(\mu_Y + (z - \mu_z)\Sigma_{ZZ}^{-1}\Sigma_{ZY}, \sigma_{YY} - \Sigma_{ZY}^T\Sigma_{ZZ}^{-1}\Sigma_{ZY})$$

If Y, Z are independent then $P(Y|Z) = P(Y)$ we can see from inspection this is true when $\Sigma_{ZY} = 0$

Let the precision matrix $\Omega = \Sigma^{-1}$ then

$$\begin{bmatrix} \Omega_{ZZ} & \Omega_{ZY} \\ \Omega_{YZ} & \omega_{YY} \end{bmatrix} \begin{bmatrix} \Sigma_{ZZ} & \Sigma_{ZY} \\ \Sigma_{ZY}^T & \sigma_{YY} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

Using the inversion formulae for 2×2 block matrices we can write

$$\Omega_{ZY} = -\omega_{Y\bar{Y}}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$$

From this we can see that the Z, Y are conditionally independent if and only if $\Omega_{ZY} = 0$

3.5.2 Joint graphical lasso

Recently, Danaher *et al* proposed the fused joint lasso model for k classes $k = 1, \dots, K$ [Danaher et al., 2014]. This method extends the lasso method for Gaussian graphical models [Friedman et al., 2008] to inference under different experimental conditions, for example between patients with and without lung cancer. The lasso model constrains the number of edges included in the model using the maximum likelihood and a penalty term based on the L1 norm. For Gaussian graphical models the inverse covariance matrix, estimated by Θ is also known as the precision matrix. It is straightforward to calculate the partial correlations from the precision matrix. The precision matrix defines the network structure where an edge exists between two variables for non-zero elements in the precisions matrix, conversely a zero entry in the precision matrix means there is no edge between them. The precision matrix has a zero entry if the two variables are conditionally independent as shown in the previous section. Similarly the partial correlations defined as $-\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$ are zero if and only if variables i, j are conditionally independent given all other variables.

The Joint Graphical Lasso (JGL) borrows information across experimental factors as well as identifying differences between regulatory networks between factors. The JGL model borrows information between conditions through a penalty term that reduces the likelihood of an edge if it is not present in all conditions. In this way, the model formulation is such that a common interaction between two genes is more likely to be found where there is a evidence for it in all conditions. The model does allow for edges to be present in a subset of conditions: due to the penalty term however, there must be stronger evidence for this interaction, with the strength of the correlation required depending on the size of the penalty term. This also potentially improves the inference in terms of reducing the number of false positives by requiring more evidence in support of an edge as well as comparing results across conditions to improve the power of detection. Edges are then included if there is evidence for them across all conditions or strong evidence in a subset of conditions. For n_k i.i.d observations in group k with sample covariance matrix s^k and $S^k = (n - 1)s^k$ the estimation of

Θ is as follows:

$$\arg \max_{\{\Theta\}} \left\{ \sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)})) - P(\Theta) \right\}$$

where

$$P(\Theta) = \lambda_1 \sum_{k=1}^K \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{k < k'} \sum_{i,j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k')}|$$

and λ_1, λ_2 are tuning parameters to be selected.

This result follows from a maximum likelihood approach to estimating the precision matrix Θ . Let $x = x_1, \dots, x_p$ be a $p \times 1$ column matrix from a multivariate normal distribution with positive definite $p \times p$ covariance matrix Σ and mean vector μ . Then the probability density function (pdf) of x is

$$f(x) \propto \frac{1}{\sqrt{\det(\Sigma)}} \exp \left\{ \frac{-(x - \mu)^T \Sigma^{-1} (x - \mu)}{2} \right\}$$

In practice the parameter μ is unknown and is replaced by the sample mean \bar{x} . Therefore, the likelihood function for n i.i.d observations is given by

$$\prod_{i=1}^n \frac{1}{\sqrt{\det(\Sigma)}} \exp \left\{ \frac{-(x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})}{2} \right\}$$

Then the log likelihood is

$$\begin{aligned} & \sum_{i=1}^n \left\{ \ln \frac{1}{\sqrt{\det(\Sigma)}} + \frac{-(x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})}{2} \right\} \\ &= \frac{n}{2} \ln \det(\Sigma^{-1}) - \frac{1}{2} \sum_{i=1}^n \{(x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})\} \end{aligned}$$

The quantity $(x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})$ is a scalar and therefore can be written as the trace of a 1×1 matrix. As $\text{tr}(AB) = \text{tr}(BA)$ where AB and BA exist, the log-likelihood becomes

$$\begin{aligned} &= \frac{n}{2} \ln \det(\Sigma^{-1}) - \frac{1}{2} \sum_{i=1}^n \text{tr}(x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x}) \\ &= \frac{n}{2} \ln \det(\Sigma^{-1}) - \frac{1}{2} \sum_{i=1}^n \text{tr}(x_i - \bar{x})(x_i - \bar{x})^T \Sigma^{-1} \end{aligned}$$

As the trace of a matrix is the sum of its diagonal elements:

$$\begin{aligned} &= \frac{n}{2} \ln \det(\Sigma^{-1}) - \frac{1}{2} \left\{ \text{tr} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \Sigma^{-1} \right\} \\ &= \frac{n}{2} \ln \det(\Sigma^{-1}) - \frac{1}{2} \{ \text{tr}(S \Sigma^{-1}) \} \end{aligned}$$

This is the log likelihood for one model, in the JGL there are K classes with n_k observations in each, and the inverse covariance matrix is $\Theta^{(k)}$ hence the combined penalised log likelihood for K classes is

$$\arg \max_{\{\Theta\}} \left\{ \sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)})) - P(\Theta) \right\}$$

Danaher *et al* noted that if the inverse covariance matrix can be written as a block diagonal matrix, inference on the non-zero subnetworks individually results in the same network as on the full matrix. A block diagonal matrix is one that has multiple blocks of non-zero elements on the diagonal, elements outside these blocks are zero. For example, writing Θ in block diagonal form would be:

$$\Theta^{(k)} = \begin{bmatrix} \theta_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_b \end{bmatrix}$$

Where we have b blocks for class k , each of these blocks may be of different dimension containing number of genes g_1, \dots, g_b with $\sum_i g_i = n_k$ the total number of genes for class k . The shrinkage parameter is used to set those values below the threshold to zero, this allows the covariance matrix to be written in block diagonal form. This is formalised in the following lemma.

Lemma 1 Suppose that the solution to the fused graphical lasso is block diagonal with known blocks. That is, the estimated inverse covariance matrix takes the form

$$\Theta = \begin{bmatrix} \theta_a & 0 \\ 0 & \theta_b \end{bmatrix}$$

Θ is a $p \times p$ matrix, θ_a is an $a \times a$ matrix, and θ_b is a $b \times b$ matrix, where $p=a+b$. In this case solving the fused graphical lasso on just the corresponding subset of features, θ_a and θ_b is equivalent to solving on Θ .

Proof. To show this we partition the full log likelihood into parts corresponding to elements in the two subsets of Θ . Since $\theta_a^1, \dots, \theta_a^k$ and $\theta_b^1, \dots, \theta_b^k$ have the same dimension for all k it suffices to show the result for a single class. The full model is

$$n(\log \det \Theta - \text{tr}(S\Theta)) - P(\Theta)$$

we split each component part separately, for $n \log \det \Theta$ we note that we can write Θ as

$$\Theta = \begin{bmatrix} \theta_a & 0 \\ 0 & I_b \end{bmatrix} \begin{bmatrix} I_a & 0 \\ 0 & \theta_b \end{bmatrix}$$

and note that $\det(AB)=\det(A)\det(B) = \det(A)\det(B)$ and $\det(I)=1$. So we have $\det(\Theta)=\det(\theta_a)\det(\theta_b)$ and

$$n \log \det \Theta = n \log \det \theta_a + n \log \det \theta_b$$

For the second term, let $nS\Theta = D$ then $\text{tr}(D) = \sum_{i=1}^p D_{ii}$ therefore $\text{tr}(D) = \sum_{i=1}^a D_{ii} + \sum_{i=a+1}^p D_{ii}$ that is $\text{tr}(S\Theta) = \text{tr} S_a\theta_a + \text{tr} S_b\theta_b$. Finally, the fused penalty term is:

$$P(\Theta) = \lambda_1 \sum_{k=1}^K \sum_{i \neq j} | \theta_{ij}^{(k)} | + \lambda_2 \sum_{k < k'} \sum_{i,j} | \theta_{ij}^{(k)} - \theta_{ij}^{(k')} |$$

It can be seen by inspection that this term can be easily split into each (i, j) element and therefore the separate blocks, as the contribution of the off diagonal elements are all zero by definition. \square

This result means that the inverse of all the blocks can be found separately, thus reducing the size of the matrices to be inverted when the covariance matrix is a block diagonal matrix. This has the potential to greatly improve computational efficiency. In addition, computational complexity is further reduced by deriving rules, for identifying zero elements, in terms of the penalty parameters (λ_1, λ_2) , S^k and the number of observations n_k . These screening rules are a pre-processing step in the JGL algorithm performed before inverting the covariance matrix.

The screening rule of the JGL determines the block diagonal structure of the covariance matrix. The screening rule, for $K > 2$, is a thresholding method that sets those values of $n_k S^{(k)}$ above the threshold to 1 and those below to zero. By setting elements of the covariance matrix to zero it is then possible to rewrite the covariance matrix in block diagonal form. The thresholding value used in the screening algorithm is the same as the shrinkage value used to invert the covariance matrix, λ_1 . In practice, a weighted correlation matrix is often used instead of the covariance matrix as this makes the selection of the thresholding value, λ_1 easier as it is constrained between $[0,1]$. The weighting is used when there are different sample sizes, n_k for the different classes.

As a preliminary result we first outline the Karush-Kuhn-Tucker (KKT) criterion that give the necessary and sufficient conditions for a solution to the JGL model [Boyd and Vandenberghe, 2009]. The KKT criterion extend Lagrange multipliers to allow for inequality as well as equality constraints. That is, they provide a set of requirements for the problem of maximising a function $f(x)$ subject to $g_i(x) \leq 0$ and $h_j(x) = 0$. The KKT criterion for a solution x^* that maximises $f(x)$ are:

$$\begin{aligned}\nabla f(x^*) &= \sum_{i=1}^m \mu_i \nabla g_i(x^*) + \sum_{j=1}^l \lambda_j \nabla h_j(x^*) \\ g_i(x^*) &\leq 0 \\ h_j(x^*) &= 0 \\ \mu_i &\geq 0 \\ \mu_i g_i(x^*) &= 0\end{aligned}$$

The penalised likelihood model can be viewed as a constrained optimisation problem in Lagrangian form as $\max f(x) - \sum_{j=1}^l \lambda_j h_j(x)$ where $\lambda_j > 0$ are tuning parameters. In the case of the JGL we have two constraints ($j = 2$) and two tuning parameters λ_1 and λ_2 . In the paper by Danaher *et. al*, for $K=2$ classes the function to maximise is:

$$\arg \max_{\{\Theta\}} \left\{ \sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)})) - P(\Theta) \right\}$$

where

$$P(\Theta) = \lambda_1 \sum_{k=1}^K \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{k < k'} \sum_{i,j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k')}|$$

Our parameter of interest is θ and $f(\theta)$ is

$$\sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)}))$$

We have two inequality constraints for the JGL model, in equation above $l = 2$ and $h_1 = \sum_{k=1}^K \sum_{i \neq j} |\theta_{ij}^{(k)}|$ and $h_2 = \sum_{k < k'} \sum_{i,j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k')}|$

Then the KKT criterion are given by

$$0 = n_1 (\Theta^{(1)})^{-1} - n_1 S^{(1)} - \lambda_1 \Gamma_1 - \lambda_2 Y$$

$$0 = n_2 (\Theta^{(2)})^{-1} - n_2 S^{(2)} - \lambda_1 \Gamma_2 + \lambda_2 Y$$

To show this we first define the subgradients for the penalty function $|\theta_{ij}^{(k)}|$ w.r.t $\theta_{ij}^{(k)}$:

$$\begin{cases} 1 & \text{if } \theta_{ij}^{(k)} > 0 \\ -1 & \text{if } \theta_{ij}^{(k)} < 0 \\ a & \text{if } \theta_{ij}^{(k)} = 0 \end{cases}$$

for some $a \in [-1, 1]$ and the subgradient of $|\theta_{ij}^{(k)} - \theta_{ij}^{(k')}|$ w.r.t $(\theta_{ij}^{(k)}, \theta_{ij}^{(k')})$, $k \neq k'$ is $(d, -d)$ where

$$d = \begin{cases} 1 & \text{if } \theta_{ij}^{(k)} > \theta_{ij}^{(k')} \\ -1 & \text{if } \theta_{ij}^{(k)} < \theta_{ij}^{(k')} \\ a & \text{if } \theta_{ij}^{(k)} = \theta_{ij}^{(k')} \end{cases}$$

for some $a \in [-1, 1]$

To find the KKT equations we need the derivatives for the penalised log-likelihood model.

$$\begin{aligned} & \arg \max_{\{\Theta\}} \left\{ \sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)})) - P(\Theta) \right\} \\ &= \frac{d}{d\Theta} \left\{ \sum_{k=1}^K n_k (\log \det \Theta^{(k)} - \text{tr}(S^{(k)} \Theta^{(k)})) - P(\Theta) \right\} \end{aligned}$$

From the standard mathematical preliminaries we have:

$$\frac{d}{d\Theta} \log \det(\Theta) = \Theta^{-1}$$

and

$$\frac{d}{d\Theta} \text{tr}(S\Theta) = S^T = S$$

The differentials for $P(\Theta)$ are given by the subgradient previously defined.

Let $\Gamma_1 = \sum_{i \neq j} \frac{d}{d\theta_{ij}^{(1)}} |\theta_{ij}^{(1)}|$ and $\Gamma_2 = \sum_{i \neq j} \frac{d}{d\theta_{ij}^{(2)}} |\theta_{ij}^{(2)}|$ and $Y = \sum_{i,j} \frac{d}{d\theta_{ij}^{(1)}} |\theta_{ij}^{(1)} - \theta_{ij}^{(2)}| = \sum_{i,j} d$, Then:

$$\nabla f(x^*) = \sum_{j=1}^l \lambda_j \nabla h_j(x^*)$$

so that:

$$n_1(\Theta^{(1)})^{-1} - n_1 S^{(1)} = \lambda_1 \Gamma_1 + \lambda_2 Y$$

and

$$n_2(\Theta^{(2)})^{-1} - n_2 S^{(2)} = \lambda_1 \Gamma_2 - \lambda_2 Y$$

Therefore we have:

$$0 = n_1(\Theta^{(1)})^{-1} - n_1 S^{(1)} - \lambda_1 \Gamma_1 - \lambda_2 Y$$

$$0 = n_2(\Theta^{(2)})^{-1} - n_2 S^{(2)} - \lambda_1 \Gamma_2 + \lambda_2 Y$$

as required.

The theorem that formalises the screening rules, Figure 3.25, for the JGL algorithm is given below along with the proof given in [Danaher et al., 2014], included for completeness.

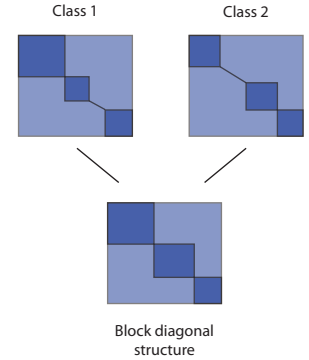


Figure 3.25: Screening rules are used to identify the block diagonal structure. Dark blue squares represent correlations passing the shrinkage parameter thresholds. This stylistic representation shows the combining of block diagonal structure using the JGL screening for two classes. The block diagonal structure is determined for each class separately and then combined to form a single block diagonal structure, with significant correlations in dark blue.

Theorem 1

For the case $K = 2$. The necessary and sufficient requirements for Θ to be the solution to the JGL model are given by the KKT criterion as:

$$0 = n_1(\Theta^{(1)})^{-1} - n_1S^{(1)} - \lambda_1\Gamma_1 - \lambda_2Y$$

$$0 = n_2(\Theta^{(2)})^{-1} - n_2S^{(2)} - \lambda_1\Gamma_2 + \lambda_2Y$$

We consider the partition C_1, C_2 of the covariance matrix into non overlapping sets, so that the covariance matrix has block diagonal form

$$\Theta^{(1)} = \begin{pmatrix} \Theta_1^{(1)} & 0 \\ 0 & \Theta_2^{(1)} \end{pmatrix}, \Theta^{(2)} = \begin{pmatrix} \Theta_1^{(2)} & 0 \\ 0 & \Theta_2^{(2)} \end{pmatrix}$$

We can show that the two following criteria (a) and (b) are equivalent:

- (a) To meet criteria above for all $i \in C_1, j \in C_2$ there exists $\Gamma_{1,ij}, \Gamma_{2,ij}, Y_{ij} \in [-1, 1]$ such that:
- $$-n_1S_{ij}^{(1)} - \lambda_1\Gamma_{1,ij} - \lambda_2Y_{ij} = 0$$
- $$-n_2S_{ij}^{(2)} - \lambda_1\Gamma_{2,ij} + \lambda_2Y_{ij} = 0$$
- (b) $|n_1S_1| \leq \lambda_1 + \lambda_2$, $|n_2S_2| \leq \lambda_1 + \lambda_2$ and $|n_1S_1 + n_2S_2| \leq 2\lambda_1$

This result means that the screening rules in (b) can be used to partition the covariance matrix into non-overlapping sets. This gives a covariance matrix with block diagonal form. From the previous results this means that each non-overlapping set (or block) determined by the rules in (b) can be inverted separately.

Proof. Without loss of generality, assume that $n_1S_1 \geq n_2S_2$. First show that (b) \Rightarrow (a). The proof is split into two cases.

Case 1:

$$n_1S_1 - n_2S_2 < 2\lambda_2$$

Case 2:

$$n_1S_1 - n_2S_2 \geq 2\lambda_2$$

Case 1:

$$\text{Let } \Gamma_1 = \Gamma_2 = \frac{-n_1S_1 - n_2S_2}{2\lambda_1} \text{ and } Y = \frac{-n_1S_1 + n_2S_2}{2\lambda_2}$$

First, note that by (b), we know that $|n_1S_1 + n_2S_2| \leq 2\lambda_1$. Therefore, $\Gamma_1, \Gamma_2 \in [-1, 1]$.

Second, note that Case 1's assumption that $n_1S_1 - n_2S_2 < 2\lambda_2$ implies that $Y \in [-1, 1]$.

Finally, we see by inspection that $-n_1S_1 - \lambda_1\Gamma_1 - \lambda_2Y = 0$, and

$$-n_2S_2 - \lambda_1\Gamma_2 + \lambda_2Y = 0.$$

Case 2:

Let $\Gamma_1 = \frac{-n_1S_1+\lambda_2}{\lambda_1}$, $\Gamma_2 = \frac{-n_2S_2-\lambda_2}{\lambda_1}$, and $Y = -1$. Then, by inspection, $-n_1S_1 - \lambda_1\Gamma_1 - \lambda_2Y = 0$, and $-n_2S_2 - \lambda_1\Gamma_2 + \lambda_2Y = 0$.

It remains to show that $\Gamma_1, \Gamma_2, Y \in [-1, 1]$. Trivially, $Y = -1 \in [-1, 1]$.

From our assumption that $|n_1S_1| \leq \lambda_1 + \lambda_2$, we know that $-1 \leq \Gamma_1$. Moreover, by the assumptions that $n_1S_1 - n_2S_2 \geq 2\lambda_2$ and $|n_1S_1 + n_2S_2| \leq 2\lambda_1$, we have that

$$\Gamma_1 = \frac{-n_1S_1+\lambda_2}{\lambda_1} \leq \frac{-n_1S_1+\lambda_2(\frac{n_1S_1-n_2S_2}{2\lambda_2})}{\lambda_1} = \frac{n_1S_1-n_2S_2}{2\lambda_1} \leq 1$$

Therefore $\Gamma_1 \in [-1, 1]$.

By the assumption that $|n_2S_2| \leq \lambda_1 + \lambda_2$, we know that $\Gamma_2 = \frac{-n_2S_2-\lambda_2}{\lambda_1} \leq 1$.

From the assumptions that

$n_1S_1 - n_2S_2 \geq 2\lambda_2$ and $|n_1S_1 + n_2S_2| \leq 2\lambda_1$, we have that

$$\Gamma_2 = \frac{-n_2S_2-\lambda_2}{\lambda_1} \geq \frac{-n_2S_2+\lambda_2(\frac{n_1S_1-n_2S_2}{2\lambda_2})}{\lambda_1} = \frac{n_1S_1-n_2S_2}{2\lambda_1} \geq 1$$

Therefore $\Gamma_2 \in [-1, 1]$.

For the second half of the proof it remains to show **(a)** \Rightarrow **(b)**.

This result follows from inspection of $0 = -n_1S_1 - \lambda_1\Gamma_1 - \lambda_gY$, as $\Gamma_1, \Gamma_2, Y \in [-1, 1]$ then we must have $|n_1S_1| \leq \lambda_1 + \lambda_g$ and similarly $|n_2S_2| \leq \lambda_2 + \lambda_g$.

Finally, adding the two equations gives $0 = -(n_1S_1 + n_2S_2) - (\lambda_1\Gamma_1 + \lambda_2\Gamma_2)$ and similarly as $\Gamma_1, \Gamma_2 \in [-1, 1]$ we therefore have $|n_1S_1 + n_2S_2| \leq \lambda_1 + \lambda_2$. \square

Lemma 2 For $K > 2$ the necessary condition for all the elements $i \in C_1$ to be completely separated from $j \in C_2$ with $i \neq j$ is $|n_kS^{(k)}| \leq \lambda_1$ for all k .

Proof. When $K > 2$ we note that, in the KKT equations, for the second part of the penalty term involving λ_2 these terms will cancel in the summation as for each pair (k, k') $k \neq k'$ we have $d - d = 0$. Therefore, the KKT criteria is

$$0 = \sum_k \{-n_kS^{(k)} - \lambda_1\Gamma_k\}$$

Where $\Gamma_k \in [-1, 1]$ for all k

From the above we can see that to satisfy the KKT criterion we must have $|n_kS^{(k)}| \leq \lambda_1$.

Conversely if $|n_kS^{(k)}| \leq \lambda_1$ for all k . Let $\Gamma_k = -n_kS^{(k)}/\lambda_1$ then

$$0 = -n_kS^{(k)} - \left(-\frac{\lambda_1 n_kS^{(k)}}{\lambda_1}\right)$$

and since $|n_kS^{(k)}| \leq \lambda_1$ then $\Gamma_k \in [-1, 1]$ \square

3.5.3 Subnetwork analysis

The lasso approach means that the resulting network inference is likely to produce disjoint sub graphs as it shrinks edges below the threshold to be exactly zero. This threshold is user-selected and consequently after identifying a subnetwork of interest we may want to reduce this *a priori* cutoff value. This would allow for the inclusion of additional genes that are correlated with the original subnetwork. Biologically we expected that the correlation between genes within a transcriptional unit would be higher than with those in a second interacting transcriptional unit, and that the strength of these connections between different transcriptional units will vary, particularly when averaged over different experimental conditions. We assumed that transcriptional units interacting with each other would be included at a lower shrinkage level. Relaxing the shrinkage penalty may therefore reveal interactions between transcriptional units. The original network of 944 genes is unlikely to include all relevant genes active under a given experimental condition. However, as there were multiple biological processes active using initially a stringent shrinkage parameter made it easier to identify these different processes (as subnetworks). These subnetworks were then expanded for an area of the network that we were interested in. Computationally it helps to focus the network on a subset of genes as opposed to allowing for a smaller shrinkage parameter at a genome-wide level. In this way, our network result is a starting point for further investigations computationally as well as experimentally, and value found in iteratively identifying and narrowing focus to networks of interest. The value of the initial network is to remove the necessity of prior information on active genes, networks or pathways.

The elements to be included in each subnetwork are determined by the block diagonal structure of the covariance matrix and inversion of these disjoint covariance matrices then determine the graph structure. We therefore explored a subnetwork of the output using the screening rule of the JGL algorithm through varying the shrinkage parameters, λ . For a set of nodes in the subnetwork the remaining genes are searched and a user specified number of genes with the highest covariance to all those in the network are returned. That is, given an initial sample covariance matrices $S^{(k)}, k = 1, \dots, K$ and shrinkage parameter λ_1 a block diagonal matrix Φ used as input into the fused JGL model with $K > 2$ is calculated as follows:

$$\phi_{(i,j)} = \begin{cases} 1 & \text{if } \exists k \text{ s.t. } |n_k S_{(i,j)}^{(k)}| > \lambda_1 \\ 0 & \text{otherwise} \end{cases}$$

Where the conditions for $\phi_{(i,j)}$ to have block diagonal form are as

outlined in Lemma 2 above. In practice, the sample correlation matrix is used for $S^{(k)}$ and where the sample sizes n_k are equal for each k , the selection of the threshold (λ_1/n_k) is easier as it is constrained between 0 and 1. Then we write Φ in block diagonal form, with b blocks:

$$\Phi = \begin{bmatrix} \phi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_b \end{bmatrix}$$

The inverse of Φ is the same as calculating the inverse of each ϕ_i $i = 1, \dots, b$ individually. Therefore for a subnetwork of interest ϕ_i , we developed a method for expanding this network by altering the shrinkage parameter λ_1 . For a user-selected number of additional genes (G), denote initial set of genes in ϕ_i as P , and s as the number of additional genes to be added to the network, the algorithm is:

Result: Identifying additional genes to include in a subnetwork expansion

Input : Nonnegative double ϵ

Input : subnetwork adjacency matrix ϕ_i

Input : Nonnegative integer G

Output : New shrinkage parameter λ^*

Output : New genes added to subnetwork, p

while $s < G$ **do**

$\lambda^* = \lambda^* - \epsilon$

Select gene(s) p such that

$\phi_M[p, P] > \lambda^*$

$s = s + |p|$

end

Algorithm 1: Subnetwork Expansion

Where ϕ_M is the maximum of $|n_k S^{(k)}|$, $k = 1, \dots, K$, and $\epsilon > 0$ is a user specified value by which the value of λ_* the shrinkage parameter is decreased on each iteration.

Because an additional gene must pass a threshold value of correlation to all genes in the subnetwork they are more likely to be connected in the expanded subnetwork. The additional selected genes along with the genes in the existing network are input into the JGL algorithm. This allows inference of only the selected expanded subnetwork, which dramatically reduces computation time.

3.5.4 Clustering for JGL

To select conditions as input into the JGL model, for a given number of clusters we group conditions that give the smallest upper bound on the block size of the covariance matrix. That is, conditions are

combined according to the similarity of the block diagonal covariance matrix after the shrinkage condition has been applied. To do this we use agglomerative clustering where the score between a new data point and existing cluster is updated at each iteration of the algorithm to allow for the change in the cluster and what would therefore be the maximum block size if these data points were included in the same JGL inference. This is for a fixed number of clusters.

3.5.5 *Network annotation and evaluation*

Given a network we then annotated the nodes and edges using various resources. These included adding gene ontology (GO) terms to each of the nodes, which can be used to colour the nodes in Cytoscape [Smoot et al., 2011], and were taken from the ENSEMBL database using BioMart [Kasprzyk, 2011]. Using the BsubCyc website [Karp et al., 2005] we were also able to gather sets of known transcriptional units. This information was then mapped onto the edges of the network, which can be used to visually identify new links within the network. We also added sigma factor information from the DBTBS [Sierro et al., 2008] and SubtiWiki websites [Michna et al., 2013]. The sigma factors are proteins that enable binding of RNA polymerase to gene promoters. As different sigma factors are active under different experimental conditions, such as heat, knowledge of the controlling sigma factor is useful in designing experiments for modifying or activating pathways. RCytoscape [Shannon et al., 2013] was used to visualise the results from the JGL algorithm.

To evaluate the result global analysis of the network was performed. To see if the number of known connections found between genes (according to the transcriptional unit information) was likely to have occurred by chance, the nodes of the network were randomly perturbed ten thousand times. This maintains the degree structure of the network. Empirical p-values for the observed number of connections for each transcriptional unit were then calculated and multiple-hypothesis corrected using Benjamini-Hochberg [Benjamini and Hochberg, 1995].

3.5.6 *Decomposing large networks*

Even with the shrinkage methods, the output from the JGL model can still contain large networks. For the *Bacillus subtilis* data set and JGL parameters used in the previous sections the largest subnetwork contained 240 genes. These networks are still difficult to visually interrogate. Therefore, we considered several different methods that can be used to decompose large networks based on edge values. These methods all assume that the edge values are qualitative variables that represent different experimental conditions or combination of

conditions an edge is present in. The idea being that we may find separate regulatory modules within a larger regulatory network that are identified by nodes connected by edges under the same conditions and that not all regulatory modules will be active under all conditions.

3.5.6.1 Using Clustering methods

The affinity propagation clustering method [Frey and Dueck, 2007] uses a similarity matrix and message passing to determine the number of clusters and the elements within those clusters. We used the number of conditions each edge is present in to create a similarity matrix. For example, with 3 conditions an edge appearing in any one condition would have similarity value $1/3$, 2 conditions $2/3$ and in all three 1. This does not however consider that, for example, a node may have similarity $1/3$ with two other nodes but these edges may be present in two different groups. Therefore, we identified where two nodes share an edge between any other node and if those edges are in different conditions a penalty is imposed, that is a negative value. Affinity propagation clustering selects exemplars of clusters and members of clusters using the similarity scores between nodes. This is done by simultaneously calculating availabilities and responsibilities for and between nodes. The availabilities refer to the availability of a node to be an exemplar and the responsibility its connection to the other nodes as assigned to exemplars. Hence, there are two factors first, the relative suitability of each node being an exemplar for node i . This is the responsibility $r(i, k)$ and is a conditional score for k being the exemplar for i given all other possible exemplars. This is calculated as the maximum availability and similarity for another node k' being the exemplar. In this way, the scoring compares the similarity of node i to candidate k and to all other nodes that includes the score of k' being an exemplar and that it should be the exemplar to node i . Second, the algorithm considers the situation that i should choose k as its exemplar given the support k must be an exemplar from the other nodes. As the algorithm iterates, the availability for some nodes will fall to zero as there is relatively little evidence that they should be exemplars. Intuitively the availability determines which nodes are exemplars and the responsibility assigns nodes to exemplars.

The availability for node i to choose node k as its exemplar is $a(i, k)$ and is initialised at zero.

$$\begin{aligned}
 a(i, k) &= 0 \\
 a(i, k) &= \min\{0, r(k, k) + \sum_{i' \neq i} \max\{0, r(i', k)\}\} \\
 a(k, k) &= \sum_{i'} \max\{0, r(i'k)\}, \text{ s.t. } i' \neq k \\
 r(i, k) &= s(i, k) - \max_{k'} \{a(i, k') + s(i, k')\}, \text{ s.t. } k' \neq k
 \end{aligned}$$

3.5.6.2 *Deterministic split of the Network*

We may split the network deterministically by separating all nodes that do not have the same conditions connecting them to a third (parent) node. Where the number of parent nodes in differing conditions exceeds a user selected threshold the two nodes are separated from all common parent nodes. However, this means that the parent nodes are no longer connected to either of the nodes. Algorithm 2 outlines the method used to deterministically separate the network. Where the inputs into the algorithm are defined as follows:

Θ^* is the adjacency matrix across all classes. There are $\sum_k 2^{(k-1)}$ combinations of the k classes. $\theta_{i,j}$ is the edge value between nodes i, j , $\theta_{i,j} \in [1, \sum_k 2^{(k-1)}]$ where $k = 1, \dots, K$ and K are the number of classes. Γ_l is the l-th adjacency matrix $\gamma_{i,j}$ is the edge value 0 or 1 between nodes i, j where 1 denotes an edge between nodes i, j in class l and 0 no edge.

Result: Deterministically split network

Input: Optional: class level η

Input: The edge value matrix, Θ^*

Input: Individual class matrices, $\Gamma_l, l=1, \dots, K$

Output: Decomposed Adjacency matrix, Γ^*

$$\Theta = \begin{cases} 1 & \text{if } \Theta^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

for *for each node (gene) i in Θ ($i=1, \dots, n$) do*

find all nodes also connected to node i:

$$a = \Theta[i,] * \Theta$$

Find all entries that have a different value to node i:

$$d = \Theta^* - \Theta^*[i,]$$

$$P = d \neq 0$$

find those with different values that are also connected to the same node, this will be our checking matrix Φ that contains those edges to be removed:

$$\Phi = \mathbf{1}(P * a \neq 0)$$

if *split by class too, then*

remove those edges with a combined class level below η , this prioritises edges with differential values and fewer classes rather than edges with different values and a high number of class combinations within them :

$$\Gamma' = \sum_{l=1}^K \mathbf{1}(\Gamma_l > 0)$$

$$C = \Gamma' + \Gamma'[i,]$$

$$\Phi_2 = \mathbf{1}(C < \eta)$$

Update check matrix so that those with different edge value and edge value less than η are known:

$$\Phi = \Phi * \Phi_2$$

end

Find the number of shared connections with node j that have different edge values to node i:

$$NP = \sum_{i=1, \dots, n} \Phi[i,]$$

Find the nodes to separate based on the number of shared connections with different edge values being greater than ω :

$$\text{sel} = \text{which}(|NP| > \omega)$$

for *each node j in sel do*

Remove all edges shared with it, node i and a third node

$$\text{nodes} = \mathbf{1}(\Phi[, \text{sel}[j]] \neq 0)$$

Initialise final separated adjacency matrix Γ^* :

$$\Gamma^* = \Gamma'$$

$$\Gamma^*[i, \text{nodes}] = 0$$

$$\Gamma^*[\text{nodes}, i] = 0$$

$$\Gamma^*[\text{sel}[j], \text{nodes}] = 0$$

$$\Gamma^*[\text{nodes}, \text{sel}[j]] = 0$$

end

end

Algorithm 2: Deterministic algorithm for splitting Large Networks

3.5.6.3 Simulation methods

Figure 3.26 shows a stylistic example of how the network decomposition is calculated. Figure 3.26 a) shows an example network where the nodes are connected under different experimental conditions as denoted by the colour of the edges between them.

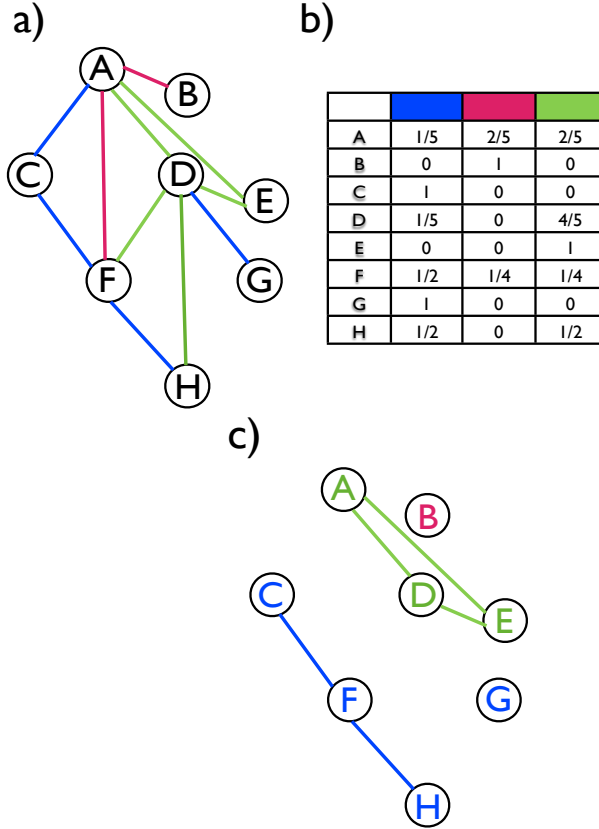


Figure 3.26: The Figure shows a stylistic example of how the simulation based network decomposition is performed. a) An example network, where nodes are connected under different example conditions as shown by the different colour edges, we use this information to decompose the network. b) As a first step, multinomial probabilities are calculated for each node based on the frequency of the edge conditions as shown in. c) By sampling from these probabilities for each node independently we can assign a single class to each node. The score to select the best fit graph is according to its similarity to the original network. The graph is further decomposed by separating nodes that have been assigned to different classes.

Multinomial probabilities of class assignment for each node are calculated based on the class value of the edges connected to it, Figure 3.26 a). We denote the probabilities for a single node (gene) in the network as P_i $i = 1, \dots, p$. Then for each P_i where we have $c = 1, \dots, \sum_k 2^{(k-1)}$ possible classes we calculate the probability of P_i being in class c as $P_{(i,c)}$,

$$P_{(i,c)} = \frac{\sum(\theta_{(i,j)} = c)}{\sum(\theta_{(i,j)} \neq 0)}, i \neq j$$

For each node independently, the class assignment is randomly generated using these probabilities as in Figure 3.26 b). We denote $C_{(i,l)}$ as the class assignment of node i in simulation l . A new adjacency matrix, Θ^* , is calculated where only edges for the selected class

assignment for each node are included.

$$\theta_{(i,j,l)}^* = \begin{cases} \theta_{i,j} & \text{if } \theta_{i,j} = C_{i,l} \text{ or } \theta_{i,j} = C_{j,l} \\ 0 & \text{otherwise} \end{cases}$$

A second optional step has been performed whereby two nodes are separated if they are connected to other nodes in different classes.

A similarity score between this and the original matrix is then calculated, the adjacency matrix that has the greatest similarity to the original matrix is selected as shown in Figure 3.26 c). That is, we aim to find the set of class assignments C_i for each node to maximise the proportion of edges in the network:

$$\arg \max_l = \frac{\sum \theta_{(i,j,l)}^* \neq 0}{\sum \theta_{(i,j)} \neq 0}$$

4

Empirical Bayes method for estimating covariances

IN THE MAJORITY OF EXPERIMENTS the number of replicates, due to time and monetary constraints, is relatively low. In differential expression studies using microarray or RNA-seq, methods have been developed that use empirical Bayes (EB) approaches to borrow information from across genes (which, by contrast to sample size, are large in number) to improve estimates. These methods have been shown to reduce the false discovery rate in differential expression analysis.

We are interested in reducing the false discovery rate for correlations between genes. As with differential expression studies, we expect the number of biological replicates to be small (< 20) however, the dimension of the correlation or covariance matrix will be large, with thousands of genes. Previous work using correlation matrices has largely focused on interpreting relationships between genes directly from the correlation matrix. As discussed in previous chapters, shrinkage methods have been used which create a sparser correlation structure making it easier to identify significant relationships.

However, we are also interested in whether we can improve the initial estimates of the correlation matrices. In our analysis, this would be a pre-processing step where these correlation matrices would then be used as input into the JGL algorithm. They could also be used in stand-alone analysis of correlations or input into other algorithms that require covariance or correlation matrices. One method, named Corpcor was introduced to improve the estimates of correlation matrices [Schäfer and Strimmer, 2005]. This method was motivated by the small n large p problem as observed in genome-wide expression problems. Analogous to the JGL model, Corpcor takes a shrinkage approach to improving the estimate over the standard sample covariance matrices that suffer when $n \ll p$. The Corpcor method uses a mixture model to combine high variance unconstrained estimates with low variance high bias constrained estimates. The mixture proportions are determined by a shrinkage parameter that is calculated analytically

based on minimising a risk function of the model.

4.1 *Exploratory data analysis*

AS INITIAL EXPLORATORY DATA ANALYSIS we looked at an existing data set available from ArrayExpress. This enabled us to assess the sample correlations for different samples sizes and whether we may be able to leverage information from across each of the correlation pairs to give a more robust estimate of the full correlation matrix. The CEL data for the experiment E-GEOD-24594 was downloaded from ArrayExpress and loaded into R [Fujiwara et al., 2011]. The data for two of the conditions was used: these were factors with E2F1 Null and E2F2 Null, these being different genotypes of the E2F transcription factors. These E2F transcription factors are known to be important regulators of the cell cycle. The expression data were normalised using Robust Multi-Array Average (rma) and standardised to zero mean and unit variance [Irizarry et al., 2003]. The data were filtered to include those probes in the top fifty percent of genes according to their variance [Bourgon et al., 2010].

The data set contains 20 replicates for each condition. This is a relatively large number of replicates for a gene expression study and we used this to compare the effect of smaller sample sizes, between 5 and 15 replicates to the full data set. We first looked at the correlations between the sample sizes using their ECDFs. Figure 4.1 shows the ECDFs for the Pearson correlations of the samples, for the full sample size of 20 and random subsets of the data containing 5,10 and 15 replicates.

Figure 4.1 shows a general reduction in the correlation values for 10 and 15 samples in comparison to 20 replicates, but the ECDFs show similar distribution properties between these three sample sizes. For the smallest sample size of 5 replicates, however, there is a noticeable difference between the ECDFs indicating that there is potentially more noise and error introduced at this level of replication in comparison to the larger sample sizes that maintain the correlation structure but with different nominal values.

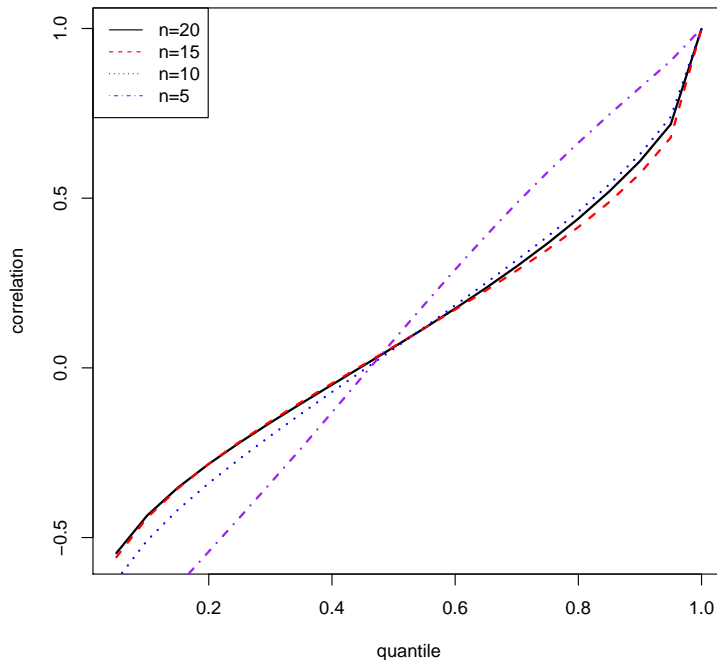


Figure 4.1: The lines are the ECDFs of the Pearson correlations. The full data set comprising 20 replicates is shown in black. The ECDFs of the correlations for random subsets of the data are also shown for 5, 10 and 15 replicates. This shows how the overall pattern to the distribution of the correlations is similar for 10 and 15 replicates but there is a large difference for 5 replicates indicating that this is too few samples to estimate correlations with.

As we wanted to estimate correlation matrices to be used with the JGL algorithm it was also interesting to see how these correlations behave within the block diagonal assumption of the JGL framework. To do this we fitted the JGL model to the full data set and looked at the rank correlation values for each gene pair. For all pairs at different sample sizes the correlation ranks are plotted in Figure 4.2 a). For all gene pairs, there was quite a lot of variability in their ranks between the different sample sizes.

As we expected that many of these correlations are not significant we further compared the ranks of the correlations that were found as significant in the JGL model. We defined significance as those genes contained within one block. Figure 4.2 b) shows the rank correlations for gene pairs in one block. It can clearly be seen that there is much less variation in the rank values for the different sample sizes in comparison to the full data set.

