THE UNIVERSITY *of* EDINBURGH

Mining Large Collections Of Gene Expression Data To Elucidate Transcriptional Regulation Of Biological Processes

Edward Curry

A Thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2010

Contents

Abstract

A vast amount of gene expression data is available to biological researchers. As of October 2010, the GEO database has 45,777 chips of publicly available gene expression profiling data from the Affymetrix (HGU133v2) GeneChip platform, representing 2.5 billion numerical measurements. Given this wealth of data, 'meta-analysis' methods allowing inferences to be made from combinations of samples from different experiments are critically important. This thesis explores the application of localized pattern-mining approaches, as exemplified by biclustering, for large-scale gene expression analysis. Biclustering methods are particularly attractive for the analysis of large compendia of gene expression data as they allow the extraction of relationships that occur only across subsets of genes and samples. Standard correlation methods, however, assume a single correlation relationship between two genes occurs across all samples in the data. There are a number of existing biclustering methods, but as these did not prove suitable for large scale analysis, a novel method named 'IslandCluster' was developed. This method provided a framework for investigating the results of different approaches to biclustering meta-analysis.

The biclustering methods used in this work involve preprocessing of gene expression data into a unified scale in order to assess the significance of expression patterns. A novel discretisation approach is shown to identify distinct classes of genes' expression values more appropriately than approaches reported in the literature. A Gene Expression State Transformation ('GESTr') introduced as the first reported modelling of the biological state of expression on a unified scale and is shown to facilitate effective meta-analysis. Localised co-dependency analysis is introduced, a paradigm for identifying transcriptional relationships from gene expression data. Tools implementing this analysis were developed and used to analyse specificity of transcriptional relationships, to distinguish related subsets within a set of transcription factor (TF) targets and to tease apart combinatorial regulation of a set of targets by multiple TFs. The state of pluripotency, from which a mammalian cell has the potential to differentiate into any cell from any of the three adult germ layers, is maintained by forced expression of Nanog and may be induced from a non-pluripotent state by the expression of Oct4, Sox2, Klf4 and cMyc. Analysis of cMyc regulatory targets shed light on a recent proposition that cMyc induces an 'embryonic stem cell like' transcriptional signature outside embryonic stem (ES) cells, revealing a cMyc-responsive subset of the signature and identifying ES cell expressed targets with evidence of broad cMyc-induction. Regulatory targets through which cMyc, Oct4, Sox2 and Nanog may maintain or induce pluripotency were identified, offering insight into transcriptional mechanisms involved in the control of pluripotency and demonstrating the utility of the novel analysis approaches presented in this work.

Acknowledgements

Many thanks to Dr. Simon Tomlinson for his supervision of my research degree, for providing guidance when needed but ultimately giving me the opportunity to figure out for myself what I should be doing. I am very grateful for having been able to learn so much about independently carrying out research and about providing constructive scientific criticism of my own and other peoples' work. Thanks also to my second supervisor Prof. Ian Chambers and my committee chair Dr. Tilo Kunath, particularly for the help they provided in the final stages of the project that involved setting out defined objectives to meet before submitting my thesis. Thanks to John, Ajay, Adrian, Laura, Florian, Sofia, Ben, Aodhan and Hina in the Stem Cell Bioinformatics group at the ISCR for their friendly support and input in discussions of my work. Further thanks to all staff and students at the ISCR and the SCRM for providing me with support during my own work and for creating an environment with so many opportunities to learn about stem cell biology and experimental research in molecular biology.

Thanks to my examiners, Dr. Andrew Harrison and Dr. Lesley Forrester, for their helpful suggestions for improving this thesis and publishing the work it describes.

I am very grateful to Prof. Hani Gabra, Prof. Bob Brown and all at the Ovarian Cancer Action research centre at Imperial College for appointing me to my first post-doctoral position before this thesis was completed, and for allowing me time to work on this project whenever it was required.

Many thanks to my parents for providing me with accommodation while I was in Edinburgh, and to Mr & Mrs Watts for providing me with accommodation in London while I was finishing off the work and writing this thesis. I will obviously always be grateful to my parents for their support of my education and encouragement during my work towards the PhD. I will also always be grateful to Vaughan sticking by me through the tough times, for so nobly putting up with being a 'PhD widow' and for still being there to meet me with open arms at the end. Thanks to Patrick, Bruce and Charles for giving me a great environment to go home to, and to the EURS, EUTC, EUSOG and EDGAS for providing distractions that help retain perspective and focus during such an absorbing endeavour as a PhD through scientific research.

Funding for this project was provided primarily by the Medical Research Council of the UK, additional funding was provided by the EuroSyStem Project.

Nomenclature

ANOVA Analysis of variance

AUC Area under curve

BGA Biclustering genetic algorithm

BIC Bayes information criterion

BicAT Biclustering analysis toolbox [Barkow et al., 2006]

BiMax Exhaustive biclustering algorithm [Prelic et al., 2006]

CC Biclustering algorithm based on local search techniques [Cheng and Church, 2000]

CDF Cumulative distribution function

ChIP Chromatin immunoprecipitation

CLT Central limit theorem

ComBiclust Combinatorial biclustering algorithm (described in Section 3.4.4)

DAVID Database for annotation, visualisation and integrated discovery [Huang et al., 2009b]

DEG Differentially-expressed gene

DNA Deoxyribonucleic acid

EBI European Bioinformatics Institute

EC Embyonal carcinoma

ECAT Embryonic stem cell associated transcript

EM Expectation maximisation

Entropy-based MNC BGA Entropy-based biclustering GA (described in Section 3.6.2)

ES Embyonic stem

Extrapolation Averaging Adaptation of RMA to large datasets [Goldstein, 2006]

FABIA Factor analysis for bicluster acquisition [Hochreiter et al., 2010]

FDR False discovery rate

FWER Family-wise error rate

GA Genetic algorithm

GDepBGA Guidegene-dependent biclustering GA (described in Section 3.6.3)

GEO Gene Expression Omnibus [Edgar et al., 2002]

GESTr Gene expression state transformation (described in Section 4.4)

GMM Gaussian mixture model

GNF Genomics Institute of the Novartis Research Foundation

GO Gene ontology [Ashburner et al., 2000]

HBLCA Heuristic biclustering method for localised co-dependency analysis (described in Section 5.1, summarised in Section 5.1.6)

ICM Inner cell mass

iPS Induced pluripotent stem

ISA Iterative signature algorithm [Bergmann et al., 2003]

IslandCluster GA approach to biclustering (described in Section 3.4.5)

K-Means Data clustering strategy with predefined number of clusters [Hartigan, 1975]

LDA Linear discriminant analysis

LIF Leukaemia inhibitory factor

LIMMA Linear models for microarray analysis [Smyth, 2004]

LiTAL Linear-tail adjusted Laplace distribution (described in Section 5.1.2)

ML Maximum likelihood

MLE Maximum likelihood estimate

MNC Multi-niche crowding GA [Cedeno et al., 1994]

MNC BGA Multi-niche crowding GA for biclustering (described in Section 3.6.1)

mRNA Messenger RNA

NCBI National Center for Biotechnology Information

OPSM Order preserving submatrix algorithm [Ben-Dor et al., 2004]

PCA Principal components analysis [Massy, 1965]

PDF Probability density function

Plaid Biclustering algorithm based on local search techniques [Lazzeroni and Owen, 2002]

qRT-PCR Quantitative reverse transcriptase polymerase chain reaction

RAM Random access memory

RefRMA Adaptation of RMA to large datasets [Katz et al., 2006]

RMA Robust multi-array average [Irizarry et al., 2003]

RNA Ribonucleic acid

ROC Receiver operating characteristic

SAM Significance analysis of microarrays [Tusher et al., 2001]

SAMBA Statistical-algorithmic method for bicluster analysis [Tanay et al., 2002]

SOM Self-organising map

TF Transcription factor

TFAS Transcription factor association score [Ouyang et al., 2009]

TranSAM Significance analysis for GESTr-transformed microarray datasets (described in Section 4.3.1)

XMotifs Motif-based biclustering algorithm [Murali and Kasif, 2003]

Chapter 1

Introduction

Over the past 20 years, technologies for simultaneous measurement of the abundance of large numbers of transcripts have been widely used in biological research. The development and adoption of standardised measurement platforms has led to the public availability of comparable datasets reporting expression levels of tens of thousands of genes in tens of thousands of biological samples, all measured in the same way. There is a wealth of transcriptional information contained in these massed datasets with billions of data points, but study of the application of pattern-mining techniques for the extraction of information relevant to a particular biological query is in its infancy. This thesis concerns the study of the application of pattern-mining techniques to large collections of gene expression data, with focused application to the elucidation of transcriptional mechanisms involved in the control of pluripotency.

This work could be considered multidisciplinary in that it draws heavily on theory from traditionally distinct fields of research. The context of this work is described via the general background information provided in the following chapter covering pluripotency and ES cell biology, transcriptomics, single study- and meta-analysis of gene expression data, clustering approaches and heuristic search techniques for function optimisation. Chapters 3-6 each introduce, describe and discuss a particular aspect of the work carried out, namely: investigation of a biclustering approach to large-scale meta-analysis of gene expression data, universal modelling of biological states of gene expression, gene expression co-dependency analysis tools, and investigation of transcriptional mechanisms of the control of pluripotency with the developed tools. A final discussion and general summary of the work is given in Chapter 7.

Chapter 2

Background

2.1 Pluripotent Stem Cells

As the work described in this thesis relates to pluripotent stem cells, at this point I clarify for the reader my interpretation of the definition of the terms 'pluripotent' and 'stem cell.'

In brief, a stem cell is a cell which can self-renew indefinitely or differentiate into a cell(s) of a specified lineage. Pluripotency is the ability of a cell to give rise to differentiated cells from any of the three primary germ layers (and primordial germ cells). Therefore, a pluripotent stem cell is a cell which can self-renew indefinitely or differentiate into any cell from any of the three primary germ layers (and primordial germ cells).

As the definition of a pluripotent stem cell refers to differentiation and lineage specification, an overview of this developmental context is provided in the remainder of Section (2.1.1). Sections (2.1.2-2.1.4) provide descriptions of specific examples of pluripotent stem cells.

2.1.1 Early Development of the Mammalian Embryo

While an animal develops continuously throughout its life cycle, it is the earliest stages of development that are most relevant to the work presented in this thesis, where there are still single cells present that can give rise to all adult tissues, and the first steps in which these cells become specified in their lineage and thus undergo restriction of their developmental potential. In these stages of development the animal is referred to as an embryo. Two stages post-fertilization are especially important here, cleavage and gastrulation. A more detailed description of the processes mentioned here is provided in [Gilbert, 2006].

Cleavage

The zygote (fertilised egg) is a large diploid cell with a high volume of cytoplasm. The first steps towards generating a complex organism (such as a mammal) from this single cell involve a number of rapid mitotic cell divisions with little or no synthesis of new organic material (other than DNA), thereby dividing the volume of the zygotic cytoplasm amongst a large number of early embryonic cells ('blastomeres'). These divisions form a 2-,4- or 8-cell structures and later, following 'compaction' by formation of cell adhesion complexes, a structure (referred to as a morula) comprising an internal group of cells surrounded by a larger, external group [Barlow and Sherman, 1972]. The embryo's 'inner cell mass' arises from the internal cells of the 16-cell morula, along with some cells dividing from the outer cells during its transition to a 32-cell stage. By the 64-cell stage, the outer cells have specialised into trophoblast cells which have become a separate compartment layer from the inner cell mass (with correspondibly

different developmental potential). Figure 2.1 (from [Gilbert, 2006]) shows the process of cleavage in a single mouse embryo, from 2-cell stage (A) through to morula (E) and blastocyst (F).

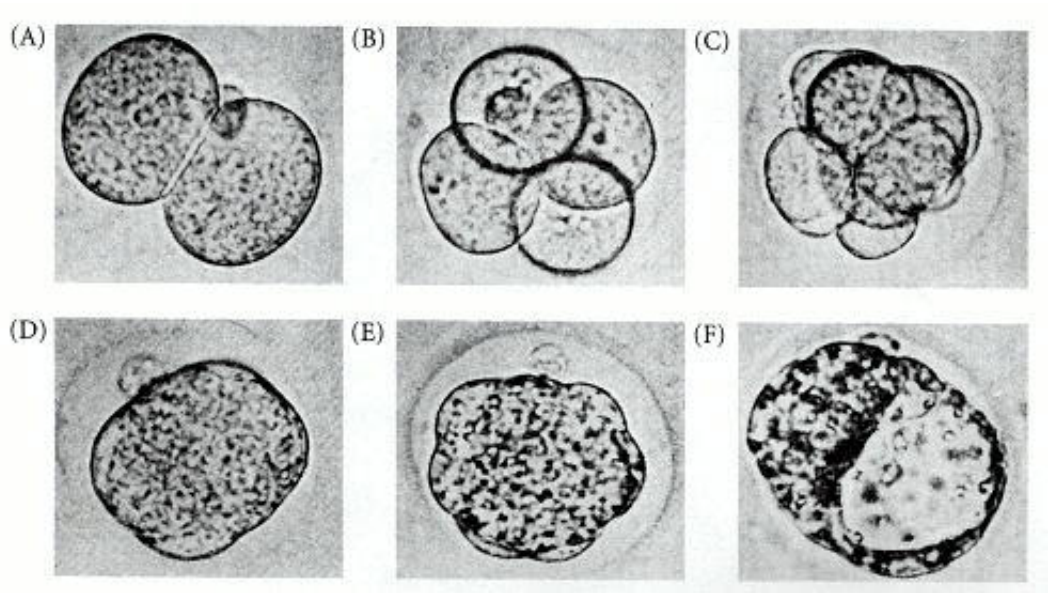


Figure 2.1: Cleavage of a mouse embryo, taken from [Gilbert, 2006]

The end result of cleavage in mammals is a 'blastocyst' comprising an outer layer (trophectoderm) of epithelial-like cells surrounding a fluid-filled cavity (blastocoel) and an inner cell mass (ICM). This is formed through secretion of fluid into the morula from the trophoblast cells during a process known as 'cavitation.' A diagram representing a blastocyst is shown in Figure 2.2 (taken from [Arnold and Robertson, 2009]).

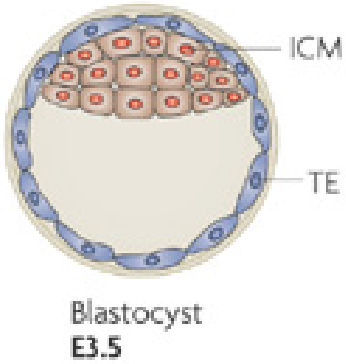


Figure 2.2: Diagram of an E3.5 mouse blastocyst, from [Arnold and Robertson, 2009]. Trophectoderm labelled TE and inner cell mass labelled ICM.

Gastrulation and Further Embryonic Development

Following formation of the blastocyst, observable structures begin to form as different cells in the embryo begin to specialise and segregate during the course of gastrulation. Through organised specification of the epiblast cells, the ICM segregates into three 'germ' layers that will later give rise to all the tissues of the adult organism. These are:

- Ectoderm: outer layer of the ICM, will eventually form skin and nervous system
- Mesoderm: middle layer, will give rise to bone, muscle, cartilage and blood
- Endoderm: interior layer of ICM, gives rise to internal organs and gut

2.1.2 Teratomas, Teratocarcinomas and Embryonal Carcinoma Cells

Teratomas are naturally occurring benign tumours that contain structures representing all three germ layers. Teratocarcinomas are their malignant counterparts, which will invade host tissue, metastasize and may continue to grow until the host organism dies. In a landmark study [Kleinsmith and Pierce, 1964] a single cell of origin was isolated that can give rise to these tumours. Given that there is a single cell responsible for generating tissues from all germ layers, it would suggest that these cells are pluripotent, and thus if these cells can be clonally expanded in culture they make a potentially useful tool to study developmental processes. For a review discussing these tumours and their potential as a model by which to study developmental processes, see [Martin, 1975].

However, it should be considered that these undifferentiated stem cells responsible for teratocarcinoma formation and propagation, known as embryonal carcinoma (EC) cells, are clearly some way removed from their counterpart pluripotent cells present in normal development. For example, these EC cells typically do not differentiate well *in vitro*, fail to contribute to chimaeras upon blastocyst injection and commonly show aneuploidy [Chambers and Smith, 2004].

2.1.3 Embryonic Stem Cells

Embryonic stem (ES) cells are pluripotent cells, derived from cells of the ICM of a blastocyst (prior to specification into the germ layers), that can be expanded indefinitely in culture. They were first derived in the mouse by Evans & Kaufman [Evans and Kaufman, 1981] and Martin [Martin, 1981], and from human embryos by Thomson et al [Thomson, 1998]. An important characteristic of ES cells is that they can contribute to chimeras when injected into blastocysts and reimplanted into the womb of a surrogate mother.

Uses of ES Cells

Considerable interest in ES cells has been shown from fields of biological and medical research, owing to their wide range of potential uses. To the end of outlining the motivation for studying ES cells, there follows a brief summary of three such areas in which ES cells have application:

Genetically Modified Animals With the ability of ES cells to contribute to the germ cells of chimeric offspring, ES cells that are genetically altered *in vitro* and subsequently implanted into a blastocyst at the appropriate stage of development can give rise to viable offspring carrying the desired genetic modification. When such (heterozygous) chimeric offspring are mated, inbred strains of animals can be cultivated that carry this genetic modification through their progeny. 'Transgenic' strains of mice arising from the introduction of genetic material from another organism have proven to be one of the most useful tools for biological research in recent years [Lewandowski, 2001].

Models of Development and Disease The potential of ES cells to be used to study developmental processes, due to their ability to be expanded in vitro and differentiated into any adult lineage, is discussed in [Rossant, 2008]. Additionally, in conjunction with their use to generate transgenic strains of animals, their application to study the development of diseases is reviewed in [Murray and Keller, 2008].

Drug Development Drug development often involves the use of cell cultures to test the effectiveness of therapies. The capacity for clonal expansion of ES cells means that derivation of target cells from ES cells via defined in vitro differentiation protocols provides an easier alternative than accessing primary tissue. For a review of some of the ways in which ES cells have been used in drug discovery, see [McNeish, 2004].

Regenerative Medicine ES cells have potential application in the field of regenerative medicine, owing to the fact that many human diseases arise from a deficiency of certain critical cell populations. As a result, the derivation and expansion of such defined cell populations in vitro from ES cells would give the potential to cure such diseases with ex vivo growth and subsequent transplantation of the relevant cells [Murray and Keller, 2008].

2.1.4 Transcriptional Control of Pluripotency

ES cells must maintian a state that is transitory in normal embryonic development by carefully balancing a large number of transcriptional events and signalling pathways that can serve as cues to drive the cells either to proliferate via self-renewal or to differentiate into more specified lineages with restricted developmental potential.

For maintainance of the pluripotent state of mouse ES cells in culture (in the absence of feeder cells or genetic modification), they require either stimulation by LIF (Leukemia Inhibitory Factor) [Smith et al., 1988] and Bmp4 (a 'Bone Morphogenic Protein') [Ying et al., 2003] or by a set of inhibitor molecules that block autonomously-generated stimuli to differentiate [Ying et al., 2008]. LIF stimulates the Stat3 signalling pathway while Bmp4 activates transcription of Id ('Inhibitor of Differentiation') genes. The '3i' ES cell culture medium [Ying et al., 2008] works by blocking the Fgf4-mediated activation of Erk signalling and Gsk3β. A diagram showing proposed mechanisms of action is shown in Fig 2.3 (from [Ying et al., 2008]).

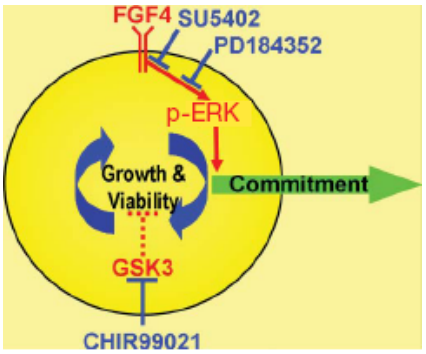


Figure 2.3: Diagram showing proposed inhibitory action of '3i' culture medium, from [Ying et al., 2008]

A number of transcription factors have been shown to be critical for efficient maintainance of ES cell identity, particularly Oct4, Nanog and Sox2. Of these, both Oct4 [Nichols et al., 1998] and Nanog [Chambers et al., 2003, Mitsui et al., 2003] are required during mouse development for specification of pluripotent cell identity. It has been suggested that the reason Sox2 is not required could be due to persistence of the maternal sox2 protein [Chambers and Tomlinson, 2009]. A summary of the balance of intrinsic and extrinsic factors required to maintain pluripotency of ES cells in self-renewal is shown in Fig 2.4 (from [Chambers and Smith, 2004]).

Forced over-expression of Nanog [Chambers et al., 2003], Klf2 [Hall et al., 2009], Esrrb [Zhang et al., 2008] but not Oct4 [Niwa et al., 2000] have been shown to drive LIF-independent self-renewal of ES cells.

Nanog is a transcription factor specifically expressed in ES cells, and binds to DNA through a single homeodomain. It was demonstrated in [Chambers et al., 2003] that over-expression of Nanog confers LIF-independent self-renewal of ES cells. It was shown in [Chambers et al., 2007] that ES cells express Nanog at varying levels, and those with lower Nanog expression have a higher propensity for differentiation. Interestingly, in this study it was also established that ES cell self-renewal is not dependent on the expression of Nanog.

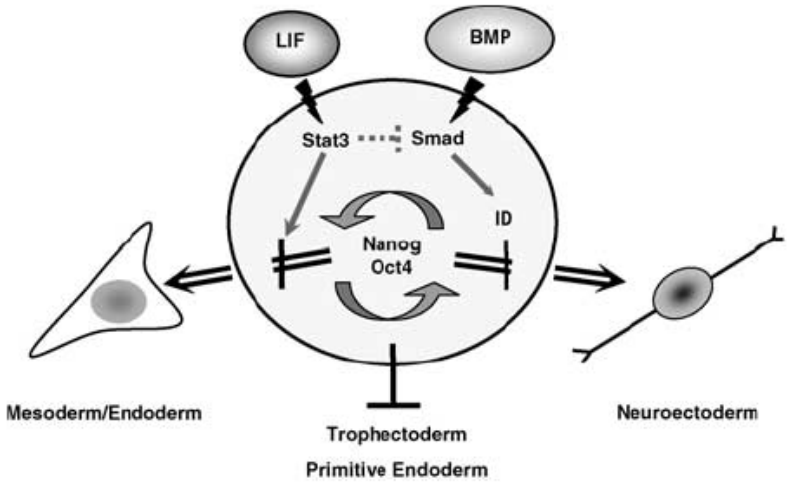


Figure 2.4: Diagrammatic representation of core ES cell self-renewal signalling, taken from [Chambers and Smith, 2004]

Oct4 is a transcription factor that interacts with DNA through both a low-affinity DNA-binding domain (specific to 'Pit,' 'Oct' and 'Unc' (POU) TFs) and a higher affinity homeodomain [Klemm and Pabo, 1996].

Sox2 interacts with DNA through its high mobility group (HMG) box domain. Sox2 is known to bind DNA co-operatively with Oct4 [Ambrosetti et al., 1997], with mechanisms of co-operative Oct4/Sox2/DNA binding proposed in [Remenyi et al., 2003] and [Williams et al., 2003].

It has been shown that Fgf4 is required for ES cells to differentiate, via activation of the Erk signalling that seems to drive differentiation [Kunath et al., 2007].

Identifying the targets of those transcription factors critical to maintainance of the pluripotent ES cell state will be a crucial step in understanding the underlying biological mechanisms involved in regulating this state and making those cell fate decisions undertaken during early embryonic development.

2.1.5 Induction Of Pluripotency

In a reversal of the unidirectional progression through states with restricted developmental potential that is encountered in normal development[1], adult cells can be shown to have been 'reprogrammed' through experimental techniques into states resembling those of the early embryo. This reprogramming has been achieved through a number of different processes:

[1] An exception to this 'rule' occurs during germ cell development [Hajkova et al., 2002]

- somatic cell nuclear transfer
- cell fusion
- cell explantation
- forced expression of defined factors

In addition to offering potential avenues for developing therapeutics such as patient-derived transplants (avoiding immune-related incompatibility and issues of transplant rejection), these *in vitro* reprogramming techniques present an opportunity to study the mechanisms by which different transcriptional programs are activated, maintained and deactivated.

Each of the methods of induction of pluripotency have their own issues, but a great deal of effort over the last few years has been put into improving the efficiency and understanding of the process of inducing pluripotency in somatic cells via induction of specified transcription factors. For this reason, the following section provides a brief overview of methods for induction of pluripotency by defined factors and theories regarding their mechanisms of operation.

From a set of 24 predominantly ES-specifically expressed genes (identified by 'digital differential expression' in [Mitsui et al., 2003]), Takahashi and Yamanaka identified 4 factors (Oct4, Sox2, Klf4 and cMyc) whose retrovirally-induced expression was sufficient to reprogram adult mouse fibroblasts into an ES-like state [Takahashi and Yamanaka, 2006]. Successful reprogramming occurrs in a very low proportion of those cells that are transfected with the viral vectors in this way, which necessitates some sort of selection criterion. Selection based on reactivation of transcription of endogenous Nanog or Oct4 appears to give rise to induced-pluripotent stem (iPS) cells that are similar to ES cells [Okita et al., 2007]. However, selection of cells based on expression of the embryonic marker SSEA1 results in cells that may not have been fully reprogrammed to an induced pluripotent state, as SSEA1 expression occurs prior to expression of Oct4 and Nanog during the reprogramming process [Brambrink et al., 2008]. Even selection based on the ES-cell expressed (but non-essential) gene Fbxo15 results in iPS cells that, while able to give rise to teratomas, are unable to contribute to chimeras and thus do not have the full potential of ES cells [Okita et al., 2007]. Some such reprogramming experiments have identified 'partially-reprogrammed' cell lines that have neither entirely silenced expression of their differentiated cell specific genes nor activated expression of pluripotency genes, possibly explained by incomplete demethylation of the endogenous pluripotency genes that has been observed in [Mikkelsen et al., 2008, Bhutani et al., 2010]. These findings seem to suggest that there are certain barriers that must be overcome before the origin cells' differentiated program can be fully silenced and the complete ES

cell transcriptional state reactivated by reversal of the pre-existing (epigenetic) silencing [Hochedlinger and Plath, 2009].

In vitro reprogramming experiments provide an opportunity to study the processes of cell fate specification (both in terms of a cell becoming more specialised or with the increase in potential seen as a 'reversal' of differentiation process) as well as a more detailed look at the transcriptional requirements of the pluripotent cell state. For example, as forced expression of Nanog in ES cells is sufficient to maintain pluripotent identity, but in somatic cells it is not sufficient to 'regain' pluripotency (unless in combination with other factors[Yul et al., 2007] (in human ES cells)), this indicates that it will be important for us to understand what barriers to Nanog's transcriptional targets are in place in differentiated cells and how these might be overcome in the reprogramming process. Additionally, for iPS cells to be used for clinical (therapeutic) purposes, the reprogramming process will need to be optimised so that it is more efficient and better-controlled, and this will be greatly assisted by a deeper understanding of the precise mechanisms involved.

2.2 Transcriptomics

The majority of the cells of an organism have the same genetic information encoding all the proteins the organism may use, but contain vastly different sets of proteins, mediated by the differences in which genes are transcribed into RNA. Measurement of the levels of the different mRNAs in a cell (or population of cells) offers a representation of the functional state of the cell given the (approximately) consistent information within.

In the advent of the sequencing of organisms' entire genomes, it has been possible to predict possible transcripts that the organism may produce in the expression of its genome, which in turn has led to the development of platforms with the ability to measure the entire state of transcription within cells. This state of transcription includes both the abundance of different trancripts within the cell and the precise state of the transcriptional apparatus (e.g. organisation of DNA into accessible or inaccessible chromatin, binding of transcriptional enhancers or repressors to DNA, etc.) that determines which genes can and will be transcribed. Analogous to the study of an organism's entire genetic content being referred to as 'genomics,' the study of the entire transcriptional state of a sample is referred to as 'transcriptomics.'

The remainder of this chapter discusses some of the transcriptomic technologies widely used in biological research, and approaches to the analysis in which the measurements they produce can be harnessed to further our understanding of the mechanims and consequences of transcriptional regulation in different biological samples.

2.2.1 Microarrays

Microarray technology involves using utilizing hybridization (or binding) properties of a large number of specifically designed 'probes' that are located on some surface in such a way that the corresponding targets can be identified after hybridization. It also relies on the labelling of prepared sample material with fluorescent dyes and quantification of fluorescence on the array by imaging and computer-based image processing. Due to the large numbers of probes used, microarrays have enabled measurements to be made on a 'whole-genome' scale.