To check that this is not due to the fact that we are looking at genes that have been included in a block (and so may have higher correlation values without meaningful structure) we also considered the correlations between randomly selected genes in different blocks, this is shown in Figure 4.2 c). Here we can see that there is greater variation between genes in different blocks compared to genes within

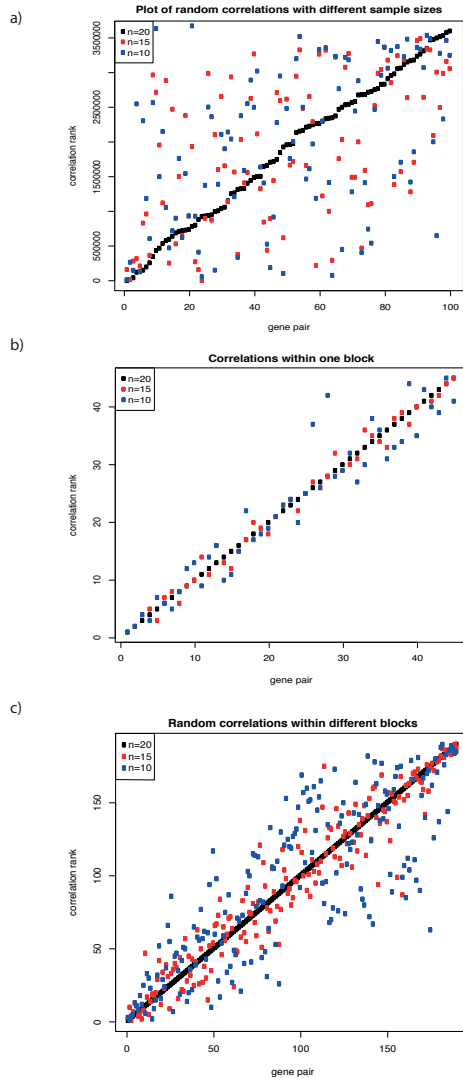


Figure 4.2: a) Randomly selected correlation pairs for different sample sizes. There is clear variation, without obvious pattern, between the correlations for each of the different sample sizes. b) After fitting the JGL model to the data and selecting a single block found by the algorithm, the plot shows correlations of genes inside this block only for different sample sizes. The model was run with $\lambda_1 = 0.85$ and $\lambda_2 = 0.05$. For correlations within a block there is close agreement between the correlation ranks for each sample size. However, there is slightly more variation with sample size 10 than 15. c) The JGL algorithm found multiple blocks (of significantly correlation genes) at the same threshold level of $\lambda_1 = 0.85$. The correlations between genes within blocks show less variation over different sample sizes as opposed to those for all genes (including those outside blocks).

the same block. This indicates that the block diagonal structure creates groups of genes with strong correlation patterns as opposed to simply higher than average correlations.

In combination, these results indicate that we may be able to share information across correlation pairs to improve the estimation of the correlation matrices. Before we introduce the empirical Bayes' model used to estimate the correlation matrices we begin with a few mathematical preliminaries.

4.2 Mathematical preliminaries

Definition 1

A symmetric $n \times n$ matrix X is positive definite if for all nonzero vectors $a \in \mathbb{R}_n$, $a^T X a$ is > 0

Definition 2

A $n \times n$ matrix X is positive semi definite if for all nonzero vectors $a \in \mathbb{R}_n$, $a^T X a$ is ≥ 0

Result 1

The covariance matrix Σ of real random vectors (x) is positive semi definite. By definition

$$\Sigma = E[(x - E(x))(x - E(x))^T]$$

so for non zero vector $a \in \mathbb{R}_n$

$$\begin{aligned} a^T \Sigma a &= E[a^T (x - E(x))(x - E(x))^T a] \\ &= E[ss^T] \geq 0 \quad , \text{where } s = a^T (x - E(x)) \end{aligned}$$

This is ≥ 0 as it is the square of two real vectors.

Result 2

A block diagonal matrix is positive definite if and only if (iff) each of its block are positive definite.

Proof. Write a matrix X in block diagonal form:

$$X = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}$$

Then for any non-zero column vector $a \in \mathbb{R}_n$

$$a^T X a = [a_1 \quad a_2] \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$a^T X a = a_1^T B a_1 + a_2^T C a_2$$

For $a^T X a > 0$ we have to have $a_1^T B a_1 > 0$ and $a_2^T C a_2 > 0$ which means that B and C have to be positive definite by definition given that a_1, a_2 are both non zero real vectors. Conversely if B and C are both positive definite then $a_1^T B a_1 + a_2^T C a_2 > 0$ meaning $a^T X a > 0$ and therefore X is positive definite. \square

Result 3

The sum of a positive definite matrix (z) and a positive semi definite matrix (S) is itself positive definite

Proof. If X is positive definite and Y is positive semi definite then for any nonzero vectors $a \in \mathbb{R}_n$

$$a^T (X + Y) a = a^T X a + a^T Y a > 0 \text{ as } a^T X a > 0 \text{ and } a^T Y a \geq 0 \quad \square$$

4.3 Empirical Bayes model

WE USED AN EMPIRICAL BAYES APPROACH to infer covariances and by simple extension correlations. The theoretical basis of this, using conjugate priors is derived in [Champion, 2003]. In this paper, the authors used either an independence prior or a flat prior with constant correlation. Using a correlation based prior makes it easier to calculate a combined value to give the value of the flat prior for the off diagonal values of the matrix as they are standardised values. The adjusted correlation matrix is later converted to a covariance matrix using the estimated variances.

The model assumes that the data are from a multivariate normal distribution. The theoretical covariance matrix for a multivariate normal distribution, Σ is positive definite, where the sample covariance matrix is positive semi definite. These sample matrices are the approximations of the theoretical matrices that are used in modelling and parameter estimation.

Bayes theorem relates the posterior distribution of the parameters given the data $p(\Theta|X)$ to the likelihood of the data $p(X|\Theta)$ and the prior distribution of the parameters $p(\Theta)$

$$p(\Theta|X) \propto p(X|\Theta)p(\Theta)$$

If we assume the data X are multivariate normal data with sample size n , covariance matrix Σ , mean μ and number of variables (genes) p , then the likelihood is proportional to

$$p(X|\Theta) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}$$

which we showed in section 3.5.2, for n i.i.d observations, can be written as

$$p(X|\Theta) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \text{tr}(S\Sigma^{-1})\right\}$$

Where $S = \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$, x_i is the sample data and we assume that the mean μ is known. We wish to obtain a Bayesian estimate for covariance matrix Σ which we denote, η . The conjugate prior for estimating a covariance matrix (η) is an inverse Wishart distribution with parameters $(\lambda z, \lambda)$. The pdf for the inverse Wishart is proportional to

$$p(\Theta) \propto |\eta|^{-((\lambda+2p+2)/2)} \exp\left\{-\frac{1}{2} \text{tr}(\lambda z \eta^{-1})\right\}$$

Where $|\eta|$ is the determinant of the matrix η . The mean is given by $\frac{\lambda z}{\lambda} = z$ and the parameter λ is related to the degrees of freedom v by $\lambda = v - p - 1$ [Champion, 2003]. To obtain the joint posterior

distribution we multiply the likelihood $p(X|\Theta)$, by the prior $p(\Theta)$. Then the joint posterior distribution $p(\Theta|X)$ is proportional to

$$p(\Theta|X) \propto |\eta|^{-n/2} \exp\left\{-\frac{1}{2} \text{tr}(S\eta^{-1})\right\} \cdot |\eta|^{-(\lambda+2p+2)/2} \exp\left\{-\frac{1}{2} \text{tr}(\eta^{-1}\lambda z)\right\}$$

$$p(\Theta|X) \propto |\eta|^{-(n+\lambda+2p+2)/2} \exp\left\{-\frac{1}{2} \text{tr}(\eta^{-1}(\lambda z + S))\right\}$$

which is an inverse Wishart distribution with parameters $(\lambda z + S, \lambda + n)$ we then estimate η , by η^0 the expected value (mean) of the distribution

$$\eta^0 = \frac{\lambda z + S}{\lambda + n}$$

As noted previously the sum of a positive definite matrix (z) and a positive semi definite matrix (S) is itself positive definite. Therefore, ideally z would be positive definite.

For z we use a block diagonal prior that is a combination of the independence and flat priors used by Champion *et. al.* Given a selected user input correlation level (this is analogous to the shrinkage parameter in the JGL model), we set all elements below this threshold to zero and use only the non-zero elements to estimate the constant value of the correlations. This means the method will either shrink correlations to zero or to the common mean value of all non-zero correlations. This is consistent with the block diagonal structure assumed in the JGL model and with our observation that we have multiple regulatory processes that contain a subset of the genes.

4.3.1 Calculating hyperparameters

In empirical Bayes methods, we estimate both the hyperparameters λ and z from the data. This means the parameters are estimated using the data rather than using a hierarchical model and assigning a prior distribution to each of the parameters or by having to choose the parameter values where little prior knowledge is available. In choosing the matrix z we are looking for an appropriate prior matrix for the covariance matrix η .

Shrinkage methods are commonly used for improving the estimates of covariance or correlation matrices. Particularly in cases where $n \ll p$ as is common in gene expression analysis, the sample sizes do not meet the assumed n large condition. They also aid interpretation by simplifying the model. For our purposes, we used the shrinkage method to generate a block diagonal form for the correlation matrix, consistent with the block diagonal assumption of the JGL model [Danaher et al., 2014]. The correlation matrix is used to determine

the elements in each block (before estimating the covariances for these). This is simply for ease of selecting the shrinkage parameters in $[0,1]$. First the sample correlation matrix is written in block diagonal form for a given level of shrinkage. For data X let Δ_x be the sample correlation matrix then for shrinkage level θ we set all values of $\Delta_x < \theta = 0$ and write Δ_x as:

$$\begin{pmatrix} \delta_1 & 0 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 & 0 \\ 0 & 0 & \delta_3 & 0 & 0 \\ 0 & 0 & 0 & \delta_4 & 0 \\ 0 & 0 & 0 & 0 & \delta_5 \end{pmatrix}$$

For each δ block at a given level of θ we extract the same elements of the shrinkage sample covariance matrix. These will be a mixture of zero elements (from the shrinkage) and the sample values. In this case, we would use a mixture prior, so z has either value γ or zero:

$$\begin{pmatrix} 1 & \gamma & 0 & 0 & \gamma \\ \gamma & 1 & \gamma & 0 & 0 \\ 0 & \gamma & 1 & \gamma & \gamma \\ 0 & 0 & \gamma & 1 & \gamma \\ \gamma & 0 & \gamma & \gamma & 1 \end{pmatrix}$$

γ is calculated from the average of the non-zero sample correlation values. One potential disadvantage of this is that we cannot guarantee that the prior matrix will be positive definite. This will depend upon the exact form the matrix takes. Often in bioinformatic applications, informative priors are generated based on the knowledge of, for example, transcription factors and their targets [Mukherjee and Speed, 2008]. Although these priors usually improve the accuracy of the inference, particularly in high dimension, they are similarly not guaranteed to be positive definite. Therefore, these priors are usually used with an MCMC algorithm, which does not specify a posterior distribution that can be evaluated analytically and, therefore, does not impose the positive definite constraint on the prior matrix that using the conjugate inverse Wishart prior does. We therefore use an alternative formulation where we assume a flat prior within each block. This is likely to have the advantage of being a positive definite matrix assuming there are no linear dependencies that would arise if, for example, all the sample variances are equal, with perfect positive correlation between them. For the flat prior z would take the form:

$$\begin{pmatrix} 1 & \gamma & \gamma & \gamma & \gamma \\ \gamma & 1 & \gamma & \gamma & \gamma \\ \gamma & \gamma & 1 & \gamma & \gamma \\ \gamma & \gamma & \gamma & 1 & \gamma \\ \gamma & \gamma & \gamma & \gamma & 1 \end{pmatrix}$$

Where γ is a constant positive correlation; we perform the procedure on the magnitude of the correlations, ignoring the sign of the correlation. We assume that the true covariance matrix is block diagonal and our prior z as constructed above will also be block diagonal. To meet the criteria that the prior matrix z is positive definite we note that a block diagonal matrix is positive definite if and only if (iff) all the blocks are positive definite. Therefore, by construction using the flat prior within each block we create a positive definite block diagonal prior. In the above we have used a single and fixed and known shrinkage level, θ . This is an equivalent assumption to the JGL model and therefore this empirical Bayes method is consistent with the JGL model, making it a suitable pre-processing step for the data that are then used as input into the JGL model.

Given the current estimate of z we then calculate λ using the following approximation suggested by [Champion, 2003]:

$$E((\rho_{ij}[\eta] - \rho_{ij}[z])^2) \simeq \frac{(1 - \rho_{ij}[z]^2)^2}{\lambda + 3} \text{ for } i \neq j,$$

where $\rho_{ij}[\eta]$ is the correlation based on η . We approximate $\rho_{ij}[\eta]$ by the sample correlations, and $\rho_{ij}[z]$ are the correlations based on z for the selected value of γ . This result follows from the distribution of the sample correlations $\rho_{ij}[\eta]$. We now give details of the derivation of this marginal distribution, the outline of which is given in [Champion, 2003]

Theorem 2 The marginal distribution of $\rho_{ij}[\eta]$ when $\eta \sim IW(z, \lambda)$ is a Normal correlation (NC) distribution with parameters $(\rho[z], \lambda + 3)$.

Proof. The authors limit their attention to 2×2 matrices, as any diagonal submatrix is similarly distributed. Therefore, the single correlation $\rho_{01}[\eta]$ is abbreviated to $\rho[\eta]$. The pdf of η can be written as:

$$p(\eta) = \frac{|\lambda z|^{(\lambda+3)/2}}{4\pi\Gamma(\lambda+2)} |\eta|^{-(\lambda+6)/2} \exp\left\{-\frac{1}{2} \text{tr } \lambda z \eta^{-1}\right\}$$

First a change of variables $\theta = \eta^{-1}$, the Jacobian for this is given as $|\theta|^{-3}$ therefore we have

$$p(\theta) = \frac{|\lambda z|^{(\lambda+3)/2}}{4\pi\Gamma(\lambda+2)} |\theta|^{(\lambda+6)/2} \exp\left\{-\frac{1}{2} \text{tr } \lambda z \theta\right\} \cdot |\theta|^{-3}$$

which gives

$$p(\theta) = \frac{|\lambda z|^{(\lambda+3)/2}}{4\pi\Gamma(\lambda+2)} |\theta|^{\lambda/2} \exp\left\{-\frac{1}{2} \text{tr } \lambda z \theta\right\}$$

A second change of variables is used $\theta_{01} = -\rho[\eta]\sqrt{\theta_{00}\theta_{11}}$ with an expansion of θ to get the pdf in terms of $\rho[\eta]$.

$$|\theta| = \theta_{00}\theta_{11} - \theta_{01}^2 = \theta_{00}\theta_{11} - \theta_{00}\theta_{11}\rho[\eta]^2$$

and

$\text{tr } \lambda z \theta = \lambda(z_{00}\theta_{00} + z_{11}\theta_{11} + 2z_{01}\theta_{01})$ and the scaling change of variables is $\sqrt{\theta_{00}\theta_{11}}$

this gives

$$\frac{|\lambda z|^{(\lambda+3)/2}}{4\pi\Gamma(\lambda+2)} [\theta_{00}\theta_{11}(1-\rho[\eta]^2)]^{\lambda/2} \exp\left\{-\frac{1}{2}(\lambda z_{00}\theta_{00} + \lambda z_{11}\theta_{11} - 2\lambda z_{01}\rho[\eta]\sqrt{\theta_{00}\theta_{11}})\right\} \cdot (\sqrt{\theta_{00}\theta_{11}})$$

Collecting terms so that we can separate out $\rho[\eta]$

$$\frac{|\lambda z|^{(\lambda+3)/2}}{4\pi\Gamma(\lambda+2)} \theta_{00}^{(\lambda+1)/2} \exp\left\{-\frac{1}{2}\lambda z_{00}\theta_{00}\right\} \theta_{11}^{(\lambda+1)/2} \exp\left\{-\frac{1}{2}\lambda z_{11}\theta_{11}\right\} (1-\rho[\eta]^2)^{\lambda/2} \exp\left\{\lambda z_{01}\rho[\eta]\sqrt{\theta_{00}\theta_{11}}\right\}$$

To get the marginal distribution of $\rho[\eta]$ integrate out θ_{00} and θ_{11} using the power series representation of the exponentials, and gamma functions $\Gamma(n) = (n-1)!$.

Integrating w.r.t θ_{00} first note that

$$\exp\left\{\lambda z_{01}\rho[\eta]\sqrt{\theta_{00}\theta_{11}}\right\} = \sum_{i=0}^{\infty} \frac{(\lambda z_{01}\rho[\eta]\sqrt{\theta_{00}\theta_{11}})^i}{i!}$$

Power series representation for an exponential is:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Collecting terms only involving θ_{00} and integrating we have

$$\int \theta_{00}^{(\lambda+i+1)/2} \exp\left\{-\frac{1}{2}\lambda z_{00}\theta_{00}\right\} d\theta_{00}$$

To evaluate this integral, we use a change of variables $u = \lambda z_{00}\theta_{00}$ that gives $d\theta_{00} = \frac{1}{\lambda z_{00}} du$, and note that if $x \sim \chi^2(k)$ then, $f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$. Therefore we can write:

$$\begin{aligned} & \int \theta_{00}^{\frac{\lambda+i+1}{2}} \exp\left\{-\frac{1}{2}\lambda z_{00}\theta_{00}\right\} d\theta_{00} \\ &= \int \frac{2^{\frac{\lambda+i+3}{2}} \Gamma\left(\frac{\lambda+i+3}{2}\right) (\lambda z_{00})^{\frac{\lambda+i+1}{2}}}{2^{\frac{\lambda+i+3}{2}} \Gamma\left(\frac{\lambda+i+3}{2}\right) (\lambda z_{00})^{\frac{\lambda+i+1}{2}}} \theta_{00}^{\frac{\lambda+i+1}{2}} \exp\left\{-\frac{1}{2}\lambda z_{00}\theta_{00}\right\} d\theta_{00} \\ &= \frac{2^{\frac{\lambda+i+3}{2}} \Gamma\left(\frac{\lambda+i+3}{2}\right)}{(\lambda z_{00})^{\frac{\lambda+i+1}{2}}} \int \frac{(\lambda z_{00}\theta_{00})^{\frac{\lambda+i+1}{2}}}{2^{\frac{\lambda+i+3}{2}} \Gamma\left(\frac{\lambda+i+3}{2}\right)} \exp\left\{-\frac{1}{2}\lambda z_{00}\theta_{00}\right\} d\theta_{00} \end{aligned}$$

Substitute $u = \lambda z_{00}\theta_{00}$, $d\theta_{00} = \frac{1}{\lambda z_{00}} du$:

$$= \frac{2^{\frac{\lambda+i+3}{2}} \Gamma(\frac{\lambda+i+3}{2})}{(\lambda z_{00})^{\frac{\lambda+i+1}{2}}} \frac{1}{\lambda z_{00}} \int \frac{(u)^{\frac{\lambda+i+1}{2}}}{2^{\frac{\lambda+i+3}{2}} \Gamma(\frac{\lambda+i+3}{2})} \exp\{-\frac{1}{2}u\} du$$

And the term inside the integral is recognisable as the pdf of a $\chi^2(k)$ distribution with $k = \lambda + i + 3$ and therefore integrates to 1. After the same calculation for integrating w.r.t θ_{11}

$$p(\eta) = \sum_{i=0}^{\infty} \frac{(\lambda z_{01} \rho[\eta])^i}{i!} \frac{[2^{\frac{\lambda+i+3}{2}} \Gamma(\frac{\lambda+i+3}{2})]^2}{(\lambda^2 z_{00} z_{11})^{\frac{\lambda+i+3}{2}}} \frac{|\lambda z|^{\frac{\lambda+3}{2}} (1 - \rho|\eta|^2)^{\lambda/2}}{4\pi\Gamma(\lambda+2)}$$

Collecting terms and rearranging:

$$= \frac{2^{\lambda+1} (1 - \rho|\eta|^2)^{\lambda/2}}{\pi\Gamma(\lambda+2)} \sum_{i=0}^{\infty} \frac{(2\lambda z_{01} \rho[\eta])^i}{(\lambda^2 z_{00} z_{11})^{\frac{i}{2}} i!} \frac{|\lambda z|^{\frac{\lambda+3}{2}}}{(\lambda^2 z_{00} z_{11})^{\frac{\lambda+3}{2}}} \left\{ \Gamma\left(\frac{\lambda+i+3}{2}\right) \right\}^2$$

Recall that by definition $\rho_{ij}[\eta] = \eta_{ij} / \sqrt{\eta_{ii}\eta_{jj}}$, and z is a 2x2 matrix with only one partial correlation $\rho_{01}[z]$. Further note that $|\lambda z| = \lambda^2 z_{00} z_{11} - \lambda^2 z_{01}^2$ then we have

$$= \frac{2^{\lambda+1} (1 - \rho[\eta]^2)^{\lambda/2} (1 - \rho[z]^2)^{(\lambda+3)/2}}{\pi\Gamma(\lambda+2)} \sum_{i=0}^{\infty} \frac{(2\rho[z]\rho[\eta])^i}{i!} \left\{ \Gamma\left(\frac{\lambda+i+3}{2}\right) \right\}^2$$

which is a normal correlation (NC) distribution $NC(\rho[z], \lambda + 3)$. \square

For a NC distribution $r \sim NC(\rho, n)$. We note that

$$\sqrt{(n)}(r - \rho) \quad \text{asymptotically distributed} \quad N(0, (1 - \rho^2)^2)$$

[Olkin and Pratt, 1958]. Therefore we have:

$$E((\rho[\eta] - \rho[z])^2) \simeq \frac{(1 - \rho[z]^2)^2}{\lambda + 3}$$

this is the equation used to estimate λ .

4.3.2 Simulated data

As mentioned above in comparison to the independence and flat prior used by Champion *et. al* we used a block diagonal prior. Using this form of prior we expect to improve the estimates as the prior better matches the data set and we denote this model as EB. The work of [Champion, 2003] used an independent prior that we label Independent and the constant non-zero prior, Flat. We compared our results with two other methods, one the Sample correlation matrix calculated

as the Pearson’s correlation matrix, and the method of Schäfer *et al*, Corpcor. For the simulated data, we used the method of Hardin to generate block diagonal matrices [Hardin et al., 2013]. We generated 100 samples for 50 genes with 5 different blocks. Correlations within blocks are set to 0.85, this is a choice motivated by the observed correlations in the previous chapter for the *Bacillus subtilis* data, that showed fairly high correlation values as shown by our choice of shrinkage value being above 0.9. We did this for three different sample sizes of 10, 15 and 20. We estimated the correlation matrices and calculated the average false and true positive rates and their standard deviations over 100 simulations. All this is done for the EB method with four different λ values, including a very small $\lambda = 0.05$ value that approximates a Flat prior and $\lambda = 1$ that is equivalent to an Independent prior. The other two methods are the Pearson’s (Sample) correlation matrix and the Corpcor method.

4.4 Results and Discussion

WITH SIMULATED DATA, the EB method had a lower false discovery rate compared to the Pearson correlation and Corpcor methods. Further, the standard deviation of the false discovery rate over different simulations was similar or lower for the EB method, indicating an estimation method that was at least as stable as the Pearson and Corpcor estimates. Table 4.1 shows there are similar FDR across all sample sizes: the largest differences are for the True positive rate (TPR) with consistently higher TPR as the sample sizes increase. We can see that using the block diagonal priors gives better result than the Flat or Independent prior as would be expected given that we have simulated block diagonal covariance matrices. The Corpcor method, which uses a common correlation structure across all pairs, with relatively large TPR rates, also results in a larger number of false positives. For all but the Flat and Independent priors, the block diagonal and Corpcor methods have the same TPR standard deviation (sd) that could be indicative of the variability in the sample data as opposed to variability in the estimation. Encouragingly however, the EB method has consistently lower FDR rates than the Pearsons (Sample) matrix and the same or lower sd, meaning that the use of the EB method has reduced error rates and provides more stable estimates.

The results show that the block diagonal prior performs better than the sample covariance matrix and the Corpcor method. Using the block diagonal prior results in better estimates than the Flat or Independent prior as would be expected. These results indicate that we can improve both the false positive and true positive rate of

	$\lambda_1 = 0.85$	$\lambda_1 = 0.05$ Flat	$\lambda_1 = 1$ Indep	$\lambda_1 = 0.8$	Pearsons	Corpcor
n=10						
FDR mean	0.04	0.05	0.03	0.04	0.07	0.26
FDR sd	0.04	0.05	0.03	0.05	0.05	0.12
TPR mean	0.88	0.82	0.80	0.89	0.89	0.89
TPR sd	0.11	0.23	0.16	0.11	0.11	0.11
n=15						
FDR mean	0.05	0.06	0.05	0.06	0.08	0.34
FDR sd	0.05	0.06	0.05	0.05	0.06	0.16
TPR mean	0.97	0.92	0.95	0.98	0.97	0.97
TPR sd	0.10	0.24	0.16	0.10	0.10	0.10
n=20						
FDR mean	0.06	0.07	0.05	0.06	0.08	0.33
FDR sd	0.05	0.05	0.04	0.05	0.05	0.14
TPR mean	0.99	0.97	0.98	0.99	0.99	0.98
TPR sd	0.11	0.20	0.19	0.11	0.11	0.11

Table 4.1: Simulation results for the empirical Bayes method with four different parameter settings, compared to the Pearsons correlation matrix and existing method Corpcor. The values in the Table are the average false discovery rate (FDR), true positive rate (TPR) over 100 simulations and the standard deviation of these, for 50 genes with three different sample (replicate) sizes, 10, 15 and 20. The EB method shows consistently and significantly lower FDR over the other methods and overlapping TPR rates to the Pearson and Corpcor methods. This indicates the increased ability of this method to identify blocks or regulatory units without artificially increasing off-diagonal (or spurious) correlations between unconnected genes.

estimating correlations that can then be used in downstream graphical analysis. This is particularly important where the sample size is low as may be expected in many experiments. Further, controlling the false discovery rate is particularly useful when the network inferences are used to drive experimental hypotheses, as we are interested in testing only the links with highest possible value.

4.5 Implementation

The algorithm for the EB method has been written in the R programming language and is available from the Bioconductor repository. The algorithm described above is for the EBsingle function within the covEB package:

<http://bioconductor.org/packages/covEB/>.

Result: An empirical Bayes estimated covariance matrix, using block diagonal prior

Input: Covariance Matrix, Σ

Input: Shrinkage threshold, η

Input: Sample size, n

Output: Covariance matrix, EBcov

From the input covariance matrix (Σ) calculate the correlation matrix (ρ)

Set all entries below input threshold (η) to zero

Calculate the block diagonal matrix

for *Each block* **do**

 | Calculate the average of the sample correlation in the block to give the estimate of γ

end

Combine each of the flat block matrices together to create one block diagonal prior

Calculate zcov, the prior covariance matrix using the prior correlation matrix z and the sample variances

Estimate hyperparameter λ , first calculate

$$k^2 = \frac{E((\rho_{ij}[\eta] - \rho_{ij}[z])^2)}{(1 - \rho_{ij}[z]^2)^2} \quad \text{where } k^2 \simeq \frac{1}{\lambda+3} \text{ so that } \lambda \simeq \frac{1}{k^2} - 3$$

$$\text{Then we set } \lambda = \begin{cases} n & \text{if } k^2 \leq 0 \\ \frac{1}{k^2} - 3 & \text{otherwise} \end{cases}$$

if $\lambda < 1$ **then**

 | $\lambda=1$

end

Calculate the EB covariance matrix, EBcov = $\frac{\lambda * zcov + S}{\lambda + n}$

Algorithm 3: EB covariance matrix estimation

5

Toxoplasma gondii and GGMs

5.1 Introduction

Toxoplasma gondii (*T. gondii*), a protozoan from the Apicomplexan group, infects nearly a quarter of the adults world-wide causing birth defects and perinatal deaths. It is an opportunistic pathogen which infects hosts with compromised immune systems and caused 15% of deaths in the AIDS epidemic. There are four different strains of *Toxoplasma gondii*, types I, II, III and type 12. Type I is most virulent in mice, whilst type II mainly affects humans, type III livestock and type 12 wild-animals. These different strains have different levels of geographical prevalence and motility within the host. For example, the type I strain has shown increased motility in comparison to type II under laboratory conditions [Harker et al., 2015]. *T. gondii* can infect all warm-blooded animals through ingestion of cysts that are shed from the definitive feline host. Once ingested the parasite can travel to multiple tissues in the host creating a secondary infection. Initially this means the parasite must pass through the intestinal wall however, it is possible that the parasite has different mechanisms that enable it to pass through different areas of the host to infect the blood, brain and other tissues. The transmission to other areas of the host causes a strong inflammatory response which influences the ability of the parasite to infect the host [Harker et al., 2015].

The parasite exists in one of three stages, the cysts from the feline host contain the sporozoite form of the parasite which are ingested by the intermediate host. Once in the intestine they form cysts which contain bradyzoites which in turn convert to tachyzoites that can move between host barriers and infecting different tissues within the host. Once the tachyzoite has migrated to a different tissue, they convert back to bradyzoites. Therefore, the tachyzoite can be viewed as the mobile version of the parasite that moves through the organism to spread the infection and the bradyzoites within the tissues represent

Motility here refers to the ability of the parasite to spread to and infect other areas of the host.

the chronic infection of the parasite [Dubey et al., 1998].

After invading the host cell the toxoplasma parasite forms a parasitophorous vacuole (PV) that surrounds and protects the parasite and a PV membrane (PVM) that acts as a transport mechanism between the parasites and the host cell [Muniz-Feliciano et al., 2013]. This enables the parasite to gain essential nutrients for its survival from the host [Laliberte and Carruthers, 2008]. A central mechanism by which the *T. gondii* parasite effects the host signalling pathways are through rhoptry proteins (ROP). Rhoptry proteins are secreted from the PV and act as pseudokinases. The ROP proteins can subvert normal cellular signalling pathways [Kim and Weiss, 2008]. This is critical for the parasite to be able to, for example, activate anti-apoptotic pathways that could otherwise lead to cell (and thus parasite) death, or deactivate pathways involved in the host's inflammatory response [Hunter and Sibley, 2012].

Analysis of *T. gondii* data has included the identification of differentially expressed genes and proteins. One of the first papers to elucidate the response of the host to toxoplasma ME49 was a genome-wide microarray analysis over different time points between 1 and 24 hours after infection. This paper outlined the immediate inflammatory response of the host and the later occurring changes to biological processes including metabolism, transcriptional regulation and cell signalling [Blader et al., 2001]. The activation of host metabolism has been speculated to be essential for the survival of the parasite as it ensures the survival of the host [Blader and Saeij, 2009].

Recently, Gene Set Enrichment Analysis (GSEA) has been expanded to create specific gene sets on *T. gondii* for functional units and processes within the cell cycle and the developmental program of the toxoplasma parasite [Croken et al., 2014]. The authors combined known gene set and annotations from sources such as KEGG and the Gene ontology with analysis of existing microarray expression data at, for example, different parts of the cell cycle. Together these data sources were combined into gene sets annotated according to the different parts of the parasites cell cycle and developmental program. In an example analysis, these gene sets were then used to identify the different processes present when comparing expression data for wild-type and mutant parasites.

Ontological and pathway analysis has similarly been used to identify differences between different strains of *T. gondii*. Using differential expression analysis of neuroepithelial cells infected with three different strains of *T. gondii*, type I, II and III, significant gene sets were compared to existing ontologies and pathways. This analysis showed different processes and pathways active for the different strains of *T. gondii*. This is consistent with the observed difference in virulence

between strains, showing different mechanisms by which the parasites invade the host [Xiao et al., 2011]. Of the three strains, Type I strain showed the most differential expression followed by Type III, with the least change in Type II. The differentially expressed genes were enriched in the central nervous system in the type I strain, nucleotide metabolism for type III whilst they found no consistent results for type II strain.

Advances have been made to infer regulatory networks in the parasite based on microRNA (miRNA). Through computational analysis, candidate miRNAs were established by comparison to the miRNAs of other organisms, human and rodent. The candidate miRNAs were further filtered by whether they were expressed in the parasite. The hypothesis was that *T. gondii* may transport some of their miRNAs into the host as part of the mechanism by which the parasite takes over host function. This is due to the similarity of the hypothesised miRNAs in the *T. gondii* and human and rodent hosts. It may also give some indication of the effects of the transcriptional networks in the host affected by the miRNAs assuming the targets of the miRNAs in the host are known [Saçar et al., 2014]. Beyond this, little progress has been made to create models of signalling and regulatory networks for hosts infected with toxoplasma.

It is known that the immune or inflammatory response triggered by the parasite invasion is like those caused by tumours. Arguably investigation into how the parasite subverts host response may provide insight into cancers[Lun et al., 2015]. To test this hypothesis, we used RNA-seq data from mouse embryonic fibroblast cells infected with two strains of toxoplasma. If the parasite caused responses in the host that are also seen in cancer cells we expect to see commonality between the biological processes active in parasite infected cells and cancer cells.

5.2 *The Hallmarks of Cancer*

In a landmark paper, Hanahan and Weinburg defined the hallmarks of cancer as the physiological traits of cancers that subvert the defence mechanisms of normal cells. These six hallmarks are 1) tissue invasion or metastasis, 2) angiogenesis 3) limitless replication potential, 4) protection against cell death, 5) provision of growth signals and 6) evasion of anti-growth signals [Hanahan and Weinberg, 2000]. The hallmarks define the multiple mechanisms through which a single mutated cell becomes cancer; the single cell typically divides and proliferates to multiple cells that together form a tumour. This tumour may metastasise or break through the basal membrane and form tumours in other parts of the organism. To facilitate an increase in tumour mass and the migration to different tissues within the organism, many tumours

induce angiogenesis. Angiogenesis is the formation of blood vessels which provides both oxygen and nutrients to support growth and migration [Nishida et al., 2006]. To be able to continually divide and replicate the cancer must have a mechanism for corrupting normal cell cycle control, that is cell division, proliferation and death. Though there are numerous ways in which cancers can subvert the normally tight regulatory mechanisms, they are also finite in number, and this has allowed the identification of important genes and pathways that influence cancer progression.

If we first consider the cell cycle of a normal cell it is possible to identify the ways in which cancer cells subvert the normal cell cycle. Normal cells have limited replication potential: there are a finite number of times that each may divide to produce further cells. In addition, they respond to cues that tell them to replicate or induce their own cell death. These three factors together control cell proliferation. Consequently, for cancer cells to grow to tumour masses, they must be able to subvert the normal cell processes [Lun et al., 2015]. They can do this in multiple ways but common between all tumours is that they divide continuously without the limitations of normal cells. Additionally, they subvert the usual signalling mechanisms that direct cell growth and death and so ensure their continued survival.

One mechanism of subversion is by the over-expression of oncogenes. These (proto) oncogenes are capable of binding to growth receptors that in turn initiate cell growth and do so independently of the normal growth factors [Polsky and Cordon-Cardo, 2003]. This means the cancer can activate cell growth without requiring the growth signals from the healthy or normal cellular environment. In conjunction with this the cancer must override the anti-proliferation signals from the normal cell and the cell death signals to allow for their continued division. Since the hallmarks of cancer were outlined in 2000 subsequent research has also identified extensive changes to metabolism that occurs during tumourigenesis [Jose et al., 2011]. As a result, it has been suggested that the six hallmarks of cancer should be extended to include metabolic reprogramming.

5.3 *Cancer metabolism*

Metabolic pathways describe the organisation of the chemical processes responsible for cellular respiration. Cellular respiration converts carbon sources from food such as glucose and glutamine into Adenosine triphosphate (ATP), the energy source for the cell. The overall conversion of glucose to ATP is comprised of multiple chemical reactions facilitated by different enzymes. Each of these chemical reactions gives rise to intermediary molecules and by-products. Central to these

metabolic processes are the gain and loss of electrons. The gain of electrons (reduction) and the loss of electrons (oxidation) occurs over multiple chemical reactions. For example, the oxidation of glucose releases electrons that are passed to NAD^+ converting it to NADH. NAD^+ and NADH are the oxidised and reduced forms of NAD respectively. The conversion of NADH to O_2 through loss of electrons, in the electron transport chain, results in energy release in the form of ATP. NADH is one of the by-products of cellular respiration. As well as continuing the reaction to form ATP these intermediaries and by-products can alternatively be used for biosynthesis; the conversion of energy into amino-acids, fatty acids, glycerol and sugars all of which provide the biomass required to create new cells. Therefore, the metabolic processes can be either catabolic (breaking down food to ATP) or anabolic (creating new biomass).

There are three metabolic pathways responsible for cellular respiration, these are glycolysis, the tricarboxylic acid (TCA) cycle and oxidative phosphorylation (OXPHOS). Metabolism of food for energy begins with glycolysis; the conversion of glucose to pyruvate that also produces 2 molecules of ATP and 2 NADH. Glycolysis is a 10-step chemical process that begins with the adding of a phosphate to glucose by a Hexokinase. The end-product of glycolysis is Pyruvate, a metabolic intermediary that in vertebrates can either be further oxidised to acetyl-CoA through a pyruvate dehydrogenase (PDH) complex or converted to lactate. The latter process does not require oxygen, and is referred to as anaerobic metabolism. In contrast, the oxidisation of pyruvate to acetyl-CoA is the linking step between glycolysis and the TCA cycle. Acetyl-CoA is the input required to the TCA and oxidative phosphorylation pathways which occur in the mitochondria to produce ATP, this is referred to as aerobic respiration [Tzamei, 2012].

In normal cells the selection of either aerobic or anaerobic metabolism is determined by the presence of oxygen. Mitochondria are the primary area of energy production for normal non-proliferating cells in the presence of oxygen as they are the most efficient at producing ATP. Aerobic respiration produces up to 36 molecules of ATP compared to the 2 molecules of ATP produced by anaerobic glycolysis. In 1923 Warburg found that in contrast to normal cells, cancer cells used the less efficient glycolysis for generating energy even when oxygen was present in the cell. Due to the presence of oxygen, the Warburg effect is also known as ‘aerobic glycolysis’ [Koppenol et al., 2011]. Figure 5.1.

There is growing evidence that the Warburg effect may be caused by tumour cells requiring biomass for cell proliferation. By comparing the intermediaries and by-products of ‘aerobic glycolysis’ and the TCA cycle and oxidative phosphorylation the commitment of glucose solely

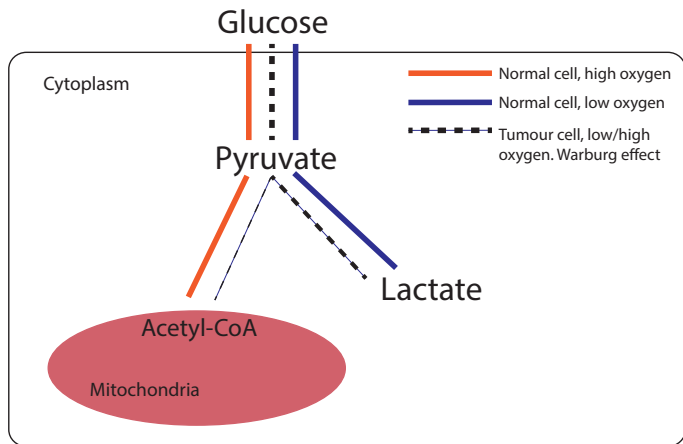


Figure 5.1: The metabolism of glucose varies between normal cells depending on oxygen status. In contrast, the Warburg effect shows increased lactate production even in the presence of oxygen.

to ATP production runs counter to the needs of the cell to produce lipids, amino acids and nucleotides for proliferation [Vander Heiden et al., 2009]. For example, whilst OXPHOS can produce 36 molecules of ATP to the 2 produced by glycolysis. Glycolysis produces more carbon and NADH than OXPHOS, both of which are required for biosynthesis.

An intermediary of the TCA cycle, citrate, can be oxidised for ATP production in the mitochondria or transported out to the cytoplasm. In the cytoplasm, citrate can be converted to acetyl-CoA and used for lipid synthesis. The conversion of citrate to acetyl-CoA is through the enzyme ATP-citrate lyase. In tumour cells, knockdown of ATP-citrate lyase resulted in a reduction of cell proliferation. This effect occurred in a glucose dependent manner; tumours with high levels of glucose metabolism showed reduced proliferation following knockdown whilst those with low levels of glucose metabolism were largely unaffected [Hatzivassiliou et al., 2005]. These examples provide evidence that the Warburg effect is the result of tumours cells manipulating metabolism to support cell proliferation. Further, there are now numerous results showing that the balance of aerobic glycolysis and oxidative phosphorylation depends on the properties of the tumour. Larger more established tumours exhibited a relatively smaller utilisation of anaerobic glycolysis and an increase in oxidative phosphorylation [Jose et al., 2011].

However, given that tumour cells have functioning mitochondria and a dependence on the mitochondria for protein synthesis, it is unlikely the sole reason for the Warburg effect is biosynthesis [DeBerardinis et al., 2007]. A second, though not mutually exclusive explanation for the Warburg effect is the regulation of reactive oxygen

species (ROS) by aerobic glycolysis [Liberti and Locasale, 2016]. ROS are produced in the mitochondria by the electron transport chain during the conversion of intermediaries such as NADH to produce ATP [Li et al., 2013b, Turrens, 2009]. An increase in glycolysis causes a decrease in the redox ratio of NAD^+/NADH . In normal cells, balance in the redox ratio is restored by the mitochondria, a process that also results in ROS production [Chiarugi et al., 2012]. A high NAD^+/NADH ratio provides multiple benefits to the cell including promotion of DNA repair, survival and biosynthesis. In tumour cells, increased rates of glycolysis increase NADH levels and maintain glycolysis; influx of NADH into the cytosol ultimately changes levels of ROS production [Locasale and Cantley, 2011]. Maintaining the correct level of ROS in the cell is critical for cell survival; ROS can promote cell proliferation but excessive levels can cause cell death [Liou and Storz, 2010].

The Warburg effect can be viewed in part as a mechanism for controlling the redox potential of the cell. The conversion of pyruvate to lactate through lactate dehydrogenase (LDH) is an alternative mechanism for converting NADH to NAD^+ to redress the redox ratio. This mechanism is used when the mitochondria is unable to maintain the redox ratio in the cell which occurs at increased levels of NADH from a higher rate of glycolysis in tumour cells. As well regulating cellular metabolism, both NAD and ROS can also influence signalling pathways through interaction with signalling proteins. Conversely, many oncogenes and tumour suppressors that influence cell cycle and cell death also influence metabolism and ROS levels in the cell [Lévy and Bartosch, 2016]. Oncogenes including *Akt* and the tumour suppressor p53 can interact with the mitochondrial membrane to influence mitochondrial ATP and ROS production [Herrera-Cruz and Simmen, 2017]. PI3K/AKT and its dependent MTOR pathway have been shown to influence glycolysis and glutamine metabolism [Csibi et al., 2013]. Master regulators of transcription HIF1 and Myc regulate metabolic pathways [DeBerardinis et al., 2008], whilst HIF1 itself is activated by ROS. Whilst the metabolic state of the host is critical for the survival of *Toxoplasma gondii* the metabolism of the parasite is also important.