Since its introduction in the 1990's[Schena et al., 1995], microarray technology has been adapted to a range of biological applications, including:

- measuring mRNA levels ('gene expression microarrays')
- identifying locations of DNA bound by transcription factors ('ChIP-on-chip')
- detecting chromosomal copy-number variation
- identifying single-nucleotide polymorphisms in genomic DNA
- identifying DNA methylation

Due to their ability to measure levels of transcription of all (protein-coding) genes in an organism's genome, gene expression microarrays have been widely used in biological research as a screening tool to identify genes (or transcripts) that may be responsible for observed phenomena, or for identifying markers of a particular process or cell type. This has resulted in a vast body of data reflecting measurements of whole-genome transcription in hundreds of thousands of biological samples.

Gene Expression Microarrays

Microarrays can be used to quantify the abundance of a large number of specified transcripts (mRNAs) through probes designed to hybridise uniquely to a particular sequence, which are positioned on the array at known locations. One of the most widely used gene expression microarray technologies has been the Affymetrix GeneChip, originally described in [Lockhart et al., 1996]. The GeneChips each feature in the order of 500,000 probe oligonucleotides that are synthesised onto the array at specified grid co-ordinates. Until the most recent platforms based on this technology, probes were designed in pairs with identical sequences aside from a central base, which on one of the probes in the pair was swapped for its complementary base. Thus one probe is a perfect match to the target sequence and the other is a mismatch probe included on the array to estimate cross-hybridisation to transcripts other than the target. Perfect match probe sequences are chosen so that the target sequences of a set of 10-20 probe pairs (perfect match & mismatch) map (ideally uniquely) to subsequences within a

consensus sequence for a transcript, derived from public annotation databases (see [Stalteri and Harrison, 2007] for more details). Each set of probe pairs is known as a probe set. The measurements from the individual probe pairs are summarised to obtain an expression level estimate for the transcript represented by the probeset.

The intensity measurements from a microarray can depend on features of the sample preparation process, the manufacture of the array, the hybridization process and the fluorescence quantification, in addition to the property of interest: the abundance of each transcript in the sample [Hartemink et al., 2001]. When comparing expression measurements from different arrays it is therefore appropriate to 'normalize' the measurements to reduce as much of the variation due to technical reasons (rather than biological differences between the samples). A demonstration of the need for normalization of Affymetrix GeneChip data is given in [Irizarry et al., 2003], along with the description of the widely-used robust multi-array average (RMA) measure of expression from GeneChips. RMA corrects for array-specific background intensity, performs quantile normalization to ensure that the distribution of intensity values across each array is the same, then uses a simple additive linear model to estimate the expression level of each transcript based on the normalized measurements from each probe in the corresponding probe set and each probe's specific hybridization affinity estimates. Owing to the normalization of probe-level measurements, RMA can only be applied to normalize measurements from different arrays of the same platform (so that the same probes are present on all arrays to be normalized). For a full description of RMA in the context of alternative GeneChip measurement strategies, see [Irizarry et al., 2003].

2.2.2 Data Warehouses

Proliferation of the use of microarrays in biological research has resulted in the publication of thousands of datasets, often releasing into the public domain the raw measurement values from these experiments. A number of databases have been created to store and make available the data from these experiments, some of which have grown into vast repositories. Two of the largest are the NCBI's GEO [Edgar et al., 2002] and the EBI's ArrayExpress [Brazma et al., 2003].

2.2.3 High-Throughput ChIP

As we currently do not have a general understanding of the ways in which TFs influence the transcription of target genes, especially when combinations of multiple TFs are considered, it is helpful to identify the locations of DNA-binding of the TFs. To achieve this, microarray technology has been adapted to work in concert with chromatin-immunoprecipitation techniques. When a biological sample is cultured with a tagged form of the TF in question, the DNA in the sample can be fragmented after cross-linking of protein-DNA binding. Using antibodies that bind to the tag on the TF,

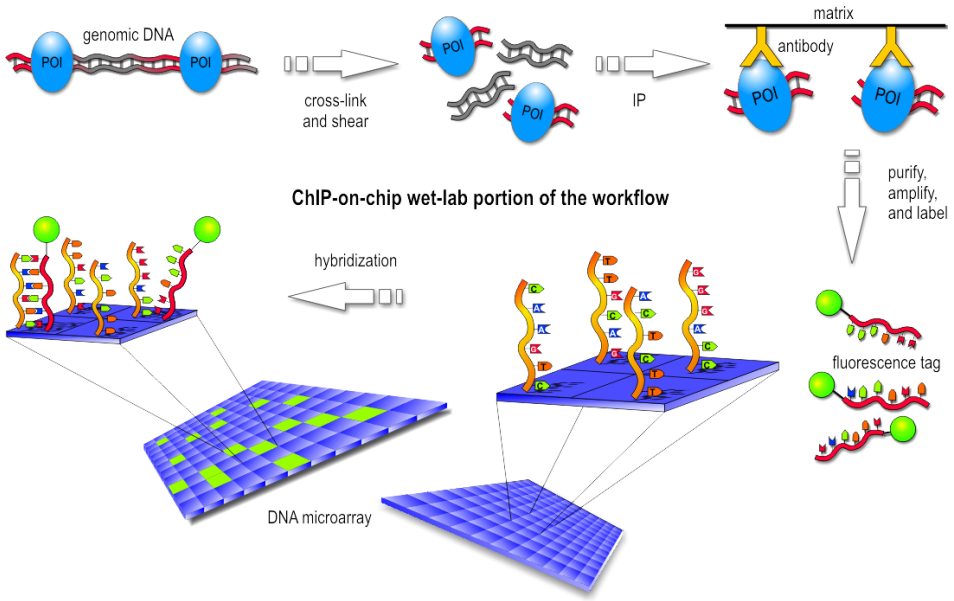


Figure 2.5: Diagrammatic representation of the chIP-on-chip assay for identifying genome-wide DNA binding sites of a transcription factor of interest (here labelled 'POI'). Protein-DNA binding is cross-linked and DNA fragmented. An antibody to the TF of interest is used to isolate DNA fragments bound by that TF, and following purification, amplification and flourescent labelling the DNA fragments are hybridised to a microarray with predicted promoter oligonucleotides fixed to known locations. The location of flourescence can be used to determine the sequences that were isolated through the antibody-purification, and thus assumed to be bound by the TF of interest.

DNA which is bound by the TF in question can be isolated, then purified after reversal of cross-linking. To identify these target DNA fragments, a 'promoter array' can be used: a microarray with fragments of promoter DNA (typically 8kb upstream to 2kb downstream of predicted transcription initiation sites) of some subset of the genes in the genome. This process is commonly referred to as 'chIP-on-chip' or 'ChIP-Chip.' A diagram representing the process is shown in Fig. 2.5.

The establishment of conclusions drawn from genome-wide DNA-binding studies such as ChIP-Chip approaches must be carefully considered [Li et al., 2008], as it is has been shown that binding of a TF to a gene's regulatory element does not necessarily result in functional regulation of the target gene's expression [Thanos and Maniatis, 1995].

2.2.4 Next-Generation Sequencing

A discussion of tools to study transcriptomics (that is, the levels of expression of the complete set of transcripts in a cell) would not be up-to-date if it failed to mention techniques based on high-throughput sequencing. Recent advances in DNA sequencing technologies have resulted in the ability to obtain hundreds of millions of sequencing reads (that can be aligned to a reference genome) in times and costs feasible for ap-

plication to profiling a number of samples prepared by an individual laboratory. A discussion of these technologies is presented in [Shendure and Hanlee, 2008].

Two applications of this technology are especially relevant here: the sequencing of genomic fragments following chIP-based isolation to identify DNA-binding sites of a TF of interest (chIP-seq), and the use of this sequencing technology to identify the abundance of all RNA transcripts in a sample (RNA-seq). The technology affords a number of advantages over microarray-based equivalents for each of the previously mentioned applications. The ability to detect any transcript (not limited to those represented by pre-specified probes) is advantageous both for measurement of unknown transcripts and DNA-binding throughout the genome. Sequencing-based transcriptional profiling (RNA-seq) offers the (theoretical) ability to quantify the absolute numbers of molecules in a sample, clearly advantageous over the essentially arbitrary measurements provided by gene expression microarrays. An example application of this technology to the study of transcription in ES cells is described in [Cloonan et al., 2008].

This technology clearly presents a great opportunity for gathering useful transcriptional data. Although the current status is that there exist data from vastly more samples profiled using microarrays in the public domain, and thus large-scale meta-analysis of gene expression data is only really appropriate for data from microarray platforms, due to the projected expansion of RNA-seq technologies referred to in [Wang et al., 2009] it may be pertinent to consider such data for incorporation into gene expression data mining analyses in the future.

2.3 Gene Expression Data Mining

The enormous number of measurements made with microarray (and other transcriptional profiling) technology presents an invaluable resource for studying the control of biological processes [Li et al., 2004], but comes with the challenge of identifying patterns in the data that can help us better understand gene function and the processes that regulate transcription.

It may be especially useful to study patterns of gene expression across wide ranges of cellular conditions as, while simple approaches may be applied to find genes with statistically significant differential expression across small sets of related experiments, greater insight into the biological functions of genes may be yielded by identifying patterns in the expression of groups of genes across many experiments [Quackenbush, 2001].

This section includes a reiteration of some motivation for analysing collections of independent microarray datasets together, followed by an outline of the wide range of

existing approaches to mining gene expression data resources to gain insight into the mechanisms of biological processes.

2.3.1 Meta-Analysis

The concept of 'meta-analysis' refers to the process of combining results from multiple, independent studies through the use of statistical techniques. In the context of using gene-expression microarray data to make inferences regarding transcriptional regulation, this involves the use of multiple datasets for a combined analysis. By considering data from multiple studies, one can theoretically reduce the impact on the overall conclusions of the particular experimental practices used in each experiment and the specific conditions in which the samples were cultivated and prepared. In this way, confidence that the observed patterns and their inferred biological implications are generally applicable can be increased. A discussion of some of the advantages of a meta-analytic approach is presented in [Ng et al., 2003] in the context of machine learning, and in [Ramasamy et al., 2008], which refers to a large collection of applications of meta-analysis of gene expression microarray data.

As most statistical tests used in analysis of gene expression data are designed for application within an individual dataset, complications may arise in trying to combine the results from repeated application of such statistical tests to multiple datasets in isolation. For example, three independent studies, [Miura et al., 2004], [Sperger et al., 2003] and [Sato et al., 2003] performed transcriptional profiling experiments on ES cells to identify sets of 'stemness' genes associated with pluripotency, but only 7 genes (out of a total of 2226 listed) were found to be common to the genelists published by the three studies. It was demonstrated in [Suarez-Farinas et al., 2005] that repeatedly applying the same statistical tests to identify genes with significant differential expression in each of a number of datasets results in overly conservative assessment of overall significance, and by combining the datasets together and performing a unified analysis of the integrated data considerably greater correspondance was shown between the individual studies. The effect of repeatedly applying statistical tests to the same genes across different datasets was described in [Suarez-Farinas et al., 2005] as the 'small intersection problem.' An admission that this problem was observed in a specific meta-analysis of microarray data [Assou et al., 2007] attempting to study ES cell transcriptional regulation provided further motivation for taking an approach in which the same analysis method may be applied to the full collection of data at once.

A number of approaches have been designed to get around this so-called 'small intersection problem,' with most involving statistical techniques specially designed for meta-analysis so that results from independent studies can be integrated, and a review of such approaches can be found in [Hong and Breitling, 2008]. A further description is given in [Moreau et al., 2003] and additional examples are presented in

[Choi et al., 2003] and [Conlon et al., 2006]. There are also a number of methods based on identifying a set of individual datasets that are 'relevant' to a particular question, based on some query input (e.g. [Hibbs et al., 2007, Caldas et al., 2009, Baughman et al., 2009]). These provide some advantage over choosing experiments purely based on annotations [Caldas et al., 2009], but still restrict the data available for making inferences. It has been demonstrated that, if possible, it is especially advantageous to collect the (multiple) datasets together and use methods that can achieve a 'unified' meta-analysis of the combined data [Suarez-Farinas et al., 2005].

The goal of such 'unified' meta-analysis methods applied to integrated collections of data is to increase the power to identify genes that are expressed at significantly different levels between defined sets of biological conditions [Engelmann et al., 2008]. However, to assist in our understanding of the transcriptional regulation of pluripotency, we are especially interested in identifying relationships in gene expression between a set of known transcription factors and any genes[2] (potentially without known roles in our processes of interest) in the genome. In order to pursue this goal with only methods that apply tests to independent datasets in isolation, multiple independent datasets are required, each involving significant differential expression of a combination of the transcription factors of interest. It would also be required that this significant differential expression be observed across cells with generally similar transcriptional programs, so that gene expression patterns associated with the expression of the TF of interest can be distinguished from gene expression patterns associated with different general biological contexts. It would therefore seem advantagous to adopt alternative strategies for the analysis of microarray data in order to identify transcriptional relationships between a set of TFs and their potential regulatory targets. This reinforces the statement that there is no single approach to analysis of microarray data, including large-scale meta-analyses, that is appropriate for every application. It is clearly important to define the biological question one wishes to answer by interrogation of the data, and use this to decide upon the analysis approach taken.

An example application of meta-analysis of gene expression data to the study of transcriptional regulatory mechanisms in ES cells is given in [Campbell et al., 2007]. In this study, a large and diverse collection of samples from adult and embryonic stem cells as well as differentiated cells were profiled using gene expression microarrays [Perez-Iratxeta et al., 2005]. The data was utilised in a series of correlation analyses to identify genes that had expression profiles correlated to that of Oct4 across a large proportion of randomly-sampled subsets of the whole collection of samples. A selection of putative targets obtained via this correlation analysis were validated using ChIP to confirm Oct4 binding to DNA proximal to the genes in question.

[2] or at least the representation of genes by probesets on a microarray platform

2.3.2 Correlation & Clustering

An intuitive approach to the task of extracting biologically relevant information from a whole set of gene expression data involves grouping together genes that share similar expression patterns, as discussed in [Quackenbush, 2001]. There exist a great many approaches to gene expression data analysis based on the principle that, if the expression of a number of genes is changing in similar ways across a group of microarrays, they are likely to be involved together in some sort of biological process(es) that are occurring. This is colloquially known as 'guilt-by-association.' One of the simplest ways of assessing similarity in expression pattern is by calculating the Pearson correlation coefficient between the expression profiles of each possible pair chosen from genes represented in the dataset.

Based on this principle, the most widely used method (and some consider it too widely used, as mentioned in [Allison et al., 2006a]) of illuminating order from a set of gene expression data is that of clustering. Its goal is to classify genes into (unspecified) groups based on their expression profiles. Clustering is generally a from of 'unsupervised learning' in which a 'distance metric' (such as the correlation coefficient) is used to group together the most similar entities, and these groupings are refined without any feedback. Some forms of supervised clustering exist, but as our prior knowledge of what the gene expression profiles should be across a number of samples is severly limited [Eisen et al., 1998], such supervised methods are not especially appropriate for our desired analysis tasks. There now follows a description of each of the two most widespread clustering methods, and a short discussion of problems associated with the clustering paradigm (and therefore common to all simple clustering methods).

Hierarchical clustering was applied to microarray data analysis in [Eisen et al., 1998] and has since become one of the most widely used methods to analyse microarray datasets. It works toward a goal of producing a 'binary tree' representation of the genes and/or samples in the dataset. For example, a binary tree for the samples in the dataset might be produced on the basis of a similarity score between each sample and the others. This consists of a recursive organisation of the elements being clustered into pairs. An example of such a hierarchical clustering of the samples from a microarray dataset is shown in Fig. 2.6. One advantage of the hierarchical clustering technique is that the tree structure enables examination of different levels of clustering, which can lead to visualisations of the data that are both intuitive and useful for exploration [Eisen et al., 1998].

Hierarchical clustering is not the only clustering approach that has been applied to the analysis of gene expression data. 'K-means' clustering, as used in [Tavazoie et al., 1999], works on the principle that a number ('k') of groups are pre-specified, and the genes are subsequently classified into one of these groups through an iterative refinement

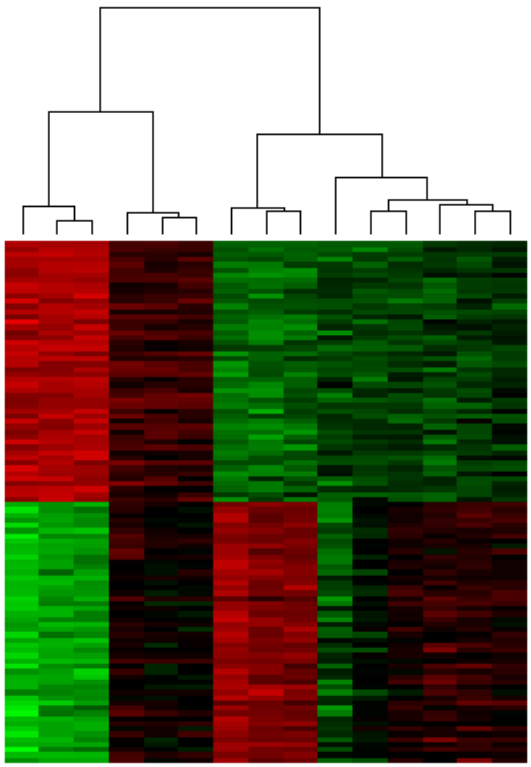


Figure 2.6: Hierarchical clustering of samples in a microarray dataset, with each sample represented by a column of the heatmap. Connected lines above the heatmap form the sample dendrogram, with the dissimilarity between two samples reflected by the height above the heatmap at which respective branches of the tree join.

process. As with hierarchical clustering, this approach is based on a distance metric. 'Self-Organising Maps' (SOMs) are also used for the clustering of gene expression data, and are essentially an extension of the k-means clustering approach, but with a predefined structure relating the clusters to one another. An application of SOMs to gene expression data analysis is described in [Tamayo et al., 1999].

The major shortcoming of the clustering paradigm when applied meta-analysis of gene expression data is that groups of genes with expression profiles that are correlated across the entirety of a large and diverse dataset may not represent groups of genes that may be working together to cause an observed biological effect. A discrepancy between these two types of groupings can arise from the fact that one gene can affect the transcription of numerous others, and numerous other genes may affect the expression of the gene in question. In addition, genes may have different roles and different transcriptional relationships under different cellular conditions. For example, Sox2 was shown in [Avilion et al., 2003] to be required in the mouse embryo for the formation of epiblast but in neural lineages it was shown in [Ferri et al., 2004] to have independent roles in neurogenesis and the proliferation of neural precursor cells. Sox2 is known to interact with Oct4 in the developing embryo (e.g. see [Masui et al., 2007]) but as Oct4 is not expressed in neural lineages the neural function of Sox2 must necessarily involve a different network of transcriptional interactions. The existence of such context-dependent transcriptional relationships imply that, especially when microarray experiments involve a large number samples covering a range of experimental conditions, potential relationships within the data are missed because the clustering method extracts only those groups of genes with ubiquitous, simple relationships (see [Zhang et al., 2007, Li, 2002]). In addition, when we are concerned with large-scale meta-analysis, as the number of samples in the dataset increases, the likelihood of finding genes with strongly-correlated expression patterns across all samples decreases. This disruption of correlation patterns will limit the potential to make confident inferences regarding relationships within the data.

2.3.3 Non Cluster-Based Approaches

While the clustering-based methods make up the majority of cases of the type of meta-analysis of gene expression data relevant here, a few alternative approaches have been developed in attempt to avoid the limitations of clustering or the complexity of biclustering.

One alternative to clustering approaches to data mining is the application of 'projection' methods (such as PCA [Massy, 1965], which is widely used in data analysis). These aim to reduce the dimensionality of the dataset into a few weighted combinations of gene profiles that explain the greater proportion of the overall variation in the dataset

(eg. [Alter et al., 2000, Lee and Batzoglou, 2003, Liebermeister, 2002]). As the resulting components are weighted combinations of the genes in the dataset, it can be difficult to translate these into any corresponding biological meaning [Hibbs et al., 2007]. An approach known as 'gene shaving' was introduced in [Hastie et al., 2000] to use PCA to identify clusters of genes with correlated expression profiles and a high level of variation across a dataset. Projection approaches have been demonstrated to be useful for classification purposes (eg. [Nguyen and Rocke, 2002]).

Visualization of gene expression data can be used as a way of identifying structure within the dataset, although this has primarily been performed in conjunction with clustering (e.g. [Eisen et al., 1998]). Visualisations of the relationships between certain elements represented in the data (e.g. genes), might have the potential for assisting in the identification of the organisational nature of different transcriptional relationships represented in gene expression data. However, those based on correlation (e.g. [Jupiter and VanBuren, 2008]) still suffer from the same critical drawbacks that affect any other approach based on the global correlation of the expression profiles, as described in Section (2.3.2).

An interesting approach to studying gene co-expression relationships is taken in [Li, 2002] and [Li et al., 2004], searching for 'liquid associations' between genes. These are essentially groups of genes whose correlation dynamics across a certain range (in the above studies, this range is the entire dataset) of samples can be explained by (i.e. is correlated with) the expression of another gene or group of genes. This approach may well be worth further study, but as yet these methods have not been applied to gene expression data collections with thousands of samples (and so the order of a billion, rather than a few million, measurements). Their feasibility for this scale of analysis task is yet to be demonstrated.

A further approach, known as 'Gene Recommender' is presented in [Owen et al., 2003] which uses a scoring system to weight individual datasets in a compendium according to proposed relevance to a given set of query genes. This scoring is based on the correlation between the query genes across each individual dataset. A number of similar gene expression data mining methods based on this principle have been presented [Hibbs et al., 2007, Baughman et al., 2009]. While in [Hibbs et al., 2007] (which looks at gene function prediction), the authors state that restricting the analysis to consider only whole datasets 'utilises the diverse data in a biologically meaningful way,' such approaches lose the ability to identify relationships in diverse biological contexts as they are restricted to those relationships consistent across individual studies and, critically, any patterns which may be evident across a group of samples comprising a subset of samples in each of a number of individual datasets will be rejected. This approach was employed to identify transcriptional regulatory candidates involved in

the oxidative phosphorylation system [Baughman et al., 2009]. Additionally, these approaches are dependent on the concept of a 'dataset' representing a concise biological context, when in fact there is no precise definition of such a concept and as a result, in practise an individual microarray dataset uploaded to a data warehouse may represent transcriptional profiling used to identify differences across many biological contexts.

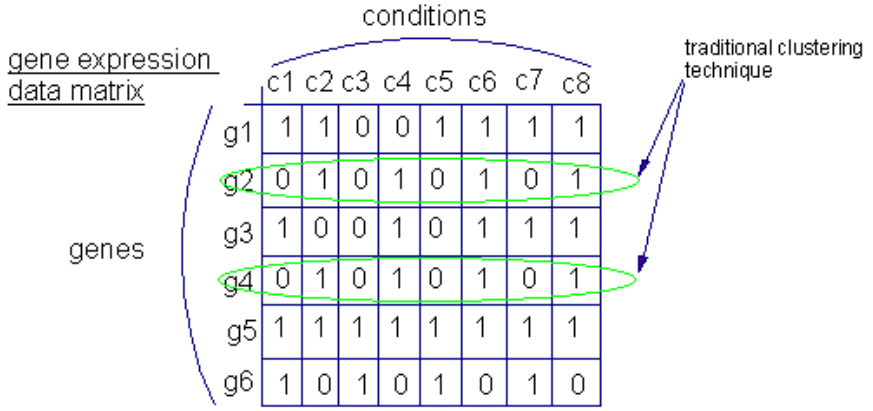
2.3.4 Biclustering

In an attempt to bypass the major problems associated with clustering, 'biclustering' formalises the approach of searching for groups of genes with expression profiles that are correlated across *some subset* of the samples in the dataset. This allows potentially more complex relationships to be elucidated [Cheng and Church, 2000], such as those involved in the transcriptional regulation of many biological processes, even when the dataset contains a large and diverse enough set of samples that a high level of correlation across the whole dataset is only observed for genes with ubiquitous transcriptional relationships. This 'two-way' clustering in which both rows and columns of a data matrix are clustered simultaneously was introduced as a concept in the 1970's [Hartigan, 1972], but received relatively little attention until Cheng & Church applied it to the analysis of gene expression data (in [Cheng and Church, 2000]), using the term 'biclustering' (introduced in [Mirkin, 1989]) to describe the approach. This section contains a description of the biclustering problem, the challenges arising from its application to the analysis of gene expression data, and a summary of the types of available (existing) methods.

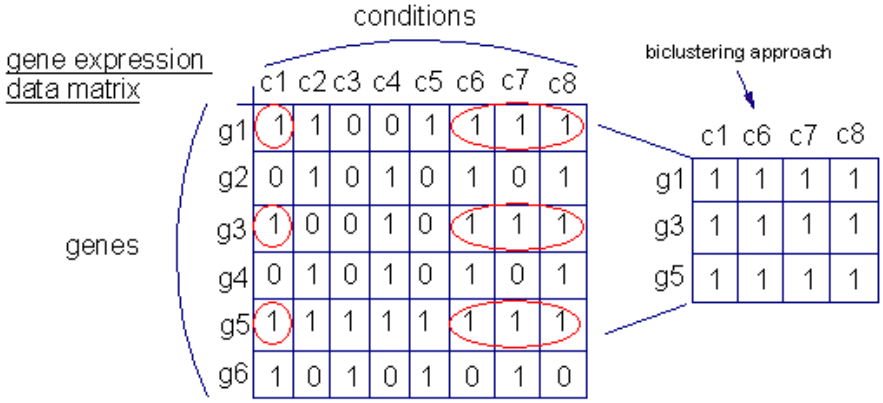
The biclustering problem formulation given here is based on that given in the [Madeira and Oliveira, 2
survey of biclustering algorithms. Assume that a matrix of normalized gene expression values has 'n' rows and 'm' columns. With the rows representing genes and columns representing samples from which the data was obtained, a bicluster is a subset of rows that exhibit similar behaviour across a subset of columns. That is, a bicluster represents a group of genes with correlated expression profiles over a subset of the samples in the dataset.

To aid visualisation of the concept of a bicluster, the artificial example shown below represents a matrix of gene expression levels for a number of genes in a number of samples ('conditions'). In Fig. 2.7(a), genes are grouped together in a 'cluster' when they share similar expression patterns across the whole range of conditions, in a similar way to the traditional clustering methods described in Section (2.3.2). Fig. 2.7(b) demonstrates the premise of biclustering, grouping genes into a 'bicluster' from a subset of conditions in which the genes share similar expression patterns.

With the condition that biclusters should be neither exclusive (as would be the case if a gene could belong to only one cluster) nor exhaustive (as would be the case if



(a) Traditional clustering method



(b) Biclustering method

Figure 2.7: Illustration of biclustering principle, with hypothetical gene expression matrices for a number of genes (g1-g6) in a number of samples (c1-c8). Say a value of 1 represents up-regulated expression of a gene relative to a background expression level and a value of 0 represents the background expression of the gene. In this case, global correlation methods will identify genes that are up-regulated in a shared, exclusive set of samples (as illustrated in (a)). Biclustering, however, enables the identification of relationships involving shared up-regulation of a set of genes in a set of samples, without considering other samples in which those genes may not display co-ordinated up-regulation (as illustrated in (b)).

every gene had to belong to at least one cluster) in regard to the data, there are many possible ways of interpreting the statement "exhibiting similar behaviour" referring to what constitutes a bicluster pattern. Correspondingly, there is a variety of the types of biclusters that existing approaches may find. Whether or not it is desirable to find biclusters of a particular type depends on the context of the problem, and so there is no general 'best approach'. However, considering the particular context is always important in determining the type of biclusters to find, as well as in determining the various optimisation techniques that may be employed by the algorithm to identify the best biclusters according to the chosen criteria. As a given bicluster may span any number of genes and any number of samples, in any data matrix there are enormous numbers of potential biclusters from which to choose the ones that represent the 'best patterns' in the data. To choose any bicluster over another, there must be some method of scoring each individual bicluster. This score may be based on the size of the bicluster, the degree of correlation of the values in the bicluster, or any other criteria (that may or may not have some biological motivation). Again, as the bicluster scoring function may be chosen to utilise characteristics of the samples or genes in the expression matrix, it can be important to consider the particular context of the problem before deciding on a scoring function. The survey of [Madeira and Oliveira, 2004] contains mathematical descriptions of a range of possible approaches.