5.4 Metabolism in *Toxoplasma gondii*

The metabolic properties of the parasite are important for survival differences in the metabolic process of the parasites can explain, at least in part, the differences in their virulence. Song *et. al* noted that whilst the different strains of toxoplasma share common genome sequence and predicted function the strains may differ in the pathways or ways in which these metabolic enzymes are used [Song et al., 2013]. They

predicted models for parasite metabolism and compared mRNA expression of metabolic enzymes between the ME49 and RH strains. The results showed an up-regulation of enzymes involved in the TCA cycle, glycolytic and pentose phosphate pathways in RH strain compared to ME49. The metabolic models predicted increase growth for the more virulent RH strain with a corresponding increase in ATP production. This increase in ATP production may be explained by the increased activity of the metabolic enzymes in RH compared to the ME49 strain. Similarly, it has been shown that the parasite expresses two different forms of lactate dehydrogenase (LDH) that convert pyruvate to lactate. LDH isoforms varied according to differentiation state suggesting different metabolic requirements of tachyzoites and bradyzoites [Yang and Parmley, 1997].

Human foreskin fibroblasts infected with ME49 showed differential gene expression analysis enriched for glycolysis but not TCA or oxidative phosphorylation. The changes in gene expression required the presence of the parasite in the host - secreted parasitic enzymes were not sufficient to induce changes to host gene expression. The results showed increased expression of the lactate dehydrogenase that converts pyruvate to lactate during glycolysis, but no change in genes involved in the TCA cycle or pentose phosphate pathway. Taken together these results showed that the ME49 parasite induces glycolysis in the host [Blader et al., 2001]. It has also been shown that several important nutrients including glucose and purine nucleotides, part of the building blocks of nucleotide bases, cannot be synthesized by the parasite and must be acquired from the host. Therefore, as well as host survival and growth, the parasite may subvert host metabolism to gain nutrients for proliferation [Blader and Saeij, 2009].

The parasite tachyzoites have similar metabolic mechanisms to their mammalian hosts [Kloehn et al., 2016]. This includes enzymes that catalyse glucose and glutamine in an internal TCA cycle within the parasite. The existence of a TCA cycle in the toxoplasma parasite was a surprising result because the parasite does not have the PDH complex which is required to convert pyruvate to acetyl-CoA, the starting point of the TCA cycle. However, studies identified an enzyme within the parasite, branched chain alpha-keto amino acid dehydrogenase (BCKDH) that takes the role of PDH in the parasite TCA cycle. The inhibition of BCKDH or other TCA cycle enzymes resulted in a loss of proliferation of the tachyzoites. Thus, highlighting the importance of metabolism and the parasitic TCA cycle in the survival of *T. gondii* [Kloehn et al., 2016].

The ability of toxoplasma to metabolise both glucose and glutamine from its host helps to ensure its survival. The depletion of either glucose or glutamine did not affect parasite growth or survival. Indi-

cating that the parasite can switch from metabolising either source in response to its availability in the host cell [Nitzsche et al., 2016]. Further, depletion of both glucose and glutamine resulted in only a partial reduction of parasite growth suggesting that the parasite is also able to metabolise amino acids or the less utilised carbon sources such as acetate or fatty acids, for energy. Consistent with this it was shown that the parasites' sole glucose transporter TgGT1 is dispensable in RH tachyzoites. Loss of glucose in the host was compensated for by an increase in glutamine uptake to ensure survival in the host, though some growth defects are observed in TgGT1 knockout parasites. Motility was restored after supplementing growth media with glutamine but not pyruvate, indicating a dependence on glycolysis but not TCA cycle [Blume et al., 2009]. However, in later work Blume *et. al* showed the toxoplasma enzyme FBPase2 (TgFBP2) is essential for parasite survival in the host. It was required in both glucose replete and depleted states indicating it may also be involved in the switching of metabolism to accommodate the availability of different carbon sources. As a mechanism of action, it was proposed that this enzyme may control the activity of parasite metabolic pathways without requiring changes in transcriptional regulation [Blume et al., 2015].

T. gondii has several advantages as a model system both within the Apicomplexan group and in comparison, to other systems for modelling intracellular parasitism and cancers. Of the parasites within the Apicomplexan group, *T. gondii* has the highest transfection efficacy, that is the introduction of the parasite or foreign DNA to eukaryotic cells. It is also amenable to the addition of reporter constructs and tags that are useful in experimentally probing and manipulating the parasite and its interaction with the host [Kim and Weiss, 2004]. Reporter constructs can be used to follow the activity of, for example, a gene in the parasite and how it interacts with the host.

We used the JGL model to analyse the regulatory networks of *T. gondii* infected host cells. By using the JGL model we negated the necessity of prior knowledge for inferring regulatory networks. We inferred the regulatory network of the hosts infected with different strains of the parasite. In this way, our focus was on the regulatory networks impacted by the parasites and the differences between the networks across two strains: ME49 (Type II) and RH (Type I). We used the empirical Bayes method outlined in the previous chapter to improve the estimates of the correlation matrices used as input to the JGL model. By annotating the network using multiple existing functional and disease ontologies we identified interesting connections in our network and evaluated these interactions using existing publicly available gene knockout genome wide expression data, where available.

5.5 Results and Discussion

5.5.1 Experimental data

THE RNA-SEQ DATA set contains samples from host mouse embryonic fibroblast (MEF) cells infected with two different *Toxoplasma gondii* strains harvested at different time points, as well as control samples of uninfected MEF cells harvested at the same time points. The two strains used were RH, a type I strain and ME49 a type II strain. These data were generated by Lalitha Sundaram a member of the Ajioka lab in the Department of Pathology at the University of Cambridge.

The data contains three biological replicates per condition, this is not enough for input into the JGL algorithm. However, as the samples are over multiple conditions for different strains, we can combine conditions over each strain to give up to 12 replicates for each of the ME49 and RH strains and 9 for the controls. There are two experimental factors for each set of infected samples. The first is multiplicity of infection (MOI) taking values of either 1.3 or 3. The second factor is time, with cells harvested at either 24 or 43 hours, after infection. Three replicates at each of these four factor combinations gives the 12 samples per strain. The control samples are uninfected cells at 0, 24 and 43 hours with three replicates at each time point.

Multiplicity of infection is the ratio of *Toxoplasma gondii* parasites to uninfected MEF cells.

By combining data to create three classes, ME49, RH and control, we have multiple perturbations of the system and potentially enough replicates to infer cell specific regulatory networks for mouse embryonic fibroblasts infected with different strains of *Toxoplasma gondii*. Although combining samples results in less specific classes, that is, the network results cannot be decomposed according to the time factors, critically these three classes from combined experimental samples can be input into the JGL algorithm. Using a joint model with this data set is an appealing modelling choice here as it has the potential to identify differences in the transcriptional response of the host to different strains of *Toxoplasma gondii*.

5.5.2 Aligning RNA-seq reads

RNA-seq reads were aligned using STAR with the current mouse annotation: the GRCm38/mm10 assembly of the mouse genome downloaded from <http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/>. As the data set for infected cells contains RNA from both the MEF cells and the *Toxoplasma gondii* parasite we would like to account for this in the alignment. There is a possibility that toxoplasma RNA-seq

reads could align to the mouse genome or vice versa if they are aligned to each genome separately. By creating a mock genome, comprising both the mouse and *T. gondii* genomes, the alignment algorithm can find the best single location for each read across both genomes. For the ME49 strains the data were aligned to such a composite genome containing both the mouse and ME49 strain *Toxoplasma gondii* genome. The RH genome was used in conjunction with the mouse genome but the annotation is currently incomplete and contains annotations for only a very small number of genes. Meaning that there were a greater number of unmapped reads for the RH data. There were also potentially false positive matches for the mouse matches if some of the RH *Toxoplasma gondii* reads mapped to the mouse genome.

Currently the RH annotation on toxo.db contains fewer than 200 genes. Therefore, for comparison purposes we also aligned the RH infected cells to a combined genome of *Mus musculus* and the GT1 strain of *T. gondii*. GT1 and RH are both type I strains of *T. gondii*, and the difference in the sequences of these two strains have been analysed. Research identified 1,394 differences in the full genomic sequence between RH and GT1 [Yang et al., 2013]. Although this research sequenced the full genome for the RH strain, RNA-seq alignment requires for these genome sequences to be annotated with the exon/gene information before they can be used in alignment programs. The differences between these two genomes take the form of either single nucleotide polymorphisms (SNPs) or insertions/deletions between the sequences, see Section 5.7.1. From an alignment perspective, there are frequently mismatches between the RNA sequences and the genome alignment and the alignment software is designed to allow for a certain number of mismatches between the reads and target sequences. This makes it possible to use the GT1 genome in place of RH despite the variation between them. Further, RNA-seq protocols provide reads that are most often partial sequences of the full exonic or gene sequences, with reads in the datasets here being 101 bases in length. The nature of the fractional input reads is also accounted for in alignment algorithms.

Altogether the above means that we expected appropriate alignment algorithms to be able to accommodate the single base pair differences between the GT1 and RH genomes when aligning reads. We used the STAR algorithm to align the RNA-seq reads to the joint genomes using their respective annotations [Dobin et al., 2013]. The STAR algorithm outputs read counts for the transcripts given in the annotation file. These transcripts are exons for the mouse genome. Therefore, we used the featureCounts functions in the R package Rsubread to combine the exon counts to counts for genes [Liao et al., 2014].

We assessed the combined genome and mouse only genome testing with the mouse only controls. The alignment of mouse only RNA to the mouse genome and the combined mouse GT1 mock genome showed good agreement, see Methods section 5.7 for more detailed analysis.

5.5.3 *Pre-processing the data*

From the raw counts, there are multiple steps through which the data are corrected for technical factors and transformed onto the continuous scale so that it is suitable for input into the JGL model. The methods for doing this are outlined in the next sections.

5.5.3.1 *Selecting input genes through expression and variance filters*

The total number of genes in the ENSEMBL annotation for *Mus musculus* is over forty thousand, including protein genes, non-coding genes and pseudogenes. In any one cell type such as embryonic fibroblasts, it is expected that only a portion of these genes will be expressed. As a result, it is common to filter datasets to exclude those that are unexpressed. We filtered the data to remove low counts per million (cpm) per gene, leaving us with 13,279 genes. This also helped to reduce the dimension of the space (size of p) that would otherwise be likely to be prohibitive in calculating the inverse correlation matrix in the JGL algorithm.

Following this we filtered the remaining genes according to the overall variance of each of the genes. Using independent filtering can increase the power of detection of following statistical tests [Bourgon et al., 2010]. We removed those genes with low variance; the combination of low variance and high correlation between two genes is not likely to be an informative relationship. We retained the top 60th quantile of expressed genes according to their variance, this is consistent with the suggestion in [Falcon and Gentleman, 2007] for using a variance filter. Overall, we used stringent filtering of the data that left 5,312 genes for input into the JGL algorithm.

5.5.3.2 *Positive semi definiteness of sample covariance matrices*

One assumption of the JGL model is that the covariance matrix should be positive semi definite. Because we are approximating the covariance matrix by the sample covariances we may have a sampling error that results in a matrix that is not positive semi definite [Anderson and Gerbing, 1984, Knol and ten Berge, 1989]. To evaluate whether the sample covariance matrix is positive definite or not we

calculated the eigenvalues of the sample correlation matrices for the ME49 and RH samples and the uninfected samples. We calculated the correlation matrices as these were the input into the JGL model - this standardisation of the covariance matrices will have no impact on whether the eigenvalues are negative or not. The summary results for these matrices are shown in Table 5.1. We can see that the minimum value for the eigenvalues are negative, and technically this means that these are not positive semi definite. However, we also note that these are very small (close to zero values) and zero eigenvalues are allowed in positive semi definite values. We could correct these eigenvalues to be zero [Vershynin, 2012, Higham, 2002, Bates and Maechler, 2017], however, this would have no impact on the output of the JGL algorithm as the small eigenvalues are rounded out to zero in the downstream calculations.

Strain	Min.	Mean	Max
ME49	-6.96e-12	1	2647
RH	-8.35e-12	1	2666
Uninfected	-8.50e-12	1	2538

Table 5.1: The eigenvalues for the sample covariance matrices are shown for the two different strains. Although negative values are present these are very small negative values, and such close-to-zero valued eigenvalues could be allowed in positive semi definite matrices.

5.5.4 Initial Data analysis, parameter selection

For initial data analysis, we have shown in the previous *Bacillus subtilis* analysis that the block structure of the correlation matrices provides a lot of information on the computational tractability and sparsity of the resulting networks. For example, since the algorithm performs matrix inversion on each block separately, the maximum block size (or number of genes in one block), acts as an upper bound on the number of genes in one subnetwork. Similarly the summary statistics of the block structure give the potential scope of the network; the number of blocks tells us the number of potential subnetworks, and the size of each of these gives the maximum number of genes that can be connected in a subnetwork. Based on the analysis of the block structure and shrinkage parameters, the values $\lambda_1 = 0.905$ and $\lambda_2 = 0.005$ were selected, see Section 5.7.6 for details.

5.5.5 Network annotation analysis

FROM A BIOLOGICAL PERSPECTIVE we began by searching PubMed for terms we expect to be associated with the genes in the network to get an overview of the accuracy of the connections within the network.

As we know that the *Toxoplasma gondii* parasite can subvert the host through similar methods to tumours, we searched for our genes and

PubMed is an online database of published peer reviewed articles. The database can be searched using keywords and phrases. Advanced queries can also be made by searching for terms in particular fields of the database such as Author or Year as well as combining search statements using logical operators such as AND or NOT.

‘toxoplasma’ search term, and also the combination of gene name and keyword ‘tumour’ as we expect this to be a more prevalent keyword in PubMed due to a larger amount of research being conducted on cancer as opposed to toxoplasma.

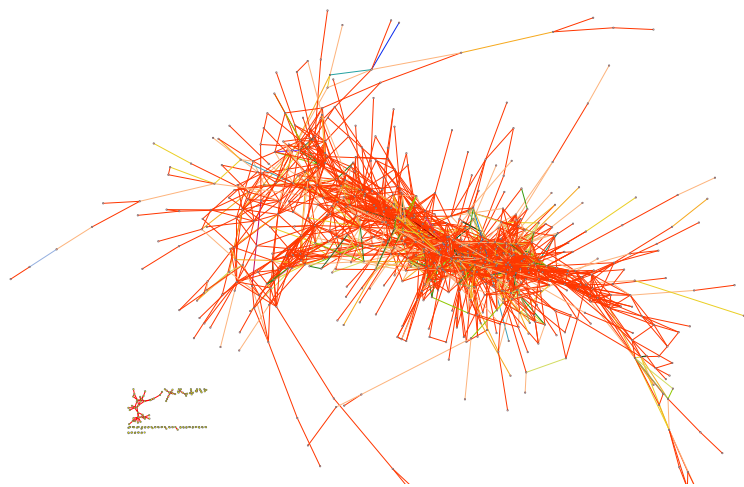
Gene names were used in a search on PubMed with the other terms of interest. Searching for gene names and toxoplasma gave 61 out of 735 genes with more than one paper found on PubMed. The maximum number of papers found for a single gene was 125 for the dihydrofolate reductase gene (*Dhfr*). A search requiring a match to both a gene name and tumour gave 619 out of 735 genes showing at least one paper. Amongst the genes showing the highest number of results were the oncogenes *Jun* and *Vegfa*, transcription factor *Myc* as well as *Hif1a*. This indicates that the results could be meaningful, as the network contains genes previously associated with toxoplasma and/or tumours in the literature.

The PubMed search treats all genes in the network individually. For the network to be informative however the connections between the genes must also be biologically meaningful. Although a lot of work has been done to understand regulatory networks in mice, there is still no resource that easily stores the experimentally validated transcriptional links. The MGI database does however provide micro RNA (miRNA) information for genes [Eppig et al., 2015]. By identifying genes that are regulated by the same miRNA we can identify genes that may be present in the same functional regulatory units. Although this does not prove that two genes are in the same transcriptional unit, it does provide evidence that supports the network result from a biological perspective. Using the data on miRNA targets available from MGI (the mouse genome informatics resource) we created an interaction matrix that adds one to the edge score between two genes each time two genes are found to be targets of the same miRNA. Edges are given a weight according to the number of miRNAs they share, consequently the network can be coloured with miRNA target information. Although we found connections between genes with shared miRNA regulators, the coverage of the miRNA is not complete and is therefore limited in its ability to validate the network output.

Figure 5.2 shows the results of annotating the edges according to their miRNA score. As expected there are large areas of orange edges that represent zero miRNAs in common but also hubs of coloured edges. This is a result we would expect to see as we know that miRNAs do not have single targets and genes acting under the same regulators, either as transcription factors or miRNA should have similar correlation profiles and potentially significant partial correlations meaning that they would be connected in a JGL output.



Figure 5.2: Overview of original network with the edges coloured according to the number of miRNA targets, taken from MGI, in common between the two genes, as shown in the legend. Orange edges indicate there are no miRNAs in common. We see one large subnetwork with hubs of colour indicating we may have functionally related genes connected in the network.



5.5.6 Empirical Bayes, aiding interpretability

With a small number of replicates, we were interested in reducing the number of false positives in the model. The empirical Bayes (EB) method introduced in the previous chapter is used to calculate a modified correlation matrix. The empirical Bayes method outputs a correlation matrix that is passed into a modified version of the JGL model. The empirical Bayes method similarly takes as input a shrinkage parameter and we used the same value in the JGL model above, that is, $\lambda_1 = 0.905$. This makes the results of the JGL algorithm with the EB matrix comparable to those of the previous section as we used the same initial shrinkage parameter to define the signal threshold in the data.

5.5.6.1 Assessing the positive definiteness of the output matrix

The original covariance matrices for each experimental factor individually gave several small but negative eigenvalues. There were 2625, 2560 and 2617 small negative eigenvalues for ME49, RH and Uninfected respectively. After recalculating the correlation matrices using the empirical Bayes method, the ME49 matrix is positive definite, with no zero values. The RH matrix has 38 negative values, the uninfected matrix however, still has 2063. This could be another indication that the initial correlation matrices for the uninfected samples are less reliable than the infected samples. However, we still had confidence in using the RH correlation matrix.

5.5.6.2 Comparing the original and EB JGL model output

The JGL model was run on the empirical Bayes estimated correlation matrices for the ME49 and RH strains only. The empirical Bayes method uses the same shrinkage value of $\lambda_1 = 0.905$ as in the previous section to define significant shrinkage values. Because the diagonal prior has impacted the values of the correlations both within and outside the defined blocks, the shrinkage parameters are altered accordingly to $\lambda_1 = 0.895$. The results found 791 genes as opposed to 735 previously. We also note that of these 791 using the Feature Type annotation from MGI we found 17 which had an annotation other than protein coding gene and 2 were annotated as small non-coding RNAs that could be miRNAs. We searched PubMed for these 791 genes and found 63 regarding ‘toxoplasma’ and 644 with search term ‘tumour’. This suggests that the additional genes included in the model using the EB matrices may be informative as opposed to false positives as we retained similar percentages of genes around 8% finding results with ‘Toxoplasma’ and 80% connecting with ‘tumour’.

There was a reduction to 2252 from 2757 edges for the RH data and this indicated a potential reduction in false positives: the network contains more genes with fewer edges, the EB matrix has not reduced the number of connected genes but has resulted in greater sparsity between the nodes as there are overall fewer edges between genes. This sparser network will be easier to navigate as there are fewer connections between genes. It may also help us to resolve the hierarchical structure of the network and identify causal or direct links by removing false positive edges from our network. For example, a subnetwork where we have an edge between each pair of genes cannot be written as a hierarchical or causal network as all genes would be affecting each other.

In contrast, the ME49 had a large increase in the number of edges from 549 to 2431 which is a comparable number to those seen for the RH strain. This suggests that there was a greater variability of correlation values within the block structures for the ME49 strain compared to the RH strain at the same shrinkage level and that using the EB procedure has moved the smaller correlations towards an overall mean that has resulted in more edges found in the model. Because we changed the shrinkage level to 0.89 from 0.905 for use with the EB correlation matrices, it is possible that this reduction in the shrinkage parameter also increased the number of edges found for the ME49 strain. However, the maximum block size plots (Figure 5.27) for the original data set support the argument that the highest correlations have reduced variability for the EB estimates. This is because if we had just reduced the shrinkage level to 0.89 using the original data we would also have a maximum block size containing over 2000 genes which is not the case using the EB matrix and shrinkage level 0.89.

The miRNA target information was also used to evaluate the edges included in the EB model Figure 5.3, the network had non-zero edges for 28 percent of the edges compared to 29 percent in the original model. Given the increase in the number of genes and large increase in the total number of edges this showed that the EB method can identify additional interactions at a comparable level of specificity. In comparison to Figure 5.2, the empirical Bayes method in Figure 5.3 shows a second network of results with similar levels of non-zero miRNA target edges. We also randomly perturbed the node names in our network and calculated the number of genes sharing miRNAs in these random networks with the same degree structure as our network. We found the empirical p-values of observing the number of edges between genes with shared miRNAs in our network or more, by chance, was three percent. This indicates that the network has a statistically significant number of genes with shared miRNAs connected to each other.

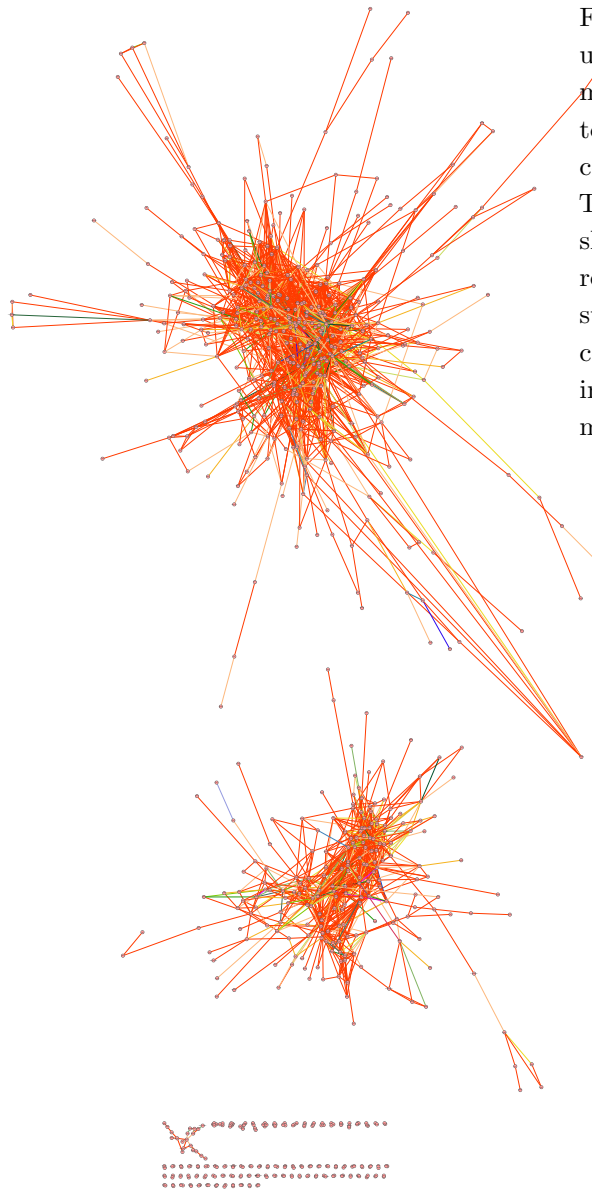


Figure 5.3: Overview of network using the empirical Bayes estimates, edges coloured according to the number of miRNAs in common between the two genes. The empirical Bayes network shows overall a sparser network result, with two smaller main subnetworks. There are hubs of colour showing connected genes in the network that have shared miRNAs.

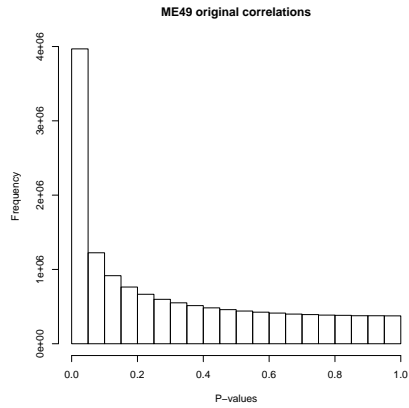
Since the empirical Bayes adjustment alters the correlation values it does not make sense to compare the results with and without the correction at the same shrinkage value. However, for both models there were a similar number of total genes included in the model with similar levels of information, as summarised by the PubMed searches and miRNA interactions. Moreover, we can see that the EB model has had the desired effect of enhancing the block diagonal structure of the input correlation matrices. This is because on a comparable network scale we now have two clear subnetworks (or blocks) from the EB estimate as shown in Figure 5.3, where the original network Figure 5.2, has one large block. Using the EB correlation matrices has resulted in a sparser network that is easier to visually interpret.

We also compared the effect of the λ_1 values used in each case that determine the block diagonal structure. This enabled us to assess the false positive rates of the input correlation matrices, although not the edges in the network output as these were the partial correlations calculated by the JGL algorithm. To do this we calculated p-values for the correlations and multiple hypothesis corrected them for each of the four matrices (ME49 and RH with and without the empirical Bayes correction). In this way we were able to establish a minimum correlation (λ_1) or critical value that would give an FDR of 5%. These values were, 0.73 for the original ME49 matrix, 0.69 for RH, 0.72 for the EB ME49 matrix and 0.77 for the EB RH matrix. The increased critical value for the RH strain from 0.69 to 0.77 showed that the EB covariance matrix now has a higher proportion of values in the null distribution. This implied a potential reduction in false positives when the shrinkage values are greater than the critical values, as they were in the parameter values we selected. Furthermore, our selection of λ values can be seen as prudent in all cases for determining the block diagonal form of the correlation matrices, as they were all larger than the critical values.

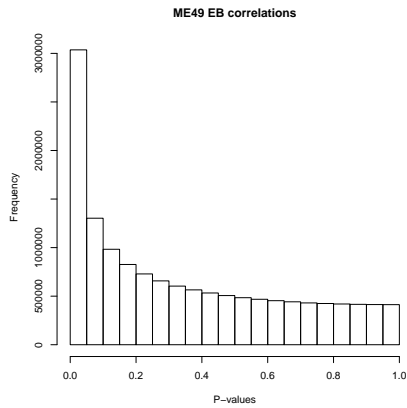
We also observed the impact of the EB estimates on the distribution of the p-values. The distribution of p-values arises from the same hypothesis test being performed multiple times. We have performed multiple hypothesis tests here for each correlation value between pairs of genes. The distribution of p-values from multiple hypothesis testing results in a mixture of values from the null and alternative distributions. In this case, the null hypothesis is that the correlations are not significantly different from zero. The theoretical distribution of the null p-values is Uniform [0,1]. Correlations that are significantly different from zero are from the alternative distribution have low p-values. This gives the peak at the left-hand side of the p-value histogram. The proportion of the null to alternative distributions can be used to multiple hypothesis correct the p-values [Pounds and Morris, 2003, Storey,

2003].

We used these plots as a diagnostic to visualise the impact of the EB estimates on the p-value distributions. The increased proportion of the ME49 null distribution means that we expected a reduced false positive rate. This is shown by the plots of the p-values for the ME49 distribution where Figure 5.4 clearly shows an increase in proportion of p-values in the null distribution following the EB correction. A similar though not as large effect is seen for the RH strain in Figure 5.5. The significance of these correlations determines the construction of the block diagonal form. There is partial overlap between the gene sets before and after the EB correction: 578 of the original 735 genes are also included in the 791 genes used by the JGL algorithm for the EB correlation matrices. Following this, though the total number of genes were comparable (735 vs 791), the JGL inference found significantly more edges between the genes for the ME49 strain using the EB input. This showed that the genes selected for the EB estimates show more conditional dependence, and more likely causal relationships. This is a stronger statement than being correlated. A significant correlation between genes meant they were input into the model. However, the small number of edges in the model showed that many of these correlations did not also result in a causal relationship, as the partial correlations calculated by the JGL model were not significant.

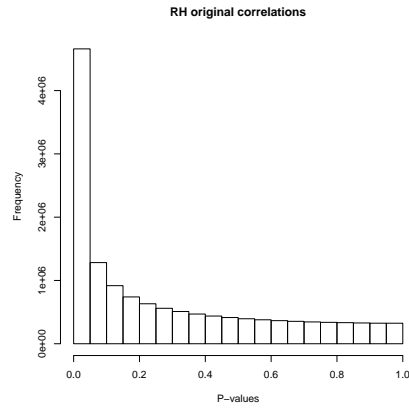


(a) P-value plot for ME49 original correlation matrix

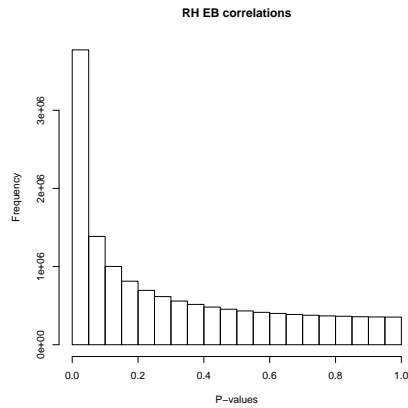


(b) P-value plot for ME49 EB correlation matrix

Figure 5.4: P-value plots for ME49 sample and EB correlation matrices. After the EB correction there is a larger proportion of correlations that would be assigned to the null distribution, indicating this could lead to a reduction in the number of false positives.



(a) P-value plot for RH original correlation matrix



(b) P-value plot for RH EB correlation matrix

Figure 5.5: P-value plots for RH sample and EB correlation matrices. After the EB correction there is a small increase in the number of correlations that would be assigned to the null distribution, this may lead to a reduction in the number of false positives.

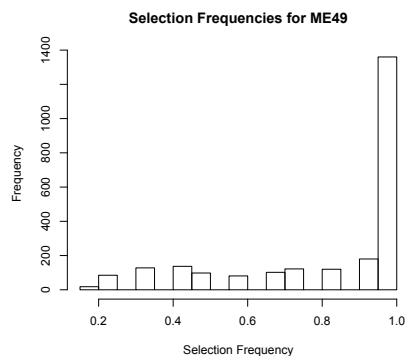
5.5.7 Evaluating the network

In the previous section analysis of the p-values for gene correlations input into the JGL model showed a high level of significance. For the model output we would also like to have a measure of the accuracy of the network results. In comparison to the *Bacillus subtilis* analysis we do not have the same transcriptional unit databases that can be used to evaluate the network edges. Therefore, the selection frequencies of the edges are used to provide an estimate of the accuracy of the networks. These selection frequencies are calculated by inferring the JGL network for each of the bootstrap samples using leave-one-out to subset the data.

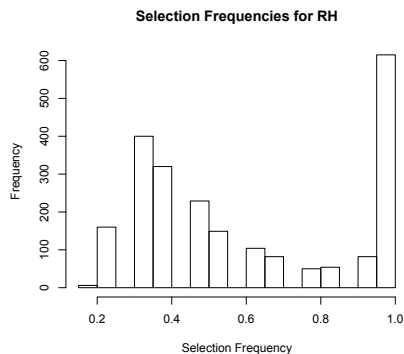
The selection frequencies for the edges in our network after bootstrap sampling and for each of the strains individually are shown in Figure 5.6. We see high selection frequencies for most of the edges for the ME49 strain as indicated by the peak at the right end of the plot. The RH network shows a more diverse distribution though there is still a large proportion of maximal selection frequencies for the edges in the network.

Bootstrap samples are used to estimate the robustness of estimated parameters. The data set used to infer the parameters are sub-setted and the parameters re-estimated based on a subset of data. This gives a confidence interval for the parameter values.

Leave-one-out method for a data set with n samples; each sample is removed once and the estimate is re-calculated using the remaining $n-1$ samples.



(a) Selection frequencies for ME49 bootstrap samples



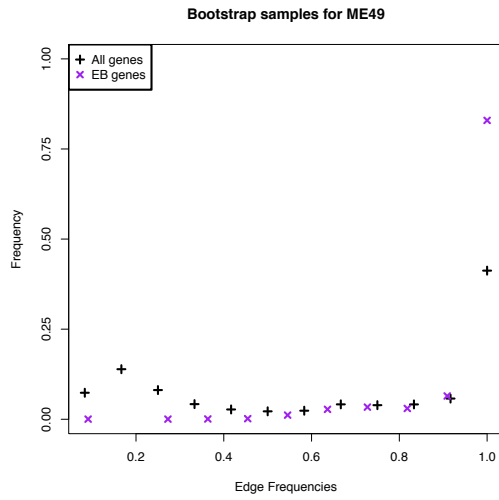
(b) Selection frequencies for RH bootstrap samples

Figure 5.6: Selection frequencies for the ME49 and RH bootstrap samples for the edges in the EB network. These sparse histograms show that the number of samples we have is likely too small to fit a distribution too. The distribution of these bootstrap selection frequencies should theoretically be a U-shaped distribution.

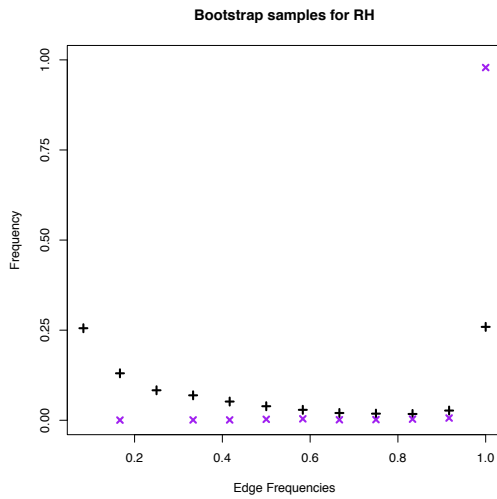
These plots are for the edges in our EB model only, and used the same shrinkage parameter as the model based on the full data set. As mentioned in earlier chapters bootstrap analysis may suffer from the low sample sizes. In this network analysis, we had a maximum of 11 samples for the ME49 network and 12 for the RH network. The Bootstrap Inference for Network CONstruction (BINCO) method uses bootstrap samples and varying shrinkage parameters to calculate FDR and select shrinkage parameters to maximise power [Li et al., 2013a]. However, this method only considers shrinkage parameters that result in a reasonable mixture of edges from both the null (insignificant) and significant set of edges. These shrinkage parameters are chosen as those that result in a U-shaped distribution of selection frequencies. This U-shape arises as a mixture of null edges that have a peak around zero and rapidly decrease to 1, while conversely significant edges will result in a peak around 1 and a left-tail of decreasing proportions towards zero. Because we had used a stringent shrinkage parameter, Figure 5.6 shows that, particularly for the ME49 strain we do not have a U-shaped distribution and that most of these edges would be classed as significant by this method.

To test this the shrinkage parameter was reduced to 0.88 from 0.895. Figure 5.7 shows the selection frequencies with the smaller shrinkage parameter in black points and those for the edges included in the EB model with $\lambda_1 = 0.895$ shown as purple points. For both strains the proportion of edges with selection frequency 1 is larger for the edges included in the EB model. In the work by Li *et. al* they found a significant reduction in FDR using a threshold method to include those edges with selection frequencies of at least 0.5. For the larger bootstrap network with shrinkage parameter 0.88 the proportion of the edges from our EB network that have a selection frequency of at least 0.5 was 0.99 for both the RH and ME49 network. Increasing this threshold to 0.8 the proportion of edges is 0.98 for the RH and 0.92 for the ME49 strain. Taken together these results indicated a robustness for the edges in our network. The same bootstrap analysis was also run with the Pearson's correlation matrix, in this case using the original shrinkage parameter $\lambda_1 = 0.905$ none of the edges found for the full data set were found in any of the bootstrap samples. Further, reducing the parameter value to $\lambda_1 = 0.89$ no edges were found for the RH strain and only one with selection frequency 0.45 for the ME49 strain. This showed how, with these low sample sizes, the EB procedure resulted in more stable correlation matrix estimates and leveraging the information from across the gene pairs enabled us to identify patterns of correlations in the data. We did not use the fitted models from the BINCO procedure as these continuous approximations did not work well with our data set: the BINCO

procedure is designed to work with a minimum of 20 samples where we had 11 for the ME49 and 12 for the RH network.



(a) Selection frequencies for ME49 bootstrap samples, $\lambda_1 = 0.88$



(b) Selection frequencies for RH bootstrap samples, $\lambda_1 = 0.88$

Figure 5.7: Selection frequencies of edges in a model with shrinkage parameter $\lambda_1 = 0.88$ for the ME49 and RH bootstrap samples. The subset of edges in the EB network. This shows a large proportion of edges included in the EB model for both strains had a selection frequency of 1. This indicates robustness in the selection of edges in the EB model.

5.5.8 *Annotating the network*

TAKING THE JGL MODEL OUTPUT using the EB correlation matrices as input we use a combination of known ontologies and supporting experimental data to evaluate the network. From a global perspective, we look for overrepresented biological ontologies in the genes included in our network. For the edges between the nodes we use publicly available experimental knockout (KO) data to provide additional evidence to support, where possible, the interactions in our network. In gene knockout experiments the organism is altered so that the target gene is rendered inactive. In knockout experiments the output for the knockout is compared to the output from a wild-type. The wild-type is a control sample where the target gene is unaltered and active as normal. By comparing, for example, genome-wide expression for wild-type and knockout samples, it is possible to identify the effect of the target gene and often form hypotheses about its function. This is done by identifying those genes (with known function) that show differential expression between the wild-type and knockout samples. Where we have experimental data in MEF cells we take this as stronger evidence than the global miRNA analysis from the previous section. The coverage for the KO analysis was limited as we required both KO experiments for genes in our network and, for the knockout to have been performed within MEF cells.

5.5.8.1 *Overrepresented terms*

We took a systems biology approach, and integrate information from additional sources to annotate the results. This gives us a wider understanding of results as well as giving a method to validate the model and further highlight potentially novel results. Annotations were taken from several different sources and for functional information we use any Gene Ontology annotation available for each gene. From a visual perspective, it is difficult to identify areas of the network sharing common ontology terms. This is because each gene will often have multiple ontologies associated with it and therefore, while there may be shared ontology terms between two genes, unless two genes share the same Gene Ontology terms they will be coloured differently. A better method to summarise the gene ontology information within the network is to perform statistical tests to determine if there is an overrepresentation of gene terms within the set of genes in the network. This is analogous to testing for overrepresentation of Gene Ontology terms within a list of differentially expressed genes.

Table 5.2 gives those GO terms that were found as being overrepresented for the full network. There are many terms found, these include

cell adhesion	homeostatic process
cell division	ribosome biogenesis
enzyme binding	extracellular space
lipid particle	cell differentiation
lyase activity	extracellular region
protein complex	biosynthetic process
Golgi apparatus	phosphatase activity
plasma membrane	endoplasmic reticulum
ligase activity	immune system process
GTPase activity	membrane organization
histone binding	DNA metabolic process
mRNA processing	chromosome segregation
ATPase activity	tRNA metabolic process
kinase activity	oxidoreductase activity
protein folding	transmembrane transport
nuclear envelope	lipid metabolic process
catabolic process	chromosome organization
protein targeting	developmental maturation
nuclease activity	protein complex assembly
helicase activity	mitotic nuclear division
response to stress	unfolded protein binding
embryo development	enzyme regulator activity
cell morphogenesis	cytoskeleton organization
peptidase activity	mitochondrion organization
nuclear chromosome	signal transducer activity
isomerase activity	circulatory system process
cell proliferation	vesicle-mediated transport
vacuolar transport	methyltransferase activity
protein maturation	cofactor metabolic process
signal transduction	cellular component assembly
cell-cell signaling	nucleocytoplasmic transport

Table 5.2: Table containing all the significantly over represented multiple hypothesis corrected gene ontology terms at five percent significance. These are for a JGL network output with two classes, the ME49 and RH *Toxoplasma gondii* infected samples and empirical Bayes matrices. Continued on next page.

cell	neurological system process
aging	cytoskeletal protein binding
growth	transcription factor binding
nucleus	structural molecule activity
cytosol	protein transporter activity
vacuole	microtubule organizing center
endosome	ubiquitin-like protein binding
lysosome	carbohydrate metabolic process
ribosome	macromolecular complex assembly
cytoplasm	nucleotidyltransferase activity
organelle	anatomical structure development
transport	small molecule metabolic process
autophagy	external encapsulating structure
nucleolus	sulfur compound metabolic process
locomotion	transmembrane transporter activity
cell death	structural constituent of ribosome
cell cycle	ribonucleoprotein complex assembly
chromosome	cytoplasmic membrane-bounded vesicle
ion binding	cellular protein modification process
nucleoplasm	cellular amino acid metabolic process
DNA binding	cellular nitrogen compound metabolic process
RNA binding	hydrolase activity, acting on glycosyl bonds
translation	protein binding transcription factor activity
cytoskeleton	transferase activity, transferring acyl groups
reproduction	generation of precursor metabolites and energy
mRNA binding	nucleobase-containing compound catabolic process
rRNA binding	nucleic acid binding transcription factor activity
intracellular	transferase activity, transferring glycosyl groups
mitochondrion	symbiosis, encompassing mutualism through parasitism
cell motility	anatomical structure formation involved in morphogenesis
lipid binding	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds

multiple cellular locations and different metabolic processes.

One very prevalent term we found is ‘cellular nitrogen compound metabolic process’, involving 202 genes out of 791 in the network. At over twenty percent of all the genes in our network this indicates that nitrogen processes are an integral part of the host response to parasite infection. This observation is supported by experimental results on Type II parasite mutants that identified genes necessary for the parasite to counteract the host production of reactive nitrogen intermediaries, a process that is essential to enable the parasite to replicate [Skariah et al., 2012].

Another valuable resource for globally assessing a network are the KEGG pathway databases [Kanehisa et al., 2016]. Using this information we tested for overrepresented pathways in the KEGG database within our network. After testing using the hypergeometric test and multiple hypothesis correction using Benjamini Hochberg we found 11 pathways significantly over represented at the 5% significance level, Table 5.3.

KEGG Pathway (Reference Number)	Adjusted p-value
Glycolysis / Gluconeogenesis (00010)	0.02
Galactose metabolism (00052)	0.02
Oxidative phosphorylation (00190)	0.01
Caffeine metabolism (00232)	0.00
Starch and sucrose metabolism (00500)	0.00
Butirosin and neomycin biosynthesis (00524)	0.00
Metabolic pathways (01100)	0.00
Ribosome biogenesis in eukaryotes (03008)	0.04
Spliceosome (03040)	0.04
Parkinson’s disease (05012)	0.01
Huntington’s disease (05016)	0.02

Table 5.3: Significant KEGG pathways in the network, reference number on the KEGG database shown in brackets. There are multiple pathways relating to metabolism as may be expected because the parasite has to gain nutrients for survival from the host.

From the KEGG analysis we found multiple pathways in metabolism, including the general result for Metabolic pathways. The parasite-host interaction involves interplay between the metabolic processes, enzymes and regulatory factors in both. The KEGG pathways include Glycolysis/Gluconeogenesis and Oxidative phosphorylation, for the host. These results showed consistency with the parasite metabolic pathways identified from analysis of different toxoplasma strains [Song et al., 2013] which included different strain metabolic responses in both Glycolysis and Oxidative Phosphorylation. The results showed a higher presence for genes in the OXPHOS pathways for the RH strain. Regulation of OXPHOS is important for parasite survival; OXPHOS produces ROS [Ray et al., 2012] and increased ROS in mice resulted in complete resistance to toxoplasma infection [Arsenijevic et al., 2000].

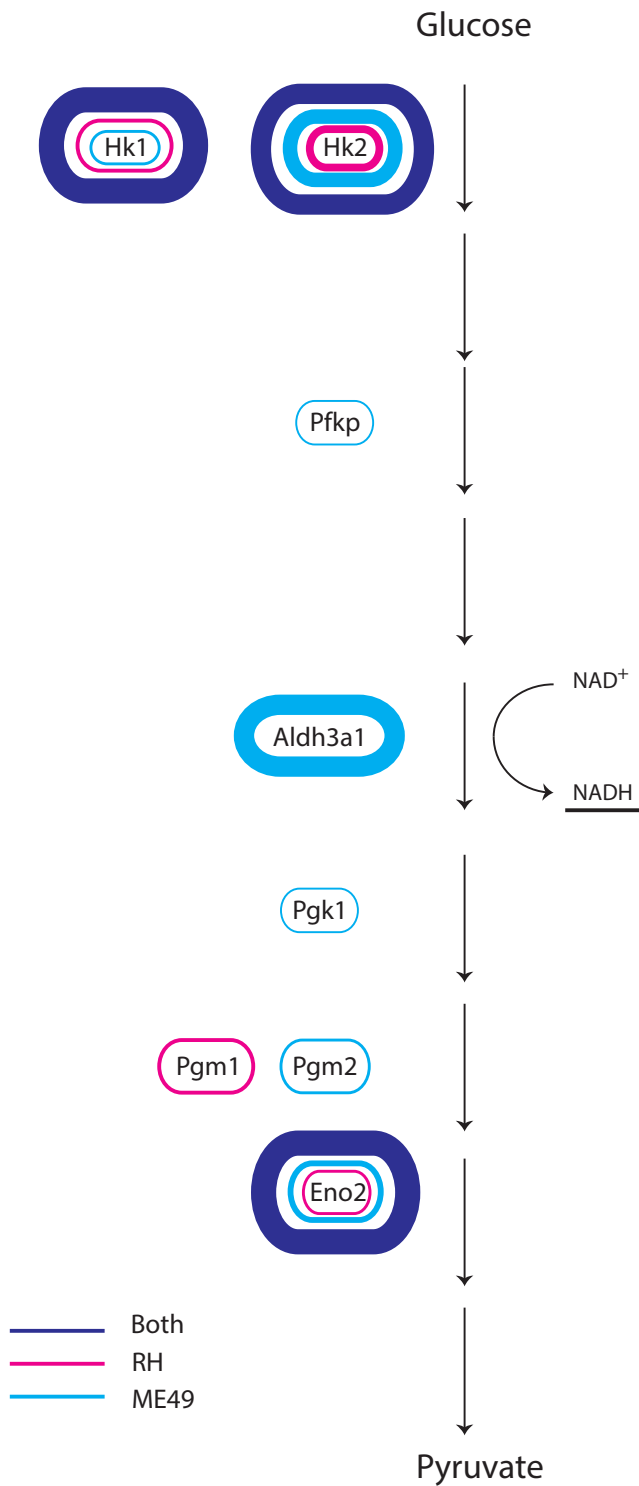


Figure 5.8: Enzymes involved in the 10 steps of glycolysis that were found in our network are shown. The circles around each gene indicate the strains they were connected in and the thickness of the line is proportional to the degree of connectivity. Reaction that creates the by-product NADH are shown.

Figure 5.8 shows the genes involved in glycolysis reactions that were present in our network. Of the 10 reactions in the glycolysis pathway most of the genes encoding the enzymes for these reactions were in our network. The rectangles around the gene show the strain each gene is in. The thickness of the lines is proportional to the degree of the gene in the network. The Hexokinases 1 and 2 were both highly connected for both strains as was Enolase 2 (*Eno2*). *Hk1* is strongly connected in the network with edges present for both strains. *Hk2* has more connections than *Hk1*. These include a mixture of edges assigned to only one or both strains. Due to the high connectivity of *Hk2* in our network, connections to this gene are shown in two of our subnetworks of interest shown later in Figures 5.9 and 5.13.

There are differences between the genes associated with glycolysis for the two strains. The *Pgm1* and *Pgm2* are two enzymes that catalyse one of the chemical reactions in the glycolysis pathway (from 3-phosphoglycerate to 2-phosphoglycerate). *Pgm1* is connected in the RH network and is more centrally connected with degree 4 connected to both *Myc* and *Jun*. The connection between *Jun* and *Myc* for the RH strain is shown in Figure 5.13. *Pgm2* is in a small subnetwork of 4 genes (with *P4ha2*, *Epm2a* and *Ugdh*) for ME49 strain. Three of the four are genes in are also involved in metabolism. The network connections showed high level of glycolysis activity though the connection of oxidative phosphorylation is less clear. We did not see a significant result for the TCA cycle. However, given the significance of OXPHOS for the RH strain this indicates regulation of aerobic respiration occurs in the host infected with the RH strain.

Normal non-proliferating cells utilise aerobic respiration of glucose or glutamine as their primary mechanism for ATP production. Our results showed that the infection of the host by Toxoplasma may induce additional metabolic pathways; the glycolysis/glucogenesis genes *Hk1*, *Hk2*, *Pgm1* and *Pgm2* were also included in the starch and sucrose metabolism. Further enrichment for the starch and sucrose metabolism is shown by the inclusion of genes *Gsub*, *Gys1*, *Gbe1* in our network. It may be expected that these genes play a role in regulating host metabolism in response to the parasite. The lactate dehydrogenase (*Ldhb*) was present in the RH network but no *Ldh* gene was found in the ME49 network. *Ldhb* is one of two isoforms that encode lactate dehydrogenase which converts pyruvate to lactate. There are five forms of the LDH proteins that consist of different combination of the LDH-M and LDH-H subunits. The two genes *Ldha* and *Ldhb* encode these two subunits. Of the five LDH forms only one is comprised entirely of the LDH-H subunit, LDH-1. Though all five LDH proteins have similar enzymatic activity, they differ in their distribution within the organism. In humans, LDH-1 is primarily

located in the heart and brain.

5.5.8.2 Supporting experimental evidence

Searching through the ArrayExpress database [Rustici et al., 2013] for knockout (KO) experiments in the MEF cell line resulted in 24 experiments that were either microarray or RNA-seq datasets. One of these was a knockout of the zinc finger protein *Zfp36* that is present in our network. This dataset (ArrayExpress accession number GSE5324) compared wild-type (WT) and KO data at different time points after treatment with actinomycin D with 5 biological replicates for each. Differential expression analysis was used to investigate supporting evidence to the interactions in our network. The strongest binding reaction found in the paper published by the authors that generated the data set, was with the Immediate Early Response 3 gene (*Ier3*) that is also present in our network and connected to *Zfp36* through the genes *Rhbdd1*, *Bhlhe40*, *Nars*, *Gja1* and *Maff* [Lai et al., 2006]. The differential expression data was calculated using the Geo2R analysis tool [Barrett et al., 2013b] to compare the wild-type and knockout data at 120 minutes as this is the longest time point available and therefore most comparable with our dataset.

The differential expression data were searched for those genes connected to *Zfp36* in the network as well as those connected to *Ier3*. We found multiple connections in our network that were also significantly differentially expressed in the *Zfp36* KO data in MEF cells. Plotting those genes connected to either *Zfp36* or *Ier3* with nodes sized according to their adjusted p-values, shown in Figure 5.9. In total 21 out of the 42 genes connected to either *Zfp36* or *Ier3* are significantly differentially expressed. Figure 5.9 shows two coloured hubs from *Zfp36* and *Ier3* as expected. Those nodes coloured green are significantly differentially expressed at the 5% level. The top 10 genes by differential expression to *Zfp36* directly connected in our network are *Ptgs2*, *Rbbp8*, *3110043O21Rik*, *Rnd1*, *Rusc2*, *Gja1*, *Herpud1*, *Nars*, *Maff* and *Ghitm*.

The combination of the knockout information with the conditional dependencies found in our network together gives a strong indication that these connections are direct regulatory interactions. The graphical results showed that the correlation between those genes connected to *Zfp36* and *Ier3* cannot be explained by any other gene included in our network. As we only included a subset of the genes in our network the KO data for *Zfp36* shows that there is also a direct connection between the activity of *Zfp36* and those significantly differentially expressed genes connected to it.

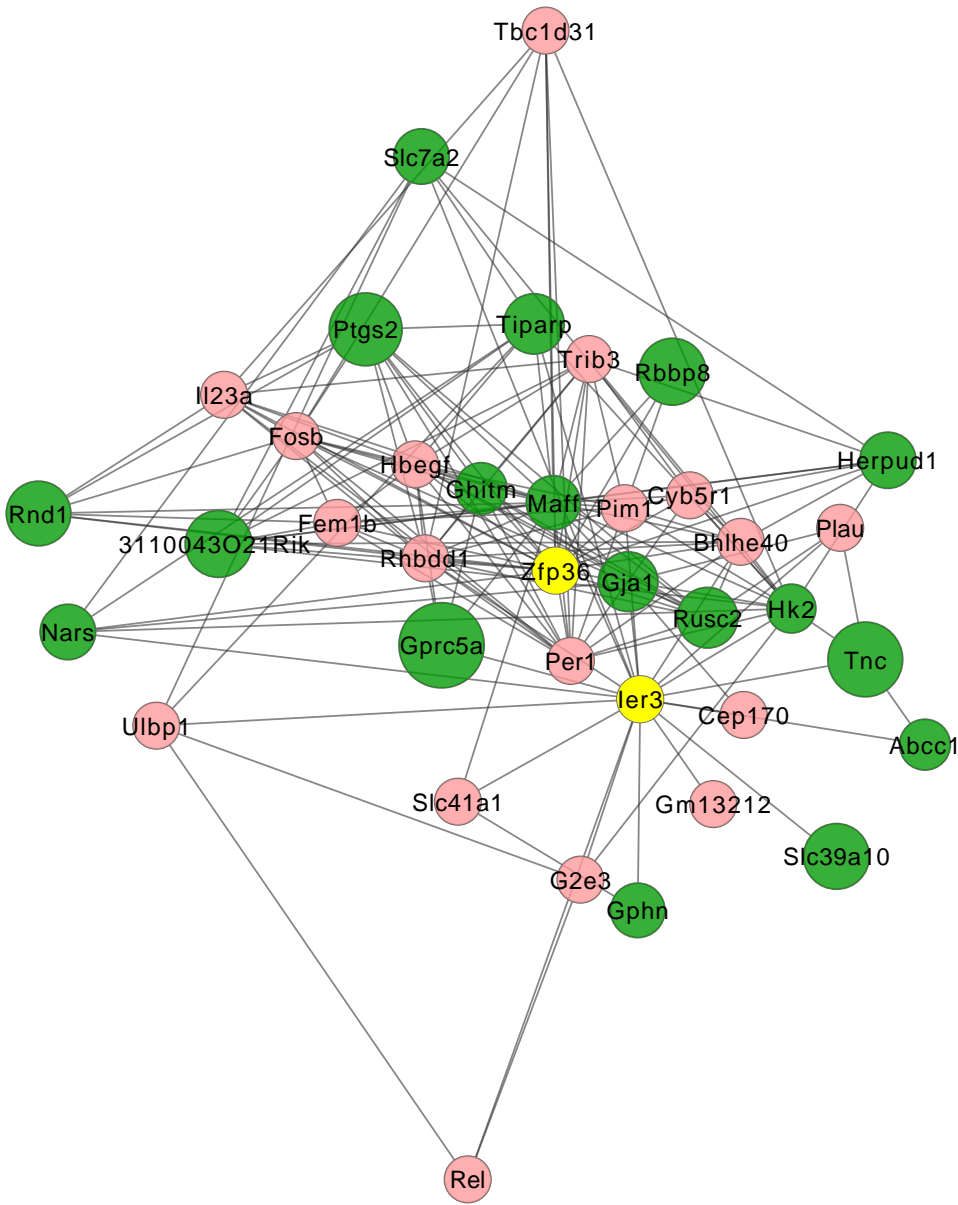


Figure 5.9: The green nodes show multiple genes that are connected to either *Zfp36* or *Ier3* in our network and that were also found as differentially expressed in the *Zfp36* KO experiment. Nodes are sized according to the adjusted $-\log(p\text{-value})$ of the differential expression, therefore the larger the node the more significant the differential expression.

5.5.9 Functional and disease Networks

The question remains how to generate testable hypotheses from our data set. Although we are globally able to identify known interactions and overrepresented functional terms, this does little to direct the researcher to the higher value targets or potentially interesting and novel connections. It is possible to visually interrogate the full network, though the complexity of the model makes this difficult. Therefore, we aimed to identify first sub networks of interest that can be reasonably assessed by a researcher.

5.5.9.1 Functional networks

The hypergeometric tests used to identify overrepresented terms in the previous section are calculated based on those genes present in the network but this does not account for the structure of the network. Therefore we combined the results from the hypergeometric tests summarised in Table 5.2 with the network results from the JGL algorithm. In this way, we moved from topological to functional or disease networks by extracting subnetworks of genes sharing ontological terms. This may also help to identify differences between the two *T. gondii* strains as we focus on subsets of the network. Visually the network output is still difficult to interpret with hundreds of genes in the largest of these. However, this approach allowed us easily to see any smaller networks of functional units and differences between the ME49 and RH strains. We only considered genes that are connected in a subnetwork containing at least three genes, to ensure that we used the information given by the network structure, which we did not have for unconnected nodes. In the following outputs, pink edges are for those edges present in the RH strain, turquoise in ME49 and blue edges for those present in both strains. As there are many significant ontology terms found in the network, over one hundred, we focused on those terms that are classified as biological processes as opposed to cellular locations or molecular functions.

We found that there were connected nodes under the two different conditions that are all ribosome biogenesis genes. These are not all the genes that are associated with this term but we focused on those that are connected as these give us potential functional networks (of connected nodes with a shared functional annotation) as opposed to single unconnected genes. For Ribosome biogenesis, Figure 5.10 there are multiple nucleolar protein (*nol*, *nop*) genes. These are connected in many cases for both strains of toxoplasma.

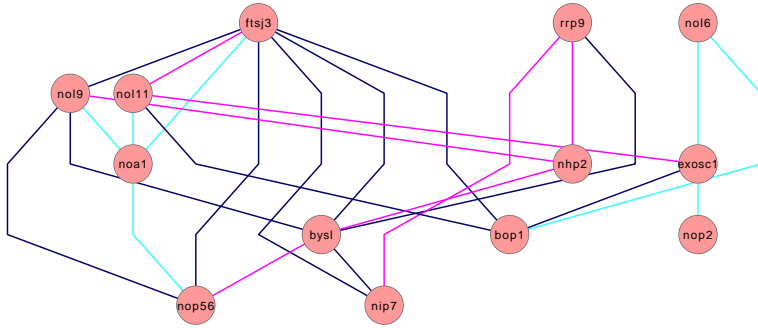


Figure 5.10: Ribosome Biogenesis subnetwork, that shows that genes with this Gene Ontology are present in both the ME49 (turquoise), RH strain (pink) and multiple genes connected under both strains (blue).

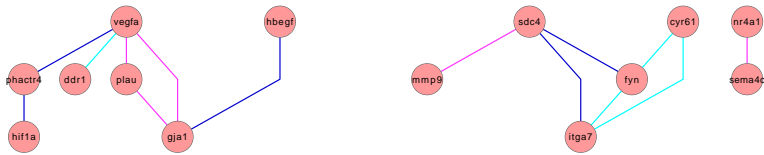


Figure 5.11: Cell Motility networks that are present for both the ME49 and RH strain. This shows higher connectivity for *Vegfa* in the RH strain and for *Fyn* in the ME49 strain.

As we know that an important part of the toxoplasma pathology is its ability to cross multiple cell and tissue barriers we looked at the Cell motility ontology network, Figure 5.11. There are several genes whose importance in *T. gondii* infection is well documented, these include *Vegfa* and *Hif1a* which are connected in both the ME49 and RH strains. Though we note that the *Vegfa* gene has higher connectivity for the RH strain and the proto-oncogene *Fyn* is similarly more connected in the ME49 cell motility network.

5.5.9.2 Disease networks

We similarly annotated the results with information from the Disease Ontology (DO) [Schriml et al., 2012]. As the disease ontology is based on human diseases and genes, here we assumed a certain level of homology of function between the same gene in different organisms and that these functions may be dis-regulated by the disease with which they are associated. The Disease Ontology uses geneRIF annotations to derive the mapping between genes and diseases <http://www.ncbi.nlm.nih.gov/gene/about-generif>. By annotating the network with the DO we are able to identify genes that have been associated with a disease. The coverage of the DO is not complete across the genome for our 791 genes: 200 genes have at least one disease associated with them, and many of them have more than one. There were 292 disease terms found within the genes of our network in total. From this information, we identified those disease terms that were attached to at least 10 genes in our network. This gave 16 diseases as shown in Table 5.4, which also shows there are substantially more genes associated with Cancer and its specialised subsets than the other diseases.

Given the relatively large number of annotation with Cancer term shown in Table 5.4 we focus on annotating the network according to this ontology term. As an overview, we first considered the full network of 791 genes. From this we saw that the network hubs contain the majority of the genes associated with cancer: these are nodes that are coloured green Figure 5.12. From here we extracted the 63 genes that are associated with cancer and those edges between them.

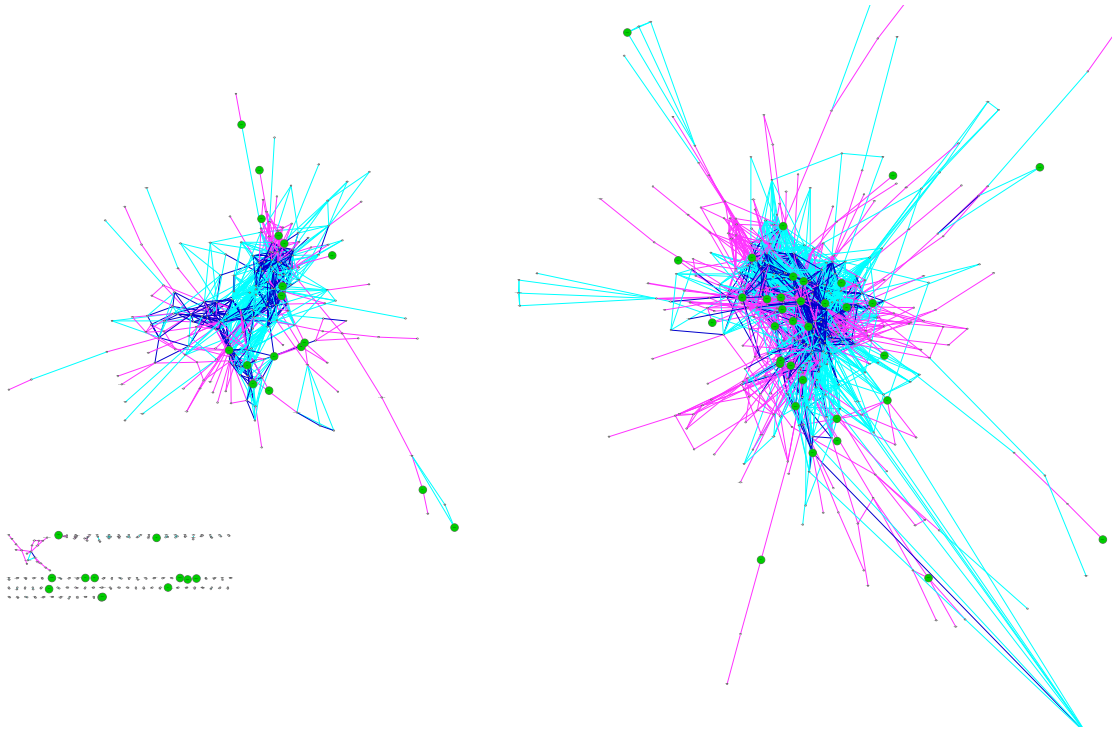
We created subnetworks associated with the cancer terms by extracting the inferred network from the JGL algorithm for only those genes annotated with the selected cancer term. Creating subnetworks of these cancer term genes we saw that there were three connected subnetworks within the larger hubs, Figure 5.13. Two of the genes within these subnetworks are known master regulators *Myc* [Dang, 2013] and Ras Homolog Family Member C (*Rhoc*) [Rajalingam et al., 2007] which are at the top of the hierarchical subnetworks. This is consistent

geneRIF: a concise annotation associating a gene with a function. geneRIF entries are required to be short and with an associated publication that details the functional annotation, thus ensuring the quality of the disease relationship.

Disease	Number genes
Atherosclerosis	16
Breast cancer	30
Cancer	63
Colon cancer	15
Diabetes mellitus	20
Embryoma	16
Endometriosis	11
Heart failure	11
Leukemia	15
Liver cancer	12
Lung cancer	12
Obesity	11
Polyarthritis	13
Prostate cancer	17
Rheumatoid arthritis	19
Schizophrenia	10

Table 5.4: Disease terms present in the *Toxoplasma gondii* network, showing the number of genes associated with each term. This is shown for only those disease terms with at least 10 genes associated with them.

with these genes having a higher degree (number of connections) to other genes, being central to the regulatory response of the host to toxoplasma infection. We also saw ligands to well-known inflammatory pathways: EGFR, *ereg*, and members of the Akt pathway, *Akt3* as well as previously noted genes, transcription factor *Jun*, growth factor *Vegfa* and signalling kinase *Jak2*.



To highlight the benefit of network inference over differential expression analysis, we identified interesting genes from our highly interconnected Cancer network Figure 5.13. The network view allowed us to select central nodes that are shown by the inference to be integral parts of the host response. These central genes have high edge degree and in contrast to the many other genes that are connected on the periphery of the network. Central genes are expected to have a more influential role the host response. The second advantage of the network view is that we could identify pathway mechanisms. Pathways can be represented by identified edges between genes and therefore cannot be identified through lists of differential expressions alone. Similarly, novel pathway connections cannot be hypothesised from ontological over representation of known gene sets.

Figure 5.12: The EB network with edges coloured according to the classes they are present in. The size and colour of the nodes is also mapped according to whether or not the genes is one of the 63 genes associated with cancer in the disease ontology. These 63 genes are shown as larger green nodes.

5.5.9.3 *Tribbles pseudokinase-3 (Trib3), a tumour repressor*

From Figure 5.13 we can clearly see that *Trib3* is highly connected particularly for the ME49 strain, this makes it a potentially interesting gene within this subnetwork as it may provide insight into the different mechanisms between the parasite strains. Although *Trib3* does not appear to have been researched with respect to toxoplasma it has been associated with malaria infection [Albuquerque et al., 2009]. Interestingly although *Trib3* was consistently shown as differentially expressed over multiple time points of infection, the authors note

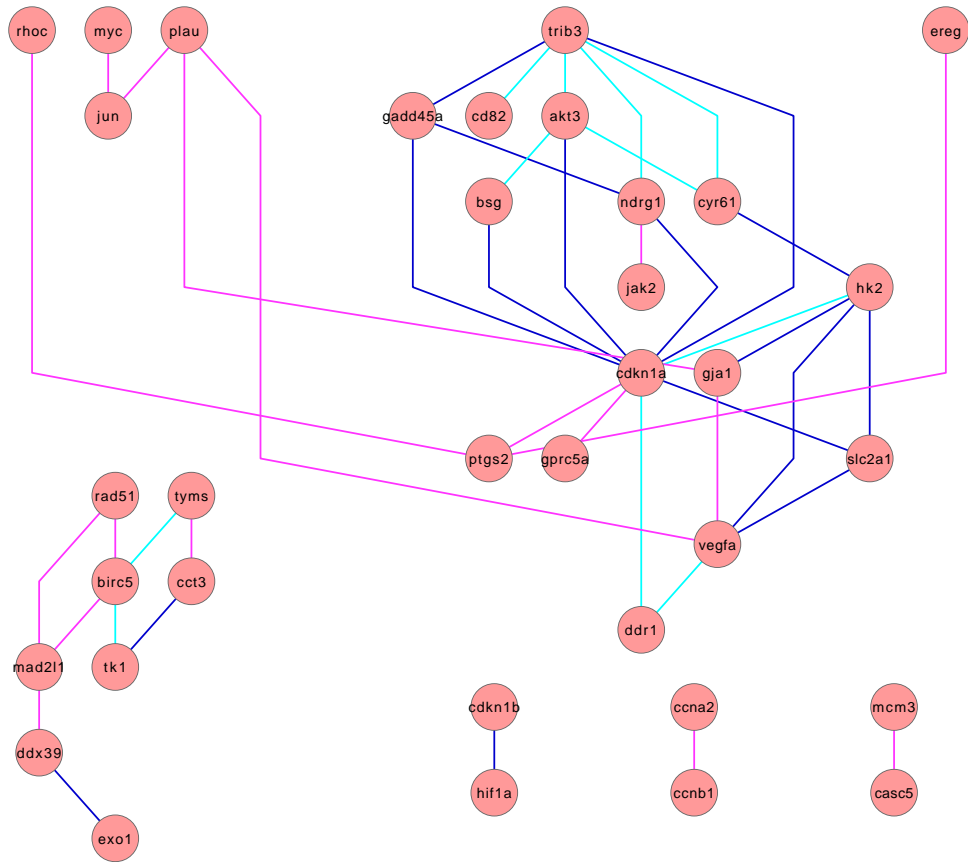


Figure 5.13: Subnetwork containing of genes associated with cancer. Turquoise for edges only in the ME49 strain, pink for edges only in the RH strain and blue for edges found in both strains.

that they were not able to identify a pathway mechanism from the list of differentially expressed genes. This highlights the difficulty in generating hypotheses of novel pathways from lists of differentially expressed genes.

Analysis of *Trib3* Knockouts in mice has shown that *Trib3* acts as a tumour repressor through AKT and the inactivation of the FOXO3 transcription factor [Salazar et al., 2015]. The paper by Salazar *et. al* included microarray analysis comparing wild type (WT) and *Trib3* Knockouts in MEF cells. As the cell type is the same as for our dataset we looked for supporting differential expression data for the connections found in our network. The paper identified the importance of *Trib3*, *Akt* and *Foxo3* therefore, we concentrated on these three genes and those genes that are connected to them in our network. The available data in ArrayExpress contains expression for 20,883 Illumina probes. When converting to gene names not all our genes were present in the data downloaded from ArrayExpress. Of those that were, 34 connections were significantly differentially expressed between the wild-type and *Trib3* KO at the 5% level and 17 were not. We show the network for *Trib3*, *Akt3* and *Foxo3* (highlighted as yellow nodes) and all genes connected directly to them in Figure 5.14. We sized the nodes in the network according to the adjusted p-value of their differential expression between *Trib3* Knockout (KO) and WT MEF cells. The nodes are also coloured in either purple or green, those with adjusted p-values below 5% are green, those above in purple. In this case, we do not use the hierarchical layout for the graph because with a large number of nodes the network image physically takes up more space and makes it hard to read, whereas a dynamic network in Cytoscape can be interactively navigated.

The direct edge between *Trib3* and *Akt3* and many of the connections around *Trib3* and *Akt3* in Figure 5.13 are turquoise: this shows that these edges were found in the ME49 network alone. As we showed earlier, using the EB method had the most noticeable effect on the summary results of the network. We therefore checked the connection of these genes in the network model using the original correlation matrix. Without using the EB correction *Trib3* is only connected to two genes in the whole network (*Pomgnt1* and *Rell1*) whilst *Akt3* is only connected to *Nifk*, *Rbms1* and *Fndc3a*. This shows that the EB method has enabled us to find interesting connections not identified from the original analysis.

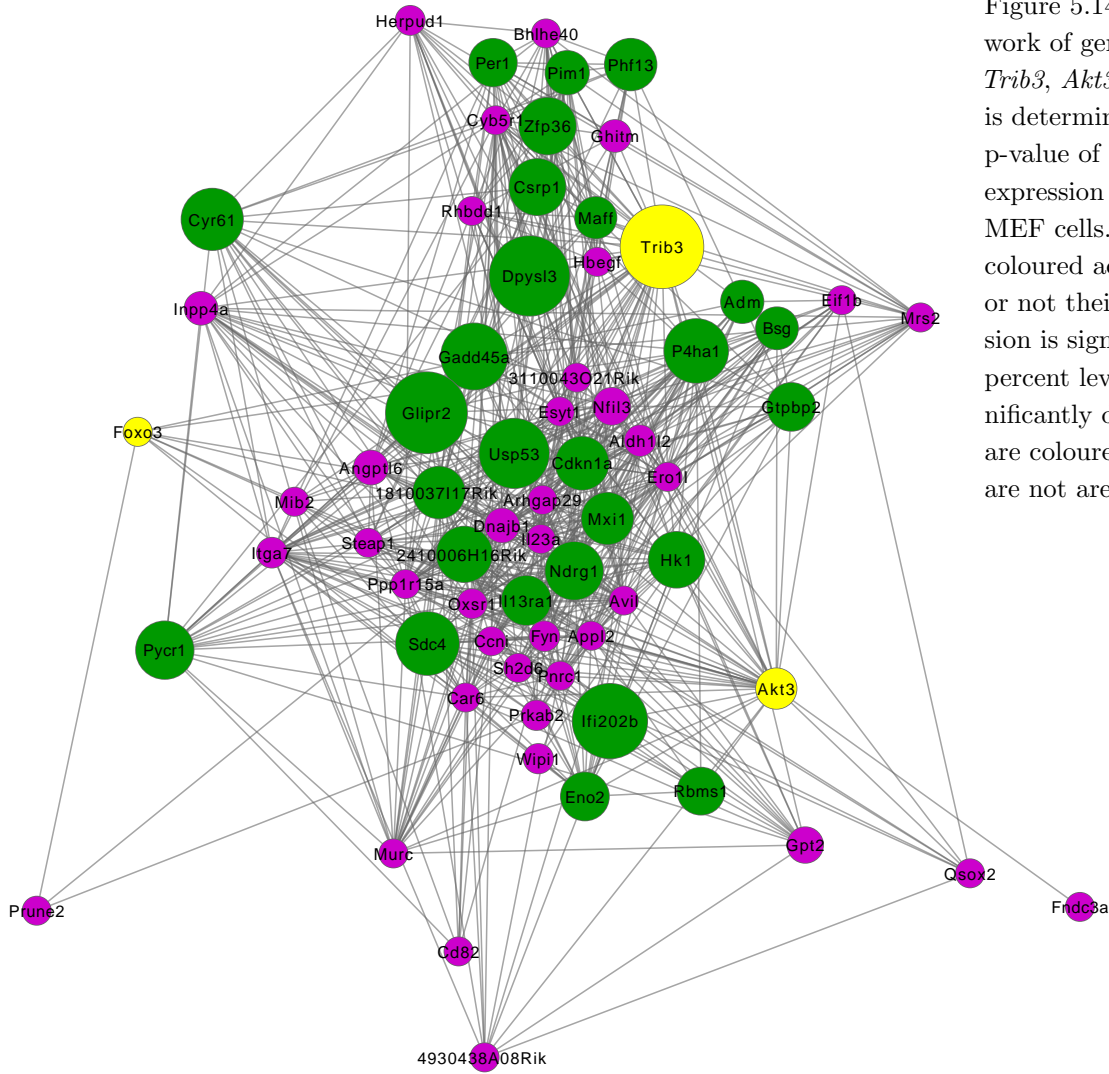


Figure 5.14: Example subnetwork of genes connected to *Trib3*, *Akt3* or *Foxo3*. Node size is determined by the adjusted p-value of the nodes differential expression in WT and *Trib3* KO MEF cells. The nodes are also coloured according to whether or not their differential expression is significant at the five percent level. Nodes that are significantly differentially expressed are coloured green those that are not are coloured purple.

5.5.9.4 *Cdkn1a*

As well as *Trib3*, Figure 5.13 shows a relatively high number of connections in the network for gene *Cdk1na*. This gene codes the protein P21, a known inhibitor of cyclin-dependent kinases (CDK) 1 and 2. P21 is a regulator of the cell cycle, and can restrict cell growth [Abbas and Dutta, 2009]. Whilst its activity is well known to be regulated by the tumour suppressor P53 [Anttila et al., 1999], it is also present in other signalling pathways including the ERBB signalling pathway. Figure 5.15 shows part of this pathway from KEGG with a part of it selected and enlarged as shown by the blue box. In this highlighted section, we see P21, as well as PI3K-AKT signalling [Lu et al., 2006], and consistent with this, in our network *Cdk1na* is connected to *Akt3*. Further, the protein encoded by *ereg* in our network is a ligand to the EGFR and ERBB4 receptor, the latter of these being another member of the highlighted ERBB signalling pathway. The part of the KEGG pathway including ERBB4 is similarly activated by ERBB2, a close homolog of *ereg*. Further in the second KEGG pathway we have highlighted we can see that *Myc* another gene in the connected cancer term sub-network can also be regulated by ERBB2. Although we were unable to identify suitable knockout experiments to verify these connections, taken together the gene descriptors and pathways information from KEGG provided some insight into the potential activity of the ERBB signalling pathway in the host response to toxoplasma. A subsequent literature search confirmed that the closely related EGF pathway has been shown to be recruited and activated by the parasite to prevent its relocation to the lysosome where the parasite can be destroyed [Muniz-Feliciano et al., 2013].

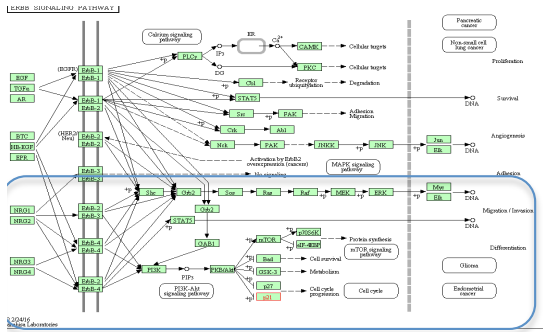
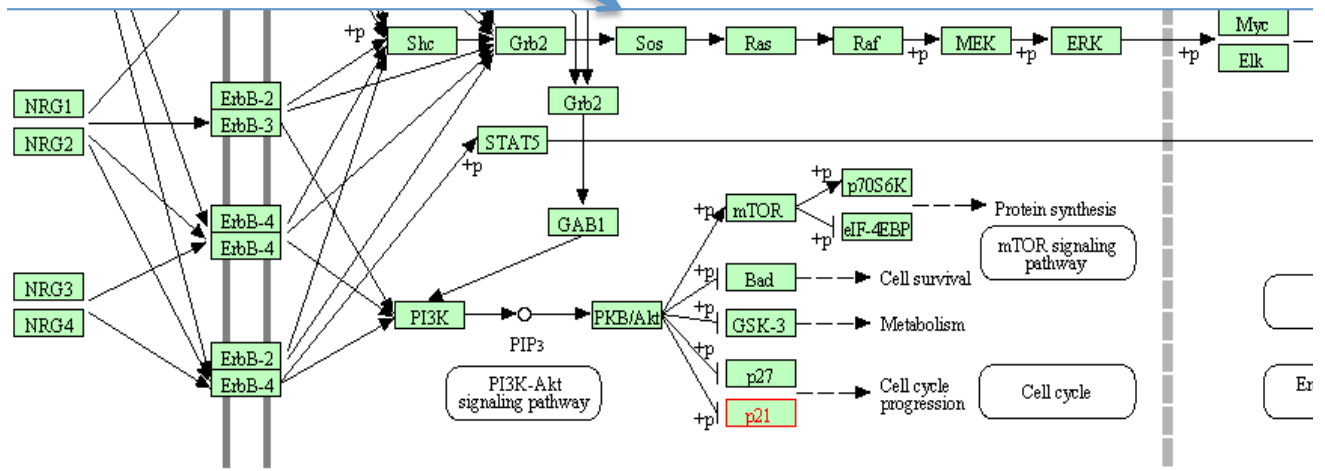


Figure 5.15: A section of the ERBB signalling pathway from KEGG. This includes P21, AKT, ERBB4 and MYC; genes and proteins found connected in our network through their shared cancer Disease ontology terms.



5.5.9.5 *G protein coupled receptor, Class c, 5a, (Gprc5a): a role in T. gondii infection?*

At the other end of the network we see *Gprc5a*, shown in Figure 5.13, connected in this sub network to *Cdkn1a*. *Gprc5a* has recently been associated with multiple diseases including carcinomas in several different tissues and chronic obstructive pulmonary disease, though its mechanisms are relatively unknown [Zhou and Rigoutsos, 2014]. The paper by Zhou *et. al* reported potential transcription factor binding sites for *Gprc5a* identified by the ENCODE project using the recognition sequence for *Gprc5a* compared to the transcription factor database predictions of JASPAR. Sites were identified for the following genes: *Jun/Fos*, *Brac1*, *Jun*, *Myc/Max*, *p53*, *Creb1*, *Fos* and *Rar/RxR* with multiple potential binding sites for some of these genes. We identified those genes connected directly to *Gprc5a* in our network, Figure 5.16. By comparing our network result to the known binding sites, we saw a connection to one of the four members of the *Fos* family, *Fosl1*. A search of the remaining network for *Fos* found that it is also connected to *Gprc5a* through *Aldh1l2*. *Fosb* also shares edges with *Gprc5a* between 8 genes, (*Rhbdd1*, *Setx*, *Tuft1*, *Ghitm*, *Socs3*, *Maff*, *Cdkn1a*, *Hbegf*). This suggests that response to toxoplasma infection in MEF cells may involve activation of *Gprc5a* through *Fos* binding. This is highlighted by the increased connectivity between *Gprc5a* and the *Fos* family of genes in our network in comparison to, for example, *Myc* and *Jun* that are also both present in the network. Previous research has also shown that *Gprc5a* knockout mice results in *Stat3* activation, and *Gprc5a* knockout mice are more resistant to apoptosis, consistent with *Gprc5a* having a role in tumour suppression [Deng *et al.*, 2010]. Furthermore it has been shown that the tumour suppressor effects of *Gprc5a* are due to *Stat3* repression, which is modulated through *Socs3* [Chen *et al.*, 2010]. Consistent with this, we found (Figure 5.16) that *Gprc5a* is directly connected to *Socs3* in our network. We can also see from Figure 5.16 that the interplay between these genes, and all those directly connected to *Gprc5a* are present in the RH strain. There are a few blue edges that indicate the interaction is present in both the ME49 and RH networks but the *Gprc5a* subnetwork seems to be a host response primarily to the RH strain.

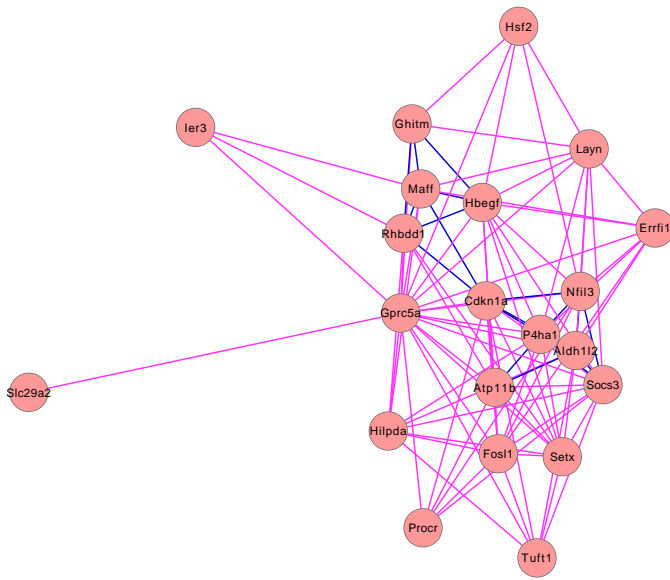


Figure 5.16: Example subnetwork containing genes connected to *Gprc5a* a G protein receptor associated with cancer. Interesting connections are to *Fosl1* a member of the *Fos* family and *Socs3*. The pink edges indicate an interaction following infection by the RH strain. Blue edges represent a few connections present for both the ME49 and RH strains.

5.6 Conclusions

The genes found in our networks show enrichment for biological processes relating to the six hallmarks of cancer, supporting our hypothesis that the *T. gondii* parasite behaves as tumour cells [Lun et al., 2015]. By combining existing annotations with the data driven JGL analysis we could identify known and novel genes and functions. From the Disease Ontology, we found significant over representation for cancer disease within our network. Given that cancer is widely studied, more so than toxoplasma it is not surprising there are more annotations for cancer than toxoplasma. We used existing knowledge on tumour progression to find supporting evidence for our hypothesis that the parasite behaves as a cancer in our network. Further prior knowledge on the physical interaction of proteins in other model systems can be used as supporting evidence for the connections in our network.

We found multiple significant Gene Ontology terms relating to the cell cycle, including cell proliferation and cell differentiation, consistent with the six hallmarks of cancer. From our list of significant Gene Ontology terms there are cellular components, *mitochondrion* and *ribosome* with associated biological processes *mitochondrion organization*, *ribosome biogenesis* and *ribonucleoprotein complex assembly*. It is known that, as well as disrupting the normal cell cycle regulation, oncogenes also increase ribosome biogenesis [van Sluis and

McStay, 2014]. Ribosome biogenesis refers to the biological process of creating ribosomes, the machinery responsible for translating mRNA into protein. Therefore, ribosome biogenesis is strongly linked to cell cycle processes, without ribosomes to translate mRNA, proteins would not be produced and cell division would not be possible. The increased transcription of ribosome DNA (rDNA) in the nucleoli, results in structurally modified nucleoli, and these visual differences between normal and cancer cells have been used to predict clinical outcomes in cancer patients. Moreover, the discovery that existing cancer drugs can inhibit ribosome biogenesis, though this was not their original purpose, has led to the specific development of cancer therapies designed to down-regulate ribosome biogenesis. Some of these studies have shown promising results, being able to selectively target rDNA while restricting off-target DNA damage [van Sluis and McStay, 2014]. Associated with this, we observe significant Gene Ontology terms for genes involved in anatomical structure formation involved in morphogenesis, the determination of shape.

From the network results, *Vegfa* (vascular endothelial growth factor- α) a pro-angiogenic factor is shown as a centrally connected gene in the cell motility subnetwork extracted from our network. This would be expected given its importance in angiogenesis, a hallmark of cancer, and the well-documented role of VEGF α in *T. gondii* infection. In terms of pathways we saw enrichment for multiple inflammatory pathways which would be expected as we know toxoplasma causes a strong inflammatory response in the host [Miller et al., 2009b].

Nitrogen metabolic processes that produce ROS also appeared in the gene ontology of multiple genes in the network. Previous work has identified proteins within the parasite that enable *T. gondii* to subvert the reactive nitrogen intermediates produced by the host in response to infection [Skariah et al., 2012] that can otherwise kill the tachyzoite form of the parasite by restricting the availability of mitochondrial and nuclear enzymes that are essential to the parasite. The difference between correlation and causation for ROS is however currently unclear [Gasparre et al., 2013, Chatterjee et al., 2011]. Nitric oxide a prevalent reactive nitrogen species intermediary acts together with ROS to damage cells. Nitric oxide can activate Hypoxia Inducible Factor 1 α (HIF1 α) under normal oxygen conditions. The increased accumulation of HIF1 α resulted in an increase in expression of HIF1 α targets [Spear et al., 2006].

HIF1 α is a well-known regulator of the inflammatory response, particularly the stress response to reduced oxygen in the cells. It is also a known contributor both to cancer and toxoplasma progression it was present in our network and has appeared in many of the selected subnetworks indicating its importance in the host response

to toxoplasma infection [Nizet and Johnson, 2009]. As a transcriptional regulator, pathways involved in the HIF1 α response include metabolism [Vander Heiden et al., 2009], cell differentiation and cell death [Nizet and Johnson, 2009]. Therefore, the presence of *Hif1 α* in our network presents several potential mechanisms of action in *T. gondii* infection.

The analysis of KEGG pathways in our network showed a clear enrichment of metabolic processes including glycolysis and oxidative phosphorylation. Therefore, where transcriptional regulators such as HIF1 α are associated with multiple pathways, our results suggest that its role in *T. gondii* infection is metabolic. There were both commonalities and differences in the metabolic genes present for the two strains, RH and ME49. In many cases, these genes have a documented function in both cancer and *T. gondii* infection. The same mechanisms of action of these genes in the two phenotypes, cancer and *T. gondii* infection, would support our hypothesis that the parasite subverts host function as tumorigenesis does.