The problem of finding a set of biclusters (whether they are exclusive or overlapping) to cover a data matrix is known to be NP-hard [Cheng and Church, 2000], as it is a generalisation of the problem of finding a minimum set of bicliques to cover a bipartite graph. 'NP-hard' is a class of computational complexity, with the significance that finding all the exact solutions for an NP-hard problem is intractable. A detailed description of NP-hardness and the theory of finding approximate solutions to such problems is provided in [Hochbaum, 1996]. The complexity of the biclustering problem lends itself to the application of heuristics, techniques from the field of artificial intelligence based on identifying a good solution to a given optimisation problem in a reasonable time, when identification of the best solution is computationally impractical. Heuristic techniques are discussed in detail in the following section.

A further point worth noting concerns normalization methods for gene expression data, which are data transformation approaches applied to measurements from gene expression microarrays in order to reduce technical variation between samples and improve the comparability of data. A range of normalization algorithms exist for data from a range of microarray technologies, with the performance of a number of the most widely used methods for data from the Affymetrix GeneChip platforms assessed in [Gyorffy et al., 2009]. Through the application of different normalization techniques to a given set of gene expression data, it may be possible for a particular biclustering algorithm to find biclusters of different types in the same dataset, corresponding to the

particular normalization technique used. Therefore, the combination of normalization method and biclustering approach to use should be considered in the context of the particular dataset and the purpose of the bicluster analysis of the data.

2.3.5 Heuristics

In computer science, a 'heuristic' approach to solving a problem involves adopting methods that aren't guaranteed to find the best possible solution or aren't guaranteed to find a solution within a desirable limit of computation time, but find a 'good' solution within 'reasonable' time in the vast majority of cases that are confronted. Finding the solutions to many real-life problems can be seen as searching for cases that result in the optimisation of given properties in a mathematical model, and many such real-life problems are particularly appropriate for the application of heuristic techniques. For example, finding the best possible bicluster in the biclustering problem defined above with a 'brute force' search technique (that evaluates every possible solution) can involve computation time exponential in the number of inputs, which would become infeasibly large for any real application of biclustering. As biclustering can be formulated as an optimisation problem, where the optimal solutions are those biclusters with the highest score according to the defined evaluation function, the theory of heuristics can be applied to biclustering to avoid the problems associated with the application of 'brute force' methods.

A 'heuristic' itself is a rule or guideline used to determine how to proceed with the search for a solution to the problem in question. In searching for the best biclusters in a dataset, the desirability of biclusters must be assessed according to some model reflecting how well a bicluster captures an idealised transcriptional relationship in the data. Owing to the fact that this idealised transcriptional relationship can only be expressed in terms of numerical trends in the measured expression values, and yet the utility of a discovered bicluster pattern in terms of guiding experimental research or answering a particular biological question may depend on unavailable prior knowledge concerning the genes and samples involved, the best bicluster according to the model may not be the best bicluster in terms of the conclusions that can be drawn or predictions that can be made on the basis of the uncovered relationships. As a consequence, heuristics are particularly appropriate in the problem of biclustering, as it simply may not be worth the considerable computational cost in order to guarantee that a given solution is the best solution according to the specified model of bicluster desirability.

Local Search Heuristics

A 'local search heuristic' is a heuristic that guides the searching procedure based on the current and 'next-step' states of the model. A local search heuristic does not take into account general properties of the solution as a whole. All 'greedy' search techniques are

examples of local search heuristics. A particularly relevant example of such a heuristic is the greedy node-deletion technique employed in [Cheng and Church, 2000]. This algorithm to find biclusters in gene expression data uses a node-based representation of the bicluster which starts with a bicluster covering the entire gene expression matrix. The algorithm proceeds by 'pruning' nodes from the bicluster until the bicluster's consistency score passes a threshold, δ. The heuristic employed determines the order to remove entries from the bicluster, identifying the node responsible for the greatest disruption in the consistency across the bicluster. This heuristic is a typical greedy search technique, at each step choosing the one that most improves the score. The advantage of such greedy search techniques is that they reduce the search space explored by the algorithm and thus speed up the optimisation process.

Problems with using local search heuristics arise when the 'solution space' is irregularly shaped - this corresponds to situations in which the best solution is not found by taking every single step in the direction of the goal. In essence, this problem is the fact that 'local optima' are not always 'global optima' - if a solution space is a landscape where a lower position corresponds to a better solution, a local search algorithm proceeds analogously to moving downhill at every step. When the algorithm moves to the bottom of a basin it will return the corresponding solution, rather than checking to see if a deeper basin lies on the other side of the current basin's lip. Examples of this behaviour occur in complex optimisation problems, where local search techniques are inapplicable [De Jong, 1975]. In the Cheng and Church biclustering algorithm, a corresponding possible scenario would be removing a node that is present in a number of good biclusters because there are a number of other nodes that are each present in mutually exclusive biclusters but have lower dissimilarity. This would therefore exclude any biclusters that include the removed node, regardless of the fact that removing a number of other nodes first could result in biclusters with much improved score.

Global Search Heuristics

A 'global search heuristic,' in contrast to the local heuristics, takes into account properties of the solution as a whole. Generally this means the approach taken is not based on constructing solutions in a stepwise manner, but forming whole solutions at each step and improving these by various methods. The advantage of global search heuristics over local search heuristics is that they can avoid the tendency to converge on locally optimal solutions that are in fact far from the best solution [Mitchell et al., 1992]. Global heuristic methods tend to be more sophisticated (and thus more complex) than local search techniques [Toern and Zilinskas, 1989].

A relevant example of a global search heuristic is that of 'simulated annealing' applied to biclustering in [Bryan et al., 2005]. Such techniques can also be referred to as 'stochastic search techniques', as they incorporate a random element into the procedure

that allows an equivalent of 'backtracking' through locally inferior solutions that lead to eventual solutions that are superior. Another example of a global search heuristic is the process of evolving solutions by means of a 'genetic algorithm' [Holland, 1975], described in detail in [Goldberg, 1989]. In a genetic algorithm, possible solutions are encoded as 'chromosomes' - as far as the algorithm is concerned, this is just a type of data structure that is usually a string of bits, integers or characters. It is essential for the process to define a function in terms of the elements of the chromosome which results in a score that reflects the desirability of the solution. This function is called the 'fitness function'. An initial 'population' of chromosomes is generated randomly, or by any quick and arbitrary method, from which an intermediate population is selected from members of this initial population (based on the fitness scores of each chromosome). A 'fitter' chromosome will have more copies passed into the intermediate population. Finally, 'genetic operators' are applied to the intermediate population in order to create the next generation of the population, for which the processes from selection to creating the next generation are repeated until either the best solutions are found or a given number of iterations (generations) is exceeded. These genetic operators include 'reproduction,' 'crossover' and 'mutation,' by which a chromosome is passed straight into the next generation, two chromosomes are randomly combined and then passed into the next generation, and every element of every chromosome is altered with a given (low) probability, respectively. There is much potential for customising the mechanisms of genetic algorithms, and potential application to biclustering (e.g. [Chakraborty and Maka, 2005]).

2.3.6 Biclustering Algorithms

Since Cheng & Church's initial paper on biclustering of gene expression data, there have been a great many different methods developed to find biclusters in gene expression data matrices. As the motivation behind the development of any bioinformatic data analysis method should be to facilitate biological research, the ultimate demonstration of effectiveness of such a method would surely be a body of biological research publications presenting significant findings identified through the application of the method. However, owing to the time taken for such a body of evidence to emerge and the documented tendency in the bioinformatics research community towards development of new analysis methods (particularly regarding clustering methods for microarray data [Allison et al., 2006b]) over the application [Hibbs et al., 2009] of methods, few biclustering algorithms are supported by results of application to biological research problems. Therefore, it is important to consider how to assess the applicability of a biclustering method to a given analysis task or biological research question.

Evaluation of Biclustering Algorithms

In the two most widely-cited surveys of biclustering methods, Madeira & Oliveira [Madeira and Oliveira, 2004] present a review of different types of model used to identify biclusters in expression data and algorithms used to find them, although present no comparison in terms of performance, while Prelic *et al* [Prelic et al., 2006] evaluate the performance of a small selection of the most popular biclustering algorithms by examining recovery of implanted biclusters in artificial data and enrichment of Gene Ontology (GO) [Ashburner et al., 2000] terms associated with the genes in each bicluster retrieved from yeast microarray datasets. An examination of these surveys (and many of the individual publications presenting novel methods) highlights the underlying cause for the confusion regarding biclustering methods: determining the degree of success with which a biclustering algorithm fulfils its desired role in biological research, through the analysis of gene expression data, is fundamentally difficult. Myers *et al* remark on the lack of appropriate evaluation frameworks for gene function prediction via analysis of gene expression data [Myers et al., 2006], pointing out critical flaws in the prevalent approaches (used in those cases where such evaluation is not ignored altogether), but their evaluation framework is based solely on the task of gene function prediction and may not be appropriate to evaluate any of the other potential applications of biclustering. This hints at the root of the evaluation problem, which is that meaningful evaluation of the application of a biclustering method to a biological research problem depends on the problem in question. As a gene expression bicluster is a pattern in expression data and not an answer to a specific biological question in itself, such meaningful evaluations ought also to take into account the ways in which the information contained in a set of biclusters might be harnessed to provide an answer to specific biological questions.

Therefore, evaluation of the *biological application* of a biclustering algorithm can be considered in two separate parts:

1. How effectively does the algorithm find the intended patterns of interest in gene expression data, within feasible limits of time and computational resources?

2. How do those patterns in the data assist in the answering of the specific biological question?

Again, this reinforces the statement made earlier that the best approach to take for a given task may depend on the particular task at hand, and in this case, identifying the most suitable biclustering algorithm for utilisation of large datasets to provide insight into transcriptional regulation of expression of key pluripotency genes would require a demonstration of comparitive successes of available algorithms in being used for similar problems.

Biclustering Algorithms for Meta-Analysis

The majority of biclustering methods available have been developed and tested for application to individual, (relatively) small microarray datasets. This is primarily due to the complexity of the biclustering problem, but as a result of this limitation, many potential pitfalls of applying this promising approach to mining very large collections of microarray data have not been studied. Some problems are mentioned in [Hibbs et al., 2007], such as 'sensitivity to noise, inability to work with data from diverse conditions' and that the methods are 'prohibitively slow.' These are all problems that have the potential to affect biclustering algorithms, and should be considered when identifying approaches which may be suitable for the task at hand, but are by no means insurmountable.

The fundamental motivation of whole-data biclustering is that we are interested in any significant patterns in relationships between the levels of transcription of groups of genes as they co-vary across samples within the data. This interest in related expression patterns of genes is retained across any subset of the samples from a collection of gene expression data on the grounds that the covarying pattern itself represents some interesting (and potentially useful) transcriptional signature shared by those samples. This is a statement of the principle that the presence of a significant bicluster in the data is effectively an association of that set of genes to a common biological process or state shared across those samples. At the very least, the identified set of genes, sharing an expression pattern across the subset of samples, represents a signature from which the biological significance of the grouping may be inferred.

Where searching for patterns across large numbers of samples, there is a higher chance of observing apparently significant expression patterns out of random fluctuations in measurement values. This issue, referred to as finding patterns in noise, must be considered when applying biclustering to large collections of data. Provided the significance of any pattern can be evaluated in terms of how unlikely that pattern could have arisen purely by chance, the scale of the datasets used for inference ought to be irrelevent. However, it may become significant when a pattern of interest occurs in such a small subset of the data that it is indistinguishable from random fluctuations in measurements that occur throughout the data. A similar problem applies in principle to any method attempting to use a large amount of data to reinforce the ability to make inferences. For example, when using a dataset-selection approach, the more datasets included in the analysis, the more important it becomes to consider the number of datasets deemed significant and the proportion of all the datasets in which a given co-expression relationship is observed. This in turn results in the same problem mentioned earlier in this paragraph, where a significant biological pattern is only present in a small set of those experiments included, and is no more likely to have been observed than one by chance. As a result of this potential problem, it is always useful to have data

concerning the relevant biological context that is not included in the analysis, so that observations made can be tested for generalisability beyond the original data matrix queried.

Finally, the complexity of the biclustering problem results in a critical limitation regarding the suitability of biclustering algorithms to the task of meta-analysis of large collections of gene expression data. As mentioned above, the time taken to identify the optimal solution to the problem of finding a given set of biclusters in a dataset through exhaustive search can increase exponentially with each additional row or column included in the dataset. The large collections of gene expression data available for the proposed meta-analysis result in data matrices containing tens of thousands of rows and tens of thousands (or even more) of columns to search through to find relevant biclusters. As the application of a biclustering approach to meta-analysis of large collections of gene expression data is yet to be reported, computational optimisation of existing biclustering algorithms has been limited to consideration of tens of thousands of genes but only rarely as many as a few hundred samples and, until two recent publications describing more efficient biclustering approaches ([Li et al., 2009] and [Huttenhower et al., 2009]), never more than a few hundred samples.

2.3.7 Data Integration

As a final comment on analysis of transcriptomic data, it should be mentioned that attempting to gain a full understanding of transcriptional regulatory processes by inference from gene expression data alone may result in incomplete conclusions (eg. [Husmeier, 2003]). In order to improve upon that part of our understanding gained from inferences from gene expression data, it would be pertinent to incorporate evidence obtained from different experimental approaches and that may capture different aspects of the behaviour of these biological systems.

The transcriptional control of biological processes can be influenced by sets of transcription factors which bind to regulatory sequences in DNA, individually or in complexes, and combinatorially affect the transcription of target genes. A key to gaining an understanding of the combinatorial actions of various TFs, and thus the transcriptional control of biological processes of interest, may lie in the use of genome-wide DNA-binding information to identify which genes may be regulated by TFs of interest (and in what combinations) in concert with gene expression data to determine the transcriptional effects of the observed combinations of the TFs binding to DNA proximal to certain 'target' genes.

To this end, a number of techniques have been developed for integrating results from gene expression datasets and TF binding data, such as from high-throughput chIP experiments as described in [Bar-Joseph et al., 2003], [Wu et al., 2007], [Chen and Stoeckert, 2007],

[Lemmens et al., 2006] and [Li et al., 2008]. Integrated computational approaches to predicting gene function or for predicting transcriptional regulators have recently been shown to succesfully provide candidates for experimental investigation in both [Hibbs et al., 2009] and [Baughman et al., 2009].

These results suggest that, when a biclustering algorithm has been used to identify co-expression relationships involving TFs of interest, additional insight into the action of these TFs and the mechanisms by which they regulate processes of interest may be gained by utilising the bicluster patterns in conjunction with genome-wide binding information (available through high-throughput ChIP experiments).

2.4 Research Objectives

In the context of the potential for the wealth of publicly available gene expression data to be utilised in the study of transcriptional regulation of biological processes (as discussed in Section (2.3)) the major objectives of the work presented in this thesis were as follows:

1. Produce a means of performing biclustering analysis on large collections of gene expression data involving thousands of samples, and to use this to investigate the impact of taking different approaches to bicluster analysis of datasets on this scale.

2. Identify data preprocessing techniques and bicluster evaluation approaches that improve upon naive biclustering approaches, in terms of the utility of output from large-scale meta-analysis of gene expression data for the study of the transcriptional control of biological processes.

3. Produce a data mining approach to identify localised transcriptional regulatory patterns of relevance to a particular biological research question, from large collections of gene expression data.

4. Utilise large-scale gene expression meta-analysis to investigate the transcriptional mechanisms of control of pluripotency by the key transcriptional regulators Oct4, Sox2, Nanog and cMyc.

Chapter 3

Development Of Biclustering For Large-Scale Gene Expression Data Mining

This chapter begins with a restatement of the motivation for finding an algorithm capable of performing biclustering on such a large scale, followed by an explanation of the unsuitability of all algorithms available at the time this work was carried out. A novel approach to the biclustering problem is presented that overcomes the main issue in applicability to large-scale meta-analysis of gene expression data: the significant increase in the time and memory required to find the desired solution(s). An approach to evaluation of such an algorithm is described and the success of this novel approach is demonstrated in terms of its ability to recover biclusters in large datasets under feasible time-constraints, contrasting with the best-performing existing alternatives.

A study was performed to investigate the consequences of taking different approaches to biclustering for meta-analysis of gene expression data from up to thousands of samples. Pitfalls discovered in the application of biclustering to meta-analysis of large collections of gene expression data are presented along with ways in which certain approaches to biclustering can be used to circumvent these problems.

3.1 Motivation

It has been demonstrated for a number of cases of whole-genome gene expression data analysis that the reliability (in the form of general, repeatable results) of conclusions attained through performing inferences based on this data can be increased by examining trends in multiple datasets. For examples, see [Suarez-Farinas et al., 2005], [Wennmalm et al., 2005], [Sohal et al., 2008] and [Cahan et al., 2005]. Intuitively, this can be explained by the fact that such in-depth assays as measure transcriptomic expression levels capture an incredibly detailed state of transcription in those cells assayed, which may reflect precise environmental (not to mention culture and sample preparation) conditions in addition to the biological state of the cells in a broader sense that is being studied. It is also true that the majority of samples with publicly available transcriptional profiles represent populations of cells, which may have varying degrees of heterogeneity. By searching for expression patterns consistent across multiple different sets of biological samples, this increases the power to look over more specific environmental effects and measurement noise [Warnat et al., 2005] & [Hong et al., 2006].

As an extension to the concept of selecting multiple datasets and searching for patterns consistent across these individual datasets, the biclustering paradigm may be applied to large collections of data from potentially heterogeneous biological samples, in order to identify coherent co-expression patterns that are consistent across subsets of all those samples represented in the data. As the number of samples used in the collection of data increases, the biological heterogeneity across these samples increases. A consequence of this is that unless the expression association is truly ubiquitous across all aspects of biology represented in the data, the chance of any desired co-expression pat-

tern being observed across all samples decreases. The availability of increasingly large and heterogeneous collections of gene expression data therefore implies that discovering 'local' structures within data, as identified through biclustering, becomes increasingly more appropriate for the meta-analysis of this data than search for 'global' patterns that are uniform and consistent across all the data.

Biclustering analysis provides the opportunity for automated discovery of localised expression patterns within a dataset, indicating transcriptional relationships that are reflected in the data available for certain genes and certain biological contexts. The ability to infer transcriptional relationships between sets of genes, and to identify biological context(s) across which these relationships are observed to be consistent, would therefore be improved by including data from as many samples and as wide a range of biological conditions as is possible. Therefore, an algorithm capable of performing biclustering across as large and diverse gene expression datasets as can be appropriately assembled was required.

3.2 Challenges

The first major challenge in applying biclustering to very large collections of gene expression data is the computational complexity of the problem. Let us generalise the definition of the biclustering problem given in Section (2.3.4) to that of finding some specified pattern across subsets of the genes and subsets of the samples in the dataset, and in order to have some practical usefulness, identifying the submatrices (genes & samples) within the whole data that reflect the best instances of this specified pattern. Without further restrictions on the structure of a bicluster, every single submatrix in the data has the potential to be a bicluster that shows the pattern of interest. In order to find which were the best biclusters in a dataset with n rows and m columns, one would have to evaluate the bicluster-pattern for each of the $O(n!m!)$ possible biclusters. Obviously, as the data matrices involved grow to contain thousands of rows and thousands of columns, this makes such exhaustive searches impractical.

This major challenge influences the design of all algorithms for biclustering of gene expression data. As described in Section (2.3.6), there have been a number of approaches taken to applying heuristic search methods to biclustering, such as those utilised in [Cheng and Church, 2000], [Bryan et al., 2005] and [Chakraborty and Maka, 2005]. BiMax [Prelic et al., 2006] and SAMBA [Tanay et al., 2002] perform exhaustive search after preprocessing of the data according to a simple model, which has the effect allowing quick reductions of the solution search space. Other techniques involve both preprocessing and heuristic search techniques, such as ISA [Bergmann et al., 2003]. The work presented in Sections (3.4-3.5) was performed in 2006-2007, shortly after the

publication of a study comparing the performance of a number of the most widely-used[1] biclustering algorithms [Prelic et al., 2006]. The [Prelic et al., 2006] study was used as a starting point for assessment of the suitability of existing biclustering algorithms for large-scale meta-analysis.

Those biclustering algorithms available at the time of performing this work were searched to find those designed for scalability and computational efficiency, so that the best candidates for application to large collections of microarray data could be identified. In testing the performance of such algorithms, it quickly became apparent that none had been designed for application to data collections involving very large numbers of samples. Even those which were reported to be efficient and apply to large datasets (that is: BiMax, *OPSM* [Ben-Dor et al., 2004] and SAMBA) were unable to process microarray datasets involving more than a few hundred samples (in fact, not more than 20, 100 and 200 samples, respectively) when the arrays had approximately 45,000 probes.

A relatively recently published paper reported a "lack of effective and efficient algorithms for reliable solving the general biclustering problem" [Li et al., 2009]. The authors presented a novel biclustering algorithm and make the claim that no other algorithms could identify biclusters in a dataset containing 1,000 samples when the number of genes in the dataset increased beyond 12,000. However, due to the fact that the work described in this chapter was performed over 2 years prior to the publication of the only feasible approaches to large-scale biclustering analysis, it was essential to develop a novel method for identifying biclusters of interest in vast data matrices before the potential benefits of the application of biclustering approaches to meta-analysis of gene expression data could be explored.

3.3 Creation Of A Large Gene Expression Dataset For Meta-Analysis

In order to perform large-scale meta-analysis of gene expression data from as comprehensive a collection of data as possible, a large number of individual microarray samples would have to be obtained. At the time of performing the work presented in the remainder of this chapter (in 2008), the GEO repository listed over $9,000$ samples' worth of data available from the Affymetrix MOE430v2 platform, which represented the largest collection of mouse gene expression data from a single platform that could be obtained. The mouse (*Mus musculus*) was chosen to be the focus of study through the course of this work, as it is in the mouse that the majority of mammalian ES cell

[1] this use has primarily been restricted to validation and comparison with other biclustering methods as there has been little reported application of biclustering approaches to biological problems

experimentation has been performed (due to the (until recently unique) ability to derive germline-competent mouse ES cells [Buehr et al., 2008]) and thus for which most data is available for study of the transcriptional control of pluripotency.

Raw data in the form of `.CEL` files were obtained from the GEO repository and a list of available files for a large data compendium was created by filtering out those files incompatible for normalization with the others or those files that represented duplicates of samples uploaded under alternative filenames. Following this filtering step, raw data from $7,990$ samples was available to be collated into one gene expression data matrix. Given that the samples came from microarrays processed in a large number of different laboratories, technical variation in the measured intensity values between each experimental set would have to be 'normalized out' by some quantile normalization process. Given the success of the RMA normalization approach applied to large collections of data (as mentioned in [Goldstein, 2006, Katz et al., 2006]), this approach would be desirable for application to the collection of data described here. However, computational limitations required a sampling-based approach (such as *Extrapolation Averaging* [Goldstein, 2006] or *RefRMA* [Katz et al., 2006]) to calculate required normalization parameters on a number of randomly-sampled subsets of the data. The `RefPlus` [Harbron et al., 2007] package in R provides a means for calculating RMA parameter sets from a given set of `.CEL` files and subsequently using those parameters to apply RMA normalization to any other set of `.CEL` files. Using the functions provided in the RefPlus R package, RMA normalization parameter sets were calculated for each of 5 randomly-sampled subsets of 100 `.CEL` files from the list of all available samples, as illustrated in Fig. 3.1. These numbers were chosen as the largest feasible for the implementations provided with the computational resources available at the time. Evidence from [Katz et al., 2006] suggests that any more than 4 subsets of 50 randomly-chosen samples ought to be sufficient for a large, diverse dataset of similar size to the one utilised here. The complete set of raw data from the $7,990$ samples was normalized by RMA using each of these parameter sets in turn, creating 5 complete data matrices. The final dataset was obtained through calculation of the average expression value (for each gene in each sample) across these 5 single parameter set normalized data matrices, as illustrated in Fig. 3.2. Annotations were obtained for all the samples in this dataset, including the GEO accession number corresponding to each sample so that full details could be looked up as required.

The dataset normalization approach described above provides an additional advantage, so long as the normalization parameter sets are kept available. These precomputed parameter sets may be applied to any additional data in isolation, allowing it to be examined in the context of the existing compendium without need to renormalize this large dataset. This allows for extendability of a large data compendium, with relatively straightforward incremental updating of the compendium as and when

new datasets become available.

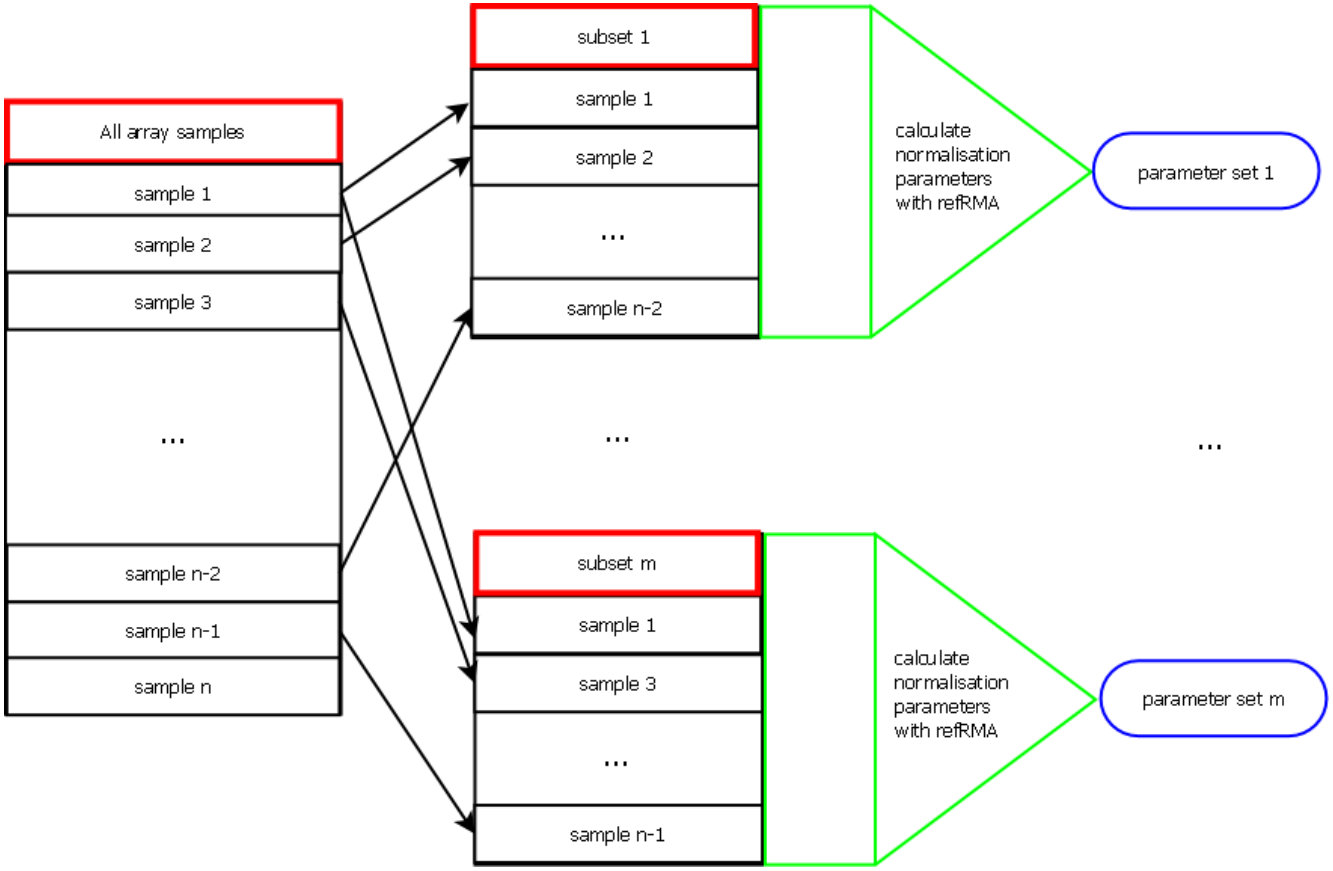


Figure 3.1: Generation of RMA normalization parameter sets for subsets of a large dataset

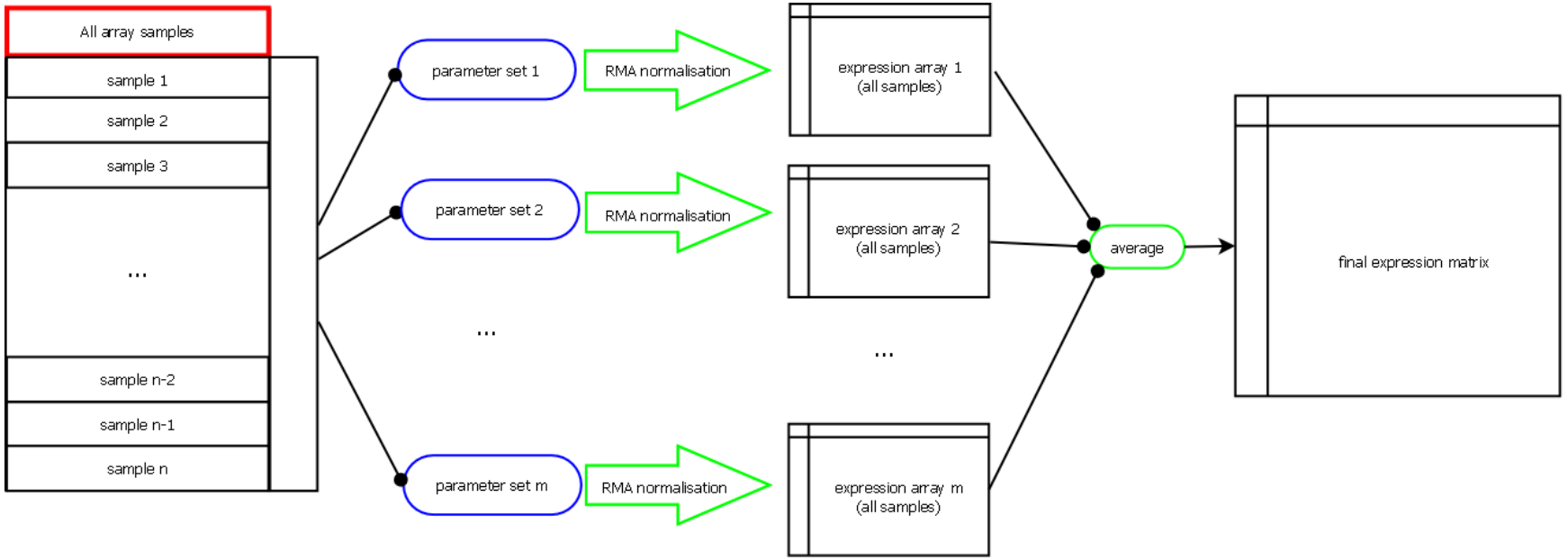


Figure 3.2: Normalization of a large dataset through extrapolation averaging with a number of pre-computed RMA normalization parameter sets

3.4 Development Of An Efficient Biclustering Approach

Given the lack of available algorithms that could be used for taking a biclustering approach to large-scale meta-analysis of gene expression data, in order to explore the possible benefits of such an approach to this type of data analysis it was necessary to develop novel biclustering methods optimised for efficiency in terms of the number of samples represented in the data matrix. This section describes a reformulation of the biclustering problem for the intended meta-analysis tasks that enable a scalably-efficient algorithm to be devised.