For *Toxoplasma gondii*, two models have been proposed to explain the dependence of parasite survival on HIF1 activation. These two models are not necessarily mutually exclusive. In the first model, the use of oxygen by the parasite leads to hypoxic conditions that result in HIF1 α expression. In the second, reactive oxygen species produced by the parasite that can induce HIF1 α are responsible. In either case, the parasite was shown to be dependent on HIF1 α activation to survive in low oxygen conditions. Both the parasite and host are capable of producing ROS intermediaries. As ROS is produced during oxygen consumption is it possible that the first model in fact links to the second [Spear et al., 2006]. Regulating ROS is important for parasite survival. Mice with increased ROS production were completely resistant to infection by toxoplasma [Arsenijevic et al., 2000].

The mechanisms by which HIF1 α influences *T. gondii* survival mirrors results found in cancerous cells. In cancer cells HIF1 has also been shown to act co-operatively with dysregulated MYC, and is known to repress the function of MYC during hypoxia (low oxygen). MYC is responsible for driving cell proliferation and growth. MYC and HIF1 were found to increase anaerobic metabolism in cancer cells through activation of *Hk2*, a HIF1 target, and activating VEGF. This resulted in energy and lactate production [Kim et al., 2007]. HK2 is responsible for the first reaction in converting glucose to energy through glycolysis. Therefore, its activation is synonymous with activation of glycolysis.

In *T. gondii* infection siRNA knockout of HK2 resulted in further reduction in parasite growth and low oxygen levels <3% compared to

oxygen at 21% [Menendez et al., 2015]. HK2 expression and relocation to the host cytoplasm was identified as an important mechanism for growth and survival of the toxoplasma parasite using the RH strain, thus further supporting the high degree of connectivity of *Hk2* in our network. As *Hk2* is similarly highly connected for the ME49 strain it is possible this mechanism of survival is important for both strains.

Akt3 and *Myc* are two other known oncogenes that influence metabolism found in our cancer sub network. AKT has been shown to induce aerobic glycolysis in leukaemia and glioblastoma cell lines, but not increase oxidative phosphorylation thereby providing further evidence of the Warburg effect [Elstrom et al., 2004]. Interestingly, *Akt2* overexpression resulted in a switch in established glioblastoma cells to aerobic glycolysis and glucose dependence but showed no effects on cell proliferation. These results suggest that in cancer cells AKT2 mainly functions to alter metabolic pathways rather than its ability to impact host cell growth or proliferation [Cheng et al., 1997].

Akt3 was connected multiple genes for both parasite strains including *Eno2* and *Cdkn1a*, with several additional connections only in the ME49 strain. The connection of *Akt3* to glycolysis pathway gene *Eno2* and the enrichment of metabolic pathways in our network similarly suggests that in *T. gondii* infection the mechanism of action of *Akt3* is through metabolic pathways with downstream effects on survival and differentiation. AKT has been shown to impact the ability of toxoplasma to differentiate from tachyzoites to bradyzoites. Infection of different cells by the type II Pru strain showed different levels of differentiation. Lactate levels were found to be a discriminatory factor in cells being either resistant or permissive to conversion to bradyzoites. Further, it was shown that altering levels of lactate through increased AKT expression altered the state of the cells, converting from a permissive to resistant state. The production of lactate was necessary to alter differentiation state; anaerobic glycolysis was required as an increase of glucose alone did not change differentiation ability [Weilhammer et al., 2012]. This is consistent with results in tumour cells where knockdown of LDH-A the lactate dehydrogenase, resulted in increased mitochondrial respiration and inhibited cell proliferation under hypoxic conditions. In our network the only lactate dehydrogenase was connected in the RH strain. If we consider the tachyzoite stage of the parasite as being analogous to proliferating tumour cells, then we might expect the RH strain to show higher levels of cell proliferation; the RH strain is more virulent than ME49 and is unable to convert to the slower replicating bradyzoite stage of the parasite [Liou and Storz, 2010]. From this perspective, inhibited cell proliferation following LDH-A knockdown in tumour cells would seem consistent with AKT expression and lactate levels causing resistance to conversion to

bradyzoites [Fantin et al., 2006]. Therefore, the connection of the RH strain only to LDH would suggest a mechanism by which tachyzoite conversion is inhibited for the type I strains in comparison to the type II strains.

Whilst our results show changes in host metabolism for both strains they also highlight differences in the mechanisms through which metabolism is affected [Molestina et al., 2008]. Previous work showed that parasite metabolism is different between the ME49 and RH strain; ME49 relying on glycolysis whilst RH utilised both glycolysis and oxidative phosphorylation. Our results similarly showed changes in glycolysis for the host cells infected with ME49 strain whilst the network for the RH strain included genes involved in both glycolysis and oxidative phosphorylation. This suggests that the parasite infection subverts the host metabolic pathways in a manner consistent with its own metabolism. The network structure of our results provides further information on the differences between the strains impact on metabolism. The connectivity of the RH strain is more profound as illustrated by the greater number of genes and connections for the RH network. For the RH strain only, *Myc* and *Jun* were connected to glycolysis gene (*pgm1*).

From a metabolic perspective, the connection of *Myc* and *Jun* to metabolism genes in our network suggests a potential mechanism through which the RH strain utilises host glutamine metabolism. Both *Myc* and *Jun* have been shown to affect glutamine metabolism in tumours [Lukey et al., 2016, Gao, 2009], whilst glutamine metabolism ensured the survival of RH tachyzoites when glucose was unavailable [Blume et al., 2009]. The oncogene *Myc* has been found active in all three strains *Toxoplasma*. The authors attributed *Myc* expression to its roles in cell survival rather than metabolism. In contrast, our results suggest an influence on metabolism given its connections *pgm1*. Notably, we do not see a connection for *Myc* in the ME49 strain. This could be due to differences in the analysis methods as well as the different time points used. However, consistent with our results it has been shown that MYC is induced by tachyzoites and the authors suggested this affect may be mediated through c-JUN [Franco et al., 2014].

While experiments using the RH strain support results from our RH network they cannot be used to validate the lack of these connections for the ME49 strain. However, overall the results for both the ME49 and RH strains we have outlined here give multiple instances of regulatory mechanisms consistent with the parasite subverting host function through signalling proteins, metabolism and ROS production as seen in cancerous cells. Moreover, many new genes were found in our network that have no previous connection to *Toxoplasma* so provid-

ing new avenues to investigate further. This helps to direct research in a cost-effective method, narrowing potential gene targets without the expense of multiple experiments.

By annotating the network results using Disease Ontology terms we could identify connections within the larger networks that could be interesting for further analysis. These included the *Trib3* and *Gprc5a* subnetworks that were identified by their association with cancer processes. For the *Gprc5a* subnetwork we found a stronger connection between *Gprc5a* and *Socs3*, the STAT3 mediator, for the RH as opposed to the ME49 strain. This is consistent with the observation that the rhopty protein ROP16 can activate STAT3 for both the type I and type II strains but the continued phosphorylation is only present for the type I ROP16. This may explain why we observed a few edges in this network for both strains but considerably more for the RH (type I) strain [Hunter and Sibley, 2012].

The network contained multiple genes that share miRNA targets, lending greater confidence to the interactions identified between them. We were also able to find publicly available gene expression data of knockout experiments in MEF cells. We found such an experiment for the gene *Zfp36* in our network and we used this to provide additional support for the interactions found in our network. Combining differential expression under knockout experiments with the conditional dependence interactions between genes provided evidence of both direct and causal relationships between genes. The knockout data is more relevant for our network as opposed to a global network as we have included only a subset of the genes. This means that we have not included all possible explanatory factors, though our filtering means we have selected the most likely according to their expression and variability in the observed data.

The supporting differential expression analysis was restricted according to the amount of experimental data that was available. The focus of the analysis has been largely to introduce a pipeline that can be used to take a relatively small set of replicates, infer networks and extract useful subnetworks that can be used to drive experimental hypotheses. As we have specifically selected very stringent model selection criteria, we have not performed global network analysis as we have a relatively small network size. Similarly, while we have identified biological interesting results, the mechanisms by which the parasite subverts the host cell are complex involving multiple processes and pathways. We have focused on a subset of these processes that were highly connected within our network. The results we have shown are not exhaustive of our network or indeed the effect of *T. gondii* infection but have provided transcriptional connections such as for the *Zfp36* network, and hypotheses that could be experimentally tested, as

with *Gprc5a*.

As previously discussed, time and monetary constraints mean that most experiments have relatively small sample sizes. Whilst technologies such as gene expression microarrays and RNA-seq can measure expression on a genome-wide scale, to produce statistically significant interactions hundreds of samples are required for the thousands of genes. This is usually performed using meta-analysis, combining information from different samples and experiments. Conversely, small scale models that focus on a few (tens of genes) usually take samples from a single, smaller-scale experiment. Table 5.6 shows examples of previous work according to the size of networks (genes), number of samples, and whether or not this study was a meta-analysis.

No. samples	No. genes	No. Interactions	Meta-analysis	Ref
386	11,032	107,157	Yes	Hu, 2013
177	17,899	155,818	No	Bae, 2013
21	9	43	No	Bansal, 2006

Table 5.5: Table of different methods and sample sizes used to infer regulatory networks. This shows that networks with large numbers of genes and interactions require large sample sizes.

From a mathematical perspective one of the main results has been our ability to infer and validate regulatory networks where only a relatively small number of replicates have been used. To date, inference of networks at a genome-wide scale meta-analysis combining hundreds of samples from different experiments have been used. In contrast, we have reduced the parameter space and used empirical Bayes method to infer networks with small sample sizes.

It is not reasonable to expect to infer a genome-wide regulatory network from a smaller number of replicates in contrast to meta-analysis. However, we have shown that smaller relevant networks can be inferred. Moreover due to the data driven nature of the model, the genes included in the final network do not need to be known *a priori*, meaning there are no constraints on the elements that can be potentially included in the model. It was initially assumed that to run these JGL models targeting only a subset of the genes, many replicates would be required. Although this would increase the power of the model our results suggest that the number of replicates required is potentially fewer than initially thought with previous analysis suggesting small sample sizes (<15) can lack accuracy in network inference [Steele and Tucker, 2008]. The paper introducing the JGL model, had close to 90 replicates, where we had less than half that. However, it must be noted that their analysis used data from individual patients with tumours. These data sets would be expected to be more heterogeneous than cell-line laboratory replicates. By using biological replicates from cell cultures the data set we have used here provides a middle ground

combining the increased variation and interpretability of the biological replicates with the decreased variability and confounding factors in comparison to *in vivo* studies.

Methods that use meta-analysis include more genes and regulatory interactions over more conditions than our methods. We expect that contained within these networks, only some of the regulatory interactions will only be present in a subset of the conditions. Therefore, there is a contribution of noise where the networks are not active in those conditions plus signal where they are active. In comparison, our approach can be viewed as taking a subset of these conditions, and only inferring regulatory interactions active in these conditions. This results in a smaller number of genes, hundreds as opposed to thousands, in the model. The success of network inference on small samples sizes will critically rely on good experimental design to select comparable yet variable replicates. In our case using time series and varying the multiplicity of infection were intuitively compatible experimental factors to use with graphical inference.

As biological networks are scale-free, for any single condition or small number of closely related conditions, we expected that our network would show a few sparse networks. This was verified by our block analysis showing how the density of the network varies with different shrinkage parameters. This is also central to the empirical Bayes method that we used to pre-process the data. Bayesian formulations of the JGL model have also been developed that provides a Bayesian method for inferring the network structure or precision matrix as opposed to the correlation matrix. That means that the incorporation of the prior is based on the network structure. This allows for informative priors that are based on biological knowledge and support edges between known interactors. One example of this approach was used to infer metabolic networks involving 17 metabolites [Peterson et al., 2013] with 24 samples in total. The Bayesian formulation is likely to be considerably more computationally demanding. This is partially reflected in the smaller scale of this model. Further, to make use of an informative prior, arguably the scope of the model is reduced because biological knowledge to inform the pairwise interactions between all nodes, metabolites in this case, is a more detailed model formulation compared to a frequentist approach.

Although we were able to infer networks from this relatively small data set, we arguably required more pre-processing of the data in comparison to, for example, a large meta-study analysis. This included using existing methods to account for technical factors such as GC content. Further, we used filtering methods to select candidate genes for input into the JGL model. As well as an expression filter that is standard practice in all RNA-seq analysis, we further

selected those genes with highest variance across all samples. This is particularly important for an experiment with small sample sizes in comparison to a meta study as we would expect fewer genes to show an activation/inhibition pattern due to the reduced number and range of perturbations. As a final pre-processing step, we used the empirical Bayes method outlined previously to generate the correlation matrices for each of the strains separately. This is used to control the false discovery rate and is particularly applicable given the number of replicates we had for each condition. We found multiple connections between *Trib3* and other genes using the EB correlations that were not found using the Pearson correlation matrix. Given that we also found experimental and literature support for the connections to *Trib3*, this indicates that the EB correlation matrix enables us to find biologically meaningful connections not present using the sample Pearson's correlation matrix.

We have also seen that the control data in this experiment was not useful for inferring networks. Although there is correlation between genes in the control network, it is not specific. That is, there is no gradual level of significant correlations as with the infected cells. For a chosen level of significance, the model would either include or exclude all of the genes. Therefore, to include an uninfected network, there are two possible options. The first is to add the control data for only those genes included in the model, potentially for each strain separately to ensure that only those genes relevant to the strain data are inferred, as opposed to all genes that pass the filter step.

A second possible option would be to perturb the system to activate it, this also requires prior knowledge either before the strain inference or by using the first set of networks to drive the experimental design. For example, cytokines activate inflammatory pathways, including IFN- γ that is central to the host response to infection by *T. gondii*. In a previous study IFN- γ responses for infected and uninfected cells have been compared through differential expression analysis that identified interference with STAT1 as an explanatory variable for the observed responses [Laliberte and Carruthers, 2008]. Using a similar approach with differential network analysis would give a more specific control, which has the advantage of potentially being more informative on the activated pathway, but loses the scope of the uninfected cell. From an experimental design perspective, it may be sensible in future experiments to include controls that are wild type cells but under some stimulation to activate pathways. This would have two potential benefits: first, the analysis may provide further understanding of pathway hierarchies and elements within them; second, comparing a control network of this nature to one with an invading host parasite may also identify differences in the networks. Conclusions from this

type of experiment could potentially indicate not only those pathways that are activated by the invading parasite but any other ways in which they are altered by the parasite. This is potentially particularly applicable to *T. gondii* infection as one class of rhoptries secreted by the parasite are kinases or pseudokinases that can interact with the signalling proteins in the hosts immune response [Hunter and Sibley, 2012].

After the EB procedure we had two networks, one for each of the strains, that had similar levels of connectivity and scope. There were approximately half the edges shared by both networks. This amount of overlap suggests there are different mechanisms or dynamics by which these two strains of toxoplasma infect and subvert the host cells. Using differential expression analysis and testing for overrepresented pathway gene sets led to the observation of considerably fewer significantly differential expressed genes and pathways for the Type II compared to Type I strains [Xiao et al., 2011]. This is broadly consistent with our results showing sparser networks for the Type II ME49 network in comparison to the Type I RH strain. However, in comparison our analysis could identify novel connections and was not constrained by the pre-defined pathways. We also used partial correlation as opposed to differential expression to classify the network results. This allowed us to identify causal interactions between genes, and this arguably gives stronger support than finding differentially expressed genes contained within the same genes sets because differential expression analysis is performed for each gene independently.

In comparison to the multiple cofactor experiment conditions used with the *Bacillus subtilis* data set, it can be argued that these conditions are more readily interpreted once combined because depending on the stage of the toxoplasma infection, there could be consistent effects on a transcriptional unit. For example, for a process that develops over time we would expect to see a positive correlation of the genes involved across the time series. That is, we may expect more homogeneity in the time series samples of infection for *Toxoplasma* compared to the *Bacillus subtilis* data set in which the cells were perturbed under multiple different and, in particular, unrelated conditions. Note that both VEG and ME49 strains are likely to only see the sporozoite and tachyzoite form of the parasite within the 43 hours duration of this data set [Jerome et al., 1998, Skariah and Mordue, 2012]. This makes the RH data set more comparable over these time points as the RH strain lacks the ability to convert from tachyzoite to the chronically infectious bradyzoite form [Lun et al., 2015].

5.7 Methods

5.7.1 Aligning reads

Reads were aligned using STAR on default settings. There are many sequence alignment tools available, with Tophat being popular [Trapnell et al., 2009]. We investigated the use of Tophat to align our reads, however, the current version of Tophat was not able to process the larger combined annotation files that would be needed. The reference genomes for *Mus musculus* (mouse) and the *Toxoplasma gondii* strain were combined using STAR. For the RH strain, we used the available RH genome from toxo.db as well as the GT1 genome also from toxo.db. The RH and GT1 strains are both type I, and comparison of their full genomes revealed 1,394 single nucleotide polymorphisms (SNPs) or insertions/deletions. SNPs are as their name suggest, changes to the genome sequence at one single position, where the sequence could in principal vary to any other the other nucleotide base. In practice one observes that most common SNPs have two alleles - i.e. out of the four possible bases ACGT, two possibilities are observed when sequences from different strains are compared. Insertions/deletions are defined where the sequences match save for a single additional base (insertion) or a missing base in the sequence (deletion) as illustrated in the stylistic Figure 5.17. Figure 5.17 gives an example sequence, sequence 1, and shows example deletion, insertions and a SNP as highlighted in red.

ACCGTTAAGA	Sequence 1
AC_GTTAAGA	Deletion
ACCGTTAATGA	Insertion
AACGGTTAAGA	SNP

Figure 5.17: Stylistic example of the insertion, deletion and polymorphism events that can occur with RNA-seq reads and must be allowed for in sequence alignment programs.

5.7.2 Converting from reads to counts

The output from STAR is a SAM file that can be converted to a BAM file using SAMtools. These BAM files are input into RsubRead this combined the counts for the exons output from STAR into genes counts. The annotation file was created using Cuffcompare from the Cufflinks package. The annotation file contained the exon read sequences for both the mouse and toxoplasma strains where relevant. These sequences were matched to their transcripts from the annotations on toxo.db or ENSEMBL for mouse. There can be multiple

transcripts that map to genes in the organisms. RsubRead combines at the transcript not gene level as this is the information passed to it from the annotation files.

5.7.3 Comparing the mouse genome to joint mouse and Toxoplasma genomes

Using an uninfected sample the results were compared by using only the mouse genome and the combined mouse and toxoplasma GT1 genome. This is to assess the accuracy of the combined genome and the combined annotation file that is used to help the alignment algorithm map reads that may cover multiple exons. These genome directories are generated by STAR using the FASTA format of the genome sequences and the annotation file associated GTF annotation files. We expected that, by chance, a few reads, which were unmatched to the mouse genome may map to the GT1 genome. Indeed, we saw a small increase of mapped reads from 64.6% to 64.8%. The reads that were aligned to exons by STAR were then summarised to give gene counts using the featureCounts function in the Rsubread package. These counts were summarised for all reads that had a minimum mapping quality score at the default value of zero. This is a relaxed quality score, meaning that for some reads with low quality scores, mapping to multiple transcripts is allowed in the summarisation step. In the Rsubread package we do not count any reads that are mapped to more than one feature (gene). However, we allowed reads mapping to multiple transcripts within the same gene to contribute towards the final count for that gene.

We extracted the read counts for the expressed mouse genes: we defined expressed genes as those that have at least 10 matching reads per million total reads in at least one of the samples, and this gave 10,482 genes. We plotted the raw read counts for these 10,482 using each of the two genomes, Figure 5.7.3. This showed a high level of similarity between the read counts given for the two genomes as would be expected.

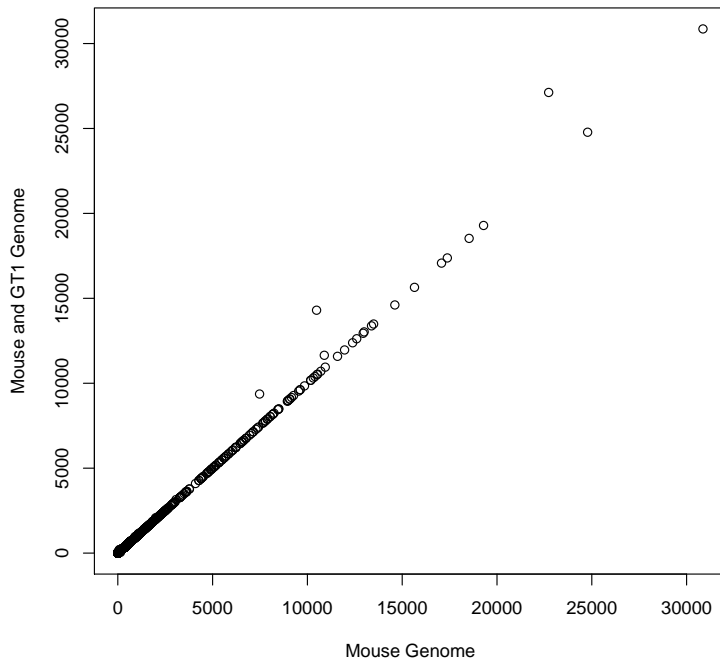


Figure 5.18: Plot comparing the raw read counts for an uninfected sample using just the Mouse genome and the Mouse GT1 combined genome. These are for genes that we defined as being expressed in the sample. This shows a good level of agreement between the two alignments.

5.7.4 Comparing the RH and GT1 strains

Although we were interested in modelling the reads from *Mus musculus* in the JGL model, because the infected samples also contained RNA from toxoplasma we wanted to be assured that no reads from toxoplasma have been mapped to the mouse genome. We had good coverage of the annotation for the ME49 strain, however the RH annotation is currently far from complete. Therefore, we used the GT1 strain to assess the accuracy of the mouse reads aligned for the RH strain. We anticipated that in using the GT1 alignment there will be an increase in the number of reads mapped but that these mainly will be previously unmapped reads (from the parasite) mapping to the GT1 genome. When summarising the counts for exon transcripts to give overall gene counts we increased the minimum read mapping quality score to 20 from the default value of zero in the Rsubread package. This had the effect of removing some reads that may have mapped to multiple exons in the same gene, by requiring a greater level of specificity through the higher mapping quality score.

Figure 5.19 shows the alignment counts for each of the RH infected samples where each sample has been mapped to a combined annotation of mouse with the RH strain and mouse with the GT1 strain.

Each of the bars is split according to the number of reads mapped to either the mouse genome or the relevant toxoplasma genome. As an overview of total reads mapped, Figure 5.19 shows no obvious difference in mappings to the mouse genome using either of the toxoplasma strains. There are, as would be expected, more reads mapped to the GT1 strain than the RH strain given the larger coverage of the GT1 strain in comparison to the RH strain. The count matrices from Rsub-read contain an additional 436 genes for the RH strain and 8136 for the GT1 strain.

The previous analysis considered the total read count across all genes. It is possible that the total read count for all genes is the same but varies for the individual genes. Therefore, using the mouse only data, we extracted those genes from *Mus musculus* genome using the read counts for the infected RH samples, this gives 45,309 genes. We used the RH data to select a subset of expressed genes for comparison purposes. Expressed was defined as having more than 10 reads per million in at least two of the samples and 13,195 genes passed this filtering step. For these genes, we plotted the raw read counts for each gene separately for the two different genomes, RH and GT1, for each sample. The results in Figure 5.20 showed that for each sample there is good agreement for the mouse genes using either the RH or the GT1 genome. There is a slight trend for higher read counts to the mouse genome for the RH strain compared to the GT1 strain (mainly evident from the higher values on the RH x-axis). This may be expected and suggested that, given the lack of a complete RH genome, some parasite RNA has mapped to the mouse genome as the best available alternative. Therefore, as we expected better accuracy from the GT1 genome, we continued the analysis using the GT1 mapping.

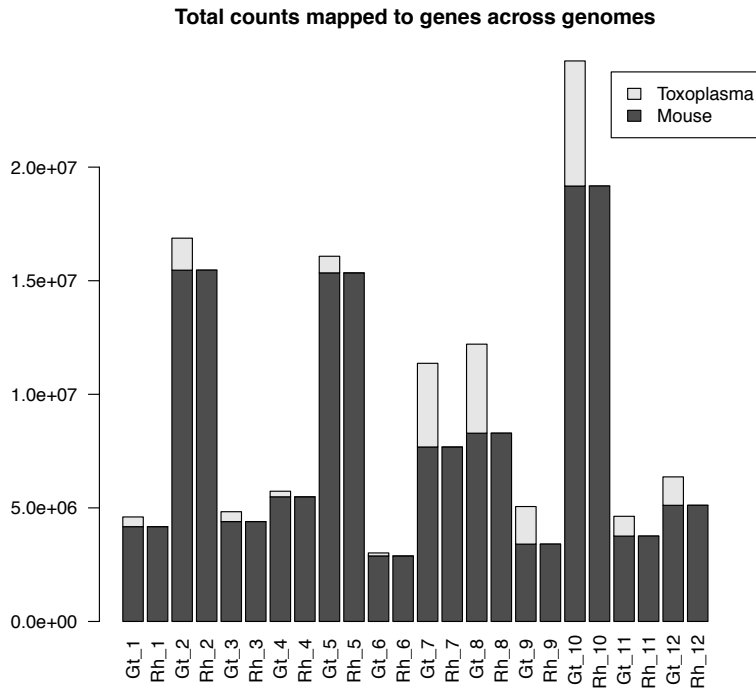


Figure 5.19: The Figures show the total read count for each of the 12 samples infected with RH strain. There are two alignments for each sample according to whether the alignment used the mouse combined with either the RH or GT1 strain. There is no obvious overall difference in the total number of reads mapped to the mouse genome under the two different annotations.

5.7.5 Comparing read counts to technical or biological factors

Before analysing the data further, we checked that there were no compounding technical or biological factors that needed to be accounted for in our model that would explain the different library sizes across the samples.

We plotted the number of reads for each sample according to the four experimental factors; Time, Strain, MOI and Lane, Figure 5.7.5. Comparing the different factors to the read count, plotted on the y-axis, there was no obvious pattern between the read counts for the samples and any of the biological (MOI,Time,Strain) or technical factors (Lane). The index in these plots is a naming convention used to label the different samples.

5.7.5.1 GC content and read length bias

Plotting the summary counts for each of the samples mapping to the mouse genome, we noticed that there are a relatively wide range of library sizes between samples, Figure 5.22. Relatively low coverage samples were seen across all three conditions, ME49 infected, RH infected and the uninfected samples, indicating this was not a result of experimental conditions or the sequence alignment. This could be

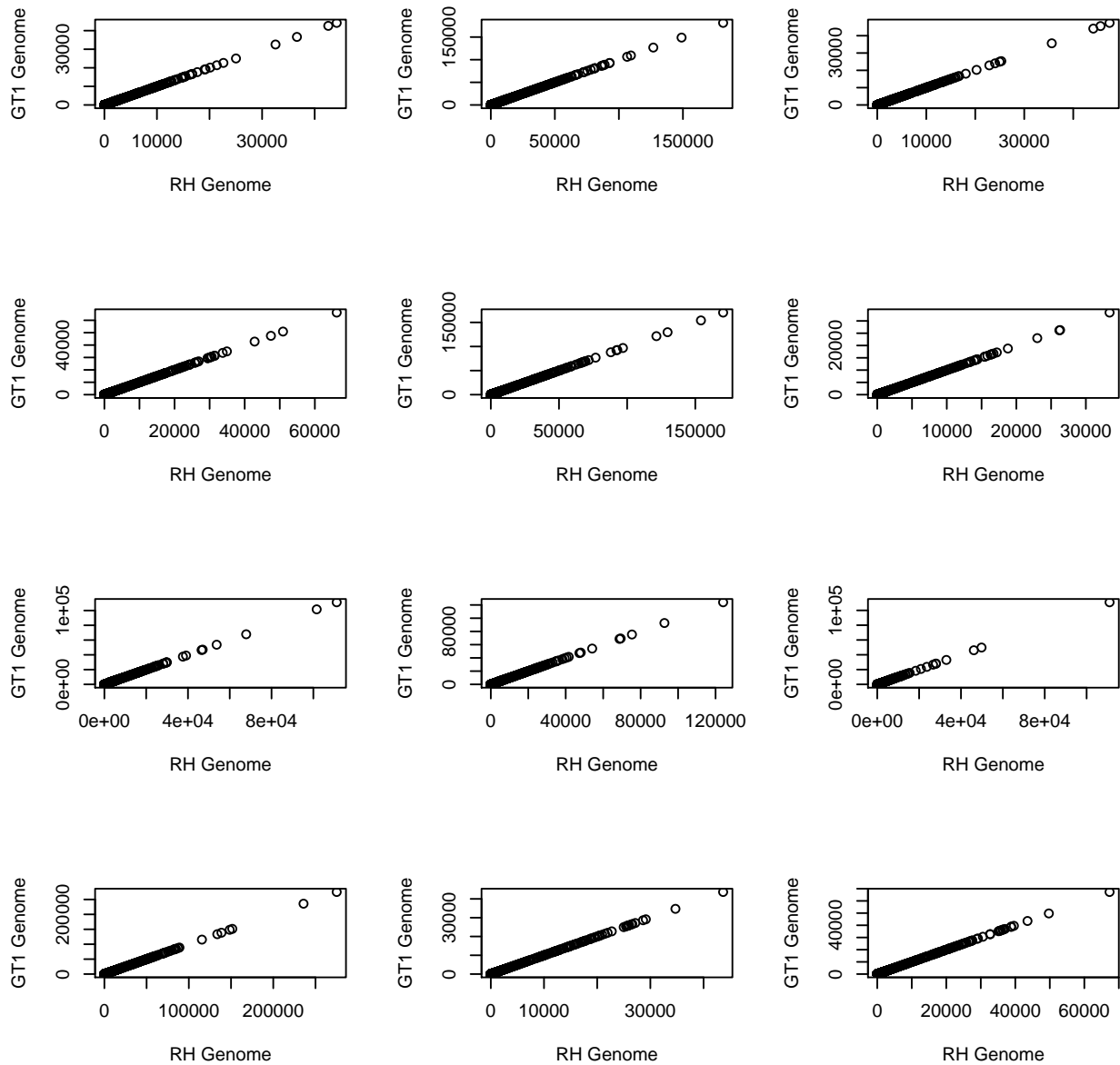


Figure 5.20: Total read count for each of the samples infected with RH strain according to their alignment; each RH infected sample has been mapped to either RH, or GT1 and combined mouse annotation. There is no obvious overall difference in the total number of reads mapped to the mouse genome under the two different annotations.

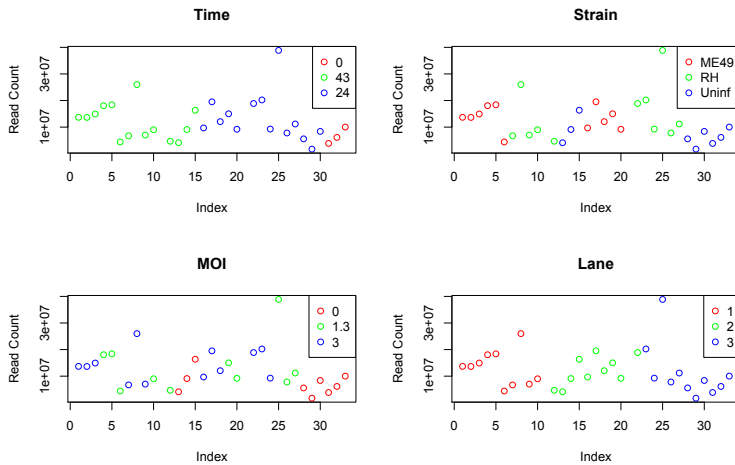


Figure 5.21: Total read count distribution for the individual samples split according to the possible biological or technical factors. There is no clear pattern between the read counts of each of the samples and the technical or biological factors.

due to a variety of technical factors, for instance, limited material or decreased Polymerase chain reaction (PCR) amplification for those samples. PCR is used to increase the amount of material available so that there is enough for sequencing. PCR involves multiple rounds of heating and cooling a preparation of DNA template, thermostable DNA polymerase and primers. In the first stage the preparation is heated to a temperature that separates the DNA strands. Once separated the preparation is cooled to a temperature that enables the primers and DNA polymerase to anneal to the separate DNA strands. The final stage of a PCR cycle involves heating the preparation to an intermediate temperature, which is optimal for the DNA polymerase and begins to create a copy of the DNA strands. These cycles are repeated allowing exponential amplification of the template DNA region flanked by primers.

It has been shown that RNA-seq counts can be biased due to the differing GC contents of genes [Benjamini and Speed, 2012]. GC content of genes is known to affect the PCR amplification rates [Mamedov et al., 2008]: for high GC content a higher temperature is needed to denature the DNA strands due to the increased thermostability of high GC content regions. There are three hydrogen bonds between a GC base pair in comparison to two for an AT pair. This additional hydrogen bond increases the stability of the gene and means higher temperatures are required to break these bonds. Low content GC content can also be biased in RNA-seq read counts due to the PCR amplification step. In this case, it is difficult for the low GC content genes to anneal to the PCR primers at higher temperatures. These PCR amplification biases are carried over to the read counts generated by alignment algorithms for RNA-seq experiments. The quality control

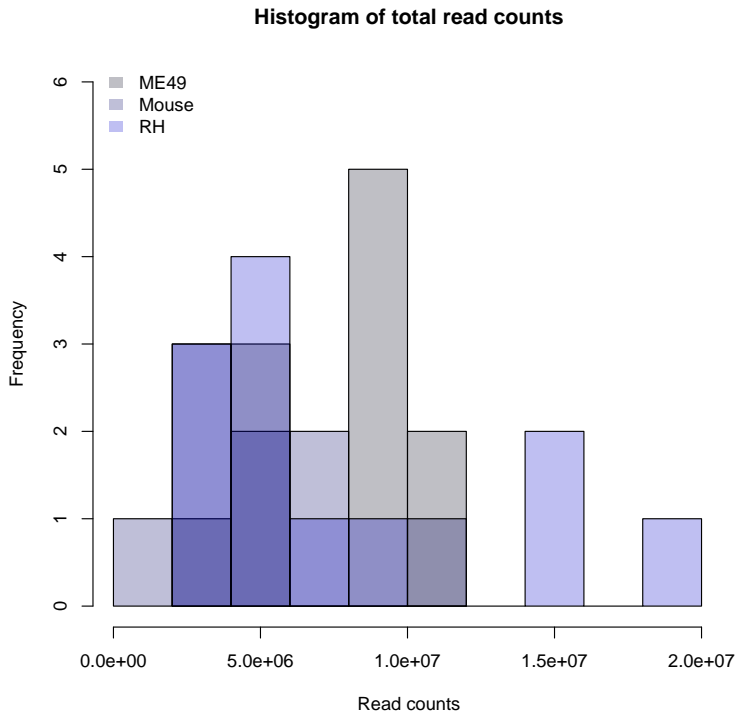


Figure 5.22: The Figures show the total read count distribution for the individual samples. This histogram shows no clear trend for the read counts to be split according to the different samples. That is, there are a similar range of read counts seen for the MEF infected with ME49 strain, RH strain and the uninfected cells (Mouse). The histograms for each Mouse, ME49 and RH strain overlap. Therefore, in some cases the colours in the histogram are the combined colour of those shown in the legend.

files, generated by fastqc, for each of the samples included plots of the theoretical and actual distribution of GC content across all reads. There were warnings of the differences between these distributions for all the samples. We used the pre-processing method of conditional quantile normalisation (cqn) to account for the impact of the GC content on the read counts [Hansen et al., 2012].

The cqn method estimates the effect of GC content on read counts using the GC content of the sequence as a covariate. This method also has the advantage of quantile normalising the overall library sizes. Using a quantile normalisation as opposed to a linear scaling of the different library sizes means that this method also considers differences in the shape of distribution of reads as well as the overall number of reads between samples. The output is expression values on a log base 2 scale which gives continuous, approximately normally distributed data that can be used as input to the JGL algorithm. The total read counts per gene output from the alignment algorithms are understandably effected by the length of the genes given the fixed length of the sequencing reads, in our case, 101bp. A common method for correcting this bias is to convert the reads per million to reads per kilobase per million mapped read (RPKM). However, we do not make this adjustment as the cqn method also takes gene lengths as a

covariate into the model and adjusts the final read counts for different gene lengths whilst also correcting for GC bias. The GC content and gene length covariates are calculated directly from the FASTA and Gene Transfer Format (GTF) annotation files.

We looked at the effects of the cqn correction by comparing the MA plots for the samples before and after the normalisation. MA plots are the average expression over replicates plotted on the x-axis against the average log fold change between two groups on the y-axis. The standard RPKM values are shown on the left hand plots, for the ME49 strain (Figure 5.23) and RH strain (Figure 5.24). These plots show that there is a baseline level of expression for the standard read counts and this can overestimate significant differences (large absolute M values) that would lead to false positives in downstream differential expression analysis. In comparison, the cqn MA plots on the right show an elliptical shape that is common with Gaussian variables. The MA plots are calculated for the ME49 compared to uninfected and RH strain compared to uninfected cells.

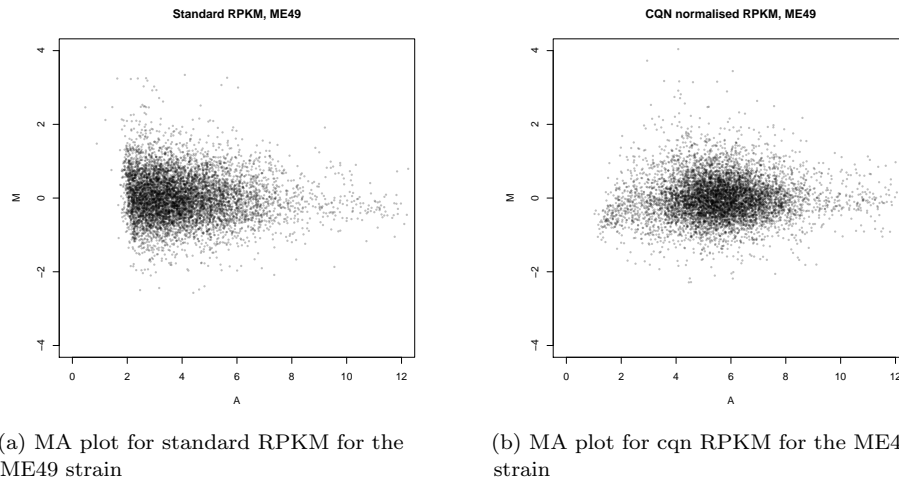
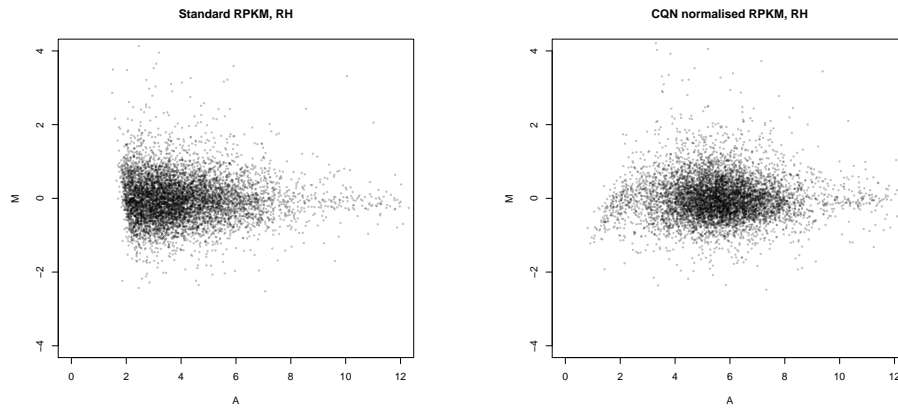


Figure 5.23: MA plots for ME49 strain before and after cqn. The cqn has removed the truncated low expression values. The distribution of the RPKM now has a central mass with a more symmetric distribution of read counts around it.

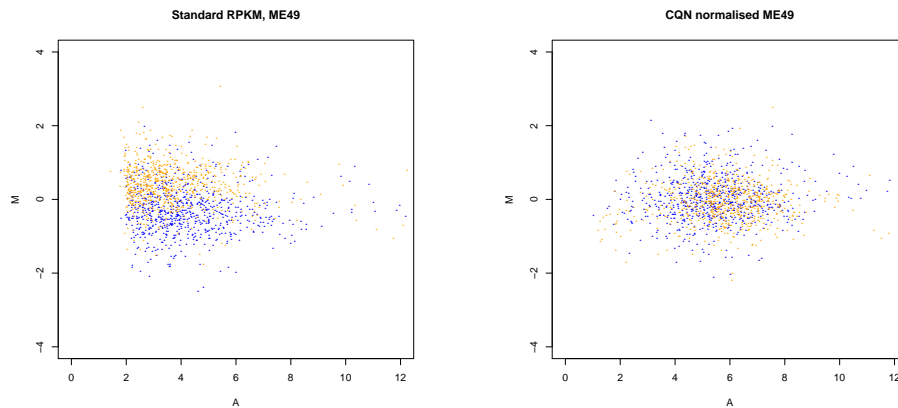


(a) MA plot for standard RPKM for the RH strain

(b) MA plot for cqn normalised RPKM for the RH strain

We also compared MA plots for a subset of genes that are at the extremes of the GC content distribution. This showed the impact of the cqn normalisation on correction for GC bias in the expression values. Low GC content genes are shown in yellow with high GC content genes in blue. The pre-normalisation MA plot clearly shows a split between the low and high GC content genes for both the RH and ME49 strains that is removed following the cqn normalisation shown in the right hand plots, Figures 5.25,5.26.

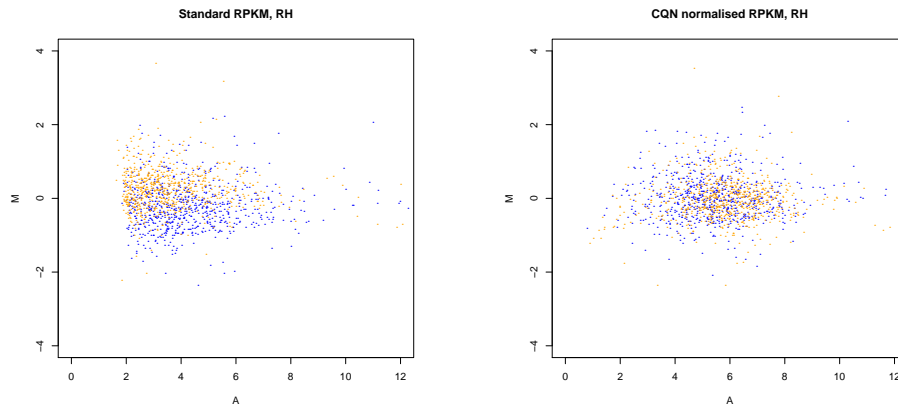
Figure 5.24: MA plots for RH strain before and after cqn. The cqn procedure removes the heavy tailed truncated distribution of read counts and results in a more symmetric distribution of read counts.



(a) MA plot for standard RPKM for the ME49 strain, yellow low GC content, blue high GC content

(b) MA plot for cqn RPKM for the ME49 strain yellow low GC content, blue high GC content

Figure 5.25: MA plots for ME49 strain before and after cqn. The cqn procedure removes the split of counts according to their GC content.



(a) MA plot for standard RPKM for the RH strain yellow low GC content, blue high GC content

(b) MA plot for cqn for the RH strain yellow low GC content, blue high GC content

Figure 5.26: MA plots for RH strain before and after cqn. Again the cqn procedure removes the split of counts according to their GC content.

5.7.6 Analysis of block size

For initial exploratory data analysis, we plotted the maximum block sizes for different shrinkage values for the three classes (ME49, RH and uninfected) individually and combined. To do this we iteratively used the first part of the JGL algorithm to calculate the block diagonal form of the covariance matrices. From these matrices, we were able to calculate the standard statistics of number of blocks and maximum block size. We calculated the block diagonal matrices for the data under different parameter values. The maximum block size is a useful tool for parameter selection particularly as it is used before the model is run. This gave us a method for model selection that does not require the time and computational expense of the suggested AIC statistic, or methods that calculate potential false positive rates on the edges, as we did not need to run the full model.

For the three classes of ME49 infected, RH infected and uninfected cells, we looked at the maximum block size under different parameters. We plotted the maximum block size for all three classes individually and combined over λ shrinkage values 0.7 – 0.95: Figure 5.27 shows that there are similar profiles for the RH and ME49 strains although they occur at different shrinkage levels. It is also clear from Figure 5.27 that whilst the infected cells have gone from a maximum block size of zero to including almost half the genes (at shrinkage value 0.9), for 0.9 and above, none of the correlations for the uninfected cells are significant, as the maximum block size is zero. The λ_1 shrinkage parameter selected was 0.91. Although 721 genes passed the screening process, the JGL algorithm found no significant partial correlations meaning there were no edges found between the genes in any of the

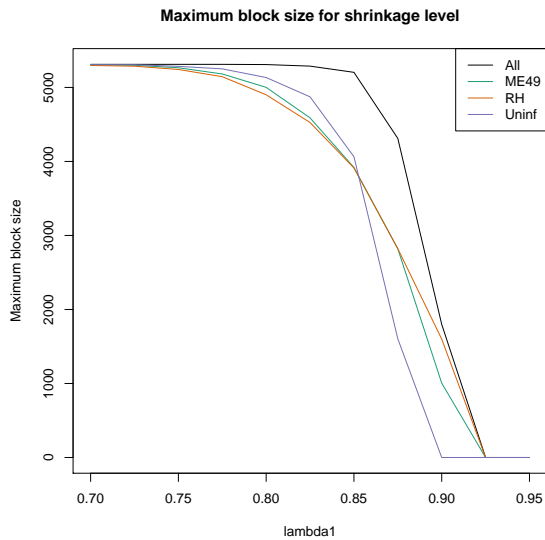


Figure 5.27: We use the maximum block size as a heuristic to select the shrinkage parameter values used with the JGL algorithm. This shows a maximum upper bound on the shrinkage value of around 0.92 above which we would not expect and genes to be included in the network. It also shows we would expect large, computationally demanding networks if the shrinkage parameter λ_1 was reduced much below 0.875.

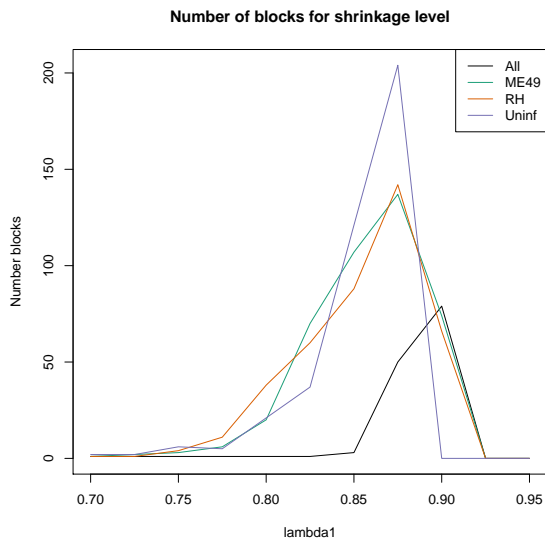


Figure 5.28: In conjunction with the maximum block size, the number of blocks gives an overview of the correlation structure for the different classes individually and combined. This information is used to inform a choice of shrinkage parameter to be used in the JGL model.

three classes. This would likely be explained as all three conditions were included in the model and no significant edges are found because values of the correlations for the uninfected are notably smaller than those for the infected cells.

To see if the uninfected cells were adversely affecting the network inference, we re-ran the JGL for only the ME49 and RH conditions, with $\lambda_1 = 0.91$ and $\lambda_2 = 0.005$, this resulted in a very small number of connections between 25 genes, with 2 edges for ME39 and 14 for the RH strain. We therefore reduced the shrinkage parameter to $\lambda_1 = 0.9$. The number of connected nodes found in the screening process is 1,394; the resulting JGL output gives 680 edges for the ME49 strain, 95 for the RH strain with 30 common edges between the two. For over a thousand genes and an inference that took over four hours to run, the small number of edges found in the model is potentially indicative of noise in the model. This could mean that by chance, expression profiles of different genes are able to partially ‘explain’ the correlations between two other genes thus reducing the partial correlations to under the threshold level of significance.

To further understand the model output we calculated the maximum block size and number of blocks over the range $\lambda_1 \in (0.9, 0.91)$ with a smaller step in values, giving a greater level of detail on the correlation structure over different shrinkage parameters. From these plots (Figures 5.29, 5.30 we selected $\lambda_1 = 0.905$, the resulting JGL model contained 735 genes with 549 edges for the ME49 strain and 2757 for the RH strain. This increase in edges at a higher shrinkage value could also support the previous observation of noise in the data set when allowing more genes into the inference.

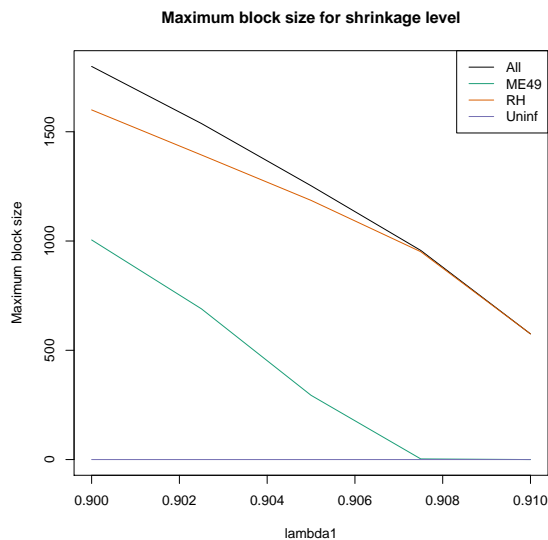


Figure 5.29: To iteratively select the parameter values, from the overview of the maximum block size we choose a subset of shrinkage values to evaluate in more detail. As the maximum block sizes for the ME49 strain is lower than for the RH strain we may choose a value below 0.906 to allow significant genes to be included for both strains.

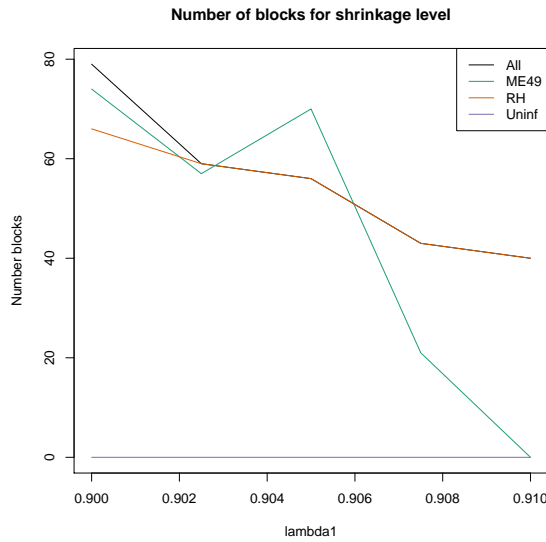


Figure 5.30: To iteratively select the parameter values, in conjunction with the maximum block size we evaluate the correlation structure in more detail for a subset of the parameter values. Although the RH strain shows a fairly steady increase over this parameter range, the ME49 strain shows a noticeable shift below 0.906, indicating shrinkage parameters below this level may result in a more informative model for this strain in particular.

Although we have significant edges in both conditions, there are substantially more for RH network in our final example. This difference in results for the two strains could be due to, differences in infection dynamics, different levels of correlation and sample sizes or differences in the accuracy of the alignments though if this was a factor we may expect more edges for the ME49 network than the RH network.

5.7.7 GO analysis of networks

We tested for over representation of Gene Ontology terms using the hypergeometric distribution. The hypergeometric distribution gave the theoretical distribution when drawing without replacement from a population of two classes. This gave the probability of drawing the observed number from one class (in our case a class is one Gene Ontology term) given the total size of the population, the total size of the class and the total number drawn. We tested all terms that are associated with at least one gene and calculate p-values for each term. These p-values are multiple hypothesis corrected using the method of Benjamini-Hochberg [Benjamini and Hochberg, 1995].

5.7.8 P-value analysis

To calculate a p-value for the correlations between genes we performed a hypothesis test. The null hypothesis is that the correlation value equals zero, that is the two genes are independent. The null distribution is a t-distribution and we calculate the p-value for each correlation pair. The p-value gives the probability of observing a value

as extreme or more extreme than the observed value assuming the correlation is not significantly different from zero.

$$H_0 : r = 0 \quad H_1 : r \neq 0$$

To test the sample correlation r , we used the standard correlation test statistic:

$$t = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

which has a t-distribution on $n - 2$ degrees of freedom under the null hypothesis H_0 , with sample size n . To see this we started with the definition of the t-distribution. The test statistic $T = \frac{U}{\sqrt{V/v}}$ where U is $N(0,1)$, V is chi-squared with v degrees of freedom, U and V are independent then T has a t-distribution on v degrees of freedom. In a standard linear regression we have $Y = \beta X + \epsilon$ where ϵ is $N(0,1)$ and we have pairs of observations $(X_i, Y_i) \quad i = 1 \dots n$, with mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ then the standard MLE of β is

$$\hat{\beta} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

In regression analysis, the test statistic

$$t = \frac{\hat{\beta}}{\sqrt{MSE / \sum_i (X_i - \bar{X})^2}}$$

has a t-distribution on $n - 2$ degrees of freedom under the null hypothesis that $\beta = 0$ and is used to test significance of $\hat{\beta}$. Where MSE is the mean square error between the observed (Y_i) and fitted values (\hat{Y}_i):

$$MSE = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}$$

We define $S_{XY} = \frac{1}{n-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$ $S_X = \sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$, the correlation $r = \frac{S_{XY}}{S_X S_Y}$. Then

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = \frac{r S_X S_Y}{S_X S_X} = \frac{r S_Y}{S_X}$$

$$\begin{aligned} MSE &= \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_i [Y_i - (\bar{Y} + \frac{S_{XY}}{S_X^2} (X_i - \bar{X}))]^2}{n-2} \\ &= \frac{\sum_i [(Y_i - \bar{Y}) - \frac{S_{XY}}{S_X^2} (X_i - \bar{X})]^2}{n-2} \\ &= \frac{\sum_i [(Y_i - \bar{Y})^2 - 2 \frac{S_{XY}}{S_X^2} (Y_i - \bar{Y})(X_i - \bar{X}) + \frac{S_{XY}^2}{(S_X^2)^2} (X_i - \bar{X})^2]}{n-2} \\ &= \frac{(n-1)[(S_Y)^2 - 2 \frac{S_{XY}}{S_X^2} + \frac{S_{XY}^2}{S_X^2}]}{n-2} \end{aligned}$$

$$\begin{aligned}
 &= \frac{(n-1)[(S_Y)^2 - \frac{S_{XY}^2}{S_X^2}]}{n-2} \\
 &= \frac{(n-1)[(S_Y)^2 - \frac{S_Y^2 S_{XY}^2}{S_Y^2 S_X^2}]}{n-2} \\
 &= \frac{(n-1)[(S_Y)^2(1-r^2)]}{n-2}
 \end{aligned}$$

Substituting these results into test statistic t, we have:

$$\begin{aligned}
 t &= \frac{r\left(\frac{S_Y}{S_X}\right)}{\sqrt{\frac{(n-1)[S_Y^2(1-r^2)]}{(n-2)((n-1)S_X^2)}}} \\
 &= \frac{r\left(\frac{S_Y}{S_X}\right)}{\sqrt{\frac{S_Y^2(1-r^2)}{S_X^2(n-2)}}} \\
 &= \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}
 \end{aligned}$$

as required.

6

Web Application

A central part of any research community is the provision of access to information, analysis tools, results and data. In bioinformatics, advanced software packages are required for many work pipelines and sharing these tools is a common part of publication. A major example of this is the R Bioconductor repository where researchers deposit R software packages and biological data as R objects that have an application to the bioinformatics community. In terms of gene expression data two of the most commonly used resources are ArrayExpress and the Gene Expression Omnibus (GEO). These databases allow researchers to upload data from their publications. It is often a requirement of publication to make the data used in analysis publicly available for verification purposes, and these databases also provide standardised formatting and quality control for datasets. The above databases include microarray and RNA-seq data. They are available online and provide tools to access the databases, that allow searching and download of data. ArrayExpress have developed cloud computing facilities and R packages for the analysis of data from ArrayExpress. GEO has an online tool for performing differential expression analysis through GEO2R using R and provides the R script used to perform the analysis.

As well as experimental data, another important area is sequence annotation data. At the simplest level this can include information such as gene names, descriptions and the sequence data itself. The Gene Ontology is the main resource used for annotations of gene function. It contains three separate ontology terms, Biological Processes, Molecular Function and Cellular Components. These ontologies have been used extensively to annotate genes in genome sequences and to identify commonalities between sets of genes - often those which have been found as differentially expressed as experimental conditions are varied. Gene set enrichment analysis (GSEA) is a closely related concept [Subramanian et al., 2005]. GSEA includes both a database of gene sets and a method for finding significantly enriched gene sets

within a list of genes. In the example above, these gene sets are categorised into eight different groups, genes that share common functions, positional genes that are closely located in the genome, those based on literature mining, genes that share regulatory motifs, computationally derived sets from cancer microarray data, Gene Ontology sets, oncogenic gene sets and immunological sets.

Closely related to gene sets are databases of known interactions between genes. These include, pathway descriptions, transcriptional units, and miRNAs and their targets. KEGG [Kanehisa et al., 2016] and Reactome [Milacic et al., 2012] are two databases containing pathway data for multiple organisms. KEGG includes pathways covering cellular processes and metabolism as well as pathways relevant to human diseases and drug categorisations. Reactome contains pathway information for multiple organisms but this does not currently include *Bacillus subtilis*. InterMine is a framework for combining various data sets into one database and providing multiple methods for accessing data in different computational languages [Smith et al., 2012]. There is a synthetic biology specific instance of InterMine, SynBioMine [unpublished, www.synbiomine.org] that includes information on *Bacillus subtilis*. BsubCyc [Caspi et al., 2014] is another *Bacillus subtilis* online resource that, similar to SynBioMine contains transcriptional unit information from DBTBS [Sierro et al., 2008]. Databases such as DBTBS provide information primarily on the transcription factors for *Bacillus subtilis*.

Using the DBTBS database, additional work has been done to add *in silico* predictions to experimentally validated transcription factors. The authors combined motif analysis with expression data to infer regulatory networks [Fadda et al., 2009]. To infer the networks, a co-expression method was used to find similarly expressed genes. The scope of this model was also limited by the amount of known regulatory information, and the available expression data - covering 1153 genes, or approximately one sixth of the *Bacillus subtilis* genome. As the authors comment, future work can expand on our knowledge of the regulatory network as more experimental data becomes available. However, this is a common constraint of methods that use prior information such as databases on regulatory motifs. The output from this method included heatmaps of the networks, and results containing interactions between the regulators but not their targets. Moreover, this information is not easily accessible. The resulting networks are presented in summary as figures in the paper, this means they are not easily interrogated by the researcher. Without access to the full network, and phenotypic or ontological information it is harder to design experiments to validate the model or any regulatory modules that may be of interest.

There are however, examples of analysis that have been made available online in a interactive and user friendly format. One of these is CoryneRegNet [Baumbach et al., 2009] this integrates inference of regulatory networks for *E. coli* and *Mycobacterium tuberculosis* with an online workflow that is designed to allow the user to query regulatory networks and predict their interspecies transfer. This is combined with functionality to integrate expression data and web based visualisation of results.

As an addition to ArrayExpress, the Expression Atlas was developed to provide differential expression data on individual genes, gene sets or cell or tissue types [Petryszak et al., 2014]. One significant part of this resource is the manual curation of the data which provides consistent experimental factor annotation across experiments and allows the combination of information from multiple experiments. The Expression Atlas does however focus on differential expression rather than differential networks or condition specific regulatory networks. Therefore, although they do include gene sets these do not give hierarchical information, or the potential differences in structures between different conditions. The Expression Atlas also does not include *Bacillus subtilis* as one of the organisms included in the curated data set. The initial data set includes *Homo sapien* and *Mus musculus* and though these sets have now been expanded to incorporate 31 different organisms in total, it does not as yet include *Bacillus subtilis*.

We similarly aimed to provide the output of the JGL model on *Bacillus subtilis* data in an interactive online resource. In addition to the example hypotheses that have been described in previous chapters, the scope of the output from the JGL algorithm and annotations mean that there are many other possible interactions that may be interesting to researchers. The analysis is designed to provide insight into regulatory networks that can then be used to further our understanding of the system and particularly on how altering parts of the cell or including circuits and inserts may impact the cell's phenotype. Consequently, we looked to implement a resource that can help in the design of synthetic circuits or provide new hypotheses on the regulatory networks. The resulting online resource we named *Bacillus subtilis* Networks (BSN).

The JGL analysis is a data driven method that can be applied to genome-wide data sets. As a result, the model can potentially include all genes in the genome and does so without any restriction on the interactions between genes, meaning it is able to find novel connections. To identify novel interactions for experimental validation, we annotated the results to indicate whether an edge between two genes is known or not in the DBTBS database. Given an edge not in the transcriptional database, the Gene Ontology information can

be used to give an idea of the functional connections between the genes. Understanding the function of transcriptional units as well as connections between them is valuable when designing experiments to validate novel edges.

The BsubCyc website enables users to navigate the known transcriptional units referenced in the database. This does not include any potentially novel interactions that require the use of network inference. Similarly, this method does not show interactions between different transcriptional units, or any condition-specific information as is given by definition in the JGL model. We planned to give the user the ability to upload their own expression data to map onto the networks as with BsubCyc. The software described here also provides unknown interactions and different combinations of networks and interactions under different experimental conditions. To facilitate the design of synthetic circuits or hypotheses to be tested experimentally we also provide PubMed searches and links, for user search terms and the genes in the networks.

STRING is an online network resource that includes protein-protein information on *Bacillus subtilis* [Szklarczyk et al., 2015]. STRING combines data using experimental evidence, text mining and computational inference to derive links between genes and allows the user to search for genes and visualise these interactions in a network view. Our analysis is based on gene expression data as opposed to protein-protein interactions this means the methodologies may give different interactions due to the different activities of genes and the corresponding protein. Additionally, BSN differs from the STRING web resource in that we can provide experimental condition information and additional tools such as network expansion and decomposition based on computational inference.

In designing BSN, we first considered those tools currently available which could be used to implement it. Our requirements were that the language or environment we used to write BSN could be hosted online, provide interactive network visualisation and can interface easily with R because the functionality we have developed is written in R.

One of the most common tools for visualising biological networks is Cytoscape [Shannon, 2003]. Cytoscape has a large set of built in functions for displaying, navigating and annotating networks. In this work, Cytoscape has been used extensively for visualising network output from the JGL network. This is possible due to the RCytoscape package which has been developed to allow Cytoscape to be controlled from R [Shannon et al., 2013]. This means that R users unfamiliar with Cytoscape can still easily use it for their end visualisation while R is used for its powerful data analysis capabilities. Cytoscape does allow for third parties to design ‘plugins’ that can be used to provide

additional functionality to Cytoscape. These plugins must be written in Java and therefore are only a viable option for those with appropriate java programming experience. Cytoscape does offer a web plugin that can be used by those with basic understanding of HTML. However, also using the functionality of R would require use of the Rserve plugins. Here again, this requires advanced programming experience and is not in general, accessible to the average R user.

In contrast, Shiny has an inbuilt Graphical User Interface (GUI) and is written in Javascript; a commonly used language for creating dynamic webpages. Javascript can be used on all computer platforms. However, coding in Shiny requires no javascript knowledge. Shiny can be used by beginner R programmers to develop web pages that have access to all the functionality of R analysis software. The Shiny framework provides a method for publishing R packages and functions onto the web. Therefore, a web user can access and use R functions without needing expertise in R. Shiny provides the basic GUI that can be accessed from R. The Shiny developers have focused on providing interactive plots. This meant that we could plot the network output from the JGL analysis and provide functionality to explore the networks interactively from a web page. Whilst the Shiny plot features are not as extensive as those in Cytoscape, we were able to provide a set of plots in our Shiny app that met all our requirements. These requirements were, the ability to select and zoom onto nodes, to view node information and to annotate networks with users uploaded data. We used Shiny to write BSN which can be hosted online to visualize and investigate the results of the JGL algorithm for three different meta-conditions of microarray gene expression data for *Bacillus subtilis*. BSN is also be hosted online using their standard free hosting service <https://jglnetworks.shinyapps.io/BacillusApp/>.

6.1 Overview of BSN

The design of BSN includes multiple tabs on the webpage which the user can utilise to select a function or type of analysis that they would like to use. There are four of these different sections outlined below:

Network Analysis: Search for genes, view information on transcriptional unit information and GO with links out to PubMed, Bsub-Cyc.org and SynBioMine. The original dataset can be searched for genes strongly correlated to a selected subset of genes and the JGL algorithm re-run with these genes.

Differential Expression: Upload files containing (differential) expression values to be overlaid onto the networks.

Large Network Decomposition: The algorithm runs simulations to find

the closest network that separates genes according to the classes edges appear in.

Analyse New Data: Uploaded gene expression data that have been mean and variance standardised can be analysed with the reference data set. This provides additional data used to infer the network.

In designing the application, we would like to be able to integrate as many useful sources of information as possible. One way we have done this is to link to existing online databases of information, that is SynBioMine and BsubCyc. These two *Bacillus subtilis* relevant sites collate a large amount of the information available on *Bacillus subtilis* and currently are updated frequently. By using these resources, we do not duplicate the databases that already exist, and remove the need to manually update the annotation data for BSN. One exception to this is that currently the transcriptional unit information must be manually updated by the maintainer of BSN. This is because the transcriptional unit information must be parsed from the individual BsubCyc webpages for each gene and then converted into a matrix of edge annotations for use with the adjacency matrix which contains the JGL network result. BSN has also been designed to allow users to upload and analyse their own data. This allows for greater coverage and specificity of data to be integrated with the existing model, which can improve the network inference.

The list below summarises the information integrated into the resource:

Partial regulatory networks based on a subset of expression data from Nicolas *et al.* These are inferred using the JGL model of Danaher *et al.*, under multiple conditions.

Transcription unit information from BsubCyc.org, this is obtained from DBTBS, which currently has 831 regulated operons. The scope for this information is over ten percent of the genome.

Sigma Factor data from DBTBS and SubtiWiki. There are 18 sigma factors currently listed on DBTBS. Across the 18 sigma factors, there are in total 778 genes listed that are regulated by these sigma factors in the *B. subtilis* genome.

Gene Ontology summary information from ENSEMBL

Links out to Gene Ontology information from SynBioMine

Links out to BsubCyc.org page for each gene that gives information on:

Gene local context: the location of the gene and surrounding genes

Transcriptional unit: experimentally validated regulatory sets

Gene Ontology terms: information on molecular function and biological processes associated with the gene

Regulation Summary Diagram: this shows the general RNA polymerase, and where known, the associated sigma factor. The diagram outlines the process of DNA transcription to RNA and translation of RNA to protein. Where the protein is known to undergo post-translation modifications this is shown as a final stage.

6.1.1 Examples of network analysis

Gene Search:

The network can be searched for a gene using the text box. If the gene is in the networks the gene is highlighted using a grey box, Figure 6.1. The mouse can also be used to select part of the network using a click and drag. For the genes within the selected area a secondary plot showing a zoom of the network is displayed below the full network, Figure 6.2. The network is annotated with the transcriptional unit information, green to indicate a known transcriptional unit connection, blue for a gene with no known transcriptional unit and red to indicate that though transcriptional unit information is known about these genes so far, they have not been shown to be in the same unit. There are also table outputs for the selected genes. The table containing information on genes provides the gene name, short description as well as links to Gene Ontology information on SynBioMine and the appropriate BsubCyc page.

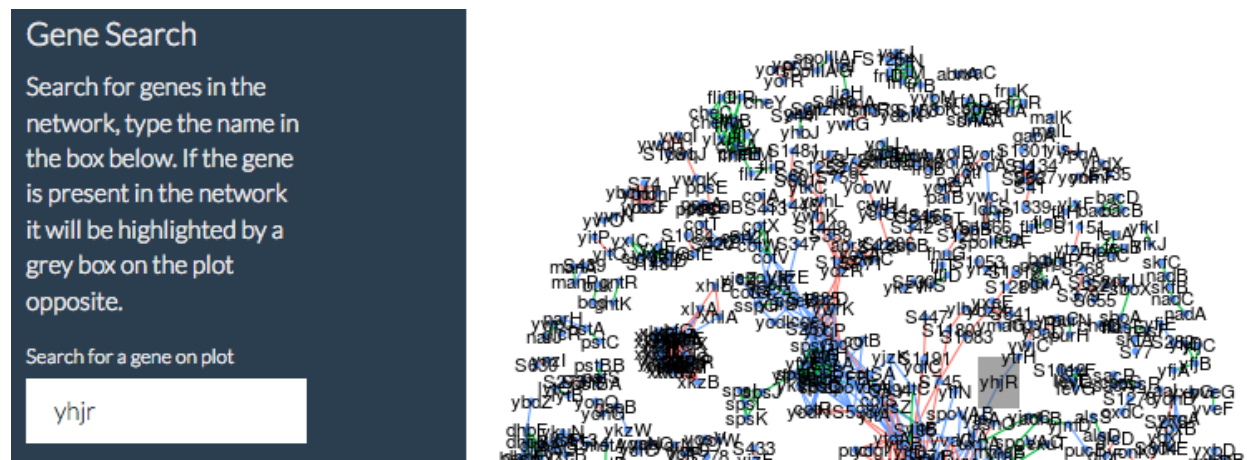


Figure 6.1: Search for gene *yhjr* in text box, result is highlighted on the network by a grey box. The grey box highlights the area of the gene of interest, the mouse cursor can then be used to select the gene.

PubMed Search:

When a search term is entered (e.g. Chemotaxis) into the query box two searches are performed for each gene on PubMed, Figure 6.3. The first is for the gene name, for each search we also combined gene name and the organism *Bacillus subtilis* (as the same gene names are often used across different organisms) to give the number of paper results for that gene/*Bacillus subtilis* combination in the gene counts column. The second search additionally has the user input search term (e.g. Chemotaxis), again a link to the PubMed results is provided along with the number of papers found on PubMed, for this combination of search term, gene name and *Bacillus subtilis*. This is useful to allow

PubMed for the gene and the search term and the gene alone.

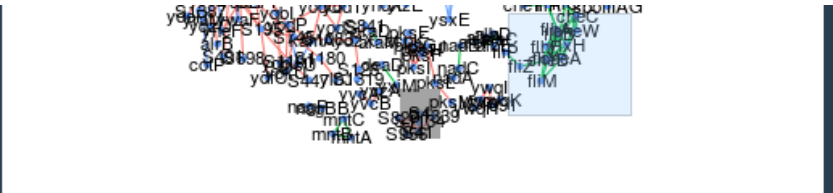
Search term in Pubmed:

Selecting Genes

Use the mouse to select genes in the plot opposite. Information on their gene ontologies and sigma factors will be displayed in the table below the plot. Additionally using the check boxes a group of genes can be selected as a sub network to expand using the button Expand Subnetwork below.

Select genes to include in Subnetwork:

- flhM
- flhY
- flhZ
- flhB
- flhA
- flhF
- ylxH
- cheB
- cheA
- cheW
- cheC



The edges in the plot are coloured according to the transcriptional information available on bsubCyc.org. Green edges are genes that are in the same known transcriptional units. Blue edges denote two genes with no known transcriptional unit information. Where two genes are connected by a red edge, either or both of the genes have known transcriptional unit information, but these two genes are not currently known to be in the same transcriptional unit.

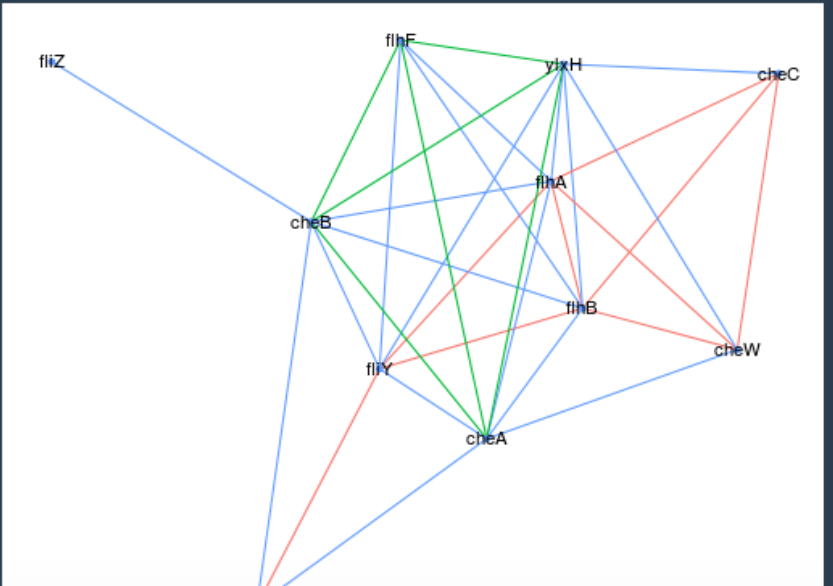


Figure 6.2: Selecting a set of genes on the network using the mouse cursor. A second image is displayed of just the selected genes. The check boxes allow the user to select any of these genes to be included in a sub network. These sub networks can then be expanded with up to 30 additional genes.

the user to see if there are any hits before moving to PubMed. It also gives an indication of how likely it is that the gene is connected to the search term through comparison of the number of hits with the gene name and the search term in comparison to the gene name alone. That is, if the number in each column is very similar, then nearly all papers mentioning the gene also mention the search term and the user can be more confident that the two are connected.

Gene	GO.Term	Sigma.Factor	Go_SynBioMine	Link_Bsubcyc	Link_PubMed	PubMed_Gene	Gene_only
cheA	phosphorelay signal transduction system [GOA00, GOA01] GO:0006928	SigD	cheA	cheA	cheA	38	46
cheB	phosphorelay signal transduction system [GOA01] GO:0006355	SigD	cheB	cheB	cheB	22	23
cheC	chemotaxis [GOA00]	SigD	cheC	cheC	cheC	21	21
cheW	chemotaxis [GOA00, GOA01] GO:0007165	SigD	cheW	cheW	cheW	16	20
fliY	ciliary or flagellar motility [GOA00, GOA01] GO:0006935	SigD	fliY	fliY	fliY	9	12

Expand Subnetwork:

The user can select a subset of genes in the network, this is by means of a set of checkboxes that automatically update with the names of those genes selected in the network. Given the selection, an example of which is shown in Figure 6.4, this function finds up to the 30 genes with the highest correlation to the set of selected genes. Taking the selected genes and the additional genes (which do not have to be in the original network), the function recalculates the JGL output just for these genes. BSN displays the networks before and after the sub network expansion, Figure 6.5.

6.1.2 Differential expression network

This function allows the user to upload their own differential expression data which is then mapped onto the network. This means that users can see, for a set of differential expression values, whether the network or set of genes is up or down regulated for the uploaded conditions. Given differential expression data the researcher often wants to gain an overview of how these genes interact with each other. This

Figure 6.3: After searching for selected genes in PubMed using the search term Chemotaxis. In addition to the standard gene links to SynBioMine and BsubCyc there is also a link out to the PubMed search page for the *Bacillus subtilis* gene and the search term (Chemotaxis). There are also summary counts on the results in the last two columns. These give the number of results on PubMed for the gene and search term and the gene alone. This is useful for two main reasons. The first is that it gives an idea of how likely the gene is to be related to the search term - that is, if the two columns have similar counts. Second, it will indicate if there are no search results, this is a useful tool for the researcher as they will not spend time searching PubMed for an uninformative result.

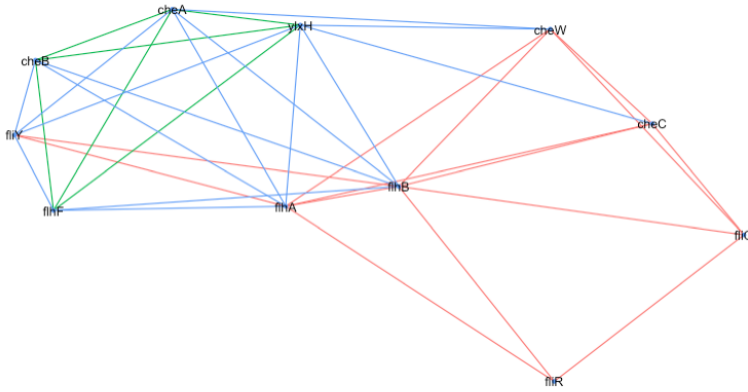


Figure 6.4: Sub network selected from the larger network for expansion. The gene names are shown along with the connections between them. These edges are coloured according to the information on transcriptional units taken from BsubCyc.org. Green edges are between two genes in the same transcriptional units. Blue edges are for a gene with no known transcriptional unit. Finally, red edges indicate that the two genes have not been linked in the same transcriptional unit but that both are connected to at least one transcriptional unit.

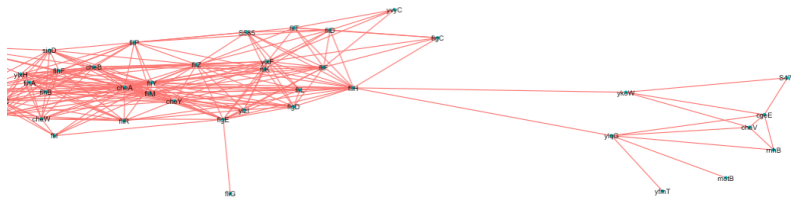


Figure 6.5: Selected sub network from Figure 6.4 after expansion. The sub network is expanded by including up to an additional 30 genes. These genes are selected as those with the highest Pearson's correlation value to all nodes in the sub network. This shows how the expansion algorithm can be used to investigate the subnetworks individually and in a computationally tractable way. Further, it can find genes that have significant interaction in the JGL model. This is because the additional 30 genes are connected to the original sub network rather than appearing as a separate sub network or as disjoint single nodes.

can be done in several different ways, for example, using gene set enrichment analysis to find functional terms common to the differentially expressed genes. Or by testing for over representation of a pathway in the set. BSN gives an alternative visual method of doing this. By mapping differential expression onto the network, users can easily identify areas of the network that are differentially expressed. This gives information not only on the regulatory networks present in the list of differentially expressed genes but also the hierarchy and network view of those genes see Figure 6.6 as an example. The legend provides information on the differential expression colour scale. From this visual perspective, it is easy to see clusters of genes with similar differential expression profiles.

The format of the file should have genes in the rows and different conditions in the columns. The file can contain multiple contrasts, that is, differential expression between multiple pairs of conditions. In this case, the column headers will be used as identifiers for each contrast: after the data are loaded, the column headers will appear as the names of the radio buttons which are used to select which of the conditions to be mapped onto the network. In this way the user is able to toggle between conditions. The example in Figure 6.6 shows that there were two columns in the dataset uploaded to BSN, these are here named Test1 and Test2. The user can select either of these and then update the network by pressing the Overlay Differential Expression button.

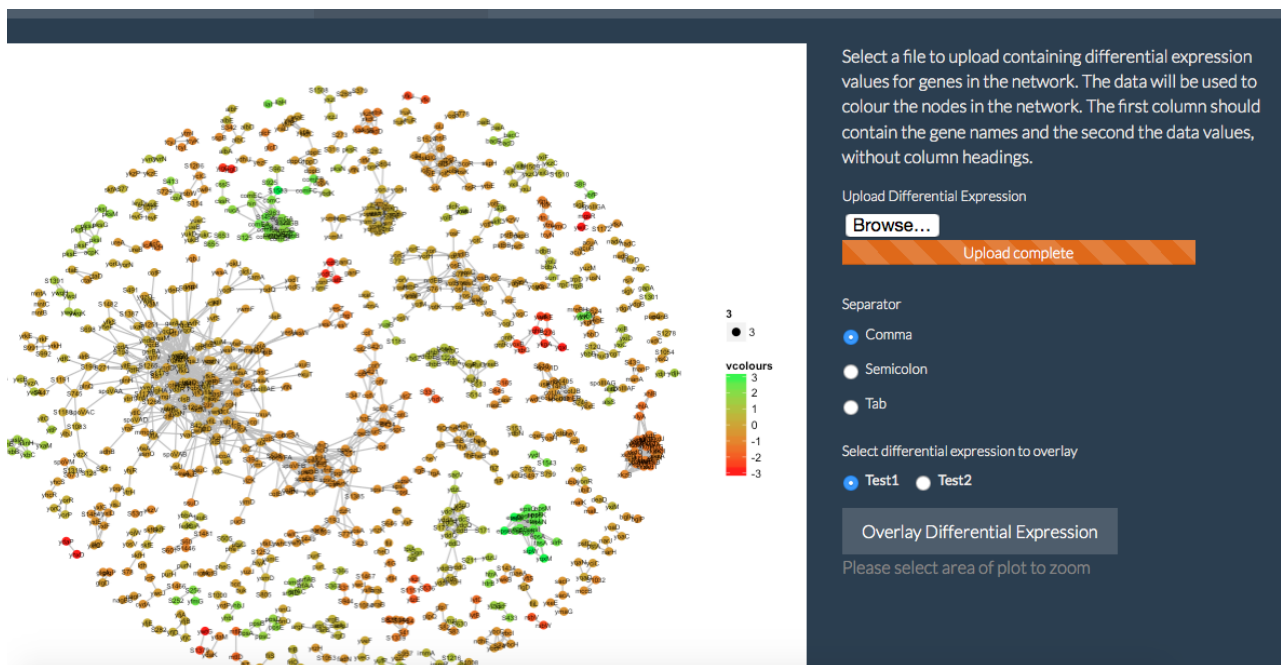
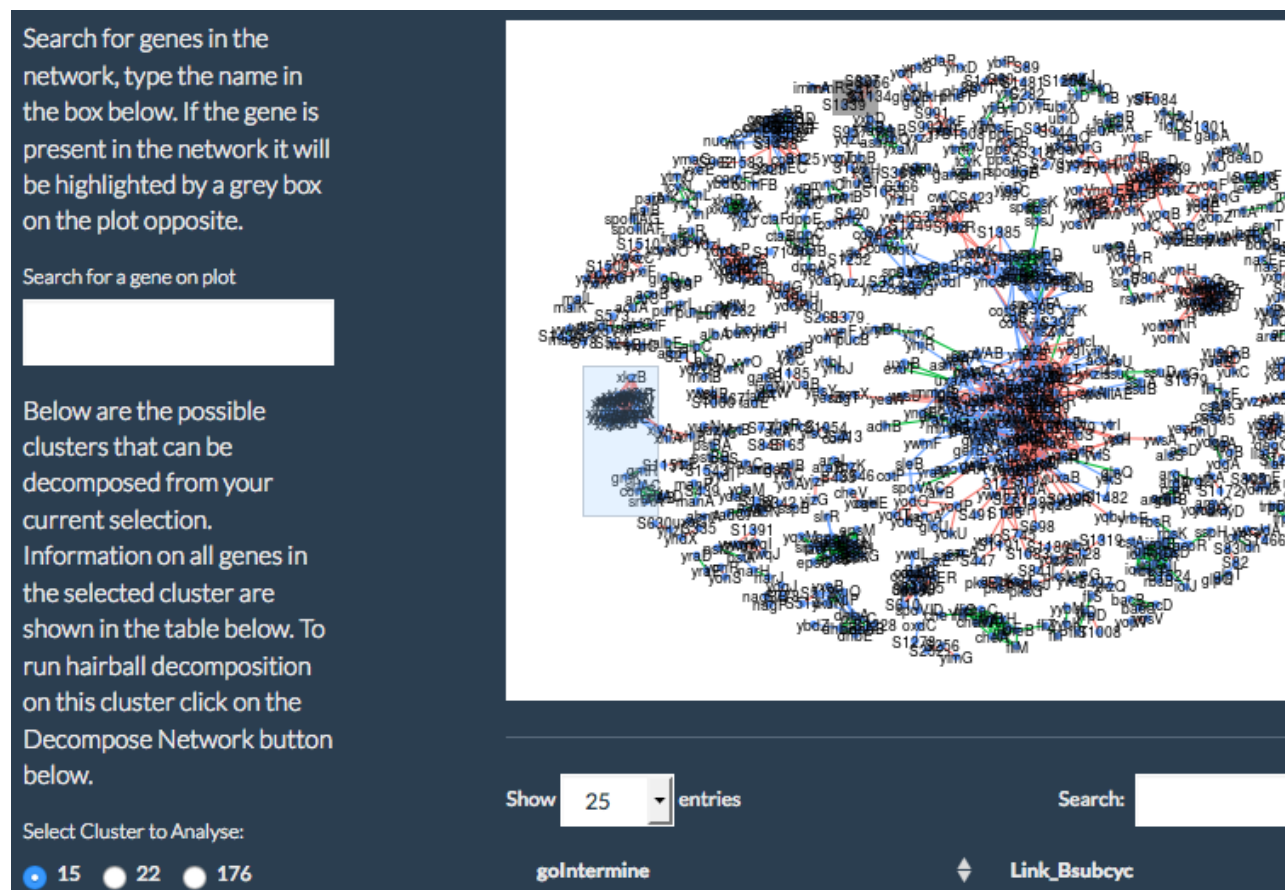


Figure 6.6: Overlaying uploaded differential expression data onto the BSN network.

6.1.3 Large network decomposition

We have implemented the Monte Carlo method for decomposing large networks as outlined in Section 3.5.6. This allows the user to select an area of interest from the network using the mouse as in Figure 6.7. It is also possible to search for a gene in the text box which, if present in the network, will be highlighted with a grey box. Once an area of the network is selected BSN will list all clusters which have at least one gene in the selected area, these appear as radio buttons to the left of the network. In our example this includes clusters 15, 22 and 176. For the cluster selected, 15, BSN lists all genes in this cluster in the table below the network. The table gives the links out to the Gene Ontology information on SynBioMine and the genes page on bsubcyc.org.



The resulting network decomposition is shown below the original network in Figure 6.8. The decomposition is run once the user presses the Network Decomposition button. This means that BSN is not slowed down during the selection of the cluster or sub network to analyse as would happen if the decomposition was run automati-

Figure 6.7: Showing the selection of an area of the network from which a single sub network or cluster can be chosen for decomposition using simulation methods based on the edge values of the sub network. The clusters contained within the highlighted portion of the network can be selected using the radio buttons at the bottom left of the screen.

cally according to the cluster currently selected by the radio buttons. Figure 6.8 shows that for this particular example sub network there were a lot of different edge conditions connecting these genes. This is because most of the genes have been separated from all other genes, meaning that connected nodes have been assigned to different edge classes.