An illustration of this reformulation of the biclustering problem in terms of a maximal-geneset biclustering principle is given below as part of the description of a proposed exhaustive combinatorial approach. The resulting search space of $O(m!)$ biclusters becomes so large that such an exhaustive approach becomes infeasible as the number of samples in the data matrix increases significantly. For this reason, elements of the combinatorial approach were incorporated into a global search heuristic framework based on genetic algorithms. The final part of this section describes the efficient genetic algorithm for large-scale meta-analysis of gene expression data.

3.4.1 Reformulation Of Biclustering Problem

Utilising a property of the desired solutions from the intended biclustering analysis, it was possible to reduce dramatically the number of potential bicluster solutions that an algorithm would have to search through. As any bicluster in a gene expression data matrix involves an expression pattern that holds across some subset of the available genes across a subset of the available samples, it would always be desired to identify the maximal subset of the genes for which that bicluster pattern holds across the given subset of samples. Therefore, provided the maximal subset of appropriate genes could be identified (easily) for any subset of the samples represented in the data, the 'gene dimension' of the bicluster search could be disregarded. Correspondingly, for any specified bicluster pattern definition that allows the maximal subset of genes fitting the bicluster pattern for any subset of samples, the search through all potential biclusters in a data matrix of n genes and m samples is reduced from $O(n!m!)$ to $O(m!)$.

3.4.2 Identifying Biologically-Relevant Gene Expression Patterns

The patterns we are hoping to reveal through biclustering analysis relate to levels of gene expression, ideally in terms of the biological consequences associated with given levels of expression for a particular set of genes. As the most comprehensive resources of genome-scale expression data at present (that is, GEO and ArrayExpress) come from gene expression microarrays that tend to measure accurately only relative (not absolute) levels of gene expression (as described in [Draghici et al., 2006]), large-scale

meta-analyses of this data intended to uncover biologically relevant information regarding relationships in the expression of certain genes in samples under certain conditions will require some way of expressing the gene expression measurements in a manner that is suitable for comparison across different samples. Typical analysis of individual microarray datasets aims to identify patterns of differential gene expression between different groups of samples. For a full meta-analysis that considers every combination of samples in a data matrix (such as the biclustering proposed here) to identify blocks of *differential* expression patterns, it would then be necessary to have all measurements in terms of a biological state of differential expression compared to some reference level. As noted in [Madeira and Oliveira, 2004], most biclustering algorithms involve a pre-processing or normalization step to convert the expression measures into a form or scale that enables such comparisons to be implicitly incorporated into the pattern-mining.

Furthermore, even in a situation where the technical considerations of gene expression measurement described above did not apply and a comprehensive and completely reliable compendium of absolute mRNA concentrations existed (such as if a 'perfect' microarray data normalization were to exist or if a perfect measurement platform were used widely enough to produce such a body of data), the fact that a particular variation in expression level of different genes can have different biological consequences motivates us to take an approach in which the variation in expression level of tens of thousands of genes can be considered simultaneously in terms of a unified scale capturing the biological significance of such variation. Therefore, the novel biclustering approaches presented in this chapter have been developed for application to a gene expression data matrix that has already been pre-processed so that the values of the matrix represent a biological 'state' of expression as opposed to some numerical intensity value.

In order to retain the focus of this chapter on the development of efficient biclustering algorithms suitable for meta-analysis on the desired scale, it is assumed here that there exist such pre-processing or normalization step as those described above. A comprehensive treatment of the study of such methods is the focus of Chapter 4 of this thesis. The assumption taken here is that we have a pre-processed gene expression dataset where the measured expression values have been converted to symbols representing a biological state of expression of the respective gene in the respective sample. This is analogous to a call of differential expression with respect to some uniformly-applied biological baseline of expression for that gene. Examples of methods for such pre-processing steps can be found in [Tanay et al., 2002, Prelic et al., 2006, Li et al., 2009].

3.4.3 Resulting Optimisation Problem

If we have a gene expression data matrix with values representing the biological expression 'state' of a particular gene (row) in a particular biological sample (column), criteria for bicluster membership that can be applied to all genes represented for any

given subset of the samples represented in the data matrix, and a measure of desirability of any bicluster, then we can search through the space of all biclusters (that is, all possible subsets of the set of all samples) to find those that are maximal in terms of the desirability measure.

The biclustering paradigm is presented as an optimisation problem where we search for optimal biclusters in a dataset according to definitions of bicluster-membership (for genes) and overall bicluster desirability. In this framework, assuming good solutions to the optimisation problem can be found, it is clear that the precise definition of bicluster desirability that is used will be crucial in determining the success of the whole approach. As the majority of publications regarding biclustering methods tend to focus on the search mechanism for finding optimal biclusters in a data matrix, rather than the effect of the scoring system used to evaluate bicluster desirability, this is an aspect of the application of biclustering to gene expression data that appears to have received relatively little attention.

The major technical focus of the remainder of this chapter deals with, in some form or another, the development of different measures of bicluster desirability and evaluation of the success of application of the resulting algorithm to tasks in meta-analysis of gene expression data. However, the following section assumes a straightforward approach (described below) to measuring bicluster desirability in order to establish the potential for using the above optimisation problem as a framework for applying biclustering to large-scale gene expression datasets.

3.4.4 Exhaustive Combinatorial Approach

As a framework from which more efficient heuristic methods could be developed, an exhaustive approach to the optimisation problem described in Section (3.4.3) was implemented, termed ComBiclust[2]. The ComBiclust implementation enabled a preliminary demonstration of the feasibility of application of the defined bicluster optimisation task to meta-analysis of gene expression data. The principal motivation for adopting an exhaustive enumeration approach was to provide a platform upon which global search heuristics could be implemented for large-scale application of biclustering, avoiding the use of greedy local search heuristics (see Section (2.3.4) for a discussion of these different approaches).

The ComBiclust algorithm consists of three components that are described below. Firstly, a data discretisation approach is taken to satisfy the pre-processing assumptions discussed in Section (3.4.2), then the algorithm proceeds the possible combinations of samples (and thus, all possible biclusters) by pairwise combination of genelists in a

[2]from 'combinatorial biclustering'

'generational' cycle until all combinations have been evaluated. Both the generational cycle and the mechanism for combining lists are described below. A summary of results of application of this algorithm is given following the algorithm description.

Discretisation

As discussed in Section (3.4.2), the aim of ComBiclust (and of all the biclustering methods introduced in this chapter) is to perform biologically-motivated meta-analysis of gene expression data by identifying sets of genes with consistent expression states across a set of samples. A number of established biclustering algorithms take the approach of classifying genes as 'upregulated', 'downregulated' or neither in each sample represented in the data matrix. The approach taken in SAMBA [Tanay et al., 2002] and *XMotif* [Murali and Kasif, 2003] is to assume that all genes' expression levels across all samples in the data matrix are normally-distributed, and so any values lying outside the range of $\mu \pm \sigma$ are significantly 'differentially expressed' according to the assumed general reference expression level: those values greater than $\mu + \sigma$ are classed as 'upregulated' and those values lower than $\mu - \sigma$ are classed as 'downregulated.' As is demonstrated in Chapter 4 of this thesis, individual genes have widely varying distributions across large collections of gene expression data and the assumption that all genes' expression values are normally distributed is clearly far from accurate. However, this approach is undoubtedly simple and effective to the extent that both SAMBA and XMotif report biologically significant biclusters in their results.

As an improvement upon discretisation based on applying a normal distribution threshold, the approach taken for ComBiclust was to use a 'cluster-based discretisation' that assigns expression states to gene expression values. This entails, for each gene, performing n k-means clusterings (see [Hartigan, 1975] for description) of the expression values for $k = 1, ..., n$. A cluster statistic is used to evaluate the cluster assignments and choose the best k: either the 'gap' statistic [Tibshirani et al., 2001] or a less computationally intensive custom clustering statistic, defined in the following paragraph. To initialize each k-means clustering, the k centroids are placed at the mean of the values separated by the $k - 1$ largest inter-value distances.

A custom clustering statistic is calculated for each gene to give a quick guideline of whether or not to assign the gene one cluster or two. It takes into account the range of values and mean value for the gene, in relation to the average range and mean of each gene in the entire dataset. This is motivated by the fact that genes with less variation across the samples are less likely to correspond to differences in the transcriptionally-regulated processes occurring in those samples, and the fact that intensity-dependent effects tend to result in wider variability at high levels of measured gene expression. The clustering statistic for a given gene, s_i, is given in Equation (3.1).

$$s_i = \left(\frac{r_i}{\bar{r}}\right) / \left(\frac{\mu_i}{\bar{\mu}}\right) \tag{3.1}$$

Where r_i and μ_i are the range of values and mean value for gene i. $\bar{r}$ and $\bar{\mu}$ are the average range and mean value of each gene across the dataset. Calculation of $\bar{r}$ is shown in Equation (3.2).

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i . \forall \; genes \; i \in \{1...n\} \tag{3.2}$$

The cluster statistics and assignments are calculated for each gene, ignoring any missing values. The missing values are then assigned to the cluster representing 'background' expression of that gene. This avoids impution of non-measured values or ignoring the gene altogether. If a matrix of Present/Absent flags is available, this can be used to mark as missing any values flagged as Absent. The discretisation process is summarised in Algorithm 1.

Input: Matrix of gene expression values, M
Output: Matrix of discretised expression levels, D
Create empty discretised matrix, D ;
foreach *Gene* i **do**
  Calculate cluster statistic s_i ;
  if $s_i < \theta$ **then**
    Set all values $M(i, ...) = 0$;
  else
    Perform k-means clustering of $M(i, ...)$ with $k = 2$;
    foreach *Sample* j **do**
      if $ClusterAssignments(i, j) = 1$ **then**
        Set $D(i, j) = 0$;
      else
        Set $D(i, j) = 1$;
      end
    end
  end
end

Algorithm 1: discretisation of gene expression matrix

Generational Cycle

Once the data has been discretised, a combinatorial approach is taken to finding the submatrices within the whole data matrix that contain consistent expression patterns. In this combinatorial approach, all possible biclusters are enumerated by considering the data in terms of 'Consensus Lists.' Such a Consensus List describes a subset of

all the samples and all those genes that contain consistent patterns across the given subset of samples. In this way, a Consensus List represents a bicluster within the data. If each column of the discretised data matrix is used to create a population of 'sample lists' (i.e. Consensus Lists containing only one sample) then all the possible biclusters can be found by combining these sample lists in every possible way.

Using a fast method of combining two Consensus Lists, a bounded procedure for evaluating all possible biclusters is possible using a generational approach to the combinations. In this generational approach, a collection of consensus lists is created and then iteratively expanded and refined using non-redundant pairwise combinations of the consensus lists in the collection and each of the sample lists. In order to keep track of the best biclusters, a `maxbiclusters` list is created that contains bicluster objects for the highest scoring biclusters according to the bicluster desirability criterion. An example of such a bicluster desirability criterion, and in fact the most widely used measure, is simply the product of the number of genes and number of samples in that bicluster. This is motivated by the simple assumption that a larger bicluster is less likely to exist in the data purely by chance. The topic of assessing bicluster desirability will be revisited later in this chapter, but the early work on development of efficient biclustering algorithms presented here used this simple method based on a naïve model of the probabilities of biclusters appearing in data by chance. Each time a new consensus list is created, if its score is higher than one of the `maxbiclusters` then that bicluster is replaced by a new bicluster corresponding to the new consensus list. This expansion/refinement to form each subsequent generation is iterated `nsamples-1` times, where `nsamples` is the number of samples in the dataset (i.e. the number of columns in the expression matrix). The approach outlined here saves the procedure from having to explore every single combination of samples when the situation arises that there are no common patterns across a particular combination of samples. When such combinations are found, they are not passed into the subsequent generation and therefore any further expansion of those consensus lists is ignored. If the generation size ever reaches zero, the iterative loop is terminated. Once the loop has terminated (either due to a zero-sized generation or due to the iterative procedure having run its course), the list of the best biclusters is returned.

Combining Genelists

The combinatorial approach to the biclustering problem described above requires a fast method for combining lists of patterns associated with the relevant genes and samples. This method is implemented as the `combine_lists` function in the novel algorithm, which uses the set inclusion operator to determine the list of genes common to both lists and then those genes with identical 'patterns' (i.e. those assigned to the same expression class) in both lists. The lists of samples from both consensus lists are combined (ignoring duplicates) and sorted. The resulting `consensus_list` object is a

list of vectors with three components: those common genes with identical patterns, the patterns associated with those genes, and the samples whose sample lists have been combined in the process of building this consensus list.

Overall Procedure

The components described through Section (3.4.4) were incorporated into a combinatorial biclustering algorithm described in Algorithm 2.

Input: Gene expression data matrix, M
Output: Set of biclusters, B, maximising some desirability score
Calculate discretised data matrix, D (as in algorithm 3.1) ;
Initialise set of optimal biclusters, B ;
Initialise set of sample lists, SL, to the individual columns of D ;
Initialise set of consensus lists, CL, to SL ;
foreach *Generation n from 1 to* $|SL|$ **do**
  Create (empty) set of new consensus lists, NL, for generation n ;
  foreach *Consensus list* $CL(i)$ **do**
    foreach *Sample list* $SL(j)$ **do**
      Combine $(CL(i), SL(j))$ to give $NL(i,j)$;
      Evaluate bicluster desirability score, $F(NL(i,j))$;
      if $\exists k.F(NL(i,j)) > F(B(k))$ **then**
        Set $B(k) = NL(i,j)$;
      end
    end
  end
  Replace CL with NL ;
end

Algorithm 2: ComBiclust: exhaustive combinatorial biclustering algorithm

Application Results

The ComBiclust algorithm was tested alongside SAMBA [Tanay et al., 2002] to give some reference for assessing the performance of the novel algorithm in terms of state-of-the-art approaches. The ComBiclust algorithm is featured here only as a starting point for development of further methods which have been evaluated more comprehensively, so only a brief summary of the evaluation results for the ComBiclust algorithm are presented here. First, artificial datasets were generated and analysed to test the algorithms' ability to recover known implanted biclusters. Second, biclustering was performed on a subset of the GNF Gene Atlas mouse gene expression compendium dataset [Su et al., 2002] so that the biological relevance of biclusters discovered by each of the algorithms in a real dataset could be examined. The consistency of the biclusters found by ComBiclust are demonstrated with expression data heatmaps shown in Figures (3.3-3.5). In each case highly-specific biclusters were found for meaningful subsets of samples and with consistent gene expression patterns. Furthermore, the

highest-scoring bicluster found in the noiseless artificial data (Figure 3.3) coincided exactly with the implanted bicluster, and the highest-scoring bicluster found in the noisy artificial data (Figure 3.4) was a component of the implanted bicluster (but also included genes that were consistently differentially expressed by chance). It may also be worth noting that in analysis of the real gene expression dataset, the bicluster of genes consistently differentially expressed in a particular subset of neural tissues (Figure 3.5) includes many genes associated with neural function (e.g. glutamate receptors, synaptosomal-associated protein and 'brain abundant' genes), whereas those shown from the application of SAMBA to the same subset of the GNF data contain fewer obviously neurally-associated genes yet contain 'housekeeping' genes such as tubulinα1 that will be expressed at a relatively high level in all samples.

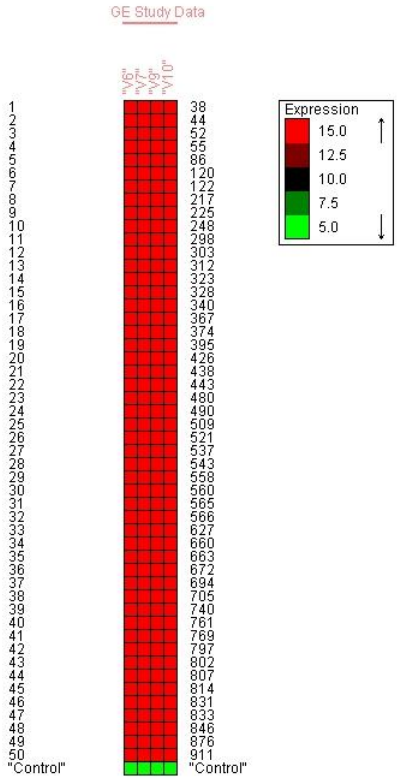


Figure 3.3: Heatmap showing hypothetical expression levels across a bicluster recovered from noiseless artificial data using ComBiclust. This was exactly the implanted bicluster.

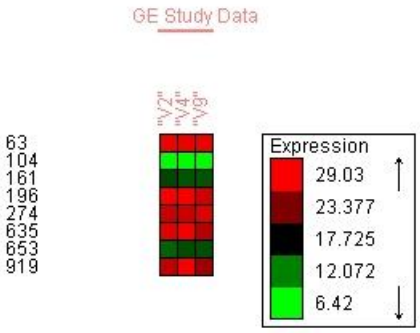


Figure 3.4: Heatmap showing hypothetical expression levels across a bicluster from noisy artificial data using ComBiclust. This was a subset of the whole implanted bicluster.

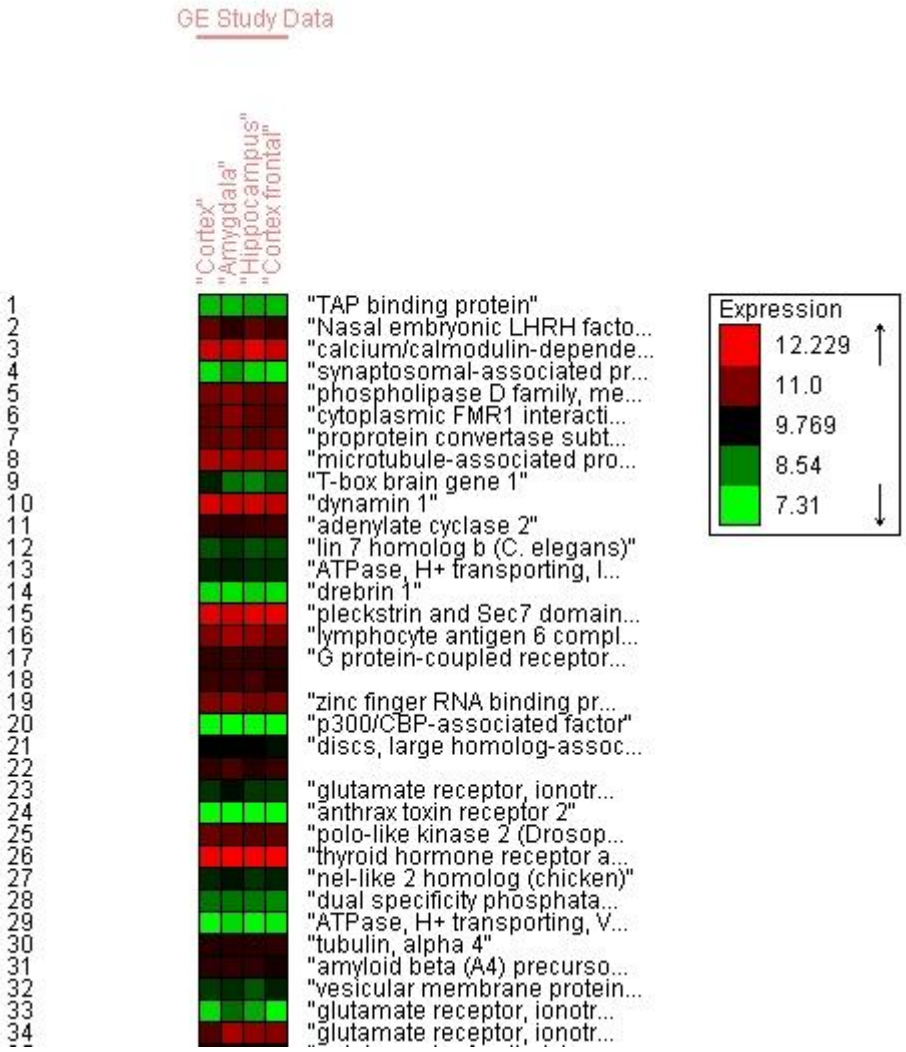


Figure 3.5: Heatmaps showing expression level of genes from a bicluster discovered in a subset of the GNF mouse gene expression dataset [Su et al., 2002] using the novel combinatorial biclustering algorithm.

3.4.5 Genetic Algorithm Approach

As introduced briefly in Section (2.3.4) (under 'Global Search Heuristics'), Genetic algorithms (GAs) are complex adaptive systems that can be used as effective function optimisers in certain domains, and are widely used for problems with especially large solution spaces that have irregular characteristics resulting in a computationally complex search task to find optimal solutions. In particular, they find most application in finding solutions to problems where local search heuristics have a tendency to result in relatively poor solutions.

In Whitley's tutorial on GAs [Whitley, 1994], he defines a GA as 'a population-based model that uses selection and recombination operators to generate new sample points in a search space.' The 'canonical' GA introduced by Holland [Holland, 1975] consists of a population of 'chromosomes' (strings of bits) that each represent a position in the search space and therefore a potential solution to the problem being tackled by the GA. These chromosomes are 'selected' to contribute to the next generation of chromosomes, which they do through reproduction operators: the details of these operators may vary, but they typically create chromosomes for the next generation's population of candidate solutions by 'crossover' and 'mutation.' In crossover, parts of each 'parent' chromosome are swapped with the other, and in mutation, each bit (element) of all the 'offspring' chromosomes is flipped with a low probability those offspring are passed into the next generation. The choice of which chromosomes are selected to be parents for producing the next generation's population is related in some way to the evaluation of success of each chromosome's corresponding solution, given by a 'fitness function.' Clearly, the precise nature of the fitness function is crucial to the success of the algorithm as, if the search mechanism implemented by the GA is successful, it will find those potential solutions that give the best values of the fitness function. A simple and effective selection mechanism is 'fitness proportional selection,' in which each chromosome contributes a number of offspring to the subsequent generation in proportion to its relative fitness score compared to the average fitness score across the whole of the current generation's population of chromosomes.

The motivation for taking a GA approach to the optimisation problem presented in Section (3.4.3) comes from the manner in which reformulation of the biclustering problem results in a natural approach to building solutions through combinations of samples, as demonstrated by the exhaustive combinatorial algorithm presented earlier in this section. As good biclusters are likely to be composed of smaller, good biclusters with similar patterns across different subsets of the large bicluster's samples, it would be likely that the GA mechanism of exploring the solution space through recombination of small, good solutions would prove to be successful. In this way, the nature of the biclustering problem presented here is similar to the ideal GA problems discussed in [Holland, 1975] and [Goldberg, 1989]. Furthermore, the fact that addition of a sample

to an existing, good bicluster can result in a far inferior (or non-existent) pattern of consistency of gene expression across the new bicluster means that 'gradient-based' local search heuristics are unlikely to perform so well due to the greater chance of adding a sample that doesn't fit the bicluster pattern (and therefore results in an inferior bicluster) than one that fits the bicluster pattern and results in a better bicluster.

Representation While GAs have already been applied to the biclustering problem, in most other GAs for biclustering (eg. [Chakraborty and Maka, 2005], [Mitra and Banka, 2006]), both genes and conditions are represented on the chromosomes and these are jointly evolved to search through the space of possible biclusters. However, as discussed in Section (3.4.1) on the reformulation of the biclustering problem, directly encoding the genes in the bicluster may be redundant as, given a subset of samples, it is relatively straightforward to find those genes (if they exist) that are consistently expressed at particular levels across those samples. By encoding only samples, the size of the chromosomes and thus the search space can be kept to a minimum while providing a framework in which those chromosomes representing small biclusters will have a tendency to combine with others sharing consistent gene expression patterns, forming progressively better solutions as the population evolves.

Fitness Function The fitness function is the mechanism by which desirability of solutions is encoded into the GA. Bearing in mind the fact that every potential solution will have to be evaluated by the fitness function in every generation of the algorithm's progress, practical constraints require the desirability of a solution to be expressed in the fitness function in such a way that the value may be calculated very quickly.

As mentioned in Section (3.4.3), the remainder of this chapter following this description and evaluation of an efficient novel biclustering algorithm focuses on theoretical and experimental work advancing definitions of bicluster desirability so that resulting biclusters found by the algorithm in large collections of gene expression data are useful for answering particular biological research questions. From the point of view of testing a novel efficient biclustering algorithm, and given the apparent success of the exhaustive combinatorial algorithm described above (in Section (3.4.4)), the initial GA for biclustering meta-analysis described here took the same, simple approach to evaluating bicluster desirability as that implemented in the exhaustive combinatorial algorithm. That is, the fitness of a chromosome x (a string with a bit for every sample in the dataset: `1` represents 'sample in bicluster' and `0` represents 'sample not in bicluster') is directly proportional to the 'volume' (the number of genes multiplied by the number of samples) of the corresponding bicluster, as shown in Equation (3.4).

$$f(x) = \left(\sum_{g \in genes} \delta_{g,x} \right) * \left(\sum x \right) \tag{3.3}$$

$$\delta_{g,x} = \begin{cases} 1 & \text{if row } g \text{ is consistent across columns for which } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Efficient Biclustering Algorithm With a data matrix that has been preprocessed according to the assumptions defined in Section (3.4.2) and with those points detailed previously in this section, we can define a GA for finding biclusters in large collections of gene expression data. The algorithm proceeds by creating an initial population of a specified number (`popsize`) of chromosomes, each a bit string of the same length as the number of columns in the data matrix, with 2 randomly chosen bits set to `1` and all other bits set to `0`. The population is evaluated to obtain fitness scores by applying the fitness function to each chromosome, chromosomes are selected to contribute to the next generation with frequency in proportion to their relative fitness score, and the next generation of chromosomes is created through applying crossover to randomly selected pairs of parent chromosomes followed by flipping each bit in the offspring chromosomes with a low probability specified by the mutation frequency parameter `mfreq`. This process of evaluation, selection and reproduction is applied iteratively over a specified maximum number (`ngens`) of generations or until algorithm convergence is detected. Convergence occurs when the fitness of the best solution in the population doesn't improve over a specified number of generations. The biclusters encoded by the best chromosomes in the final population are returned as lists of the samples encoded in the chromosome and the genes that have consistent patterns across the subset of the data defined by those samples. A summary of this process is given in Algorithm 3.

Input: Gene expression data matrix, M

Output: Set of biclusters, B achieving good desirability scores

Calculate discretised data matrix, D (as in algorithm 3.1);

Initialise *population* to matrix of 0's (`popsize` rows and same number of columns as D);

foreach *Chromosome (row of population) i* **do**

    Select two randomly-chosen elements to set to 1;

end

foreach *Generation g* **do**

    foreach *Chromosome i* **do**

        Calculate $Fitnesses(i) = f(Chromosome_i)$ ($f(x)$ as Equation (3.4);

    end

    Create *pop*2 by sampling rows of *population* with $freq. \propto \frac{Fitnesses(i)}{mean(Fitnesses)}$;

    foreach *pair of Chromosomes (i,j)* $\in$ *pop2* **do**

        if *random#* $<$ *xfreq* **then**

            Create random integer $crossoverPoint$ between 1 and $|Chromosome_i|$;

            Swap elements of $Chromosome_i$ and $Chromosome_j$ with indices $> crossoverPoint$;

        end

        foreach *element of Chromosomes (i,j)* **do**

            if *random#* $<$ *mfreq* **then**

                Flip bit representing that element;

            end

        end

        Insert resulting chromosome pair into two rows of *population*;

    end

end

Create (empty) list of solutions S;

foreach *Chromosome i* **do**

    Calculate $Fitnesses(i) = f(Chromosome_i)$ ($f(x)$ as Equation (3.4);

    if $\exists k. f(Chromosome_i) > f(S(k))$ **then**

        Set $S(k) = Chromosome_i$;

    end

end

foreach *Solution* S_k **do**

    Create bicluster B_k with empty lists of *biclusterSamples* and *biclusterGenes*;

    foreach *Element i* **do**

        if $B_{k_i} = 1$ **then**

            Add i to *biclusterSamples* for B_k;

        end

    end

    foreach *Gene g* **do**

        if $mean(D(g, biclusterSamples)) > threshold$ **then**

            Add g to *biclusterGenes* for B_k;

        end

    end

It was discovered through application of this fast genetic algorithm to discovery of biclusters in large collections of gene expression that the population had a tendency to converge rapidly on biclusters involving only two or three samples. In an attempt to avoid this premature convergence, the biclustering approach given in Algorithm 3 was adapted to incorporate an 'Island Model' of population evolution. The island model is described in [Whitley, 1994], and involves the separation of the population of chromosomes into subpopulations or 'demes.' By limiting exchange of solutions from one deme to another, premature convergence of the algorithm's population of chromosomes can be avoided. The biclustering genetic algorithm effectively avoided premature convergence when the population of chromosomes was divided into 4 demes that exchanged one chromosome every 10 generations of the GA iteration. The resulting genetic algorithm for biclustering was called 'IslandCluster' on account of this island-based population isolation within the GA mechanism.

3.5 Evaluation Of Efficient Biclustering Approach

IslandCluster is a novel biclustering method developed to enable the application of a biclustering analysis approach for meta-analysis of large collections of gene expression data. Meta-analysis of gene expression data based on biclustering represents an approach to biological data analysis that might provide opportunities to gain insight into the transcriptional mechanisms governing control of biological processes, but has not previously been studied. Following the development of this analysis method, it is clearly essential to demonstrate that the IslandCluster method performs as intended. As the impracticality of application of existing biclustering methods to large-scale meta-analysis of gene expression data is due to the computational complexity of the task in terms of the number of samples represented in the dataset, this is clearly an issue that must be dealt with by any algorithm proposed for this task. Therefore, including the criteria described in Section (2.3.4), there are three principal areas in which such an algorithm must succeed if it is to facilitate the types of analysis described above:

1. **Efficiency**: the method must be able to perform its analysis of large datasets in a practically useful time frame, and that this holds true for datasets of the future that may have data from dramatically more samples

2. **Effective bicluster discovery**: the method must be able to discover biclusters in the data, regardless of how fast its execution times may be

3. **Biological significance of discovered biclusters**: the biclusters in the data must relate to some biological signature within the data, as opposed to some random (or systematic but biologically irrelevant) measurement variation

This section describes the experimental frameworks used to evaluate the IslandCluster algorithm's performance in terms of each of the above criteria for success. Firstly, the feasibility of IslandCluster as a method to discover biclusters in datasets of increasing size is compared with those for existing biclustering algorithms. Secondly, artificial dataset testing is used to demonstrate successful recovery of the desired bicluster patterns by IslandCluster. Thirdly, biological significance of biclusters discovered by IslandCluster in real data is demonstrated through functional enrichment analysis with comparison to performance of existing biclustering algorithms. Finally, this section concludes with a discussion of the significance of these performance evaluation results for the IslandCluster algorithm.

3.5.1 Computational Efficiency

The amount of data available in repositories of gene expression data in the public domain is ever-increasing, at a near-exponential rate [Ball et al., 2003] as the application of high-throughput measurement assay technologies becomes more widespread. With the potential advantages of having more data representing a wider range of biological

conditions comes the computational (not to mention analytical) challenge of finding relevant patterns and sub-structures within a vast array of measurements. As discussed in Section (2.3.4), the potential of biclustering methods for application to analysis of such large datasets is hindered by the complexity of the biclustering problem. A few publications have proposed biclustering approaches with claims for applicability on a large scale (such as [Tanay et al., 2002], and [Li et al., 2009]), although in order to perform the biclustering meta-analysis proposed in this chapter it was necessary to employ a method that could analyse, as one, collections of whole-genome transcriptional data involving thousands to tens of thousands of samples (and for future potential, even more than this).

To demonstrate the potential of any biclustering approach to be used as a basis for large-scale meta-analysis, variously sized subsets of samples from a large collection of gene expression microarray data were constructed so that the efficiency of performance of each biclustering approach could be assessed in terms of the feasibility of executing biclustering analysis on a range of dataset sizes, with a focus on the number of samples represented in the dataset. This focus on the number of samples as opposed to number of genes in the dataset comes from the fact that, no matter how many measurements are taken for each sample, the number of unique genes is constant for a given organism (and even the number of uniquely transcribed sequences is likely to remain constant at a similar order of magnitude). Therefore, as technologies advance and the use of high-throughput transcriptome measurement platforms increases, the number of samples available for meta-analysis will continue to increase while the number of genes remains (approximately) the same.

Experimental Procedure

In order to evaluate the potential of different algorithms to be applied to large-scale meta-analysis tasks, it was necessary to determine whether or not each algorithm could feasibly complete execution on datasets of a range of sizes. As described above, the particular concern is the growth of execution time as the number of samples in the dataset increases. Therefore, one very large data matrix was assembled from nearly 8,000 microarray samples (from the Affymetrix MOE430v2 platform) downloaded from the NCBI's Gene Expression Omnibus [Edgar et al., 2002] and compiled into one dataset according to the description given in Section (3.4). A number of submatrices with increasing numbers of columns were sampled from this whole data matrix: submatrices were created with 10, 20, 50, 100, 200, 500, 1000, 1500, 2000, 3000, 5000 and 7990 columns. All submatrices included all 45,101 probesets included on the microarray platform.

Each algorithm was used in turn to perform biclustering analysis on each of the submatrices. Those biclustering methods with available implementations that appear to be

most widely used, and those in the literature reporting to be fast enough for large-scale application were used for comparitive purposes. This list of algorithms included: XMotif [Murali and Kasif, 2003], Bimax [Prelic et al., 2006], ISA [Bergmann et al., 2003, Ihmels et al., 2004], OPSM [Ben-Dor et al., 2004] and SAMBA [Tanay et al., 2002]. All algorithms were run on the same computer with a quad-core processor and 24GB of RAM, running Centos Linux. The results of all of these algorithm execution runs are presented in Fig. 3.6.

Subsequent to the development of the IslandCluster algorithm and the majority of the following work presented in this thesis, two biclustering algorithms have been published reporting applicability to such large-scale analysis: Qubic [Li et al., 2009] and *COALESCE* [Huttenhower et al., 2009].

Results

It is clear from the indication of completion of biclustering runs by each of the tested algorithms given in Fig. 3.6 that only IslandCluster could return output from datasets involving thousands of samples within feasible analysis times.

The implementation of Bimax provided in the `R` package `Biclust` returns a set number of biclusters discovered in combinations of the genes and samples across only as large a subset of the dataset as is required to identify the specified number of biclusters. This implementation was able to return a set number (100) of biclusters for all datasets evaluated. However, the implementation of Bimax provided in the tool *BicAT* [Barkow et al., 2006], which performs exhaustive biclustering analysis of the input dataset, did not terminate within 24hrs when run on a dataset comprising only 100 samples. This method would not therefore be appropriate for investigation of large-scale biclustering meta-analysis.

No. of Samples	10	20	50	100	200	500	1000	1500	2000	3000	5000	7990
IslandCluster	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bimax	Yes	Yes	No	No	No	No	No	No	No	No	No	No
ISA	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No
Plaid	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No
Xmotif	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No
CC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No
OPSM	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No
SAMBA	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No

Figure 3.6: Table indicating success or failure of different biclustering methods at processing collections of gene expression data made with increasing numbers of samples (from left to right in table). Successful processing of a dataset by a method is indicated with a green box, failure to process dataset indicated with a red box. It is obvious that any methods unable to process these datasets will be unsuitable for large-scale gene expression data meta-analysis.

The results presented here agree with the report given in [Li et al., 2009] that no existing algorithms in 2009 were capable of performing biclustering on datasets with over 12,000 genes and 1,000 samples. However, the IslandCluster algorithm was able to return a result for datasets with 45,101 probesets and nearly 8,000 samples, demonstrating that it was clearly efficient enough to perform large-scale meta-analysis of gene expression data, and was thus the first biclustering method to achieve this property.

3.5.2 Artificial Dataset Testing

As discussed above, in order to demonstrate a method's potential applicability to a task it is essential to prove that the implementation successfully discovers the patterns it is intended to find. For pattern-mining methods this is most commonly demonstrated using artificially-generated datasets with known structure, containing deliberately implanted patterns of interest in known locations within the dataset. Such datasets enable a situation ideal for validation purposes in that the 'truth' (i.e. the perfect solution) is known, so the success of a method can be easily measured by comparing the output from the method with the ideal solutions and summarising the overlaps. Further analysis of true positive, false positive, true negative and false negative rates can also be performed if necessary.

The drawback of such artificial dataset testing stems from the fact that the datasets tend to be generated from the same model of the real data that is used (either explicitly or implicitly) by the method to identify patterns of interest. In the case here, this means assuming that there exist bicluster patterns within the data and that these have the same (or at least a similar) definition to the biclusters the algorithm is supposed to find. Additionally, whatever model is used to generate the artificial data, the model will be inaccurate in some way and there is always a chance that equivalent results would not be observed in a real application of the method. For these reasons, a thorough evaluation of a pattern-mining method's performance must also include an assessment of some related desirable property of the output from its application to real data, but artificial dataset testing provides an important basis for *proving* that the implementation of the algorithm succeeds in finding the patterns it is supposed to find, regardless of whether or not the discovery of such patterns solves the task at hand.

In order to demonstrate the novel algorithm's ability to discover biclusters in data matrices, artificial gene expression datasets were created with implanted biclusters in known locations, and the recovery of these implanted biclusters through use of the novel biclustering method was examined.

Experimental Procedure

A simple program was developed to construct data matrices with specified ranges of 'background' values for each row. A number of artificial biclusters were implanted into the data by setting the data matrix values for a specified number of randomly sampled rows across a specified numbers of randomly sampled columns to a specified multiple of the other values. A record of the rows and columns modified for each bicluster is provided in the program's output. Finally, uniformly-distributed random 'noise' is applied to the matrix by multiplying each value in the data matrix by a factor randomly sampled from a specified range. Using this program, 50 datasets with 5000 rows and 50 columns were created with 1 to 5 biclusters varying in dimensions from 500 rows by 2 columns to 2000 rows by 20 columns. Noise across these datasets ranged from 0 (resulting in discrete levels for each row corresponding to whether a value is in a bicluster or not) to 1 (potentially eradicating the effect of bicluster implanting).

These datasets were analysed with each of a panel of biclustering algorithms: *Island-Cluster*, ISA [Bergmann et al., 2003, Ihmels et al., 2004], Bimax [Prelic et al., 2006], Cheng & Church's algorithm [Cheng and Church, 2000], XMotif [Murali and Kasif, 2003] and Plaid [Lazzeroni and Owen, 2002]. Algorithm implementations provided in the `R` packages `Biclust` and `isa2` were used for the evaluation. The biclusters returned by each of the algorithms were tested against each of the implanted biclusters from the corresponding dataset. An F-measure was obtained for each {recovered bicluster,implanted bicluster} pair to summarise the ability of the recovered bicluster to specifically and accurately match the implanted bicluster. The F-measure incorporates both precision and recall, as shown in Equations (3.5-3.7) which give the formula for the F-measure for overlap of two sets A and B. In this case, A and B represent the genes of the recovered and implanted biclusters, respectively. To get an overall measure of the overlap between a recovered bicluster and an implanted bicluster, the F-measure for the genes was multiplied by the F-measure for the samples.

$$precision = \frac{|A \cap B|}{|A|} \tag{3.5}$$

$$recall = \frac{|A \cap B|}{|B|} \tag{3.6}$$

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{3.7}$$

As each recovered bicluster could represent any one of the implanted biclusters (or none at all), the best F-measure for each recovered bicluster with any of the implanted biclusters in the respective dataset was taken as representative of that bicluster's success.

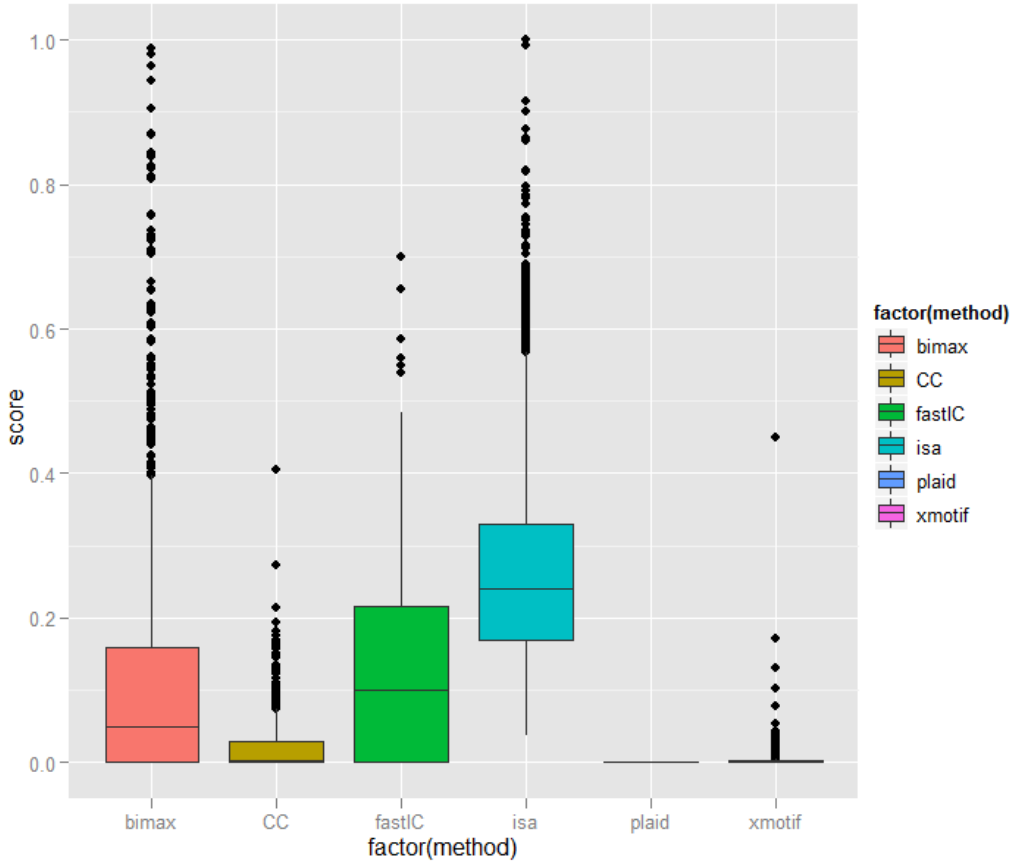


Figure 3.7: Distribution of overall implanted bicluster recovery scores for biclusters identified by each of a panel of algorithms in a set of 50 artificially-generated gene expression datasets. IslandCluster method is labelled in the plot as 'fastIC'.

Results

With scores for each bicluster returned by each biclustering algorithm from each artificial dataset, the success of the algorithms could be compared overall and for each dataset. As various properties of the datasets were varied, such as the number of biclusters implanted and the level of noise obscuring the implanted biclusters, the effect of these properties on the success of each algorithm could also be demonstrated. Fig. 3.7 shows the overall distribution of scores for biclusters found by each algorithm, demonstrating that IslandCluster recovers the implanted patterns with a similar degree of effectiveness to that of BiMax and better than the remaining algorithms, with the exception of ISA, which performs the best. Fig. 3.8 shows the distribution of bicluster scores for each dataset as the level of noise increases, which seems to indicate that the ability of most of the algorithms to recover implanted patterns is reduced with increasing noise in the data.

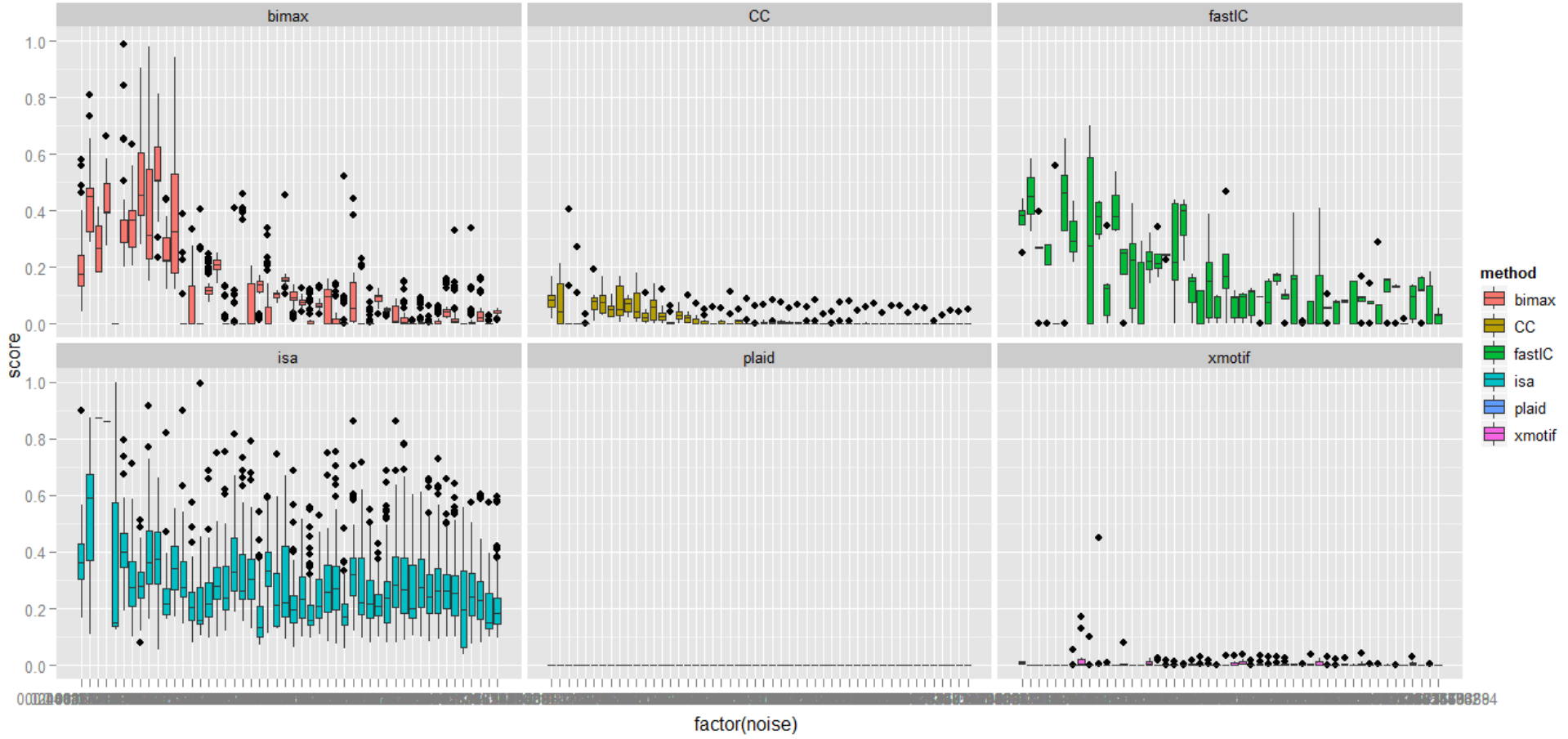


Figure 3.8: Distribution of implanted bicluster recovery scores separated into individual artificial datasets, shown in a separate panel for each algorithm with the noise across the dataset increasing from left to right across each panel. Recovery scores for IslandCluster are shown in top right panel, labelled 'fastIC'.

These results demonstrate that, despite being sufficiently scalable to return biclusters from datasets with thousands of samples, the IslandCluster algorithm recovers implanted biclusters in artificial datasets better than many existing biclustering algorithms. It did not perform quite as well as ISA, which is consistently among the best performing algorithms in a number of published comparative evaluations [Prelic et al., 2006, Ihmels et al., 2004], but it still achieves a high degree of success when compared with existing algorithms reported to perform effectively in [Prelic et al., 2006],[Cheng and Church, 2000],[Murali and Kasif, 2003] and [Ihmels et al., 2004]. From this it can be inferred that the fast IslandCluster algorithm is not only able to perform biclustering analysis on collections of gene expression data on a scale impossible with existing biclustering methods, but is also able to find the intended localised differential expression patterns within data matrices. However, it remains to be demonstrated that such patterns in real gene expression data correspond to biological signatures.

3.5.3 Biclusters in Collections Of Real Data

While it is useful to be able to demonstrate that an algorithm finds the desired patterns when these are implanted into artificially-constructed datasets, this does not show that the given algorithm represents a successful approach to performing the desired analysis task on real data. In addition to showing that the algorithm finds the patterns of interest effectively, it is also necessary to demonstrate that those patterns of interest, when found in real data, provide some useful insight into a real problem. That is, it remains essential to provide evidence that the approach to the given analysis task, involvding the application of the algorithm in question, is of some worth (in a real situation).

Given the nature of the systems being studied in this work (i.e. mammalian transcription), in particular the complexity involved and the amount that is not known about the system, an evaluation approach similar to that taken with artificial datasets is infeasible: the body of knowledge is not sufficiently complete to be able to confirm that a transcriptional association between two genes predicted by a method but not previously reported as having strong (or indeed any) experimental evidence is a failure in the application of the method. Bearing this in mind, it is necessary to find a way of evaluating the success of the algorithm at performing the desired meta-analysis task.

In the case studied here, there is a wide range of possible applications (in terms of inferring useful information regarding biological systems) of any gene expression patterns identified through biclustering, and this further complicates the evaluation process through a lack of availability of a clear application scenario wherein a failure to infer any known information corresponds to an obvious failure of the algorithm. The goal of evaluation therefore becomes, in a general sense, to provide evidence that the

discovered biclusters have some biological significance, and in more specific cases this can be demonstrated through the successful application of information inferred from discovered biclusters to particular prediction or association tasks.

A group of potential approaches to suggesting biological significance of biclusters involves identifying statistically significant sets of genes that are implicated in the same biological processes. Such a signature within a bicluster suggests that the gene expression pattern revealed by the bicluster represents the transcriptional regulation of some (set of) biological process(es). As a relatively comprehensive resource of annotations of biological processes associated to individual genes, the Gene Ontology (GO) [Ashburner et al., 2000] represents a tool with which this annotation-based evaluation could be performed. The use of statistical enrichment of GO terms to identify biological patterns in gene lists is widespread, with examples of tools for performing such analysis including: *DAVID* [Huang et al., 2009b], *FuncAssociate* [Berriz et al., 2003] and *GOstats* [Falcon and Gentleman, 2007]. Following the precedent set in the [Prelic et al., 2006] review of biclustering algorithms, many biclustering methods use the proportion of biclusters enriched for any GO terms at a given statistical significance threshold as a measure of success of an algorithm's ability to discover biologically significant (and some go as far as to say 'relevant') gene expression patterns. However, as pointed out in [Myers et al., 2006], there are situations in which such assessment of biological significance can be flawed: not least because it is not usually interesting to know that a particular gene list is statistically enriched for genes associated with some especially broad category such as 'transcription.' For example, it may be very well knowing that you have associated a transcription factor, say, to a number of other transcription factors, but unless these are involved in regulating the transcriptional activation of related biological processes then this doesn't indicate any (functionally) significant association and therefore biologically relevant or significant signature represented by the gene list. Additionally, a particular problem was demonstrated in [Myers et al., 2006] regarding the dataset used in the [Prelic et al., 2006] (and many subsequent) GO evaluations. Gene function prediction analysis based on the yeast microarray dataset from [Gasch et al., 2000] has been shown to be heavily dependent on a single GO category, 'ribosome biogenesis.' As the comparative evaluation presented in [Prelic et al., 2006] did not consider which GO terms were being enriched, it is possible that the ribosomal signature from this dataset may bias the evaluation results. While a true demonstration of success of a bioinformatics pattern-mining algorithm ought to come from the use of the algorithm as part of a research project to find answers to a particular biological question, as a preliminary indication that the novel biclustering algorithm could find biologically significant gene expression patterns in large datasets it seemed that demonstration of the consistent presence of relevant biological process signatures within bicluster gene lists would suffice. This allows for comparison with other biclustering or meta-analysis gene-association methods using the same datasets. The IslandCluster al-

gorithm was used, along with a number of methods included for comparitive purposes, to identify gene expression patterns within real datasets. The gene lists arising from the discovered patterns were evaluated for enrichment of relevant GO terms, and the distribution of such enrichments over a set of genelists for each method was calculated. A full description of these analysis methods is given below, followed by presentation of the results.

Experimental Procedure

For assessment of the IslandCluster algorithm's ability to discover biologically significant biclusters in large collections of data, bicluster analysis was performed on the dataset described in Section (3.4), containing data from 7,990 microarrays measuring whole-genome gene expression levels in mouse samples from a wide range of biological contexts and conditions. For each bicluster found by the algorithm, those genes (best) fitting the bicluster expression pattern were tested for enrichment of relevant GO term associations using the 'conditioned' testing provided by the GOstats tool and described in [Falcon and Gentleman, 2007]. The distribution of enrichment scores across each of the biclusters is presented in Fig. 3.9. The scores given in Fig. 3.9 represent $-\log$ p-values, so the greater the score the higher the enrichment. Only GO term enrichments with p-values smaller than 0.01 are included in this analysis.

In order to compare the performance of IslandCluster with that of other biclustering algorithms, smaller datasets had to be used as no existing algorithm could perform biclustering on the whole dataset used in the above analysis. Therefore, a smaller subset of this compendium of mouse microarray data was used, with only 500 samples selected. Biclustering analysis was performed on this dataset using ISA, Plaid, Bimax, Cheng & Church's algorithm and IslandCluster. For each algorithm, the proportion of biclusters showing enrichment of any GO terms from the *biological process* ontology to different minimum significance thresholds is shown in Fig. 3.10, similar to the figure provided in [Prelic et al., 2006].

Results

The enrichment of GO biological process terms across biclusters discovered by the IslandCluster algorithm in the 7,990 microarray mouse gene expression dataset is demonstrated in Fig. 3.9. It is clear from Fig. 3.9 that the biclusters discovered by the IslandCluster algorithm in large gene expression datasets generally appear to represent some biological signature extracted from the data, given that all biclusters enrich some GO categories with a p-value less than 0.01 and a large number of biclusters enrich at least some GO categories with a p-value of less than $1 * 10^{-16}$, which equates to a $-\log$ score of greater than approximately 53.