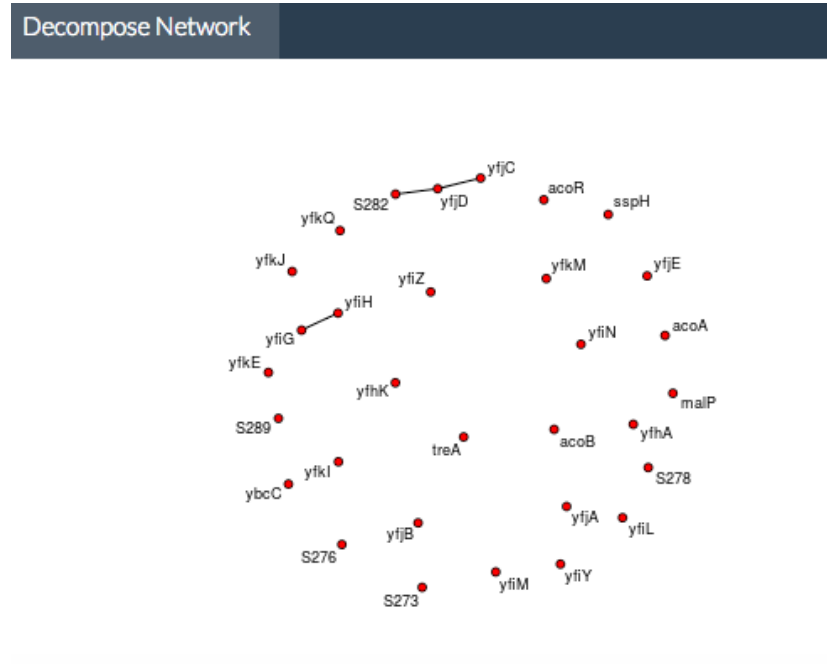
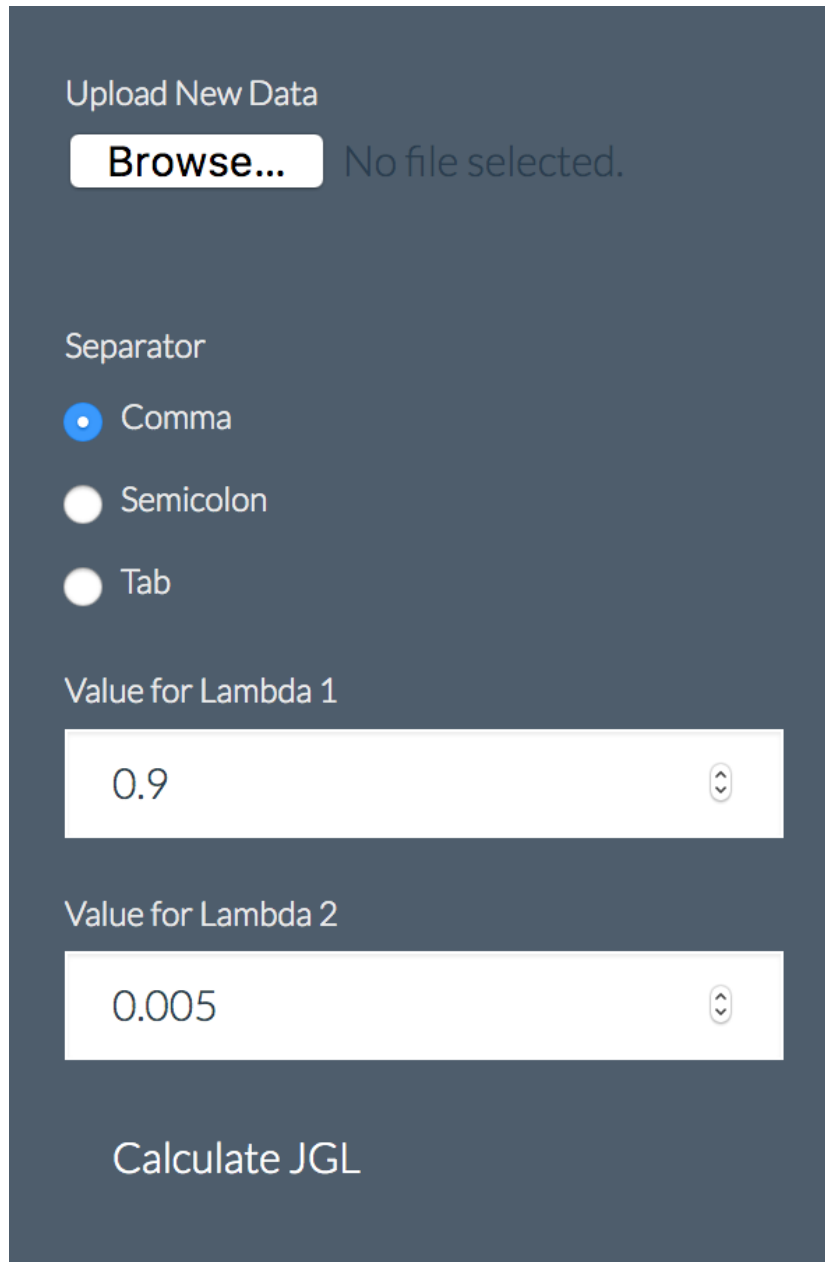


Figure 6.8: The selected region from Figure 6.7 after using the Monte Carlo method to split a cluster that was previously one network.

6.1.4 Analyse new data

BSN allows the user to upload their own data on *Bacillus subtilis* for integration with the results already available. As shown in Figure 6.9 the file format is uploaded from the user's personal computer and, consistent with the differential expression data upload, must be separated by either comma, semicolon or tab. The format of the file should have genes in the rows and different conditions in the columns. From here BSN will find all those genes in the dataset that match those in the current network. The JGL algorithm is then run on the intersection of the genes present in both the BSN network and the uploaded dataset. This means that BSN is currently useful for identifying differences in the network structures between different conditions, however it has not as yet been extended to allow inclusion of all available genes in the uploaded data set, see Section 6.3 for further discussion.



The image shows a dark-themed web interface for uploading data and configuring parameters. At the top, it says "Upload New Data". Below this is a "Browse..." button and the text "No file selected.". Underneath, there is a "Separator" section with three radio button options: "Comma" (selected), "Semicolon", and "Tab". Below the separator options are two input fields for "Value for Lambda 1" (containing "0.9") and "Value for Lambda 2" (containing "0.005"). At the bottom of the interface is a "Calculate JGL" button.

Figure 6.9: BSN allows the user to upload their own data to be added to the network. This file can be comma, tab or semicolon separated. The user can also select the two shrinkage parameters to be used with the JGL algorithm.

6.1.5 Exporting data

On the Network Analysis tab the button ‘Generate Report’ writes the results contained within the summary table to a csv file that is downloaded to the user’s computer. All the images generated by BSN can be downloaded to the user’s computer by right clicking and selecting ‘Save Image As’. This opens the standard ‘Save File’ dialog that allows the user to give a filename and location to the image, as default all these images are saved as png files.

6.2 Code outline

An overview of BSN is shown in Figure 6.10. The main data is the JGL gene network which was the output from the JGL model we derived earlier based on the *Bacillus subtilis* data. BSN allows user input, in the form of gene or sub network selection or as data uploads to BSN. There are three main algorithms made available through BSN, these are the sub network expansion, network decomposition and the original JGL R package. These are all shown in Figure 6.10 under *App Function*, the output is either another graphical network of gene interactions or a table containing the gene names and information such as Gene Ontology terms.

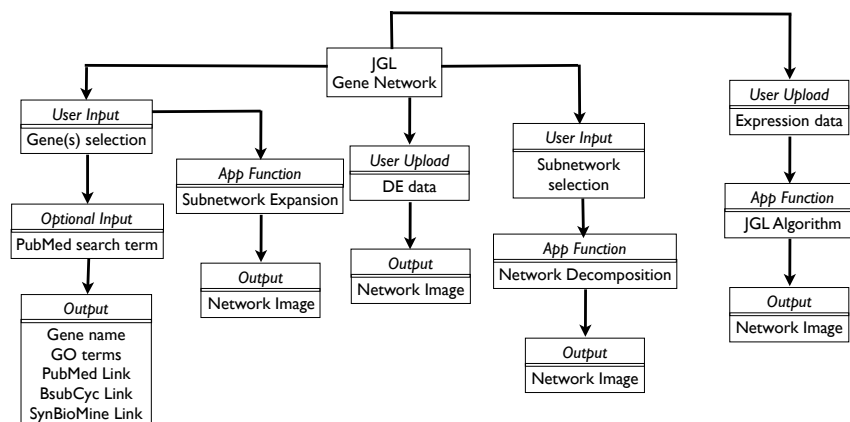


Figure 6.10: An overview of BSN. The main data is the JGL gene network. There are also user inputs within BSN, *User Input* as well as data uploads: *User upload*. The functionality available within BSN is outlined as are the outputs. The output is either tables of information on the genes or network views of the interactions between the genes.

6.3 Future developments

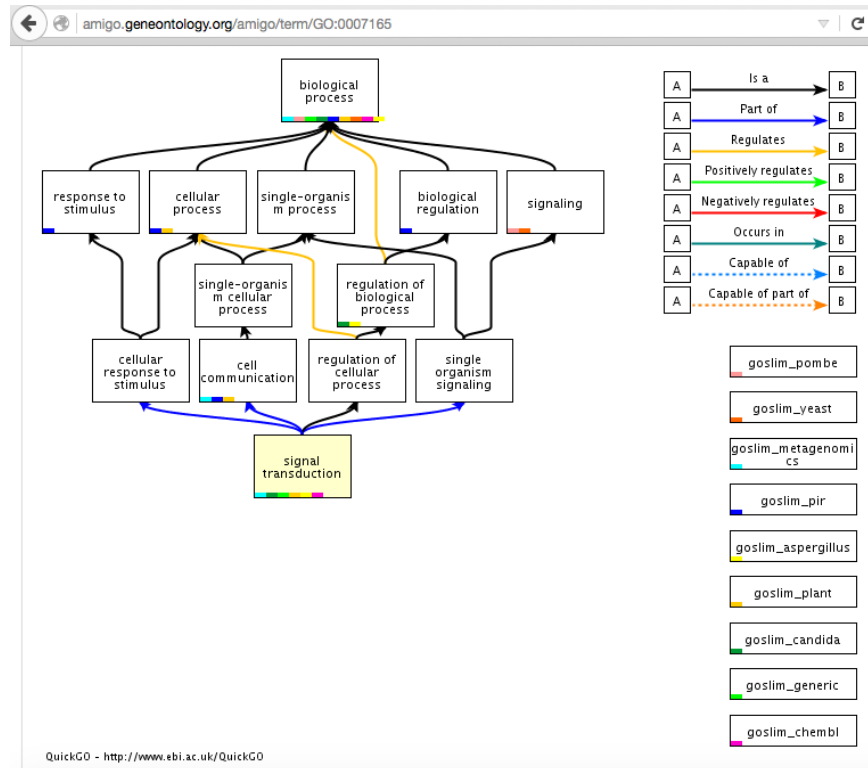
There are different ways in which BSN can be developed in the future. These include increasing the data that have been analysed and adding in more condition-specific information. Analysis of the *Bacillus subtilis* data set from Nicolas *et. al* in Section 3.3.2 showed that there were potentially another two JGL networks, based on different samples that could be added to BSN. This would allow a larger proportion of the genome to be available to the user for analytical purposes.

Another option would be to add additional options for the analysis of user uploaded data specifically to enable them to select appropriate shrinkage parameters therefore making the analysis more tractable. This would be beneficial because it is difficult to know without prior knowledge or previous experience of the JGL model what shrinkage parameters to use. For instance, it would be possible to add an overview of the block sizes at different shrinkage levels as with the *Toxoplasma gondii* analysis. This could also be extended to provide summary statistics of the block sizes, that is the number of blocks, and their average size in addition to the maximum block size. It may also be possible to give an indication of how long the analysis would take, but this, is not solely determined by the maximum block size, it also depends on how quickly the inverse is found - the speed of convergence will depend on the sparsity of the individual blocks and whether they have an easily inverted form.

BSN could also be extended to use the empirical Bayes method for estimating the correlations. Again, this method would be most useful with guidance on the shrinkage parameters to use, and this could be provided by the analysis of the block sizes at given levels of shrinkage which can also be performed on the empirical Bayes correlation matrices. Analogous to the baseline expression in the Expression Atlas, it would also be useful to have baseline networks for different cell or tissue types. As a standard, the data used in our analysis is filtered to include only those genes which we think are expressed before any correlations are calculated. As seen with the *Toxoplasma* analysis, the untreated cells for the filtered data set show similar correlations across all the expressed genes. Obviously, the subset of genes that passes the filter will be different for different cell types. It may be possible to generate the larger networks, with block sizes covering thousands of genes, to infer the baseline networks. This would also allow for comparison of different networks, for instance between an untreated or healthy cell populations and a condition or perturbation of interest.

There is potential to include additional annotation resources, for example in our gene table we could add a link from assigned Sigma

factors to their corresponding page on DBTBS, which would give information on all known regulons of the Sigma factor. We could also develop BSN to check for multiple Sigma factors being assigned to the same gene and display all possible Sigma factors rather than just the first. We could also add functionality which allows the user to select the edge annotation: this is standard functionality in Cytoscape and would be useful to add into BSN. For example, in the differential expression mapping the user may want to switch between transcriptional unit information on the edges and condition information. This would then provide an easy way to see the up or down regulation of known regulatory units and then change the annotation view to see the conditions in which this is true according to the output from the JGL model. Ideally an algorithm for allowing a hierarchical representation of the network in R would be beneficial. Unlike Cytoscape, currently these algorithms are not present in R and so we are not able to make use of this layout in the network output.



Gene Ontology information is a valuable resource however, one current area of research is how to summarise or simplify the ontological information. This is because the ontologies are in network form themselves: this means there are general parent terms with more specific child terms and a gene may be associated with either just a parent

Figure 6.11: An example of the graph structure for Gene Ontology terms. Terms can be parents or children of other ontology terms.

term or a parent and multiple child terms. An example of the directed acyclic graph structure of the Gene Ontologies is shown in the Figure 6.11, it can be seen that the term ‘Signal Transduction’ is a child term that is part of ‘cellular response to stimulus’, ‘cell communication’ and ‘single organism signaling’. When comparing ontologies across sets of genes, it is usual to search for overrepresented terms. This gives an overview of common terms and indicates, cellular components, biological processes and molecular functions that are unexpectedly prevalent in the gene sets.

From a network perspective however, it is still difficult to summarise the ontological information. We used the GO slim terms that aim to simplify the gene ontologies [Blake et al., 2015]. However, when colouring nodes according to their GO slim terms, Cytoscape is not able to find commonalities between the different annotations. The colouring methods available in Cytoscapes vizmapper (visualisation mapper) are either randomised or rainbow scale. Using the rainbow scale the colours will be according to the sorted terms which means that the colour is essentially according to the first few terms within the GO slim set. This means that whilst there may be common terms between two connected genes, they may have very different node colours if the common terms are not at the beginning of the annotation. Therefore, we could also develop the functional and disease analysis methods used with the *Toxoplasma gondii* experiment into extra annotation methods in BSN. This would mean changing the annotation colouring according to a single term of interest. Two useful options would be to give a search box where the user can input any ontological term, and the second that would give a list of terms to choose from all of which have been found to be statistically overrepresented in the network.

Bibliography

Tarek Abbas and Anindya Dutta. p21 in cancer: intricate networks and multiple activities. *Nature Reviews Cancer*, 9(6):400–414, 6 2009. ISSN 1474-175X. DOI: 10.1038/nrc2657. URL <http://www.nature.com/doifinder/10.1038/nrc2657>.

Sónia S Albuquerque, Céline Carret, Ana Grosso, Alice S Tarun, Xinxia Peng, Stefan HI Kappe, Miguel Prudêncio, and Maria M Mota. Host cell transcriptional profiling during malaria liver stage infection reveals a coordinated and sequential set of biological events. *BMC Genomics*, 10(1):270, 2009. ISSN 1471-2164. DOI: 10.1186/1471-2164-10-270.

James C. Anderson and David W. Gerbing. The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2):155–173, 1984. ISSN 00333123. DOI: 10.1007/BF02294170.

M A Anttila, V M Kosma, J Hongxiu, J Puolakka, M Juhola, S Saarikoski, and K Syrjänen. p21/WAF1 expression as related to p53, cell proliferation and prognosis in epithelial ovarian cancer. *British journal of cancer*, 79(11-12):1870–8, 1999. ISSN 0007-0920. DOI: 10.1038/sj.bjc.6690298. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2362791&tool=pmcentrez&rendertype=abstract>.

Richard A. Armstrong. When to use the Bonferroni correction. *Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*, 34(5):502–508, 2014. ISSN 14751313. DOI: 10.1111/opo.12131.

Mario L Arrieta-Ortiz, Christoph Hafemeister, Ashley Rose Bate, Timothy Chu, Alex Greenfield, Bentley Shuster, Samantha N Barry, Matthew Gallitto, Brian Liu, Thadeous Kacmarczyk, Francis Santoriello, Jie Chen, Christopher D A Rodrigues, Tsutomu Sato, David Z Rudner, Adam Driks, Richard Bonneau, and Patrick Eichenberger. An experimentally supported model of the *Bacillus subtilis* global

transcriptional regulatory network. *Molecular Systems Biology*, 11(12): 1–17, 2015. ISSN 1744-4292. DOI: 10.15252/msb.20156236.

D Arsenijevic, H Onuma, C Pecqueur, S Raimbault, B S Manning, B Miroux, E Couplan, M C Alves-Guerra, M Goubern, R Surwit, F Bouillaud, D Richard, S Collins, and D Ricquier. Disruption of the uncoupling protein-2 gene in mice reveals a role in immunity and reactive oxygen species production. *Nature Genetics*, 26(4):435–439, 2000. ISSN 1061-4036. DOI: 10.1038/82565.

Harm Van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Response to "The Reality of Pervasive Transcription". *PLoS Biology*, 9(193588):7–10, 2011. DOI: 10.1371/journal.pbio.1001102.

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, 1 2011. ISSN 1471-0064. DOI: 10.1038/nrg2918. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3140052&tool=pmcentrez&rendertype=abstract>.

Christian L. Barrett and Bernhard O. Palsson. Iterative reconstruction of transcriptional regulatory networks: An algorithmic approach. *PLoS Computational Biology*, 2(5):429–438, 2006. ISSN 1553734X. DOI: 10.1371/journal.pcbi.0020052.

Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, 41(D1):991–995, 2013a. ISSN 03051048. DOI: 10.1093/nar/gks1193.

Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, 41(D1):1–5, 2013b. ISSN 03051048. DOI: 10.1093/nar/gks1193.

David P. Bartel. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233, 1 2009. ISSN 00928674. DOI: 10.1016/j.cell.2009.01.002. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867409000087>.

Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 4 2005. DOI: 10.1038/ng1532. URL <http://www.nature.com/ng/journal/v37/n4/full/ng1532.html>.

Douglas Bates and Martin Maechler. Matrix: Sparse and Dense Matrix Classes and Methods, 2017. URL <https://cran.r-project.org/package=Matrix>.

Jan Baumbach, Tobias Wittkop, Christiane Katja Kleindt, and Andreas Tauch. Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nature Protocols*, 4(6):992–1005, 2009. ISSN 1754-2189. DOI: 10.1038/nprot.2009.81. URL <http://www.nature.com/doifinder/10.1038/nprot.2009.81>.

Stephen P. Bell and Anindya Dutta. DNA Replication in Eukaryotic Cells. *Annual Review of Biochemistry*, 71(1):333–374, 2002. ISSN 0066-4154. DOI: 10.1146/annurev.biochem.71.110601.135425. URL <http://www.annualreviews.org/doi/10.1146/annurev.biochem.71.110601.135425>.

Y Benjamini and Y Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 57(1): 289–300, 1995. URL <http://medcontent.metapress.com/index/A65RM03P4874243N.pdf><http://www.jstor.org/stable/10.2307/2346101>.

Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):1–14, 2012. ISSN 03051048. DOI: 10.1093/nar/gks001.

Nicolas Bisson, D Andrew James, Gordana Ivoisev, Stephen a Tate, Ron Bonner, Lorne Taylor, and Tony Pawson. Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. *Nature biotechnology*, 29(7):653–8, 7 2011. ISSN 1546-1696. DOI: 10.1038/nbt.1905. URL <http://www.ncbi.nlm.nih.gov/pubmed/21706016>.

Ira J. Blader and Jeroen P.J. Saeij. Communication between *Toxoplasma gondii* and its host: impact on parasite growth, development, immune evasion, and virulence. *APMIS*, 117(5-6):458–476, 5 2009. ISSN 09034641. DOI: 10.1111/j.1600-0463.2009.02453.x. URL <http://www.ncbi.nlm.nih.gov/pubmed/2810527><http://doi.wiley.com/10.1111/j.1600-0463.2009.02453.x>.

Ira J Blader, Ian D Manger, and John C Boothroyd. Microarray Analysis Reveals Previously Unknown Changes in *Toxoplasma gondii*-infected Human Cells *. *The Journal of Biological Chemistry*, 276(26): 24223–24231, 2001. DOI: 10.1074/jbc.M100951200.

Alexandre Blais and Brian David Dynlacht. Constructing transcriptional regulatory networks. *Genes & Development*, 19(13): 1499–1511, 7 2005. ISSN 0890-9369. DOI: 10.1101/gad.1325605. URL <http://genesdev.cshlp.org/content/19/13/1499.shorhttp://www.genesdev.org/cgi/doi/10.1101/gad.1325605>.

J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, S. Burgess, T. Buza, C. Gresham, F. McCarthy, L. Pillai, H. Wang, S. Carbon, H. Dietze, S. E. Lewis, C. J. Mungall, M. C. Munoz-Torres, M. Feuermann, P. Gaudet, S. Basu, R. L. Chisholm, R. J. Dodson, P. Fey, H. Mi, P. D. Thomas, A. Muruganujan, S. Poudel, J. C. Hu, S. A. Aleksander, B. K. McIntosh, D. P. Renfro, D. A. Siegele, H. Attrill, N. H. Brown, S. Tweedie, J. Lomax, D. Osumi-Sutherland, H. Parkinson, P. Roncaglia, R. C. Lovering, P. J. Talmud, S. E. Humphries, P. Denny, N. H. Campbell, R. E. Foulger, M. C. Chibucos, M. Gwinn Giglio, H. Y. Chang, R. Finn, M. Fraser, A. Mitchell, G. Nuka, S. Pesseat, A. Sangrador, M. Scheremetjew, S. Y. Young, R. Stephan, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, M. D. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, G. T. Hayman, S. J. Wang, V. Petri, P. D’Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J. M. Cherry, M. C. Costanzo, J. Demeter, S. S. Dwight, S. R. Engel, B. C. Hitz, D. O. Inglis, P. Lloyd, S. R. Miyasato, K. Paskov, G. Roe, M. Simison, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, D. Li, E. Huala, J. Argasinska, C. Arighi, A. Auchincloss, K. Axelsen, G. Argoud-Puy, A. Bateman, B. Bely, M. C. Blatter, C. Bonilla, L. Bougueleret, E. Boutet, L. Breuza, A. Bridge, R. Britto, C. Casals, E. Cibrian-Uhalte, E. Coudert, I. Cusin, P. Duek-Roggli, A. Estreicher, L. Famiglietti, P. Gane, P. Garmiri, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, C. Hulo, R. Huntley, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, A. Macdougall, M. Magrane, M. Martin, P. Masson, P. Mutowo, C. O’Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Shypitsyna, A. Stutz, S. Sundaram, M. Tognolli, C. Wu, I. Xenarios, J. Chan, R. Kishore, P. W. Sternberg, K. Van Auken, H. M. Muller, J. Done, Y. Li, D. Howe, and M. Westerfeld. Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015. ISSN 13624962. DOI: 10.1093/nar/gku1179.

Martin Blume, D. Rodriguez-Contreras, Scott Landfear, Tobias Fleige, D. Soldati-Favre, Richard Lucius, and Nishith Gupta. Host-derived glucose and its transporter in the obligate intracellular pathogen *Toxoplasma gondii* are dispensable by glutaminolysis. *Proceedings of the National Academy of Sciences*, 106(31):12998–13003, 8 2009. ISSN 0027-8424. DOI: 10.1073/pnas.0903831106. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.0903831106>.

Martin Blume, Richard Nitzsche, Ulrich Sternberg, Motti Gerlic, Seth L. Masters, Nishith Gupta, and Malcolm J. McConville. A *Toxoplasma gondii* gluconeogenic enzyme contributes to robust central carbon metabolism and is essential for replication and virulence. *Cell Host and Microbe*, 18(2):210–220, 2015. ISSN 19346069. DOI: 10.1016/j.chom.2015.07.008. URL <http://dx.doi.org/10.1016/j.chom.2015.07.008>.

Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21):9546–51, 5 2010. ISSN 1091-6490. DOI: 10.1073/pnas.0914005107. URL <papers2://publication/uuid/69A203EF-BB54-4F53-B483-181AA4F41BC4><http://www.ncbi.nlm.nih.gov/pubmed/20460310><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2906865>.

Stephen Boyd and Lieven Vandenbergh. Convex Optimization. page 716, 2009. ISSN 10556788. DOI: 10.1109/TAC.2006.884922.

Roy J Britten and Eric H Davidson. Gene Regulation for Higher Cells : A Theory. *Science*, 165(3891):349–357, 1969.

Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10(6):2763–2788, 2009. ISSN 14220067. DOI: 10.3390/ijms10062763.

Michael J. Buck and Jason D. Lieb. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004. ISSN 08887543. DOI: 10.1016/j.ygeno.2003.11.004.

Joerg Martin Buescher, Wolfram Liebermeister, Matthieu Jules, Markus Uhr, Jan Muntel, Eric Botella, Bernd Hessling, Roelco Jacobus Kleijn, Ludovic Le Chat, François Lecointe, Ulrike M\”{a}der, Pierre Nicolas, Sjouke Piersma, Frank R\”{u}gheimer, Dörte Becher, Philippe Bessieres, Elena Bidnenko, Emma L Denham, Etienne Dervyn, Kevin M Devine, Geoff Doherty, Samuel Drulhe,

Liza Felicori, Mark J Fogg, Anne Goelzer, Annette Hansen, Colin R Harwood, Michael Hecker, Sebastian Hubner, Claus Hultschig, Hanne Jarmer, Edda Klipp, Aurélie Leduc, Peter Lewis, Frank Molina, Philippe Noirot, Sabine Peres, Nathalie Pigeonneau, Susanne Pohl, Simon Rasmussen, Bernd Rinn, Marc Schaffer, Julian Schnidder, Benno Schwikowski, Jan Maarten Van Dijl, Patrick Veiga, Sean Walsh, Anthony J Wilkinson, Jörg Stelling, Stéphane Aymerich, and Uwe Sauer. Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. *Science (New York, N.Y.)*, 335(6072):1099–103, 2012. ISSN 1095-9203. DOI: 10.1126/science.1206871. URL <http://www.sciencemag.org/content/335/6072/1099.abstract>.

Jennifer E F Butler and James T Kadonaga. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, 16(20):2583–2592, 2002. ISSN 0890-9369. DOI: 10.1101/gad.1026202.The. URL <http://www.ncbi.nlm.nih.gov/pubmed/12381658>.

Elisa Cabiscol, Jordi Tamarit, and Joaquim Ros. Oxidative stress in bacteria and protein damage by reactive oxygen species. *International Microbiology*, 3(1):3–8, 2010. ISSN 1618-1905. DOI: 10.2436/im.v3i1.9235. URL <http://130.206.88.107/revistes224/index.php/IM/article/view/4c457c0c498c3.002>.

Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, Sandrine L Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano, and Antonio Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279): 318–25, 1 2010. ISSN 1476-4687. DOI: 10.1038/nature08712. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4011561&tool=pmcentrez&rendertype=abstract>.

Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):459–471, 2014. ISSN 03051048. DOI: 10.1093/nar/gkt1103.

Colin J. Champion. Empirical Bayesian estimation of normal variances and covariances. *Journal of Multivariate Analysis*, 87(1):60–79,

10 2003. ISSN 0047259X. DOI: 10.1016/S0047-259X(02)00076-3. URL <http://linkinghub.elsevier.com/retrieve/pii/S0047259X02000763>.

A. Chatterjee, S. Dasgupta, and D. Sidransky. Mitochondrial Subversion in Cancer. *Cancer Prevention Research*, 4(5):638–654, 5 2011. ISSN 1940-6207. DOI: 10.1158/1940-6207.CAPR-10-0326. URL <http://cancerpreventionresearch.aacrjournals.org/cgi/doi/10.1158/1940-6207.CAPR-10-0326>.

Katherine C Chen, Laurence Calzone, Attila Csikasz-nagy, Frederick R Cross, Bela Novak, and John J Tyson. Integrative Analysis of Cell Cycle Control in Budding Yeast. 15(August):3841–3862, 2004. DOI: 10.1091/mbc.E03.

Yulong Chen, Jiong Deng, Junya Fujimoto, Humam Kadara, Taoyan Men, Dafna Lotan, and Reuben Lotan. Gprc5a deletion enhances the transformed phenotype in normal and malignant lung epithelial cells by eliciting persistent Stat3 signaling induced by autocrine leukemia inhibitory factor. *Cancer Research*, 70(21):8917–8926, 2010. ISSN 00085472. DOI: 10.1158/0008-5472.CAN-10-0518.

J Q Cheng, D a Altomare, M a Klein, W C Lee, G D Kruh, N a Lissy, and J R Testa. Transforming activity and mitosis-related expression of the AKT2 oncogene: evidence suggesting a link between cell cycle regulation and oncogenesis. *Oncogene*, 14(23):2793–2801, 1997. ISSN 09509232. DOI: 10.1038/sj.onc.1201121.

Alberto Chiarugi, Christian Dölle, Roberta Felici, and Mathias Ziegler. The NAD metabolome — a key determinant of cancer cell biology. *Nature Publishing Group*, 12(11):741–752, 2012. ISSN 1474-175X. DOI: 10.1038/nrc3340. URL <http://dx.doi.org/10.1038/nrc3340>.

Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140):140, 1 2007. ISSN 1744-4292. DOI: 10.1038/msb4100180. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2063581&tool=pmcentrez&rendertype=abstract>.

Hyonho Chun, Xianghua Zhang, and Hongyu Zhao. Gene regulation network inference with joint sparse Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 8600(September): 00–00, 2014. ISSN 1061-8600. DOI: 10.1080/10618600.2014.956876. URL <http://www.tandfonline.com/doi/abs/10.1080/10618600.2014.956876>.

Michael B Clark, Paulo P Amaral, Felix J Schlesinger, Marcel E Dinger, Ryan J Taft, John L. Rinn, Chris P Ponting, Peter F Stadler, Kevin V Morris, Antonin Morillon, Joel S Rozowsky, Mark B Gerstein, Claes Wahlestedt, Yoshihide Hayashizaki, Piero Carninci, Thomas R Gingeras, and John S. Mattick. The Reality of Pervasive Transcription. *PLoS Biology*, 9(7):e1000625, 7 2011. ISSN 1545-7885. DOI: 10.1371/journal.pbio.1000625. URL <http://dx.plos.org/10.1371/journal.pbio.1000625>.

Jesse C Cochrane and Scott a Strobel. Riboswitch effectors as protein enzyme cofactors. *RNA (New York, N.Y.)*, 14(6):993–1002, 2008. ISSN 1355-8382. DOI: 10.1261/rna.908408.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szczesniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 12 2016. ISSN 1474-760X. DOI: 10.1186/s13059-016-0881-8. URL <http://www.ncbi.nlm.nih.gov/pubmed/26813401><http://genomebiology.com/2016/17/1/13>.

Francis H. C. Crick. Central Dogma of Molecular Biology, 1970. ISSN 0028-0836. URL <http://dx.doi.org/10.1038/227561a0>.

Matthew McKnight Croken, Weigang Qiu, Michael W White, and Kami Kim. Gene Set Enrichment Analysis (GSEA) of *Toxoplasma gondii* expression datasets links cell cycle progression and the bradyzoite developmental program. *BMC genomics*, 15(1):515, 1 2014. ISSN 1471-2164. DOI: 10.1186/1471-2164-15-515. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4092224&tool=pmcentrez&rendertype=abstract>.

Alfred Csibi, Sarah-maria Fendt, Chenggang Li, George Poulgiannis, Andrew Y. Choo, Douglas J Chapski, Seung Min Jeong, Jamie M. Dempsey, Andrey Parkhitko, Tasha Morrison, Elizabeth P. Henske, Marcia C. Haigis, Lewis C Cantley, Gregory Stephanopoulos, Jane Yu, and John Blenis. The mTORC1 Pathway Stimulates Glutamine Metabolism and Cell Proliferation by Repressing SIRT4. *Cell*, 153(4):840–854, 5 2013. ISSN 00928674. DOI: 10.1016/j.cell.2013.04.023. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867413004650>.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(2):373–397, 11 2014. ISSN 13697412. DOI: 10.1111/rssb.12033. URL <http://arxiv.org/abs/1111.0324>.

Chi V Dang. MYC, Metabolism, Cell Growth, and Tumorigenesis. *Cold Spring Harbor Perspectives in Medicine*, 3(8):a014217–a014217, 8 2013. ISSN 2157-1422. DOI: 10.1101/cshperspect.a014217. URL <http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a014217>.

Eric H Davidson, Jonathan P Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, Ochan Otim, C Titus Brown, Carolina B Livi, Pei Yun Lee, Roger Revilla, Alistair G Rust, Zheng Jun Pan, Maria J Schilstra, Peter J C Clarke, Maria I Arnone, Lee Rowen, R Andrew Cameron, David R McClay, Leroy Hood, and Hamid Bolouri. A genomic regulatory network for development. *Science (New York, N.Y.)*, 295(5560):1669–78, 3 2002. ISSN 1095-9203. DOI: 10.1126/science.1069883. URL <http://www.ncbi.nlm.nih.gov/pubmed/11872831>.

M. H. de Smit and J. van Duin. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences*, 87(19):7668–7672, 1990. ISSN 0027-8424. DOI: 10.1073/pnas.87.19.7668. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.87.19.7668>.

Ralph J DeBerardinis, Anthony Mancuso, Evgueni Daikhin, Ilana Nissim, Marc Yudkoff, Suzanne Wehrli, and Craig B Thompson. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19345–50, 2007. ISSN 1091-6490. DOI: 10.1073/pnas.0709747104. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2148292&tool=pmcentrez&rendertype=abstract>.

Ralph J. DeBerardinis, Julian J. Lum, Georgia Hatzivassiliou, and Craig B. Thompson. The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. *Cell Metabolism*, 7(1):11–20, 2008. ISSN 15504131. DOI: 10.1016/j.cmet.2007.10.002.

Jiong Deng, Junya Fujimoto, X. F. Ye, T. Y. Men, C. S. Van Pelt, Y. L. Chen, X. F. Lin, Humam Kadara, Qingguo Tao, Dafna Lotan, and R. Lotan. Knockout of the Tumor Suppressor Gene Gprc5a in Mice Leads to NF- κ B Activation in Airway Epithelium and Promotes Lung Inflammation and Tumorigenesis. *Cancer Prevention Research*, 3(4):424–437, 4 2010. ISSN 1940-6207. DOI: 10.1158/1940-6207.CAPR-10-0032. URL <http://cancerpreventionresearch.aacrjournals.org/cgi/doi/10.1158/1940-6207.CAPR-10-0032>.

Youping Deng, Hongwei Wang, Ryuji Hamamoto, David Schaffer, and Shiwei Duan. Functional Genomics , Genetics , and Bioinformatics. 2015:10–12, 2015. DOI: 10.1155/2015/184824.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. ISSN 13674803. DOI: 10.1093/bioinformatics/bts635.

Adam Driks. Bacillus subtilis Spore Coat. *American Society for Microbiology*, 63(1):1–20, 1999.

NC Duarte and Scott a Becker. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777–1782, 2007. ISSN 0027-8424. DOI: 10.1073/pnas.0610772104. URL <http://www.pnas.org/content/104/6/1777.short>.

J. P. Dubey, D. S. Lindsay, and C. A. Speer. Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts. *Clinical Microbiology Reviews*, 11(2): 267–299, 1998. ISSN 08938512. DOI: PMC106833.

Rebecca L Elstrom, Daniel E Bauer, Monica Buzzai, Robyn Karnauskas, Marian H Harris, David R Plas, Hongming Zhuang, Ryan M Cinalli, Abass Alavi, Charles M Rudin, and Craig B Thompson. Akt Stimulates Aerobic Glycolysis in Cancer Cells. *Cancer Research*, 473(64):3892–3899, 2004.

The Encode and Project Consortium. Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. 447(June), 2007. DOI: 10.1038/nature05874.

Janan T. Eppig, Judith A. Blake, Carol J. Bult, James A. Kadin, Joel E. Richardson, A. Anagnostopoulos, R. P. Babiuk, R. M. Baldarelli, J. S. Beal, S. M. Bello, J. Berghout, O. Blodgett, N. E. Butler, L. E. Corbani, S. L. Cousins, H. Dene, H. J. Drabkin, K. L. Forthofer, P. Hale, L. Hutchins, M. Knowlton, M. Law, J. R. Lewis, M. McAndrews, D. S. Miers, H. Montencko, L. Ni, H. Onda, W. Pittman, J. M. Recla, D. J. Reed, B. Richards-Smith, D. Sitnikov, C. L. Smith, M. Tomczuk, L. L. Washburn, and Y. Zhu. The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, 43(D1):D726–D736, 2015. ISSN 13624962. DOI: 10.1093/nar/gku967.

Ayla Ergün, Carolyn a Lawrence, Michael a Kohanski, Timothy a Brennan, and James J Collins. A network biology approach to prostate cancer. *Molecular Systems Biology*, 3(82), 2 2007. ISSN 1744-4292. DOI: 10.1038/msb4100125. URL <http://msb.embopress.org/cgi/doi/10.1038/msb4100125>.

M. D. Ermolaeva, O White, and S L Salzberg. Prediction of operons in microbial genomes. *Nucleic acids research*, 29(5):1216–21, 3 2001. ISSN 1362-4962. DOI: 10.1093/nar/29.5.1216. URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.5.1216><http://www.ncbi.nlm.nih.gov/pubmed/11222772><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC29727>.

Abeer Fadda, Ana Carolina Fierro, Karen Lemmens, Pieter Monsieurs, Kristof Engelen, and Kathleen Marchal. Inferring the transcriptional network of *Bacillus subtilis*. *Molecular bioSystems*, 5(12):1840–52, 2009. ISSN 1742-2051. DOI: 10.1039/b907310h. URL <http://www.ncbi.nlm.nih.gov/pubmed/20023724>.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, 2007. ISSN 13674803. DOI: 10.1093/bioinformatics/btl567.

Valeria R. Fantin, Julie St-Pierre, and Philip Leder. Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 9(6):425–434, 2006. ISSN 15356108. DOI: 10.1016/j.ccr.2006.04.023.

Alessandro Fatica and Irene Bozzoni. Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, 15(1):7–21, 12 2013. ISSN 1471-0056. DOI: 10.1038/nrg3606. URL <http://dx.doi.org/10.1038/nrg3606><http://www.nature.com/doifinder/10.1038/nrg3606>.

Magdalena Franco, Anjali J. Shastri, and John C. Boothroyd. Infection by *Toxoplasma gondii* specifically induces host c-Myc and the genes this pivotal transcription factor regulates. *Eukaryotic Cell*, 13(4):483–493, 2014. ISSN 15359778. DOI: 10.1128/EC.00316-13.

R Franklin and R Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171:740–741, 1953.

Brendan J Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. (February), 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–41, 7 2008. ISSN 1468-4357. DOI:

10.1093/biostatistics/kxm045. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3019769&tool=pmcentrez&rendertype=abstract>.

N I R Friedman, Michal Linial, and Iftach Nachman. Using Bayesian Networks to Analyze Expression Data. 7(1998):601–620, 2000.

Masaya Fujita and Richard Losick. The master regulator for entry into sporulation in *Bacillus subtilis* becomes a cell-specific transcription factor after asymmetric division. *Genes and Development*, 17(9): 1166–1174, 2003. ISSN 08909369. DOI: 10.1101/gad.1078303.

Kenichiro Fujiwara, Inez Yuwanita, Daniel P. Hollern, and Eran R. Andrechek. Prediction and genetic demonstration of a role for activator E2Fs in Myc-induced tumors. *Cancer Research*, 71(5):1924–1932, 2011. ISSN 00085472. DOI: 10.1158/0008-5472.CAN-10-2386.

Gennaro Gambardella, Maria Nicoletta Moretti, Rossella De Cegli, Luca Cardone, Adriano Peron, and Diego Di Bernardo. Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14):1776–1785, 2013. ISSN 13674803. DOI: 10.1093/bioinformatics/btt290.

Et Al Gao, Ping. c-Myc suppression of miR-23 enhances mitochondrial glutaminase and glutamine metabolism. *Nature*, 458(7239):762–765, 2009. ISSN 1476-4687. DOI: 10.1038/nature07823.c-Myc.

Giuseppe Gasparre, Anna Maria Porcelli, Giorgio Lenaz, and Giovanni Romeo. Relevance of mitochondrial genetics and metabolism in cancer development. *Cold Spring Harbor perspectives in biology*, 5(2): 1–17, 2013. ISSN 1943-0264. DOI: 10.1101/cshperspect.a011411. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3552507&tool=pmcentrez&rendertype=abstract>.

Yoav Gilad, Scott A. Rifkin, and Jonathan K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, 24(8):408–415, 8 2008. ISSN 01689525. DOI: 10.1016/j.tig.2008.06.001. URL <http://linkinghub.elsevier.com/retrieve/pii/S0168952508001777>.

Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B R Azevedo, Rebecca A Zufall, and Eran Elhaik. On the Immortality of Television Sets : “ Function ” in the of ENCODE. *Genome biology Evolution*, 5 (3):578–590, 2013. DOI: 10.1093/gbe/evt028.

Marc Güell, Eva Yus, Maria Lluch-Senar, and Luis Serrano. Bacterial transcriptomics: what is beyond the RNA hori-zome? *Nature reviews. Microbiology*, 9(9):658–669, 2011. ISSN 1740-1526. DOI: 10.1038/nrmicro2620.

Cecilia Guerrier-Takada, Katheleen Gardiner, Terry Marsh, Norman Pace, and Sidney Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 PART 2):849–857, 1983. ISSN 00928674. DOI: 10.1016/0092-8674(83)90117-4.

John B. Gurdon. Transplanted Nuclei and Cell Differentiation. *Scientific American*, 219(6):24–35, 1968.

Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, 1 2000. ISSN 00928674. DOI: 10.1016/S0092-8674(00)81683-9. URL <http://www.ncbi.nlm.nih.gov/pubmed/10647931><http://link.springer.com/10.1007/s00262-010-0968-0><http://linkinghub.elsevier.com/retrieve/pii/S0092867400816839>.

Kasper D Hansen, Rafael a Irizarry, and Zhijin Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics (Oxford, England)*, 13(2):204–16, 4 2012. ISSN 1468-4357. DOI: 10.1093/biostatistics/kxr054. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3297825&tool=pmcentrez&rendertype=abstract>.

Johanna Hardin, Stephan Ramon Garcia, and David Golan. A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 7(3):1733–1762, 9 2013. ISSN 1932-6157. DOI: 10.1214/13-AOAS638. URL <http://projecteuclid.org/euclid.aoas/1380804814>.

K S Harker, N Ueno, and M B Lodoen. Toxoplasma gondii dissemination: a parasite’s journey through the infected host. *Parasite immunology*, 37(3):141–9, 3 2015. ISSN 1365-3024. DOI: 10.1111/pim.12163. URL <http://www.ncbi.nlm.nih.gov/pubmed/25408224>.

Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402 (December):47–52, 1999.

Georgia Hatzivassiliou, Fangping Zhao, Daniel E. Bauer, Charalambos Andreadis, Anthony N. Shaw, Dashyant Dhanak, Sunil R. Hingorani, David A. Tuveson, and Craig B. Thompson. ATP citrate lyase inhibition can suppress tumor cell growth. *Cancer Cell*, 8(4): 311–321, 2005. ISSN 15356108. DOI: 10.1016/j.ccr.2005.09.008.

Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–, 2 2014. ISSN 1940-9818. DOI: 10.2144/000114133. URL <http://>

www.biotechniques.com/BiotechniquesJournal/2014/February/Library-construction-for-next-generation-sequencing-Overviews-and-challenges/biotechniques-349889.html.

P Hegde, R Qi, C Gay, S Dharap, JE Hughes, E Snestrud, N Lee, and J Quackenbush. A concise guide to cDNA microarray analysis. *BioTechniques*, 29(3):548–562, 2000.

Tina M Henkin. Riboswitch RNAs : using RNA to sense cellular metabolism Riboswitch RNAs : using RNA to sense cellular metabolism. pages 3383–3390, 2008. DOI: 10.1101/gad.1747308.

Maria Sol Herrera-Cruz and Thomas Simmen. Cancer: Untethering Mitochondria from the Endoplasmic Reticulum? *Frontiers in Oncology*, 7(May):9–13, 2017. ISSN 2234-943X. DOI: 10.3389/fonc.2017.00105. URL <http://journal.frontiersin.org/article/10.3389/fonc.2017.00105/full>.

Nicholas J. Higham. Computing the nearest correlation matrix - A problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. ISSN 02724979. DOI: 10.1093/imanum/22.3.329.

Chung-Chau Hon, Jordan A. Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J. L. Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M. Poulsen, Jessica Severin, Marina Lizio, Hideya Kawaji, Takeya Kasukawa, Masayoshi Itoh, A. Maxwell Burroughs, Shohei Noma, Sarah Djebali, Tanvir Alam, Yulia A. Medvedeva, Alison C. Testa, Leonard Lipovich, Chi-Wai Yip, Imad Abugessaisa, Mickaël Mendez, Akira Hasegawa, Dave Tang, Timo Lassmann, Peter Heutink, Magda Babina, Christine A. Wells, Soichi Kojima, Yukio Nakamura, Harukazu Suzuki, Carsten O. Daub, Michiel J. L. de Hoon, Erik Arner, Yoshihide Hayashizaki, Piero Carninci, and Alistair R. R. Forrest. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 543(7644):199–204, 2017. ISSN 0028-0836. DOI: 10.1038/nature21374. URL <http://www.nature.com/doifinder/10.1038/nature21374>.

Nicholas J Hudson, Brian P Dalrymple, and Antonio Reverter. Beyond differential expression: the quest for causal mutations and effector molecules. *{BMC} Genomics*, 13(1):356, 7 2012. ISSN 1471-2164. DOI: 10.1186/1471-2164-13-356. URL <http://www.biomedcentral.com/1471-2164/13/356/abstract>.

Christopher a. Hunter and L. David Sibley. Modulation of innate immunity by *Toxoplasma gondii* virulence effectors. *Nature Reviews Microbiology*, 10(11):766–778, 2012. ISSN 1740-1526. DOI: 10.1038/nrmicro2858. URL <http://dx.doi.org/10.1038/nrmicro2858>.

- Trey Ideker and Nevan J Krogan. Differential network biology. *Molecular systems biology*, 8(565):565, 1 2012. ISSN 1744-4292. DOI: 10.1038/msb.2011.99. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3296360&tool=pmcentrez&rendertype=abstract>.
- Francesco Iorio, Timothy Rittman, Hong Ge, Michael Menden, and Julio Saez-Rodriguez. Transcriptional data: a new gateway to drug repositioning? *Drug discovery today*, 00(00), 8 2012. ISSN 1878-5832. DOI: 10.1016/j.drudis.2012.07.014. URL <http://www.ncbi.nlm.nih.gov/pubmed/22897878>.
- Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003. ISSN 1465-4644. DOI: 10.1093/biostatistics/4.2.249. URL <http://biostatistics.oxfordjournals.org/>.
- D.-E. Jeong, S.-H. Park, J.-G. Pan, E.-J. Kim, and S.-K. Choi. Genome engineering using a synthetic gene circuit in *Bacillus subtilis*. *Nucleic Acids Research*, 43(6):e42–e42, 2015. ISSN 0305-1048. DOI: 10.1093/nar/gku1380. URL <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1380>.
- H Jeong and R Albert. The large-scale organization of metabolic networks. 760(1990):651–654, 2000.
- M. E. Jerome, J. R. Radke, W. Bohne, D. S. Roos, and M. W. White. *Toxoplasma gondii* bradyzoites form spontaneously during sporozoite-initiated development. *Infection and Immunity*, 66(10):4838–4844, 1998. ISSN 00199567.
- Michael I. Jordan. Graphical Models. *Statistical Science*, 19(1): 140–155, 2 2004. ISSN 0883-4237. DOI: 10.1214/088342304000000026. URL <http://projecteuclid.org/euclid.ss/1089808279>.
- Caroline Jose, Nadège Bellance, and Rodrigue Rossignol. Choosing between glycolysis and oxidative phosphorylation: A tumor’s dilemma? *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1807(6):552–561, 6 2011. ISSN 00052728. DOI: 10.1016/j.bbabi.2010.10.012. URL [10.1016/j.bbabi.2010.10.012http://linkinghub.elsevier.com/retrieve/pii/S0005272810007206](http://linkinghub.elsevier.com/retrieve/pii/S0005272810007206).
- Alexander Jung, Gabor Hannak, and Norbert Goertz. Graphical LASSO based model selection for time series. *IEEE Signal Processing Letters*, 22(10):1781–1785, 2015. ISSN 10709908. DOI: 10.1109/LSP.2015.2425434.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016. ISSN 0305-1048. DOI: 10.1093/nar/gkv1070. URL <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1070>.

Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology*, 9(10): 770–80, 10 2008. ISSN 1471-0080. DOI: 10.1038/nrm2503. URL <http://www.ncbi.nlm.nih.gov/pubmed/18797474>.

Peter D Karp, Christos a Ouzounis, Caroline Moore-Kochlacs, Leon Goldovsky, Pallavi Kaipa, Dag Ahrén, Sophia Tsoka, Nikos Darzentas, Victor Kunin, and Núria López-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33(19):6083–9, 1 2005. ISSN 1362-4962. DOI: 10.1093/nar/gki892. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1266070&tool=pmcentrez&rendertype=abstract>.

Arek Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*, (Database 2011), 1 2011. ISSN 1758-0463. DOI: 10.1093/database/bar049. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3215098&tool=pmcentrez&rendertype=abstract>.

Ahmad S Khalil and James J Collins. Synthetic biology: applications come of age. *Nature reviews. Genetics*, 11(5):367–79, 5 2010. ISSN 1471-0064. DOI: 10.1038/nrg2775. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2896386&tool=pmcentrez&rendertype=abstract>.

Boris Kholodenko, Michael B Yaffe, and Walter Kolch. Computational approaches for analyzing information flow in biological networks. *Science signaling*, 5(220):re1, 4 2012. ISSN 1937-9145. DOI: 10.1126/scisignal.2002961. URL <http://www.ncbi.nlm.nih.gov/pubmed/22510471>.

J.-w. Kim, Ping Gao, Y.-C. Liu, Gregg L Semenza, and Chi V Dang. Hypoxia-Inducible Factor 1 and Dysregulated c-Myc Cooperatively Induce Vascular Endothelial Growth Factor and Metabolic Switches Hexokinase 2 and Pyruvate Dehydrogenase Kinase 1. *Molecular and Cellular Biology*, 27(21):7381–7393, 11 2007. ISSN 0270-7306. DOI: 10.1128/MCB.00440-07. URL <http://mcb.asm.org/cgi/doi/10.1128/MCB.00440-07>.

Kami Kim and Louis M Weiss. Toxoplasma gondii: the model apicomplexan. *International journal for parasitology*, 34(3):423–32, 3 2004. ISSN 0020-7519. DOI: 10.1016/j.ijpara.2003.12.009. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3086386&tool=pmcentrez&rendertype=abstract>.

Kami Kim and Louis M Weiss. Toxoplasma: the next 100years. *Microbes and infection / Institut Pasteur*, 10(9):978–84, 7 2008. ISSN 1286-4579. DOI: 10.1016/j.micinf.2008.07.015. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2596634&tool=pmcentrez&rendertype=abstract>.

Sunny Sun Kin Chan. What is a Master Regulator? *Journal of Stem Cell Research & Therapy*, 03(02):10–13, 2013. ISSN 21577633. DOI: 10.4172/2157-7633.1000e114. URL <http://www.omicsonline.org/what-is-a-master-regulator-2157-7633.1000e114.php?aid=12636>.

Teresia Kling, Patrik Johansson, José Sanchez, Voichita D. Marinescu, Rebecka Jörnsten, and Sven Nelander. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*, 43(15), 2015. ISSN 13624962. DOI: 10.1093/nar/gkv413.

J Kloehn, M Blume, SA Cobbold, EC Saunders, MJ Dagley, and MJ McConville. Using metabolomics to dissect host-parasite interactions. *Current Opinion in Microbiology*, 32:59–65, 8 2016. ISSN 13695274. DOI: 10.1016/j.mib.2016.04.019. URL <http://dx.doi.org/10.1016/j.mib.2016.04.019><http://linkinghub.elsevier.com/retrieve/pii/S1369527416300546>.

Tom Knight. Idempotent Vector Design for Standard Assembly of Biobricks. *MIT Libraries*, pages 1–11, 2003. DOI: <http://hdl.handle.net/1721.1/21168>. URL <http://dspace.mit.edu/handle/1721.1/45138>.

Dirk L. Knol and Jos M F ten Berge. Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, 54(1): 53–61, 1989. ISSN 00333123. DOI: 10.1007/BF02294448.

Hideki Kobayashi, Mads Kærn, Michihiro Araki, Kristy Chung, Timothy S Gardner, Charles R Cantor, and James J Collins. Programmable cells : Interfacing natural and engineered gene networks. 2004.

Willem H Koppenol, Patricia L Bounds, and Chi V Dang. Otto Warburg’s contributions to current concepts of cancer metabolism. *Nature Reviews Cancer*, 11(5):325–337, 5 2011. ISSN 1474-175X.

DOI: 10.1038/nrc3038. URL <http://www.nature.com/doi/10.1038/nrc3038>.

Roberta Kwok. Five hard truths for synthetic biology. *Nature*, 463 (January):288–290, 2010. URL <http://www.nature.com/news/2010/100120/full/463288a.html><http://europepmc.org/abstract/MED/20090726>.

Wi S Lai, Joel S Parker, Sherry F Grissom, Deborah J Stumpo, and Perry J Blackshear. Novel mRNA Targets for Tristetraprolin (TTP) Identified by Global Analysis of Stabilized Transcripts in TTP-Deficient Fibroblasts. *Molecular and Cellular Biology*, 26(24): 9196–9208, 12 2006. ISSN 0270-7306. DOI: 10.1128/MCB.00945-06. URL <http://mcb.asm.org/content/26/24/9196>. [shorthttp://mcb.asm.org/cgi/doi/10.1128/MCB.00945-06](http://mcb.asm.org/cgi/doi/10.1128/MCB.00945-06).

Julie Laliberte and Vb Carruthers. Host cell manipulation by the human pathogen *Toxoplasma gondii*. *Cellular and molecular life sciences*, 65(12):1900–1915, 2008. DOI: 10.1007/s00018-008-7556-x.Host. URL <http://www.springerlink.com/index/827u887508h4h75t.pdf>.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Showkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward

Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kasprzyk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2 2001. ISSN 0028-0836. DOI: 10.1038/35057062. URL <http://www.nature.com/doifinder/10.1038/35057062>.

Peter Langfelder, Paul S. Mischel, and Steve Horvath. When Is Hub

Gene Selection Better than Standard Meta-Analysis? *PLoS ONE*, 8 (4), 2013. ISSN 19326203. DOI: 10.1371/journal.pone.0061505.

I. Lee. A Probabilistic Functional Network of Yeast Genes. *Science*, 306(5701):1555–1558, 2004. ISSN 0036-8075. DOI: 10.1126/science.1099511. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1099511>.

Rosalind C Lee, Rhonda L. Feinbaum, and Victor Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, 12 1993. ISSN 00928674. DOI: 10.1016/0092-8674(93)90529-Y. URL <http://linkinghub.elsevier.com/retrieve/pii/009286749390529Y>.

Celine Lefebvre, Gabrielle Rieckhof, and Andrea Califano. Reverse-engineering human regulatory networks. *Wiley interdisciplinary reviews. Systems biology and medicine*, 4(4):311–25, 2012. ISSN 1939-005X. DOI: 10.1002/wsbm.1159. URL <http://www.ncbi.nlm.nih.gov/pubmed/22246697>.

P Lévy and B Bartosch. Metabolic reprogramming: a hallmark of viral oncogenesis. *Oncogene*, 35(32):4155–4164, 8 2016. ISSN 0950-9232. DOI: 10.1038/onc.2015.479. URL <http://www.nature.com/doifinder/10.1038/onc.2015.479>.

Shuang Li, Li Hsu, Jie Peng, and Pei Wang. Bootstrap inference for network construction with an application to a breast cancer microarray study. *Annals of Applied Statistics*, 7(1):391–417, 2013a. ISSN 19326157. DOI: 10.1214/12-AOAS589.

Xinyuan Li, Pu Fang, Jietang Mai, Eric T Choi, Hong Wang, and Xiao-feng Yang. Targeting mitochondrial reactive oxygen species as novel therapy for inflammatory diseases and cancers. *Journal of hematology & oncology*, 6(1):19, 2013b. ISSN 1756-8722. DOI: 10.1186/1756-8722-6-19. URL <http://link.springer.com/article/10.1186/1756-8722-6-19/fulltext.html>.

Y. Li and S. a. Jackson. Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3: Genes|Genomes|Genetics*, 5(6): 1075–1079, 6 2015. ISSN 2160-1836. DOI: 10.1534/g3.115.018127. URL <http://g3journal.org/cgi/doi/10.1534/g3.115.018127>.

Joe C. Liang, Ryan J. Bloom, and Christina D. Smolke. Engineering Biological Systems with Synthetic RNA Molecules. *Molecular Cell*, 43 (6):915–926, 2011. ISSN 10972765. DOI: 10.1016/j.molcel.2011.08.023. URL <http://dx.doi.org/10.1016/j.molcel.2011.08.023>.

Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014. ISSN 14602059. DOI: 10.1093/bioinformatics/btt656.

Maria V Liberti and Jason W Locasale. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends in Biochemical Sciences*, 41(3): 211–218, 3 2016. ISSN 09680004. DOI: 10.1016/j.tibs.2015.12.001. URL <http://dx.doi.org/10.1016/j.tibs.2015.12.001><http://linkinghub.elsevier.com/retrieve/pii/S0968000415002418>.

Wendell Lim, Connie Lee, and Chao Tang. Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell. *Molecular cell*, 49(2):202–212, 2013. DOI: 10.1016/j.molcel.2012.12.020.

MaGeou-Yarh Liou and Peter Storz. *Reactive oxygen species in cancer*, volume 44. 2010. ISBN 1071576100. DOI: 10.3109/10715761003667554.Reactive.

JW Locasale and LC Cantley. Metabolic flux and the regulation of mammalian cell growth. *Cell metabolism*, 14(4):443–451, 2011. DOI: 10.1016/j.cmet.2011.07.014.Metabolic. URL <http://www.sciencedirect.com/science/article/pii/S1550413111003457>.

S Lu, C X Ren, Y Liu, and D E Epner. P13K-Akt signaling is involved in the regulation of p21(WAF/CIP) expression and androgen-independent growth in prostate cancer cells. *International Journal of Oncology*, 28(1):245–251, 2006.

Michael J Lukey, Kai Su Greene, Jon W Erickson, Kristin F Wilson, and Richard A Cerione. The oncogenic transcription factor c-Jun regulates glutaminase expression and sensitizes cells to glutaminase-targeted therapy. *Nature Communications*, 7:11321, 2016. ISSN 2041-1723. DOI: 10.1038/ncomms11321. URL <http://www.nature.com/doifinder/10.1038/ncomms11321>.

Zhao-Rong Lun, De-Hua Lai, Yan-Zi Wen, Ling-Ling Zheng, Ji-Long Shen, Ting-Bo Yang, Wen-Liang Zhou, Liang-Hu Qu, Geoff Hide, and Francisco J Ayala. Cancer in the parasitic protozoans *Trypanosoma brucei* and *Toxoplasma gondii*. *Proceedings of the National Academy of Sciences of the United States of America*, 112(29):8835–42, 7 2015. ISSN 1091-6490. DOI: 10.1073/pnas.1502599112. URL <http://www.ncbi.nlm.nih.gov/pubmed/26195778>.

John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):

34, 2011. ISSN 1741-7007. DOI: 10.1186/1741-7007-9-34. URL <http://www.biomedcentral.com/1741-7007/9/34>.

T.G. Mamedov, E. Pienaar, S.E. Whitney, J.R. TerMaat, G. Carvill, R. Goliath, A. Subramanian, and H.J. Viljoen. A fundamental study of the PCR amplification of GC-rich DNA templates. *Computational Biology and Chemistry*, 32(6):452–457, 12 2008. ISSN 14769271. DOI: 10.1016/j.compbiolchem.2008.07.021. URL <http://linkinghub.elsevier.com/retrieve/pii/S1476927108000881>.

J. R. Managbanag, Tarynn M. Witten, Danail Bonchev, Lindsay A. Fox, Mitsuhiro Tsuchiya, Brian K. Kennedy, and Matt Kaeberlein. Shortest-path network analysis is a useful approach toward indentifying genetic determinants of longevity. *PLoS ONE*, 3(11), 2008. ISSN 19326203. DOI: 10.1371/journal.pone.0003802.

Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6286–91, 4 2010. ISSN 1091-6490. DOI: 10.1073/pnas.0913357107. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2851985&tool=pmcentrez&rendertype=abstract>.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 8 2012. ISSN 1548-7105. DOI: 10.1038/nmeth.2016. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3512113&tool=pmcentrez&rendertype=abstract>.

Florian Markowetz. How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS computational biology*, 6(2):e1000655, 3 2010. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1000655. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2829042&tool=pmcentrez&rendertype=abstract>.

Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21(21):4026–4032, 2005. ISSN 13674803. DOI: 10.1093/bioinformatics/bti662.

Florian Markowetz, Dennis Kostka, Olga G Troyanskaya, and Rainer Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–i312, 7 2007. ISSN 1367-4803,

1460-2059. DOI: 10.1093/bioinformatics/btm178. URL <http://bioinformatics.oxfordjournals.org/content/23/13/i305>.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucl. Acids Res.*, 34 (suppl_1):108–110, 2006. ISSN 1362-4962. DOI: 10.1093/nar/gkj143.

Matthew T Menendez, Crystal Teygong, Kristin Wade, Celia Florimond, and Ira J. Blader. siRNA Screening Identifies the Host Hexokinase 2 (HK2) Gene as an Important Hypoxia-Inducible Transcription Factor 1 (HIF-1) Target Gene in *Toxoplasma gondii* - Infected Cells. *mBio*, 6(3):00462–15, 7 2015. ISSN 2150-7511. DOI: 10.1128/mBio.00462-15. URL <http://mbio.asm.org/lookup/doi/10.1128/mBio.00462-15>.

Raphael H Michna, Fabian M Commichau, Dominik Tödter, Christopher P Zschiedrich, and Jörg Stülke. SubtiWiki—a database for the model organism *Bacillus subtilis* that links pathway, interaction and expression information. *Nucleic acids research*, pages 1–7, 10 2013. ISSN 1362-4962. DOI: 10.1093/nar/gkt1002. URL <http://www.ncbi.nlm.nih.gov/pubmed/24178028>.

Marija Milacic, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D’Eustachio, and Lincoln Stein. Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers*, 4(4):1180–1211, 2012. ISSN 20726694. DOI: 10.3390/cancers4041180.

Bryan W Miller, Garnet Lau, Chris Grouios, Emanuela Mollica, Miriam Barrios-Rodiles, Yongmei Liu, Alessandro Datti, Quaid Morris, Jeffrey L Wrana, and Liliana Attisano. Application of an integrated physical and functional screening approach to identify inhibitors of the Wnt pathway. *Molecular systems biology*, 5:315, 1 2009a. ISSN 1744-4292. DOI: 10.1038/msb.2009.72. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2779086&tool=pmcentrez&rendertype=abstract>.

Catherine M Miller, Nicola R Boulter, Rowan J Ikin, and Nicholas C Smith. The immunobiology of the innate response to *Toxoplasma gondii*. *International journal for parasitology*, 39(1):23–39, 1 2009b. ISSN 1879-0135. DOI: 10.1016/j.ijpara.2008.08.002. URL <http://www.ncbi.nlm.nih.gov/pubmed/18775432>.

Robert E. Molestina, Nadia El-guendy, and Anthony P. Sinai. Infection with *Toxoplasma gondii* results in dysregulation of the host cell

cycle. *Cellular Microbiology*, 10(5):1153–1165, 2008. ISSN 14625814. DOI: 10.1111/j.1462-5822.2008.01117.x.

Sach Mukherjee and Terence P. Speed. Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318, 2008. ISSN 1091-6490. DOI: 10.1073/pnas.0802272105. URL <http://www.pnas.org/content/105/38/14313.abstract>.

Luis Muniz-Feliciano, Jennifer Van Grol, Jose Andres C Portillo, Lloyd Liew, Bing Liu, Cathleen R. Carlin, Vern B. Carruthers, Stephen Matthews, and Carlos S. Subauste. Toxoplasma gondii-Induced Activation of EGFR Prevents Autophagy Protein-Mediated Killing of the Parasite. *PLoS Pathogens*, 9(12):1–15, 2013. ISSN 15537374. DOI: 10.1371/journal.ppat.1003809.

Vivek K Mutalik, Joao C Guimaraes, Guillaume Cambray, Colin Lam, Marc Juul Christoffersen, Quynh-Anh Mai, Andrew B Tran, Morgan Paull, Jay D Keasling, Adam P Arkin, and Drew Endy. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature methods*, 10(4):354–60, 2013. ISSN 1548-7105. DOI: 10.1038/nmeth.2404. URL <http://www.ncbi.nlm.nih.gov/pubmed/23474465>.

N. Nandagopal and M. B. Elowitz. Synthetic Biology: Integrated Gene Circuits. *Science*, 333(6047):1244–1248, 2011. ISSN 0036-8075. DOI: 10.1126/science.1207084. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1207084>.

Pierre Nicolas, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, Elodie Marchadier, Mark Hoebeke, Stéphane Aymerich, Dörte Becher, Paola Bisicchia, Eric Botella, Olivier Delumeau, Geoff Doherty, Emma L Denham, Mark J Fogg, Vincent Fromion, Anne Goelzer, Annette Hansen, Elisabeth Härtig, Colin R Harwood, Georg Homuth, Hanne Jarmer, Matthieu Jules, Edda Klipp, Ludovic Le Chat, François Lecoq, Peter Lewis, Wolfram Liebermeister, Anika March, Ruben a T Mars, Priyanka Nannapaneni, David Noone, Susanne Pohl, Bernd Rinn, Frank Rügheimer, Praveen K Sappa, Franck Samson, Marc Schaffer, Benno Schwikowski, Leif Steil, Jörg Stülke, Thomas Wiegert, Kevin M Devine, Anthony J Wilkinson, Jan Maarten van Dijl, Michael Hecker, Uwe Völker, Philippe Bessières, and Philippe Noirot. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science (New York, N. Y.)*, 335(6072):1103–6, 3 2012. ISSN 1095-9203. DOI: 10.1126/science.1206848. URL <http://www.ncbi.nlm.nih.gov/pubmed/22383849>.

Naoyo Nishida, Hirohisa Yano, Takashi Nishida, Toshiharu Kamura, and Masamichi Kojiro. Angiogenesis in cancer. *Vascular Health and Risk Management*, 2(3):213–219, 2006. ISSN 11766344. DOI: 10.2147/vhrm.2006.2.3.213.

Richard Nitzsche, Zagoriy Vyacheslav, Richard Lucius, and X Nishith Gupta. Metabolic Cooperation of Glucose and Glutamine Is Essential for the Lytic Cycle of Obligate Intracellular Parasite. *The Journal of Biological Chemistry*, 291(1):126–141, 2016. DOI: 10.1074/jbc.M114.624619.

Victor Nizet and Randall S Johnson. Interdependence of hypoxic and innate immune responses. *Nature reviews. Immunology*, 9(9):609–17, 2009. ISSN 1474-1741. DOI: 10.1038/nri2607. URL <http://www.nature.com.ezproxy.library.wisc.edu/nri/journal/v9/n9/abs/nri2607.html>.

Vincent Noireaux, Yusuke T Maeda, and Albert Libchaber. Development of an artificial cell, from self-organization to computation and self-reproduction. *Proceedings of the National Academy of Sciences*, 108(9):3473–3480, 3 2011. ISSN 0027-8424. DOI: 10.1073/pnas.1017075108. URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1017075108>.

Ingram Olkin and J.W. Pratt. Unbiased estimation of certain correlation coefficients. *The annals of mathematical statistics*, 29(1):201–211, 1958. ISSN 00034851. DOI: 10.2307/2237306.

Mark S. Paget. Bacterial sigma factors and anti-sigma factors: Structure, function and distribution. *Biomolecules*, 5(3):1245–1265, 2015. ISSN 2218273X. DOI: 10.3390/biom5031245.

Alexander F Palazzo and Eliza S Lee. Non-coding RNA : what is functional and what is junk ? 6(January):1–11, 2015. DOI: 10.3389/fgene.2015.00002.

Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 10 2009. ISSN 1471-0056. DOI: 10.1038/nrg2641.

D Pe’er, a Regev, G Elidan, and N Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S215–S224, 2001. ISSN 1367-4803. DOI: 10.1093/bioinformatics/17.suppl1.S215.

Dana Pe’er and Nir Hacohen. Principles and strategies for developing network models in cancer. *Cell*, 144(6):864–73, 3 2011. ISSN 1097-4172. DOI: 10.1016/j.cell.2011.03.001. URL

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3082135&tool=pmcentrez&rendertype=abstract>.

Christine Peterson, Marina Vannucci, Cemal Karakas, William Choi, Lihua Ma, and Mirjana Maletic-Savatic. Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. *Statistics and Its Interface*, 6(4):547–558, 2013. ISSN 19387989. DOI: 10.4310/SII.2013.v6.n4.a12. URL <http://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0006/0004/a012/>.

R. Petryszak, T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, J. McMurry, J. C. Marioni, J. Malone, K. Megy, G. Rustici, A. Y. Tang, J. Taubert, E. Williams, O. Mannion, H. E. Parkinson, and A. Brazma. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 42(D1):D926–D932, 2014. ISSN 0305-1048. DOI: 10.1093/nar/gkt1270. URL <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1270>.

David Polsky and Carlos Cordon-Cardo. Oncogenes in melanoma. *Oncogene*, 22(20):3087–3091, 2003. ISSN 0960-8931. DOI: 10.1097/00008390-199309002-00008.

Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003. ISSN 13674803. DOI: 10.1093/bioinformatics/btg148.

Priscilla E M Purnick and Ron Weiss. The second wave of synthetic biology: from modules to systems. *Nature reviews. Molecular cell biology*, 10(6):410–22, 2009. ISSN 1471-0080. DOI: 10.1038/nrm2698. URL <http://www.ncbi.nlm.nih.gov/pubmed/19461664>.

Arjun Raj and Alexander van Oudenaarden. Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008. DOI: 10.1016/j.cell.2008.09.050.

Krishnaraj Rajalingam, Ralf Schreck, Ulf R. Rapp, and Štefan Albert. Ras oncogenes and their downstream targets. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1773(8):1177–1195, 8 2007. ISSN 01674889. DOI: 10.1016/j.bbamcr.2007.01.012. URL <http://linkinghub.elsevier.com/retrieve/pii/S0167488907000286>.

P.D. Ray, B.W. Huang, and Y. Tsuji. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell Signal*, 24(5):981–990, 2012. DOI: 10.1016/j.cellsig.2012.01.008.Reactive.

David N Reshef, Yakir a Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science (New York, N.Y.)*, 334(6062):1518–24, 12 2011. ISSN 1095-9203. DOI: 10.1126/science.1205438. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3325791&tool=pmcentrez&rendertype=abstract>.

John L Rinn, Michael Kertesz, Jordon K Wang, Sharon L Squazzo, Xiao Xu, Samantha A Brugmann, L Henry Goodnough, Jill A Helms, Peggy J Farnham, Eran Segal, and Howard Y Chang. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*, 129(7):1311–1323, 6 2007. ISSN 00928674. DOI: 10.1016/j.cell.2007.05.022. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867407006599>.

J David Robertson. Membrane Structure. *The Journal of Cell Biology*, 91(3):189–204, 1981.

Guillermo Rodrigo, Javier Carrera, and Alfonso Jaramillo. Genetdes: Automatic design of transcriptional networks. *Bioinformatics*, 23(14):1857–1858, 2007. ISSN 13674803. DOI: 10.1093/bioinformatics/btm237.

Simon Rogers, Mark Girolami, Walter Kolch, Katrina M Waters, Tao Liu, Brian Thrall, and H Steven Wiley. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics (Oxford, England)*, 24(24):2894–900, 12 2008. ISSN 1367-4811. DOI: 10.1093/bioinformatics/btn553. URL <http://www.ncbi.nlm.nih.gov/pubmed/18974169>.

Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Mirosław Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Jon Ison, Maria Keays, Natalja Kurbatova, James Malone, Roby Mani, Annalisa Mupo, Rui Pedro Pereira, Ekaterina Pilicheva, Johan Rung, Anjan Sharma, Y Amy Tang, Tobias Ternent, Andrew Tikhonov, Danielle Welter, Eleanor Williams, Alvis Brazma, Helen Parkinson, and Ugis Sarkans. ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic acids research*, 41(Database issue):987–90, 1 2013. ISSN 1362-4962. DOI: 10.1093/nar/gks1174. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531147&tool=pmcentrez&rendertype=abstract>.

Müşerref Duygu Saçar, Caner Bağcı, and Jens Allmer. Computational Prediction of MicroRNAs from *Toxoplasma gondii* Potentially Regulating the Hosts' Gene Expression. *Genomics, proteomics & bioinformatics*, 12(5):228–238, 2014. ISSN 2210-3244. DOI:

10.1016/j.gpb.2014.09.002. URL <http://www.ncbi.nlm.nih.gov/pubmed/25462155>.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(22 April 2005):523–529, 2005. ISSN 0036-8075. DOI: 10.1126/science.1105809.

Julio Saez-Rodriguez, Leonidas G Alexopoulos, Jonathan Epperlein, Regina Samaga, Douglas a Lauffenburger, Steffen Klamt, and Peter K Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology*, 5(331), 1 2009. ISSN 1744-4292. DOI: 10.1038/msb.2009.87. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2824489&tool=pmcentrez&rendertype=abstract>.

M Salazar, M Lorente, E García-Taboada, E Pérez Gómez, D Dávila, P Zúñiga-García, J María Flores, a Rodríguez, Z Hegedus, D Mosén-Ansorena, a M Aransay, S Hernández-Tiedra, I López-Valero, M Quintanilla, C Sánchez, J L Iovanna, N Dusetti, M Guzmán, S E Francis, a Carracedo, E Kiss-Toth, and G Velasco. Loss of Tribbles pseudokinase-3 promotes Akt-driven tumorigenesis via FOXO inactivation. *Cell death and differentiation*, 22(1):131–44, 2015. ISSN 1476-5403. DOI: 10.1038/cdd.2014.133. URL <http://www.ncbi.nlm.nih.gov/pubmed/25168244>.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):Article32, 1 2005. ISSN 1544-6115. DOI: 10.2202/1544-6115.1175. URL <http://www.ncbi.nlm.nih.gov/pubmed/16646851>.

Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Y.-W. W. Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 1 2012. ISSN 0305-1048. DOI: 10.1093/nar/gkr972. URL <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr972>.

R. A. Sclafani and T. M. Holzen. Cell Cycle Regulation of DNA Replication. *Annual Review of Genetics*, 41(1):237–280, 2007. ISSN 0066-4197. DOI: 10.1146/annurev.genet.41.110306.130308. URL <http://www.annualreviews.org/doi/10.1146/annurev.genet.41.110306.130308>.

Alexander Serganov and Dinshaw J Patel. Ribozymes , riboswitches and beyond : regulation of gene expression without proteins. 8 (September):776–790, 2007. DOI: 10.1038/nrg2172.

P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13 (11):2498–2504, 11 2003. ISSN 1088-9051. DOI: 10.1101/gr.1239303. URL <http://www.genome.org/cgi/doi/10.1101/gr.1239303>.

Paul T Shannon, Mark Grimes, Burak Kutlu, Jan J Bot, and David J Galas. RCytoscape: tools for exploratory network analysis. *BMC bioinformatics*, 14(1):217, 1 2013. ISSN 1471-2105. DOI: 10.1186/1471-2105-14-217. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3751905&tool=pmcentrez&rendertype=abstract>.

Rob Shields. Cultural Topology: The Seven Bridges of Königsburg, 1736. *Theory, Culture & Society*, 29(4-5):43–57, 2012. ISSN 0263-2764. DOI: 10.1177/0263276412451161. URL <http://journals.sagepub.com/doi/10.1177/0263276412451161>.

Nicolas Sierro, Yuko Makita, Michiel de Hoon, and Kenta Nakai. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic acids research*, 36(Database issue):93–6, 1 2008. ISSN 1362-4962. DOI: 10.1093/nar/gkm910. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247474&tool=pmcentrez&rendertype=abstract>.

Vijai Singh. Recent advancements in synthetic biology: Current status and challenges. *Gene*, 535(1):1–11, 2014. ISSN 03781119. DOI: 10.1016/j.gene.2013.11.025. URL <http://dx.doi.org/10.1016/j.gene.2013.11.025>.

Sini Skariah and Dana G Mordue. Identification of *Toxoplasma gondii* genes responsive to the host immune response during in vivo infection. *PLoS one*, 7(10):e46621, 1 2012. ISSN 1932-6203. DOI: 10.1371/journal.pone.0046621. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3468626&tool=pmcentrez&rendertype=abstract>.

Sini Skariah, Robert B Bednarczyk, Matthew K McIntyre, Gregory A Taylor, and Dana G Mordue. Discovery of a novel *Toxoplasma gondii* conoid-associated protein important for parasite resistance to reactive nitrogen intermediates. *Journal of immunology (Baltimore, Md. : 1950)*, 188(7):3404–15, 2012. ISSN 1550-6606. DOI: 10.4049/jimmunol.1101425. URL <http://www.jimmunol.org/content/188/7/3404.full>.

Richard N. Smith, Jelena Aleksic, Daniela Butano, Adrian Carr, Sergio Contrino, Fengyuan Hu, Mike Lyne, Rachel Lyne, Alex Kalderimis, Kim Rutherford, Radek Stepan, Julie Sullivan, Matthew Wake-ling, Xavier Watkins, and Gos Micklem. InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28(23):3163–3165, 2012. ISSN 13674803. DOI: 10.1093/bioinformatics/bts577.

Michael E Smoot, Keiichiro Ono, Johannes Ruscseinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, 2 2011. ISSN 1367-4811. DOI: 10.1093/bioinformatics/btq675. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3031041&tool=pmcentrez&rendertype=abstract>.

Carl Song, Melissa a Chiasson, Nirvana Nursimulu, Stacy S Hung, James Wasmuth, Michael E Grigg, and John Parkinson. Metabolic reconstruction identifies strain-specific regulation of virulence in *Toxoplasma gondii*. *Molecular systems biology*, 9(708):708, 2013. ISSN 1744-4292. DOI: 10.1038/msb.2013.62. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4039375&tool=pmcentrez&rendertype=abstract>.

Wade Spear, Denise Chan, Isabelle Coppens, Randall S Johnson, Amato Giaccia, and Ira J Blader. The host cell transcription factor hypoxia-inducible factor 1 is required for *Toxoplasma gondii* growth and survival at physiological oxygen levels. *Cellular Microbiology*, 8(2):339–352, 2 2006. ISSN 1462-5814. DOI: 10.1111/j.1462-5822.2005.00628.x. URL <http://doi.wiley.com/10.1111/j.1462-5822.2005.00628.x>.

Emma Steele and Allan Tucker. Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. *Journal of biomedical informatics*, 41(6):914–26, 12 2008. ISSN 1532-0480. DOI: 10.1016/j.jbi.2008.01.011. URL <http://www.ncbi.nlm.nih.gov/pubmed/18337190>.

John Storey. THE POSITIVE FALSE DISCOVERY RATE : A BAYESIAN INTERPRETATION AND THE q-VALUE. *The Annals of Statistics*, 31(6):2013–2035, 2003.

J. M. Stuart. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255, 2003. ISSN 0036-8075. DOI: 10.1126/science.1087447. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1087447>.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, and Benjamin L Ebert. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005. ISSN 0027-8424. DOI: 10.1073/pnas.0506580102.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 12 2007. ISSN 0090-5364. DOI: 10.1214/009053607000000505. URL <http://projecteuclid.org/euclid.aos/1201012979>.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian Von Mering. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015. ISSN 13624962. DOI: 10.1093/nar/gku1003.

The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, 489(7414):57–74, 2012. DOI: 10.1038/nature11247.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, 5 2009. ISSN 1367-4811. DOI: 10.1093/bioinformatics/btp120. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672628&tool=pmcentrez&rendertype=abstract>.

Julio F Turrens. Mitochondrial formation of reactive oxygen species. *The Journal of physiology*, 552(Pt 2):335–44, 2009. ISSN 0022-3751. DOI: 10.1113/jphysiol.2003.049478. URL <http://www.ncbi.nlm.nih.gov/pubmed/14561818>.

Iphigenia Tzamelis. The evolving role of mitochondria in metabolism. *Trends in Endocrinology & Metabolism*, 23(9):417–419, 9 2012. ISSN 10432760. DOI: 10.1016/j.tem.2012.07.008. URL <http://dx.doi.org/10.1016/j.tem.2012.07.008><http://linkinghub.elsevier.com/retrieve/pii/S1043276012001361>.

A. Valouev, Ds D.S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9):829–834, 2008. ISSN 1548-7091. DOI: 10.1038/nmeth.1246.Genome-Wide. URL <http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.1246.html>.

Harm van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLoS Biology*, 8(5):e1000371, 5 2010. ISSN 1545-7885. DOI: 10.1371/journal.pbio.1000371. URL <http://dx.plos.org/10.1371/journal.pbio.1000371>.

Jan Maarten van Dijl and Michael Hecker. Bacillus subtilis: from soil bacterium to super-secreting cell factory. *Microbial cell factories*, 12:3, 2013. ISSN 1475-2859. DOI: 10.1186/1475-2859-12-3. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564730&tool=pmcentrez&rendertype=abstract>.

Marjolein van Sluis and Brian McStay. Ribosome biogenesis: Achilles heel of cancer? *Genes & cancer*, 5(5-6):152–3, 2014. ISSN 1947-6019. URL <http://www.ncbi.nlm.nih.gov/pubmed/25061498>.

M. G. Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science*, 324(5930):1029–1033, 5 2009. ISSN 0036-8075. DOI: 10.1126/science.1160809. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1160809>.

Roman Vershynin. How Close is the Sample Covariance Matrix to the Actual Covariance Matrix? *Journal of Theoretical Probability*, 25(3): 655–686, 2012. ISSN 08949840. DOI: 10.1007/s10959-010-0338-z.

Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome Networks and Human Disease. *Cell*, 144(6):986–998, 3 2011. ISSN 00928674. DOI: 10.1016/j.cell.2011.02.016. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867411001309>.

J. Wang, X. Zhou, P. L. Bradley, S.-F. Chang, N. Perrimon, and S. T.C. Wong. Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening. *Journal of Biomolecular Screening*, 13(1):29–39, 2007. ISSN 1087-0571. DOI: 10.1177/1087057107311223. URL <http://jbx.sagepub.com/cgi/doi/10.1177/1087057107311223>.

Kai Wang, Masumichi Saito, Brygida C Bisikirska, Mariano J Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A Margolin, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature Biotechnology*, 27(9):829–837, 2009a. ISSN 1087-0156. DOI: 10.1038/nbt.1563. URL <http://www.nature.com/nbt/journal/v27/n9/abs/nbt.1563.html>.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 1 2009b. ISSN 1471-0064. DOI: 10.1038/nrg2484. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract>.

J.D Watson and F.H.C Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

Dina R. Weillhammer, Anthony T. Iavarone, Eric N. Villegas, George A. Brooks, Anthony P. Sinai, and William C. Sha. Host metabolism regulates growth and differentiation of *Toxoplasma gondii*. *International Journal for Parasitology*, 42(10):947–959, 9 2012. ISSN 00207519. DOI: 10.1016/j.ijpara.2012.07.011. URL <http://linkinghub.elsevier.com/retrieve/pii/S0020751912001968>.

Wade C. Winkler, Ali Nahvi, Adam Roth, Jennifer A. Collins, and Ronald R. Breaker. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, 428(6980):281–286, 2004. ISSN 0028-0836. DOI: 10.1038/nature02362. URL <http://www.nature.com/doi/10.1038/nature02362>.

P J Wittkopp and G Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet*, 13(1):59–69, 2012. ISSN 1471-0064 (Electronic)1471-0056 (Linking). DOI: 10.1038/nrg3095. URL <http://www.ncbi.nlm.nih.gov/pubmed/22143240>.

Yin Xia, Tianxi Cai, and T. Tony Cai. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2):247–266, 2015. ISSN 14643510. DOI: 10.1093/biomet/asu074.

Jianchun Xiao, Lorraine Jones-Brando, C. Conover Talbot, and Robert H. Yolken. Differential effects of three canonical *Toxoplasma* strains on gene expression in human neuroepithelial cells. *Infection and Immunity*, 79(3):1363–1373, 2011. ISSN 00199567. DOI: 10.1128/IAI.00947-10.

Ninghan Yang, Andrew Farrell, Wendy Niedelman, Mariane Melo, Diana Lu, Lindsay Julien, Gabor T Marth, Marc-Jan Gubbels, and Jeroen P J Saeij. Genetic basis for phenotypic differences between different *Toxoplasma gondii* type I strains. *BMC genomics*, 14(1): 467, 1 2013. ISSN 1471-2164. DOI: 10.1186/1471-2164-14-467. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3710486&tool=pmcentrez&rendertype=abstract>.

Shumin Yang and Stephen F. Parmley. *Toxoplasma gondii* expresses two distinct lactate dehydrogenase homologous genes during its life cycle in intermediate hosts. *Gene*, 184(1):1–12, 1997. ISSN 03781119. DOI: 10.1016/S0378-1119(96)00566-5.

H Zhou and I Rigoutsos. The emerging roles of GPRC5A in diseases. *Oncoscience*, 1(12):765–776, 2014. ISSN 2331-4737; 2331-4737. DOI: 10.18632/oncoscience.104.

Zhong Zhou, Yi Shen, Muhammad Riaz Khan, and Ao Li. LncReg: A reference resource for lncRNA-associated regulatory networks. *Database*, 2015(17):1–7, 2015. ISSN 17580463. DOI: 10.1093/database/bav083.