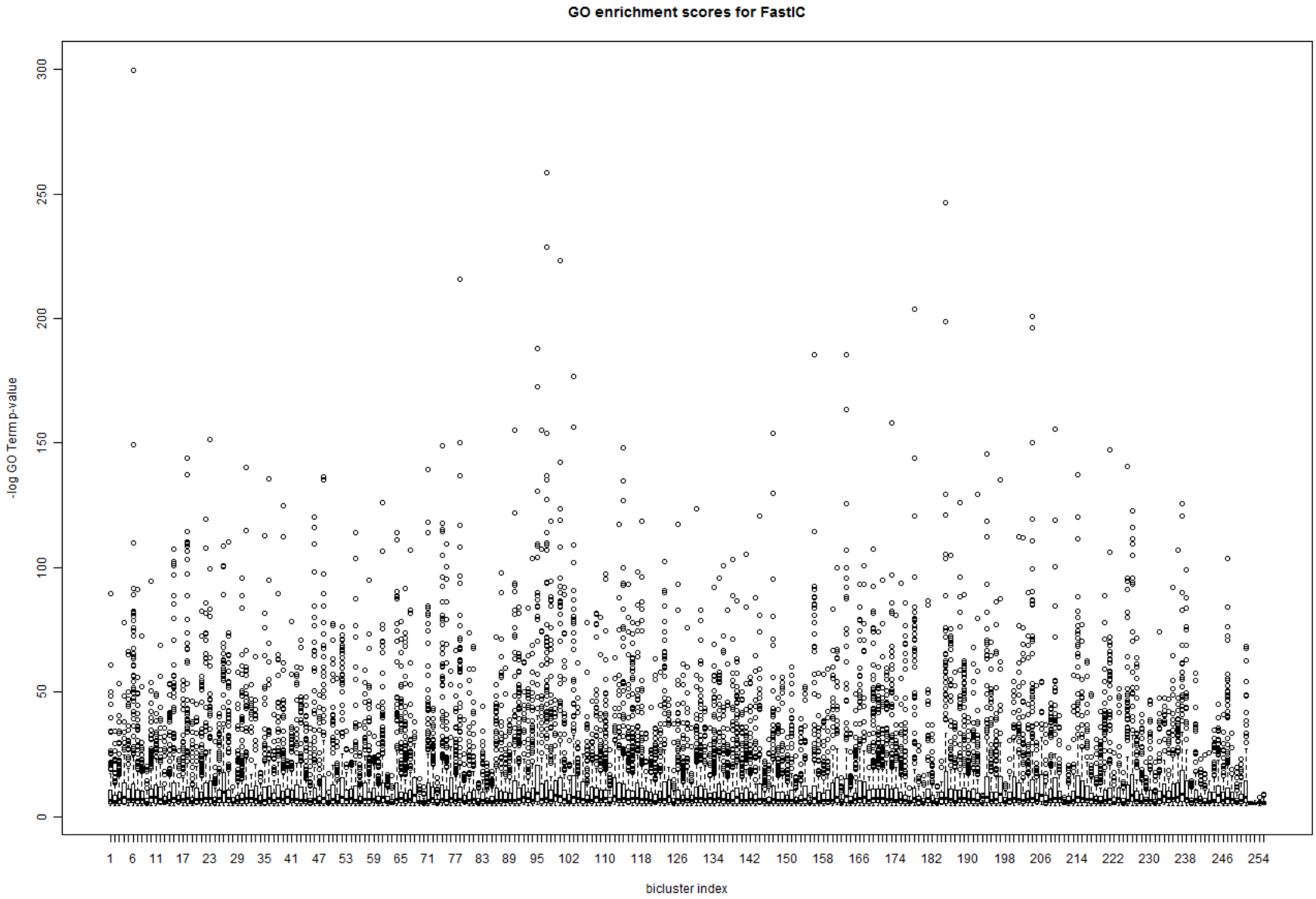


Figure 3.9: Enrichment score distributions for individual biclusters (with less than 75% overlap) identified in the 7,990 sample gene expression dataset described in Section (3.3) using the IslandCluster method. Enrichment scores are given as $-\log$ p-values, so a greater score implies a more significant enrichment. Scores are shown for all GO terms with a conditional enrichment significance p-value of lower than 0.01, as calculated by GOstats [Falcon and Gentleman, 2007].

To provide comparative evaluation against existing biclustering algorithms widely cited in the literature, Fig. 3.10 shows the proportions of biclusters, discovered in the same 500-sample subset of the dataset described in Section (3.3) by each of the algorithms mentioned through Section (3.4.5), with genelists enriched for GO biological process terms to a range of minimum significance thresholds.

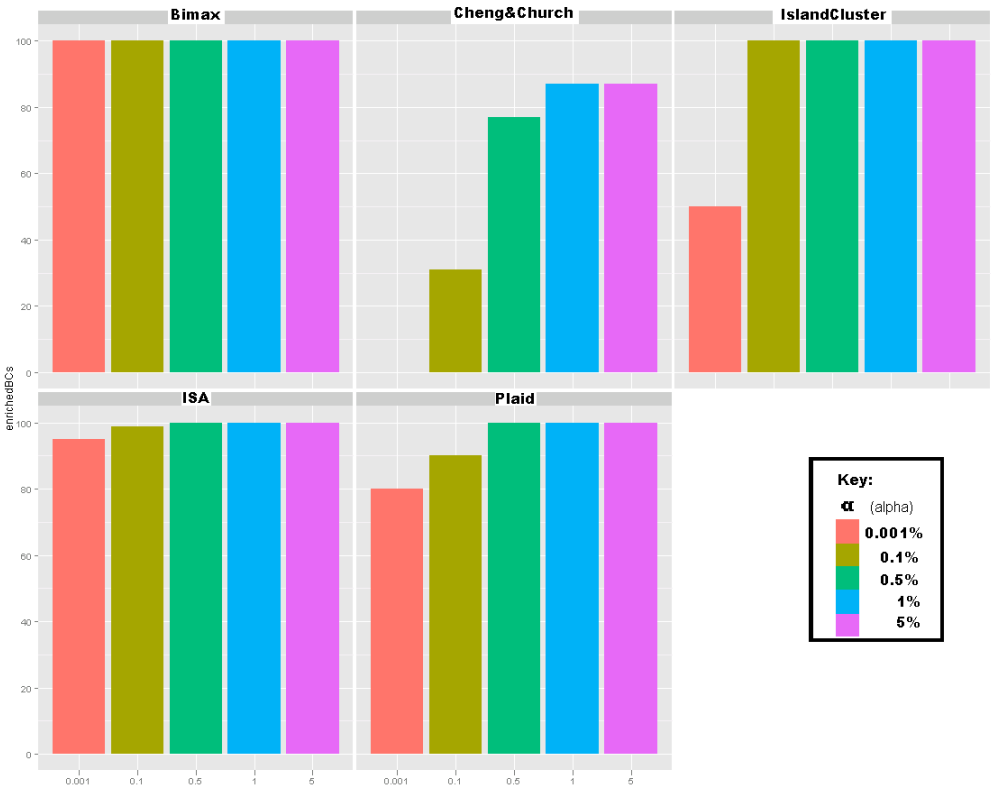


Figure 3.10: Proportions of biclusters enriched for any GO terms to a range of minimum significance thresholds. Each panel reflects the results for biclusters discovered using a different biclustering method.

It was noted that the calculated enrichments were highly significantly correlated with the number of genes in the bicluster, regardless of the method used to calculate the biclusters, with Pearson $\rho = 0.51$ and significance p-value from correlation test $p < 2.2 * 10^{-16}$. The typical numbers of genes in biclusters were found to vary between the biclustering methods used, as illustrated in Fig. 3.11.

As the GO enrichments were found to be highly dependent on the numbers of genes in biclusters, and the distribution of numbers of genes in biclusters varied across different methods, this might bias the comparison presented in Fig. 3.10 toward methods that return biclusters with large numbers of genes (i.e. Bimax and ISA). Therefore, a similar comparative evaluation was performed for only biclusters with 250 or fewer genes. The results of this comparative evaluation are shown in Fig. 3.12. The results presented

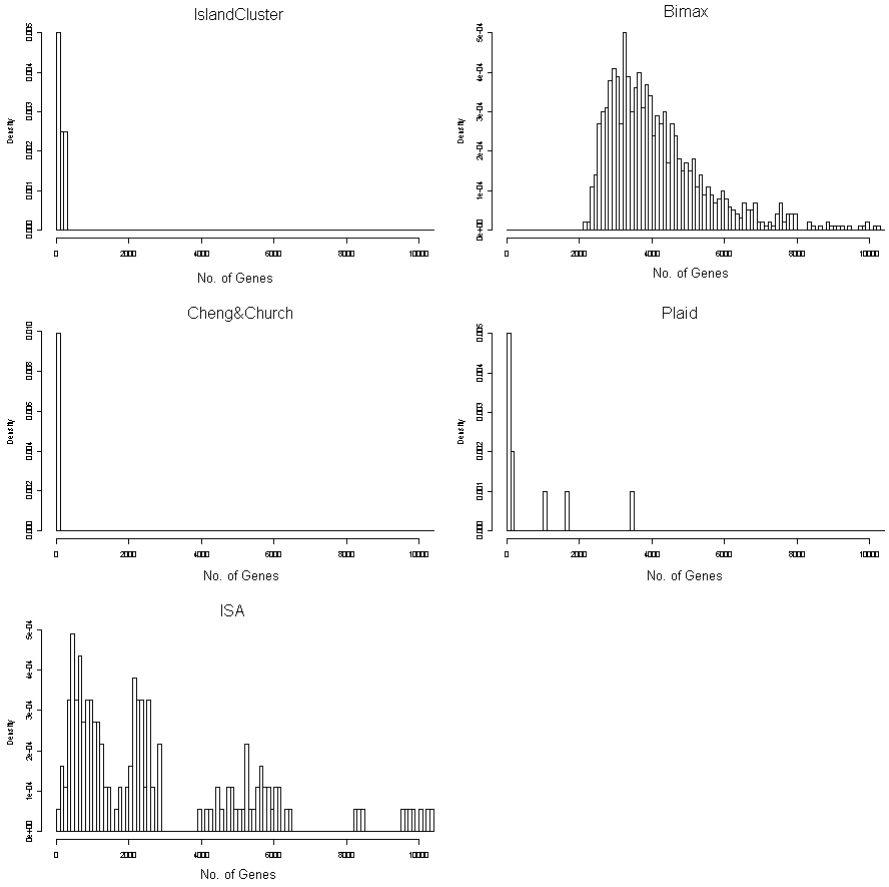


Figure 3.11: Histograms showing distribution of numbers of genes in biclusters found by each biclustering method when applied to the 500 sample subset of the large gene expression data matrix. Histogram bar heights represent the relative proportion of biclusters discovered by the algorithm in question that include a number of genes within the specified range. The horizontal axis indicates the numbers of genes in the biclusters, ranging from 0 (left) to 10,000 (right) in each panel.

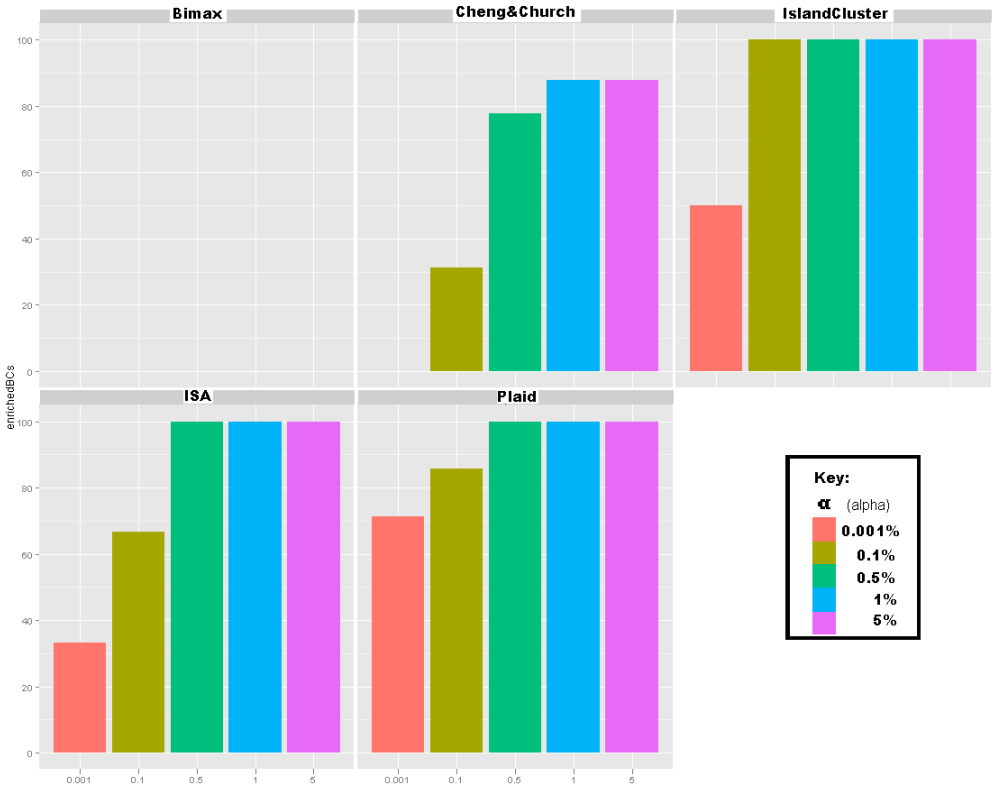


Figure 3.12: Proportions of biclusters enriched for any GO terms to a range of minimum significance thresholds. Each panel reflects the results for biclusters discovered using a different biclustering method and containing no more than 250 genes.

here indicate that in terms of recovering biclusters that represent functional biological signatures within a gene expression dataset, the IslandCluster algorithm seems to perform at least as well as a number of the most widely-used existing biclustering methods.

3.5.4 Discussion: Implications Of Efficient Biclustering Method

The immediate impact of this demonstration of a novel biclustering algorithm capable of being used for large-scale meta-analysis of gene expression data is that, for the first time, this application of the biclustering paradigm could be studied. It became possible, as a consequence of the development of the IslandCluster algorithm, to identify meta-analysis tasks in which using a biclustering approach to discover locally-significant patterns in gene expression data could provide useful insight into biological research problems.

This approach presented above could be used as a starting point for refinement and optimisation of the biclustering approach when applied to specific meta-analysis tasks, taking into account specific structures identified within the data and their relevance (or lack of it) to particular biological processes involving an element of transcriptional regu-

lation. Given that the datasets represent such complexity as the whole transcriptome of a given organism's cells in a vast collection of conditions, there ought to be a great deal of scope in adapting the subtle pattern-mining paradigm of biclustering to identification of specific locally-significant patterns in the data that can answer particular biological questions. It is only the development of methods capable of performing this analysis on such a large scale that enabled the potential of specifically adapted (biclustering-based) pattern mining approaches to be realised. A clearly useful application of such an efficient approach involves the assessment of how biclustering-based approaches to large scale gene expression meta-analysis may yield results with application to particular biological investigations, and the reasons why certain biclustering-based approaches may not provide results with application to particular biological questions.

3.6 Further Development Of Biclustering Approaches For Improved Meta-Analysis Of Gene Expression Data

As referred to in Section (2.3), different methods of analysis will best suit different applications. The efficient biclustering algorithm presented above constitutes a general approach for identification of biclusters in very large gene expression datasets, where the discovered biclusters represent groups of genes that are consistently expressed at a relatively high level over a particular group of samples. When the goal of the analysis to be performed is something other than finding any such groups of genes consistently expressed across any (given) group of samples, this novel approach may be adapted to find some other patterns of interest based on the biclustering principle that will be more useful in helping to provide insight into the particular biological question being studied. For example, the definition of the bicluster evaluation function (i.e. the GA's fitness function) allows for relatively straightforward encoding of alternative ideas of ideal bicluster patterns for the algorithm to identify within the data.

The following sections of this chapter present alterations of the general method represented by the IslandCluster algorithm, the motivation behind each alteration and the consequences of the corresponding development of the approach in the form of comparative analysis of the biclusters discovered using the method with and without the alteration. With each of these approaches, the desired meta-analysis task remains, as a general concept, inference of possible transcriptional relationships between genes. In this area the utility of the relationships it might be possible to infer as a result of the discovered biclusters is considered, and is often the motivation behind the specified alteration to the method. The first alteration considered presents improvement in the GA mechanism underlying the algorithm, while the subsequent alterations concern the definition of ideal bicluster solutions and affect the properties of the discovered biclusters.

3.6.1 Modification Of Genetic Algorithm Approach

The canonical GA [Holland, 1975] provides a mechanism for finding a good solution to a difficult function optimization problem. While this mechanism is often successful [Whitley, 1994], when applied to problems with many good solutions a canonical GA will typically find only one of these good solutions [Cedeno et al., 1994]. In the case of the IslandCluster algorithm described in Section (3.4.3), the 'Island' population model goes some way to ensuring a diversity of solutions across the distinct subpopulations. However, to guarantee finding multiple distinct biclusters in a dataset, biclusters discovered in a previous run of the GA must be 'masked out' so that any subsequent runs identify distinct solutions. This approach has two obvious drawbacks:

1. As a component (bicluster) is removed from the data, this prevents any biclusters

with overlapping regions from being discovered

2. The execution time for discovering biclusters increases with the number of biclusters discovered, presenting a clear trade-off between completeness of discovery of biclusters and time taken for analysis to complete

If the goal of the biclustering analysis to be performed is to identify distinct biclusters simultaneously, the problem can be considered an attempt to find optima of all significant modes of a multi-modal function. That is, the bicluster discovery becomes a multimodal function optimization problem. Multi-modal search spaces (such as the sample-set space involved in the biclustering problem as defined in Section (3.4.3)) can exhibit the property that there may be a number of distinct, good solutions.

Multi-Niche Crowding

A family of alternative GA methods based on the 'crowding' principle has been developed to avoid these problems encountered in the application of GAs to multi-modal function optimization problems. The principle of crowding, as introduced by De Jong in [De Jong, 1975], provides an alternative approach to the replacement of solutions in the current generation with offspring solutions destined for the next generation. Rather than replacing all solutions in the current generation with all the offspring solutions (as in the canonical GA), for each offspring in turn a random sampling of a specified number of solutions is taken from the current generation and the most similar of these to the given offspring solution is replaced. This replacement of similar individuals in the population will delay the convergence of the population onto only one optimum of the function being optimized, by maintaining stable subpopulations. However, when this is combined with the fitness-proportional selection of the canonical GA, one of the subpopulations will almost inevitably dominate the population [Cedeno et al., 1994] and thus the ability to identify different local optima of the multi-modal search space is not achieved.

Such multi-modal optimization is the goal of the 'multi-niche crowding' (MNC) method developed by Cedeno [Cedeno, 1995] This method aims to evolve subpopulations by encouraging reproduction and replacement to occur among similar solutions. First, the selection process uses 'crowding selection' in which a random solution is selected for reproduction: a 'mating pair' is selected as the most similar from a group of solutions randomly sampled (with replacement) from the population. Second, the replacement procedure known as 'worst among most similar' is used, in which a number of separate *crowding factor groups* are created through random sampling (with replacement) across the population. A candidate for replacement is chosen from each crowding factor group by picking the solution in the group that is most similar (in terms of the eventual output corresponding from the solution) to the offspring in question,

and the candidate with the lowest fitness is replaced by the offspring. The improved performance of the MNC GA when compared to other methods of multi-modal function optimization is demonstrated in [Cedeno et al., 1995, Cedeno and Vemuri, 1996]. The MNC GA approach was adopted to produce a biclustering algorithm, termed the MNC BGA. The MNC BGA is based on the bicluster evaluation performed by IslandCluster, but better suited to the multi-modal search involved in finding multiple distinct biclusters in a large gene expression dataset.

3.6.2 Revisiting Bicluster Definition 1: Probabilistic Biclustering

In some sense, the enrichment of GO terms in biclusters discovered using the IslandCluster algorithm demonstrates some biological significance of the discovered biclusters through co-annotation of the gene. However, a simple analysis of the distribution of 'dimensions' (the numbers of genes and samples represented by the bicluster) of the discovered biclusters reveal startling observations that also apply to the existing biclustering algorithms used for comparison in the evaluations in Section (3.5). The majority of the biclusters found by any of these methods tend to cover a relatively large proportion of the genome, making any specific functional inference regarding the regulatory activity of any of the included genes practically impossible. The result of this analysis of a set of biclusters discovered by the MNC BGA is shown in Fig. 3.13. A further problem comes from the fact that in any of the biclustering methods described so far in this thesis, there is no prioritisation or ranking of the genes belonging to a bicluster. If a bicluster includes over 1000 equally-ranked genes, as is the case for 63% of the biclusters shown in Fig. 3.13, it will not have much utility in suggesting suitable candidates to follow up with experimental investigation. Such utility would be greatly increased with some measure of how well a gene characterises the transcriptional pattern represented by a bicluster.

One possible explanation for this problem where existing biclustering approaches tend to return large unranked genelists relates to the naivety of the model used to evaluate bicluster desirability. While based on the principle of entropy (from information theory) that a bicluster that is less likely to appear by chance is a more informative bicluster, the assumption that the bicluster 'volume' (number of samples $*$ number of genes) is in directly inverse proportion to the probability that those samples and genes would be present in a random bicluster purely by chance is clearly far from accurate. Inaccuracies in the estimation of probability of a bicluster's genes and samples being included in the bicluster by chance might lead to difficulties in the interpretation of biclustering results. Consequently, it was hypothesised that improving these probability estimates might lead to improvements in the utility of biclusters discovered with the MNC BGA. The remainder of Section (3.6.2) therefore considers the development of a more sophisticated model for the information content of a bicluster and the effect of incorporating such a model into the MNC BGA.

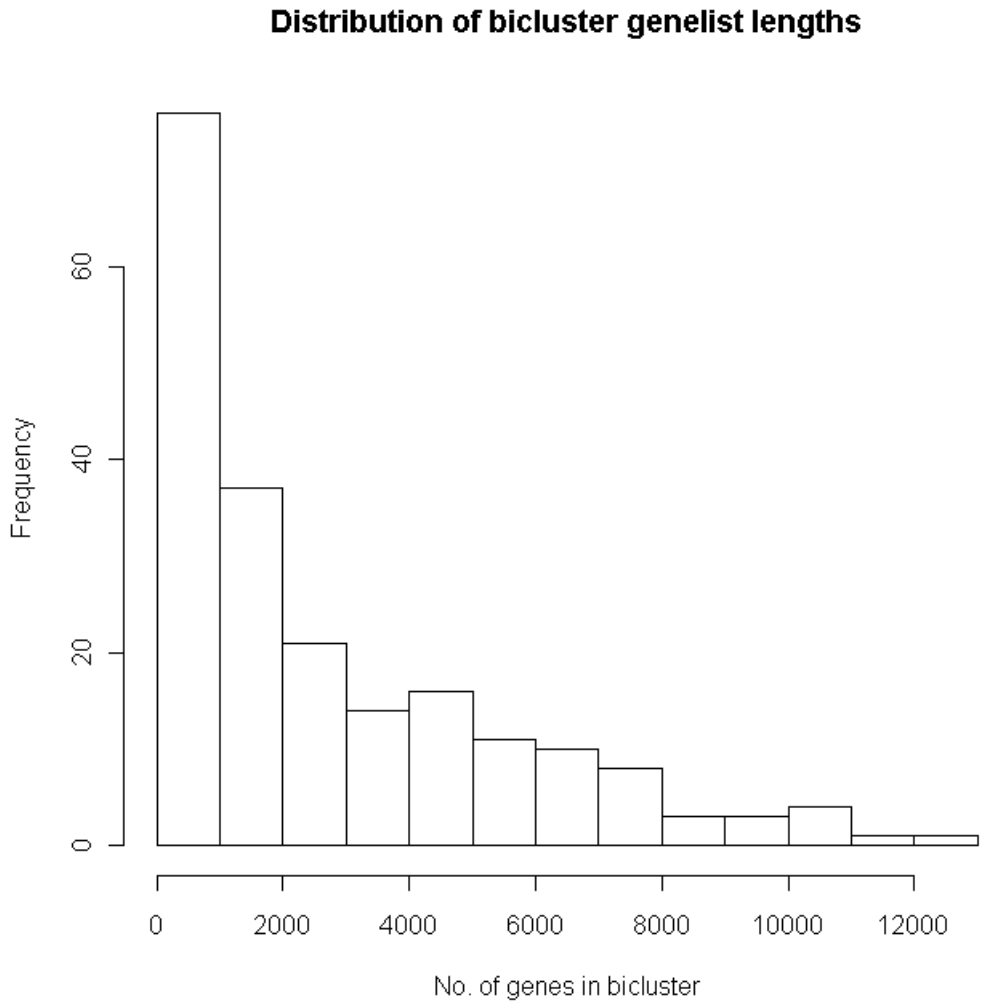


Figure 3.13: Histogram showing the distribution of numbers of genes across a set of 205 biclusters discovered using the MNC BGA, applied to the dataset described in Section (3.4). The height of the bars represent the number of biclusters with genelist length lying in the given range.

Recently a biclustering algorithm called FABIA was proposed in [Hochreiter et al., 2010]. This method for bicluster discovery considers ranking biclusters according to information content, described as the proportion of variance in the whole dataset explained by the bicluster, and therefore makes some progress towards the objectives described here in Section (3.6.2). However, the FABIA approach does not involve explicit models of the different distribution patterns of expression of different genes, and thus interpretability of the measures of bicluster information content is restricted to ranking of the biclusters and can not be used to infer the significance an observation of a particular gene belonging to a particular bicluster. Furthermore, the FABIA method has not been used to perform bicluster analysis on datasets from more than a few hundred samples, and is unsuitable for meta-analysis of gene expression data on the scale considered in this thesis.

Entropy-Based Biclustering

A number of observations regarding general properties of biclusters discovered using the IslandCluster and MNC BGA methods led to specific areas in which the naive model for estimating bicluster desirability, as implemented in IslandCluster and the MNC BGA, could be improved. For a start, as the number of samples in the dataset is typically significantly smaller than the number of genes, the presence of any additional sample fitting the bicluster pattern will in these cases be less likely to occur by chance than an additional gene. Therefore, the number of samples and number of genes in the bicluster ought to be appropriately weighted in terms of significance to the calculation of bicluster desirability, in accordance with the dimensions of the dataset.

In the biclusters discovered by the MNC BGA that utilised this naive model of bicluster probability, it was observed that some genes were more likely to occur by chance in a randomly selected bicluster pattern than others. This situation arises due to the fact that different genes had different proportions of 'high' and 'low' expression states in the dataset. Inspection of the distributions of expression values for a panel of genes (as in Fig. 3.14) indicates that this observation of different proportions of 'high' and 'low' expression states for different genes reflects the underlying expression values. Each gene would be equally likely to occur in any bicluster by chance if and only if all genes had the same proportion of 'high' and 'low' expression states across the dataset. As this could not accurately reflect the underlying distribution of expression values of all genes, as illustrated by Fig. 3.14, it seems pertinent to address the probabilistic modelling issues arising from the situation in which some genes may be more likely than others to occur by chance in any bicluster.

A more accurate model for bicluster-pattern occurrence by chance would therefore be based on the probabilities for each particular gene and each particular sample considered separately, as demonstrated in Equation (3.8), where $D_{g,s}$ represents the element of

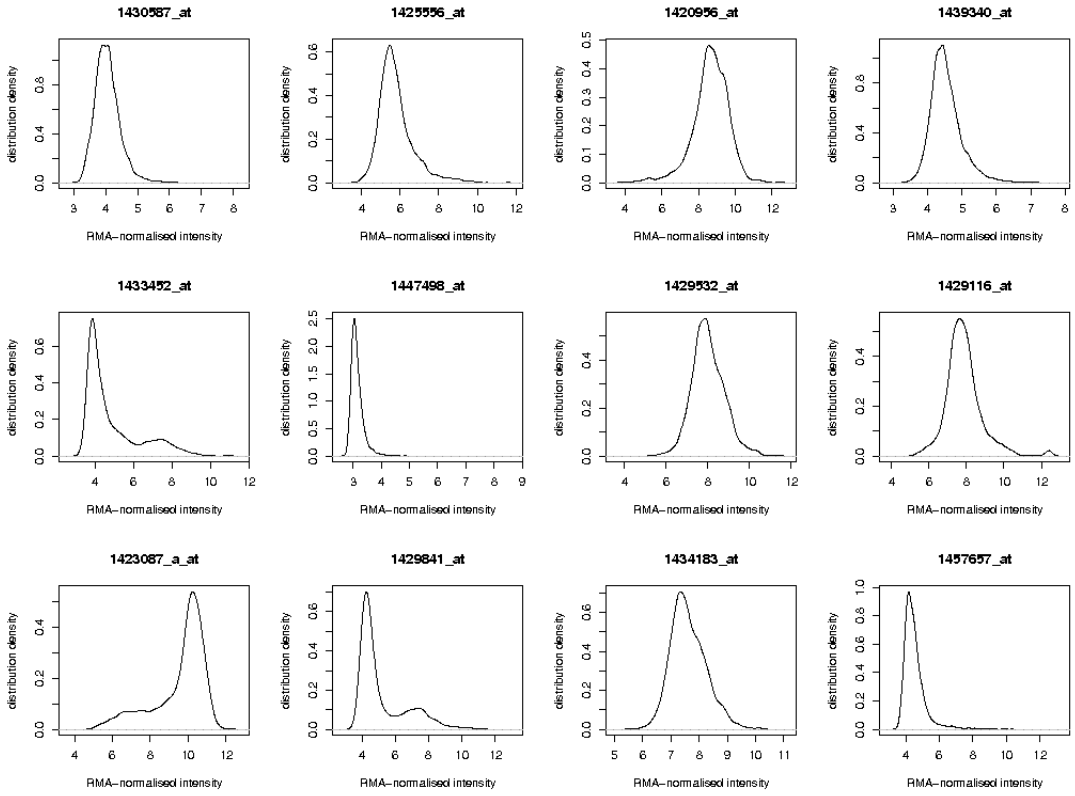


Figure 3.14: Distributions of expression levels of each of a panel of genes measured across a large number of microarray samples. The height of the graph in each panel represents the relative abundance of samples in which the gene had the normalized intensity value denoted by position along the horizontal axis.

the discretised matrix of expression levels calculated from the dataset indexed by gene g and sample s, and *genes* and *samples* represent the genes and samples, respectively, included in the bicluster pattern *bicluster*.

$$P(bicluster) = \prod_{g \in genes} \left(\prod_{s \in samples} P(D_{g,s} = hi) \right) \tag{3.8}$$

If we consider that the overall distribution of expression levels across all genes will be approximately consistent in each of the samples in the dataset, due to the quantile-normalization process applied to ensure each sample's measurements are comparable to any other's (this may not always be the case, but will typically be so when the gene expression data come from microarrays, especially when these data are obtained from different experimental datasets), then the probability of a particular gene being highly expressed in any randomly chosen sample can be assumed to be the same. As a result, the expression for bicluster probability can be simplified from that given in Equation (3.8) to Equation (3.9) in which D_g represents the row of the discretised matrix of expression levels corresponding to all measurements of gene g.

$$P(bicluster) = \prod_{g \in genes} P(D_g = hi)^{|samples|} \tag{3.9}$$

As we are interested in the information content of a given bicluster, this is in proportion to the negative logarithm of the probability of the observation occuring by chance. This simplifies the bicluster probability into a natural expression for bicluster desirability based on the information content, this expression is given in Equation (3.10).

$$-\log(P(bicluster)) = -|samples| \sum_{g \in genes} \log\left(P(D_g = hi)\right) \tag{3.10}$$

This definition of an entropy-based bicluster score is clearly flexible in terms of the definitions used for $P(D_g = hi)$. If the input matrix D is discretised, a simple estimation for the probability that a given gene will be found in a high expression state in any randomly chosen sample (i.e. $P(D_g = hi)$) can be obtained by dividing the number of samples in which the given gene is assigned a high expression state by the total number of samples in the dataset.

By incorporating the features discussed above into a probabilistic measure of bicluster desirability, the evaluation of bicluster score moves from the simplistic expression given in Equation (3.11) to the expression in Equation (3.12) that incorporates features of the datatset being processed to help identify biclusters that carry more information (from an information theory perspective) in that they are less likely to occur by chance. In Equation (3.12) S represents the set of all samples.

$$f(x) = |samples| * |genes| \tag{3.11}$$

$$f(x) = -|samples| * \sum_{g \in genes} \log \left(\frac{\sum_{s \in S} \delta_{g,s}}{|S|} \right) \tag{3.12}$$

$$\delta_{g,s} = \begin{cases} 1 & \text{if } D_{g,s} = hi \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

An additional feature of this more sophisticated entropy-based bicluster scoring is that it provides a natural means of scoring and ranking the genes within a bicluster, based on the negative logarithm of their probability of appearing in the bicluster. This is essentially a measure of the bicluster-specificity of a gene expression pattern (i.e. a gene that is expressed consistently highly across the samples of a bicluster but not elsewhere will have a higher rank in that bicluster than a gene expressed consistently across a large subset of the samples in the dataset, of which the bicluster samples represent only a relatively small component). As discussed earlier in Section (3.6.3), other biclustering algorithms seem not to offer the ability to rank genes within a bicluster, yet this might be particularly important in improving the utility of biclusters where the number of genes in a bicluster is greater than a practically useful number for the intended inference task.

The following section demonstrates the results in practical terms of incorporating this more sophisticated measure of bicluster probability by comparing the biclusters discovered by the MNC BGA using the naive model for bicluster desirability and the more sophisticated entropy-based model described in this section. A further demonstration of the improvement afforded by this model is provided through evaluation of the success of application of the discovered biclusters to meta-analysis tasks.

Bicluster Properties: Comparing Naive And Entropy-Based Biclusters

As a means of assessing the practical impact of adopting the improved estimator of bicluster information content in the entropy-based biclustering framework, the MNC BGA was used to discover biclusters in real datasets of gene expression data using fitness functions based on each of the bicluster desirability expressions given in Equations (3.11 & 3.12). The resulting sets of biclusters discovered by each algorithm were analysed in order to demonstrate the improvement arising from the adoption of the more sophisticated entropy-based method, in terms of a number of desirable bicluster characteristics.

In a collection of biclusters discovered in a particular dataset, certain properties stand out as being particularly desirable or undesirable. The inclusion in a bicluster

of genes that would be shared across many biclusters involving distinct subsets of the samples in the dataset is generally less informative than genes that are more specific to the bicluster in question. Additionally, a bicluster has more statistical 'support' if the observed patterns are consistent across a greater number of samples. For these reasons, distributions were obtained for the numbers of samples and the pairwise proportional overlap across a set of biclusters discovered by the algorithm incorporating each method of evaluating bicluster desirability.

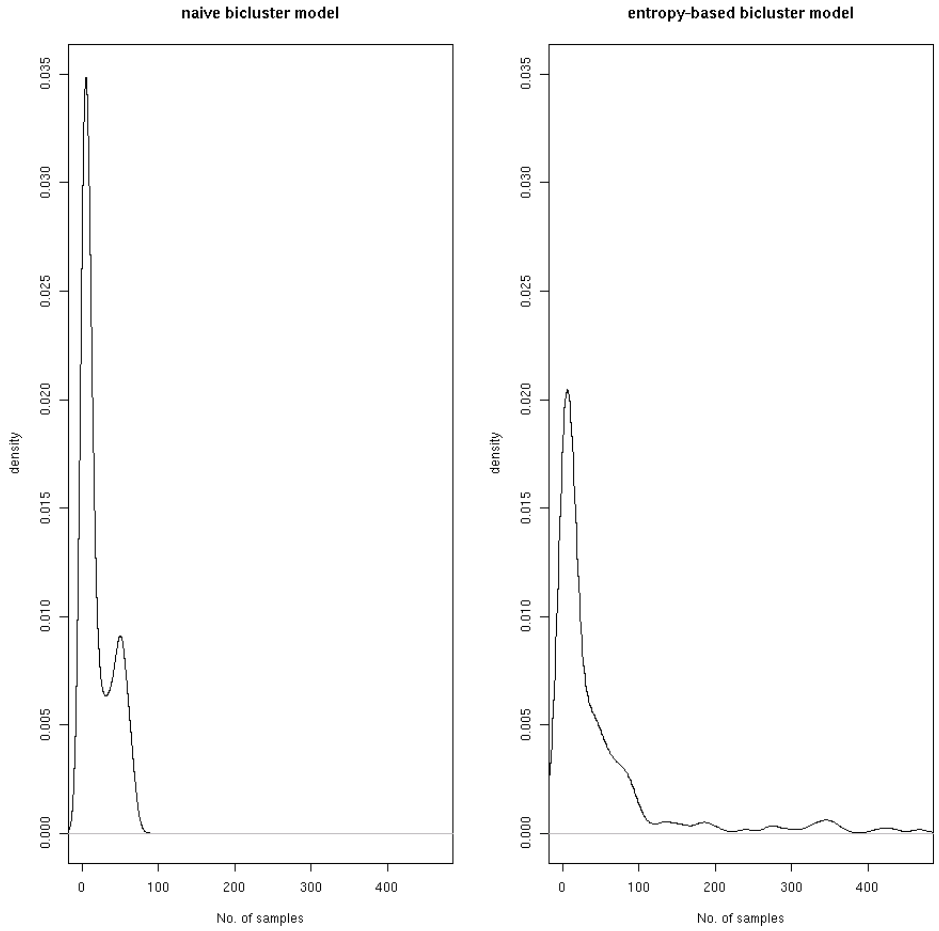


Figure 3.15: Distributions of numbers of samples in biclusters for sets of biclusters discovered by the MNC BGA using the naive bicluster desirability model (left panel) and the more sophisticated entropy-based model (right panel)

From the plots shown in Fig. 3.15 it is clear that the incorporation of the more sophisticated model for bicluster entropy into the novel biclustering algorithm results in a general increase in the number of samples in the discovered biclusters. Especially noticeable is the number of biclusters involving hundreds of samples: the more sophisticated entropy-based algorithm finds a number of such large biclusters (with more significant support from the data) whereas the naive algorithm finds none.

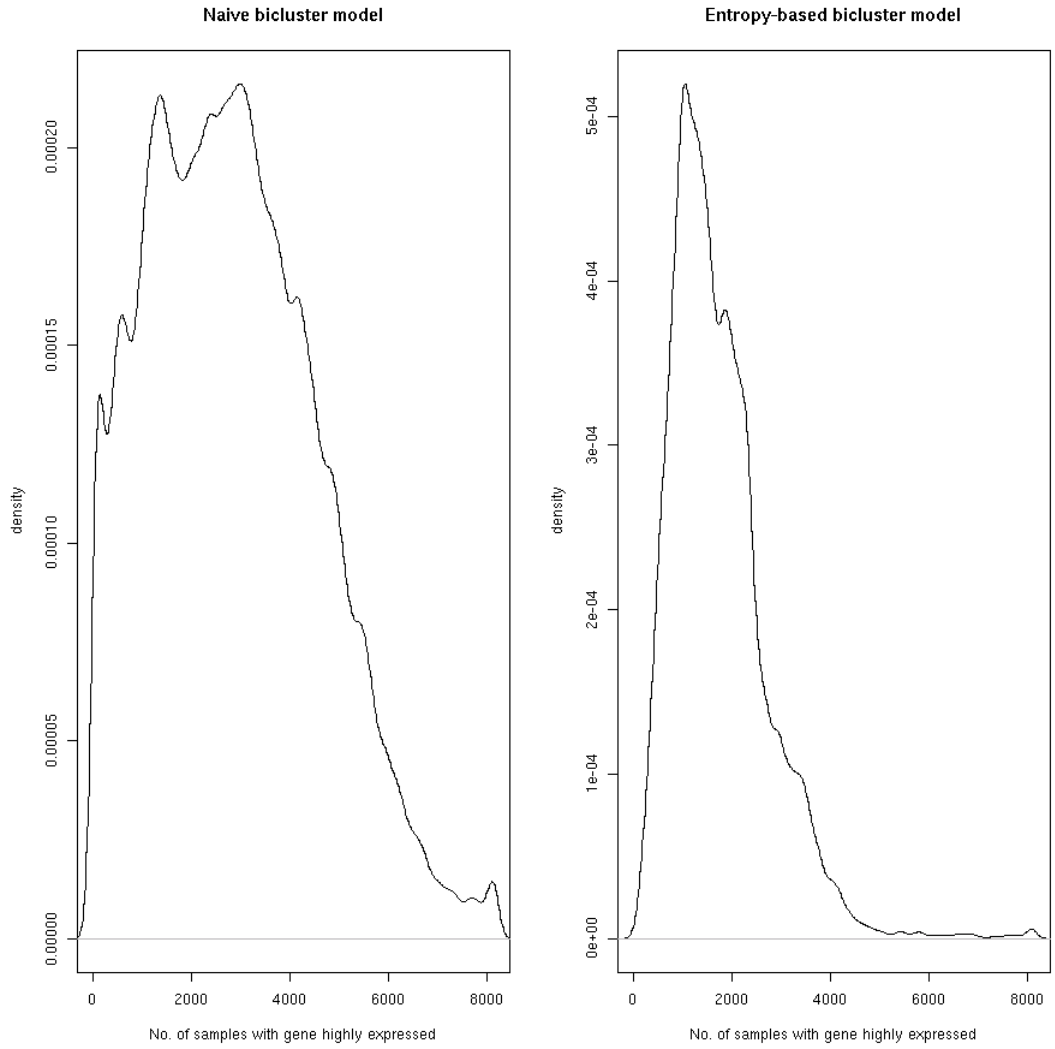


Figure 3.16: Distribution of the numbers of samples with high expression of bicluster genes, for all genes in a number of biclusters discovered by the MNC BGA using the naive bicluster desirability model (left panel) and for top ranking 200 genes in a number of biclusters discovered by the MNC BGA using the entropy-based bicluster desirability model (right panel).

The plots in Fig. 3.16 demonstrate that a significant proportion of the genes in biclusters discovered with the naive algorithm are expressed across the majority of all the samples in the dataset. However, due to the ranking provided by the entropy-based algorithm, bicluster genes can be prioritised to those that tend to be expressed across a smaller proportion of the samples in the dataset. This would imply that the biclusters discovered using the entropy-based method can represent more specific expression components (i.e. 'local' data patterns) within the dataset than the expression patterns represented by the naive MNC BGA biclusters.

Examining the impact of incorporating the more sophisticated entropy-based measure of bicluster desirability on the general properties of the biclusters discovered using the MNC BGA reveals that, in the terms laid out in the previous two paragraphs, the resulting set of biclusters will be more reliable and convey more useful information for inferring context-specific transcriptional relationships between genes. This demonstrates that the entropy-based model of bicluster desirability presented in this section represents an improvement over the naive approach that is commonly adopted amongst other biclustering algorithms (as described in Sections (3.4.3 & 3.4.5)). However, this comparison of bicluster properties is not really a measurement of success of the biclustering algorithm as a tool for meta-analysis of gene expression datasets. For a more conclusive demonstration of the improvement afforded to the biclustering process by adopting this more sophisticated scoring approach it would be necessary to show an improvement in performance of the discovered biclusters at some practical meta-analysis task for which application of the algorithms is intended.

Evaluation For Meta-Analysis

In order to assess the improvement in practical applicability of results afforded by the adoption of the more sophisticated entropy-based bicluster desirability score into the MNC GA biclustering algorithm, it was necessary to evaluate the utility of the biclusters in terms of their successful application to meta-analysis tasks. As a result of the availability of reference data, the tasks chosen were based on either function prediction through co-association or prediction of genomic binding of a transcription factor (TF).

One potentially useful application of meta-analysis of gene expression data is in the prediction of the regulatory activity of particular TFs. If we believe that a particular TF is involved in a certain biological process of interest, it may help in the understanding of the molecular mechanisms involved in that process if we can identify genes whose expression is regulated by the TF in question. This can be applied to the evaluation of biological signifiance of biclusters if we have lists of genes proximal to genomic regions bound by particular TFs. A statistical enrichment in a bicluster of the target genes of a given TF provides an indication that the bicluster represents some transcriptional

regulatory component involving the TF in question. To this end, for a number of TFs for which genome-scale binding data is available, biclusters were identified in the submatrix of the dataset described in Section (3.3) corresponding to those samples with high expression of the TF in question (thus ensuring that the biclusters all include the TF). For each of these sets of biclusters, overlap of each bicluster's genelist to the corresponding TF's list of bound genes were calculated. The average number of overlaps in a list of up to 200 genes taken from an individual bicluster are shown in Fig. 3.17 for each TF used, for both the naive MNC BGA and the BGA adopting entropy-based bicluster desirability evaluation. A genelist length of 200 was arbitrarily chosen as a means of demonstrating the impact of the ranking provided by the entropy-based measure of bicluster desirability, especially in terms of the highest-ranking observations. For the case of the entropy-based bicluster, the top ranking genes from each bicluster were always taken in order to demonstrate the impact of a ranking on the biological signature represented by a bicluster.

These plots show that adopting the more sophisticated entropy-based biclustering approach generally results in a greater chance of discovering biclusters that are strongly statistically enriched for the presence of genes bound by a particular TF. While it is true that the lack of such a signature doesn't imply that the corresponding gene list does not represent some biologically significant transcriptional signature, the consistent appearance of more biclusters having greater such enrichments does provide evidence to suggest that the entropy-based GA method of bicluster discovery is more likely to identify real transcriptional signatures within a large dataset and thus be a more useful tool for predicting transcriptional activity than the naive biclustering GA.

As mentioned in Section (3.5.3), the Gene Ontology (GO) can be used as a tool for assessing functional signatures statistically over-represented in a given gene list. While the approach of examining enrichment of any GO terms in gene lists has some limitations (as discussed in Section (3.5.3)) for evaluating biological relevance of biclusters, it can serve as a potentially useful means of indicating whether discovered biclusters seem to match some functional biological signature. In conjunction with a demonstration of improved overlap to genome-binding targets of TFs, an increase in enrichment of GO terms within the same bicluster gene lists provides a further indication of improved biological significance of a given set of biclusters. Therefore, the biclusters discovered using the MNC BGA incorporating each of the naive and entropy-based methods of bicluster desirability scoring were tested for all statistically enriched GO term annotations. The distribution of GO term enrichment values for each of the biclusters discovered by each of these methods is shown in Fig. 3.18. GO term enrichments were calculated using the conditional enrichment testing implemented in the `GOstats` [Falcon and Gentleman, 2007] R package available through Bioconductor.

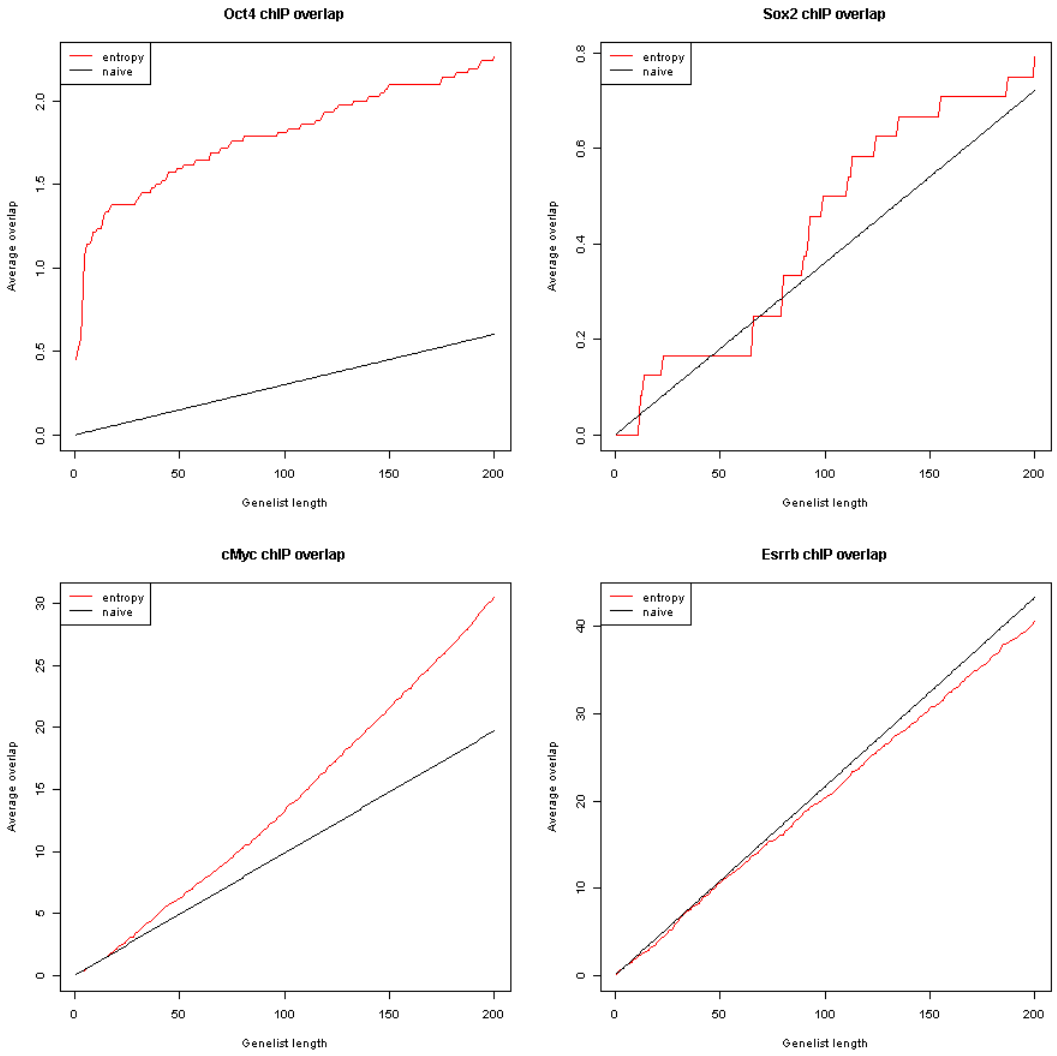


Figure 3.17: Average overlap of bicluster genelists of varying length with DNA-binding targets for each of a number of TFs. Ranking for genelists from biclusters discovered using the MNC BGA with entropy-based bicluster model is provided by the bicluster model, and corresponding overlaps are shown in red. Overlaps for biclusters discovered using MNC BGA with naive bicluster model are representative of expected overlap for a randomly sampled genelist of the given length from any of the relevant biclusters. These overlaps are shown in black.

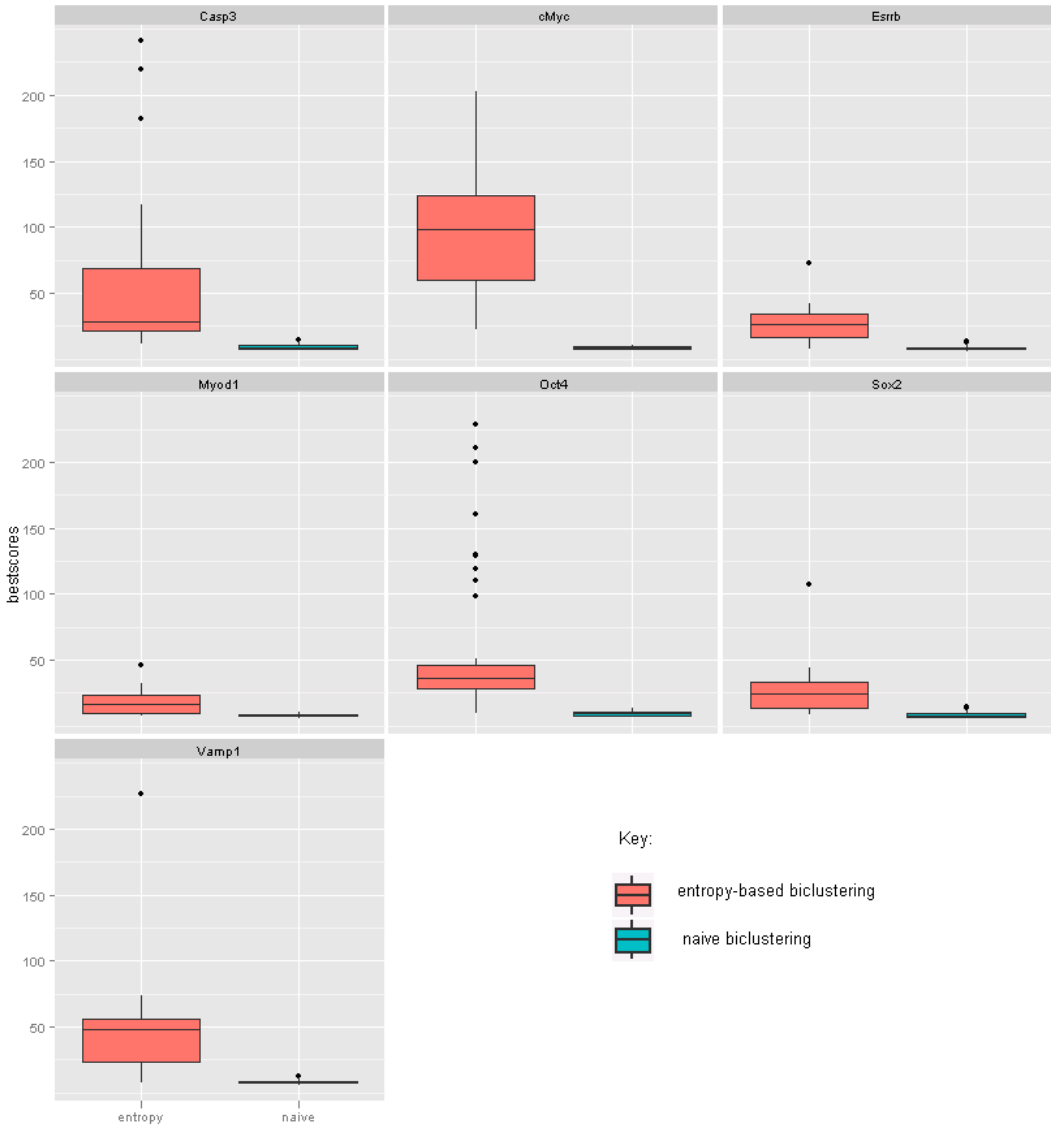


Figure 3.18: Enrichments of GO terms in bicluster genelists for MNC BGA using entropy-based (red) and naive (blue) models of bicluster desirability. Each panel shows -log GO term enrichment p-value distribution for biclusters found in a different subset of the data collection, corresponding to the samples with a high level of expression of the indicated gene.

The generally increased levels of enrichment of GO terms in biclusters discovered by the entropy-based GA compared to those from the naive GA method provides further evidence that the more sophisticated entropy-based approach results in improved discovery of biologically significant biclusters. In conjunction with the results of analysis of enrichment in biclusters for TF binding targets presented above, this evaluation of biclustering meta-analysis demonstrates that adopting the more sophisticated entropy-based model of bicluster desirability presented in this section results in an improved method for discovering biologically significant biclusters in large collections of gene expression data. Furthermore, enrichment analysis of the biclusters discovered using the entropy-based MNC BGA demonstrate practical utility of adopting this biclustering approach for the meta-analysis tasks of gene function prediction and prediction of genomic binding of a TF.

3.6.3 Revisiting Bicluster Definition 2: Localised Co-dependency Analysis

In many questions involved in biological research, we are interested in identifying information in relation to particular genes of interest. The methods mentioned so far in this chapter involve searching in a general sense for any transcriptional patterns, regardless of the particular biological context or genes involved. These methods could potentially be used to identify patterns of particular interest to a given biological question by searching through all the patterns discovered to find those that best suit the context in question. However, such an approach would necessarily imply that the patterns discovered are all of the same structure, as there is no consideration of the particular biological question in the specification of the desired patterns. There is therefore an implicit assumption that the general pattern (in this case, the concept of a bicluster under whichever definition is used) is a universally applicable piece of information that can help provide an answer to any potential question for which the meta-analysis of gene expression data might be intended.

This universal approach to bicluster analysis was found to be particularly problematic when applied to large, heterogeneous collections of gene expression data for the purposes of investigating mechanisms of transcriptional regulation. If a gene of interest is relatively specifically expressed in a few cell or tissue types, the most significant bicluster patterns in the data involving that gene tend to correspond to the cohort of genes with characteristic expression in the biological context represented by the bicluster samples. This observation is reflected in the dominance of sample-dependent bicluster patterns discovered by all widely-used biclustering methods, demonstrated in [Chia and Karuturi, 2010]. The sample-dependent bicluster pattern referred to here corresponds to a bicluster representing any set of samples across which a given set of genes show highly correlated expression. Simply identifying sets of genes with consis-

tent expression specific to a particular biological context does not provide any insight into transcriptional regulatory mechanisms over and above those that drive the predominant transcriptional profile that characterises the biological context in question. To investigate transcriptional regulatory mechanisms within any biological context, it is therefore essential to consider the overall similarity of transcriptional program of the bicluster samples so that gene expression patterns reflecting transcriptional regulation can be separated from the general transcriptional profile characterising the biological context of each bicluster sample. The remainder of Section (3.6.3) describes this concept, termed 'localised gene expression co-dependency analysis,' which is based on the dissociation of transcriptional regulatory effects from characteristic transcriptional profiles. An implementation of this analysis approach in the MNC BGA framework is described and the results of application to large scale gene expression meta-anlaysis are demonstrated.

Localised Co-dependency Analysis: Guidegene-Dependent Biclusters

The localised gene expression co-dependency analysis introduced in the previous paragraph involves the identification of gene expression patterns relating to changes in expression of a gene of interest in a particular biological context, by querying a large collection of gene expression datasets. For this task it would be especially useful to find a collection of subsets of samples with consistently high expression of the genes of interest and a collection of associated subsets of samples each with generally similar expression programs to a corresponding subset in the first collection but with significantly lower expression of the genes of interest. As this search for gene expression patterns occurring across subsets of the samples in a dataset it is essentially a biclustering problem, a strategy was devised to utilise the MNC BGA to implement gene expression co-dependency analysis. By implementing gene expression co-dependency analysis with biclustering it should be possible to extract from large collections of gene expression data information regarding mechanisms of transcriptional regulation of a particular gene or set of genes of interest, with biclusters separating co-dependency observations into biological contexts across which they are observed. The implication of taking this approach to biclustering would be for evaluation of bicluster desirability to incorporate not only the unlikelihood of a given bicluster pattern being discovered in the data by chance, but also the perceived relevance of the bicluster's gene expression patterns to the genes being investigated. This perceived relevance of a bicluster must distinguish a gene expression pattern related to the particular gene or set of genes of interest from general differences in the characteristic gene expression programs across different biological contexts represented in the dataset.

The biclustering approach used to perform gene expression co-dependency analysis is referred to here as 'guidegene-dependent biclustering[3].' With a guidegene-dependent bicluster, any genes that share a similar pattern of expression with the gene of interest within the biological context of the bicluster would be especially interesting as they would be more likely to be regulated by a shared transcriptional mechanism than any randomly-selected gene specifically expressed in a cell type that always expresses the gene of interest (as is typically the result of applying any of the biclustering methods mentioned so far in this chapter to large and diverse collections of gene expression data).

It is proposed here that guidegene-dependent biclustering could provide an improved method for inferring transcriptional relationships involving particular genes of interest by querying large collections of gene expression data. To investigate the impact of taking this approach to biclustering based meta-analysis, a modified MNC BGA was created implementing guidegene-dependent biclustering. This guidegene-dependent BGA (GDepBGA) was used to identify biclusters in the same datasets as were presented in the evaluation of the entropy-based BGA described in Section (3.6.2).

Guidegene-Dependent Biclustering Algorithm

The MNC biclustering BGA could be adapted to discover guidegene-dependent biclusters through incorporation of an appropriate fitness function, as the principles of exploring the bicluster search space with a GA (as described earlier in this chapter) remain. The guidegene-dependent bicluster, in contrast to the naive or entropy-based biclusters defined earlier in this chapter, would have to be evaluated not only on the basis of the samples in the bicluster and the consistency or specificity of each gene's expression across the bicluster but also on the expression patterns of those genes across samples defined as relevant by their overall similarity to the bicluster samples and their contrasting level of expression of the 'guide' gene. Therefore, the ensuing bicluster scoring function must incorporate the notion of relevance of non-bicluster samples in determining guidegene-dependent expression patterns and an assessment of *appropriate* variation of each bicluster gene across such samples.

Given that the principles of bicluster information content relating to entropy, as calculated in the scoring function given by Equations (3.11 & 3.12), apply to the concept of guidegene-dependent biclusters, this is taken as the basis for calculating desirability of guidegene-dependent biclusters. In addition to this, the contribution of each gene in the bicluster to the score is scaled by a term estimating the relative co-dependence of that gene's expression with the query gene in the biological context represented by the bicluster and the samples used for comparison.

[3]This name comes from the fact that the transcriptional patterns represented by the biclusters aim to capture the co-dependency of expression level of each bicluster gene with the given 'guide' gene

The co-dependence of expression of a bicluster gene with the guidegene could be estimated from the consistency of that gene's adherence to a pattern of high expression across the bicluster samples and low expression across the comparison samples (with low expression of the guidegene), scaled somehow by the relevance of those comparison samples to the bicluster.

The relevance of non-bicluster samples to be used as comparison-samples is determined by the overall similarity of the sample in question to those in the bicluster and in terms of the contrast of the expression level of the guidegene in the sample and the (presumably high) expression level across the bicluster. Assuming that the comparison samples used have a low level of expression of the guidegene, the relevance of the comparison can be expressed in terms of the inverse of the overall dissimilarity (according to some distance metric) between that sample and the bicluster samples. Taking a simple approach to the calculation of such similarity-based relevance, the Euclidean distance between the column of the data matrix representing the sample and each of those columns representing the bicluster samples can be calculated and normalized to a (0,1) range by linear transformation from the range defined by the minimum and maximum such pairwise sample distances observed from the whole set of possible pairs of samples. It is desired that very similar samples should have dramatically greater significance than those comparisons involving more distant samples. In order to achieve this effect, a scoring term was developed that weighted a gene's comparison observations used to determine the guidegene-dependence of expression by the negative logarithm of the mean normalized Euclidean distance of the sample used to obtain that observation and each of the bicluster samples. Of course, it is worth noting that a variety of such scaling functions might be appropriate here, attaching a different degree of significance of the similarity of comparison samples to bicluster samples. Therefore, at this stage, this method of evaluating guidegene-dependence is essentially arbitrary, although chosen on the basis of preliminary testing of alternatives and exploration of the results.

The resulting guidegene-dependent bicluster scoring function is defined by Equations (3.14-3.17). In Equation (3.14) S_c represents the set of (non-bicluster) samples used for evaluation of contrast of expression, coincident with that of the guidegene, to the bicluster samples. S_b represents the bicluster samples and S in Equation (3.15) represents the set of all samples in the dataset.

$$f(x) = -|S_b| \sum_{g \in genes} \log(P(g)) \sum_{c \in S_c} (\delta_{g,c} \log(H(c, S_b))) \tag{3.14}$$

$$P(g) = \frac{\sum_{s \in S} \delta_{g,s}}{|S|} \tag{3.15}$$

$$H(x, samples) = \frac{1}{|samples|} \sum_{s \in samples} \sqrt{\sum_{g \in G} (D_{g,x} - D_{g,s})^2} \tag{3.16}$$

$$\delta_{g,s} = \begin{cases} 1 & \text{if } D_{g,s} = hi \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

A similar analysis as that presented in Section (3.6.2) was performed and is described below for the guidegene-dependent BGA to evaluate the specificity of biclustering results to the query guidegene and the biological significance of those discovered biclusters.

Bicluster Properties: Comparing Guidegene-Dependent Biclusters And Entropy-Based Biclusters

As discussed in Section (3.6.2), certain properties are especially desirable in collections of biclusters discovered by a given algorithm. Of particular interest in a guidegene-dependent biclustering framework is the specificity of results to each query gene. Unlike the other biclustering algorithms mentioned in this chapter, the guidegene-dependent biclustering algorithm can return different genelists for the same set of bicluster samples, depending on the guide gene used to evaluate the expression patterns. To illustrate this effect on overall biclustering results, a rank-based overlap analysis was performed on genelists from biclusters arising from analysis of the dataset used in Section (3.6.2) with the guidegene-dependent BGA described above, using the guidegenes Oct4, Sox2, cMyc, Myod1, Esrrb, Casp3 and Vamp1. Average rank scores were calculated for all genes across the set of biclusters discovered by the GDepBGA, for each of guidegenes in turn. Standard deviation of the averaged rank scores for each of the bicluster sets was calculated for every gene. The distribution of these standard deviations of average rank scores for genes being associated through biclusters to each of the guidegenes is shown in Fig. 3.19. For comparison, the equivalent rank-based overlap scores were calculated between biclusters discovered by application of the entropy-based BGAs presented earlier in this chapter to different 'guidegene-expressing subsets' of the whole dataset are shown alongside.

The plot in Fig. 3.19 show that there is less correspondence between genelists from biclusters discovered using the GDepBGA with different guidegenes than those from the entropy-based MNC BGA across the same set of guidegenes. This indicates that the transcriptional information represented by biclusters discovered by the guidegene-dependent BGA are more specific to the gene of interest provided as a query than even the sophisticated entropy-based MNC BGA. Measurement of the specificity of bicluster genes to a particular guidegene revealed a similar pattern when using an alternative score to indicate high rank in biclusters specifically from one guidegene (data not shown). This specificity suggests that the gene expression patterns identified using the GDepBGA are more likely to be relevant to a particular biological question

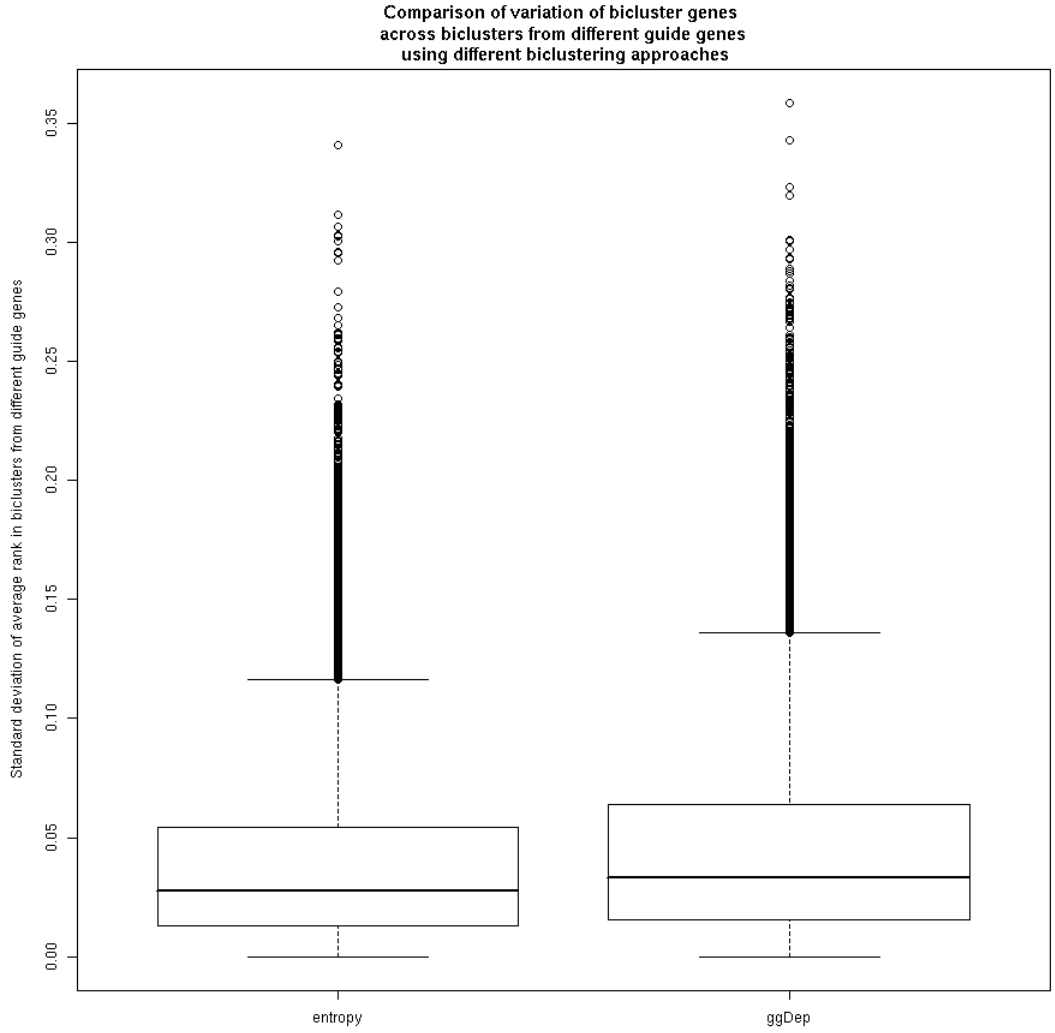


Figure 3.19: Distributions of guidegene specificity scores for genes in biclusters discovered by each of the entropy-based MNC BGA (left-hand box) and the GDepBGA (right-hand box). The specificity score for a gene represents the standard deviation across the set of averaged rank scores for that gene appearing in the set of biclusters corresponding to each guidegene. A higher value indicates more variability in average bicluster genelist ranks for the given gene across the different guidegenes, and so indicates a more specific association to an individual guidegene.

than those represented by biclusters discovered through the traditional biclustering framework, even those containing the gene of interest.

Provided that these query-specific biclusters represent transcriptional associations with biological significance to at least as clear an extent as those biclusters discovered using the best alternative biclustering strategies (in this case represented by the entropy-based MNC BGA, as demonstrated in Section (3.6.2)), this specificity and relevance to a particular query makes the GDepBGA a clearly superior tool for prediction of the transcriptional activity of particular genes of interest (when considered against other biclustering-based analysis approaches). An evaluation of this essential biological significance of guidegene-dependent biclusters is given below.

Evaluation For Meta-Analysis

In order to evaluate the biological significance and utility (for answering biological questions) of biclusters discovered using the GDepBGA, the analysis approaches used in Section (3.6.2) to evaluate the success of the entropy-based MNC BGA were applied to output from applying the guidegene-dependent BGA to the dataset used above, with a selection of guidegenes used as queries to identify different sets of patterns within the data. Statistical enrichment of GO terms within these biclusters are shown in Fig 3.20 alongside corresponding enrichments achieved through application of the entropy-based and naive MNC BGA. For a subset of the guidegenes (Oct4, Sox2, cMyc and Esrrb) lists of predicted targets were available from high-throughput chIP data. Overlaps to each gene's binding-targets from the genelists from guidegene-dependent biclusters using the appropriate TF as a guidegene are shown in Fig 3.21 alongside corresponding enrichments of each TF's binding targets within biclusters discovered using the entropy-based BGA on each TF-expressing subset of the whole dataset.

These collections of enrichments demonstrate that the GDepBGA discovers biclusters representing gene expression patterns within the data that, while specific to a particular gene of interest, are highly statistically enriched for the present of consistent biological process signatures and for relevant DNA-binding targets of particular TFs. This combination of features suggests a clear advantage of the guidegene-dependent biclustering approach described above over other biclustering-based approaches, in terms of meta-analysis of large collections of gene expression data for prediction of transcriptional regulatory mechanisms involving particular genes of interest (or any process for which such genes might be biological markers). It is proposed that the specificity demonstrated by the GDepBGA, whilst still capturing significant biological signatures within the data, indicates the utility of the concept of localised gene expression co-dependency analysis introduced in this section.

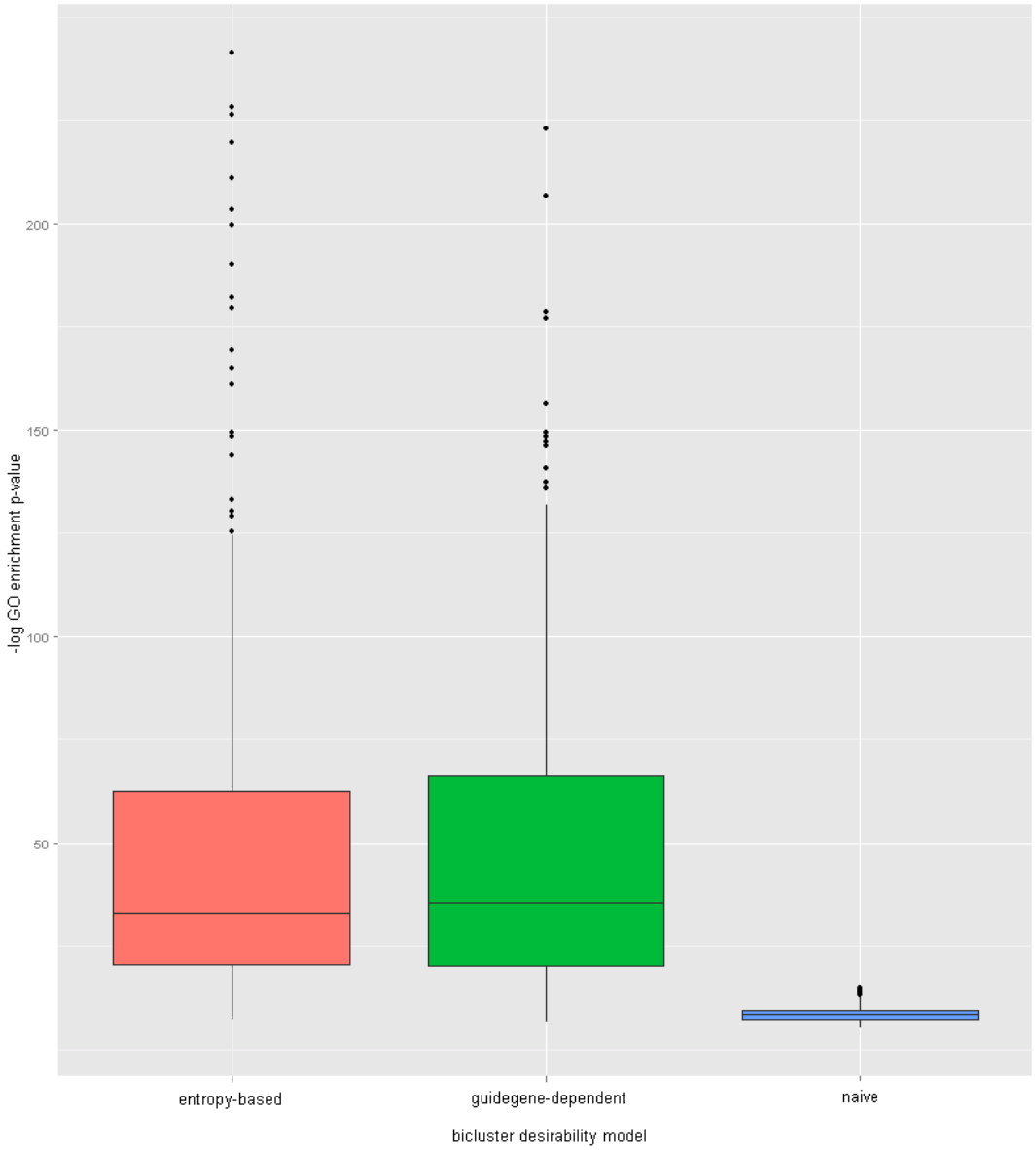


Figure 3.20: Enrichments of GO terms in bicluster genelists for MNC BGA using entropy-based (red) and naive (blue) models of bicluster desirability compared to GDep-BGA (green). Vertical axis represents -log GO term enrichment p-value distribution for biclusters found by the corresponding method. Around 200 biclusters were analysed for each method.

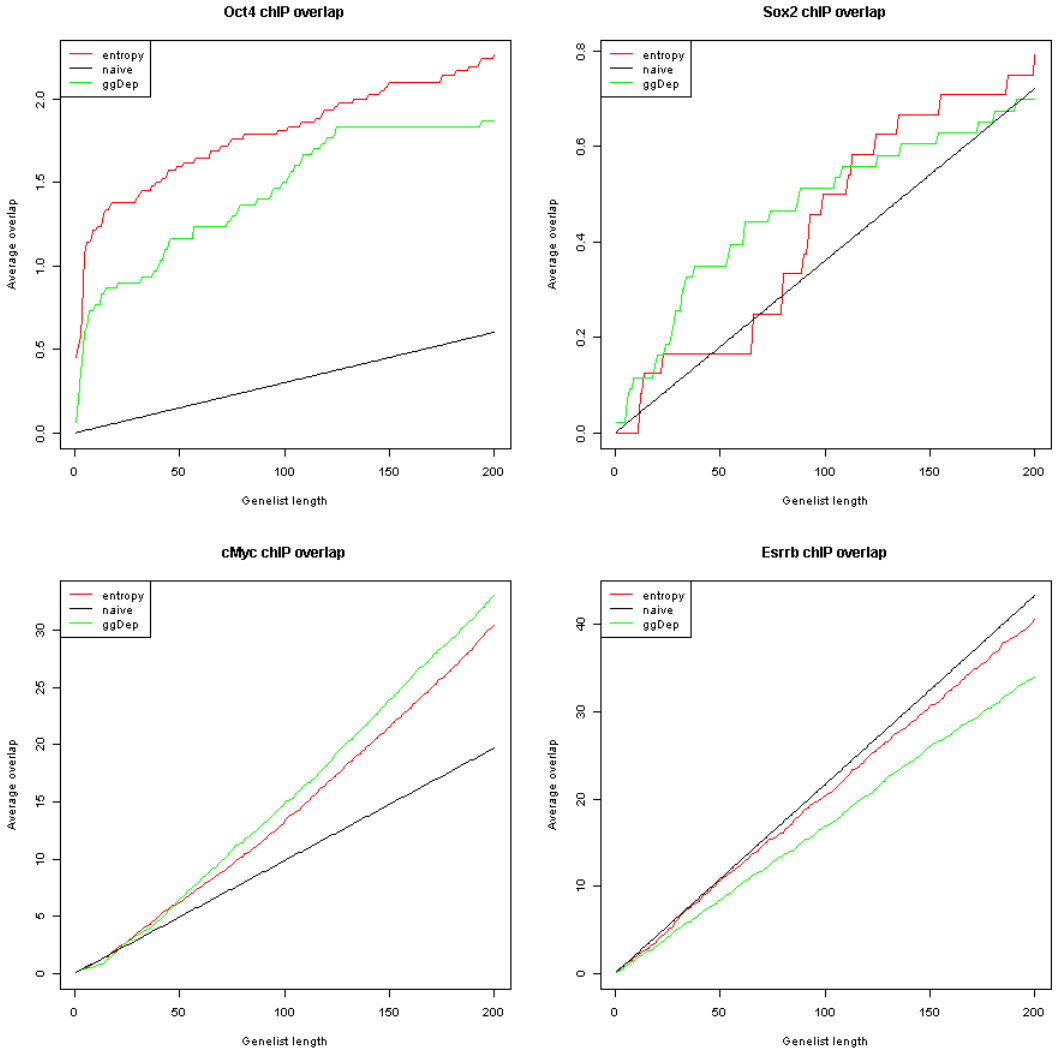


Figure 3.21: Average overlap of bicluster genelists of varying length with DNA-binding targets for each of a number of TFs. Ranking for genelists from biclusters discovered using the MNC BGA with entropy-based and guidegene-dependent bicluster models is according to the bicluster model, and corresponding overlaps are shown in red and green, respectively. Overlaps for biclusters discovered using MNC BGA with naive bicluster model are representative of expected overlap for a randomly sampled genelist of the given length from any of the relevant biclusters. These expected overlaps are shown in black.

3.7 Discussion

This chapter describes the motivation behind application of a biclustering approach to the meta-analysis of large collections of gene expression data, principally being that as the diversity within the collection of data increases the overall correlation in expression between genes with all but the most universal transcriptional relationships tends to disappear. In addition, the necessity for development of novel biclustering approaches for application to meta-analysis of very large collections of gene expression data was demonstrated through analysis of the growth of execution times of existing biclustering algorithms with the size of meta-analysis datasets, with the results presented in Section (3.5.1) illustrating the infeasibility of existing biclustering algorithms' application to gene expression datasets involving thousands of samples. As the chances of a particular transcriptional regulatory signature being differentially present in the data and reliably recovered through some pattern-mining method increase with the size of the dataset, it would clearly be advantageous to be able to analyse data from as many samples as possible. As there are thousands of gene expression datasets publicly available the observed lack of extendability of the feasible scale of biclustering possible with existing algorithms is clearly a severe limitation in terms of the potential for useful application of the biclustering paradigm to meta-analysis of gene expression data.

A reformulation of the standard biclustering problem was developed so that the complexity of the task when applied to mining gene expression datasets for transcriptional patterns was reduced. Based on this reformulation, which is presented in Section (3.4.1), novel biclustering algorithms were developed. The exhaustive biclustering algorithm presented in Section (3.4.4) was shown to identify consistent gene expression patterns through biclustering, although was infeasible for application to large-scale meta-analysis. To this end, a genetic algorithm was developed to explore efficiently the combinatorial search space of possible biclusters arising from the biclustering problem as defined in Section (3.4.1). This novel biclustering algorithm (IslandCluster) was shown through application to artificial datasets and large collections of gene expression data to successfully discover biclusters with a comparative level of biological significance to those discovered using existing state-of-the-art algorithms but to be able to achieve this on the desired scale for meta-analysis, which (at the time of development of IslandCluster) was impossible with any existing method.

The development of IslandCluster enabled the application of biclustering to meta-analysis on a previously unreported scale. This in turn enabled, for the first time, the study of application of biclustering to such large-scale meta-analysis of gene expression data. Through evaluation of the properties of biclusters discovered by IslandCluster, modifications were proposed for improving the practical impact of this biclustering approach to meta-analysis of gene expression data in terms of answering specific questions

in biological research. A theoretical advance regarding the entropy-based estimation of bicluster desirability was developed. This framework is presented in Section (3.6.2) and the practical benefits of incorporating such an approach to estimation of bicluster desirability into a biclustering algorithm are illustrated through various examples in the application of these algorithms to meta-analysis of gene expression data (also in Section (3.6.2)).

It was noted that the traditional biclustering approach results in a prevalence of biclusters representing expression profiles generally characteristic of the biological context reflected in the bicluster. This results in a lack of ability to distinguish specific transcriptional regulatory relationships from general context-dependent expression programs. To avoid this problem, and to identify transcriptional relationships between genes even within an individual biological context, the concept of localised gene expression co-dependency analysis was developed. Crucial to this concept is the consideration of the overall similarity of transcriptional programs between samples with observed expression patterns involving a gene or set of genes of interest. Guidegene-dependent biclustering was developed as a means of implementing localised gene expression co-depedendency analysis, adapting the biclustering paradigm to the prediction of transcriptional relationships involving particular genes of interest. The development of this concept into an alternative bicluster definition and corresponding estimator of bicluster desirability is presented in Section (3.6.3). This novel biclustering concept led to the development of a guidegene-dependent biclustering algorithm that was applied to the meta-analysis of gene expression data. The success of this application of the GDepBGA, as demonstrated in Section (3.6.3), indicates that a tool has been developed as a result of the work presented in this chapter that provides a means to estimating transcriptional regulatory information regarding particular genes of interest through the application of biclustering to meta-analysis of gene expression data on a previously unreported scale. The scale of the datasets analysed in this work result in this constituting a novel application of biclustering. The practical considerations identified in this work as necessary for biclustering algorithms to be successfully applied to this task represent unique insight offered into large scale gene expression data mining. The concept of localised gene expression co-dependency analysis as a means of inferring transcriptional relationships from large and diverse collections of gene expression data has not previously been considered, thus its introduction in this thesis creates considerable opportunity for further work on gene expression data mining for inference of transcriptional regulatory mechanisms.

As discussed in Section (3.4.2), the novel algorithms presented in this chapter (along with many of the existing algorithms mentioned) require some preprocessing of raw gene expression measurements into a scale indicating the relative expression level of each gene in each sample in the dataset. The evaluations of application of these algo-

rithms presented throughout this chapter demonstrate that the approaches taken here were successful to some degree in facilitating discovery of biologically significant (and relevant) biclusters. However, the precise nature of these preprocessing methods and the manner in which they transform the underlying data is interesting. The study of such methods for the biological interpretation of gene expression data may yield possible improvements in the applicability of biclustering meta-analysis of gene expression data to real problems in biological research. This is the topic of the next chapter.

Chapter 4

Development of a Framework for Biological Interpretation of Gene Expression Data

The biological implications of the expression of a given gene at a given level may be difficult to assess from a single measurement without prior knowledge relating to the gene in question. For example, can any regulatory activity be inferred from establishing that a constitutively-expressed 'housekeeping' gene is expressed with a particular concentration of mRNA? As a result, standard analyses of gene expression datasets tend to be based on the identification of genes that are differentially-expressed between some samples of interest and a control set (as described in [Dudoit et al., 2002, Slonim and Yanai, 2009]). For any pattern mining approaches not based on a straightforward comparison between defined groups of samples to be applied to meta-analysis of large collections of gene expression data, a differential expression framework is insufficient. To facilitate application of a wider range of pattern mining techniques to gene expression meta-analysis, this chapter concerns the study, development, evaluation and application of methods for transformation of gene expression measurements in large collections of data into a standardised scale, offering straightforward biological interpretation of any measurement taken in isolation of all others.

First, such methods for transformation of gene expression data are discussed in terms of providing the ability to discover absolute gene expression patterns on a large scale. Existing approaches involving discretisation of data are discussed and evaluated alongside the novel discretisation approach described in Section (3.4.4). Owing to the limitations inherent to all discretisation approaches, continuous data transformations are introduced and attempts to use 'spike-in' datasets to calibrate data are presented. A novel framework for modelling the biological state of expression corresponding to measured intensity levels is presented and the results of application of this method to real data are also presented. This novel data transformation is evaluated in terms of its measurement of non-biological variation of expression and its impact on improving the performance of a biclustering algorithm (both in recovering implanted patterns in artificial datasets and increasing the biological significance of biclusters discovered in real data). An application of this data transformation outside the context of large-scale data mining is discussed, regarding the interpretation of individual (small-scale) gene expression datasets. Finally, the implications of the development of these novel data transformations are discussed.

4.1 Facilitating Large-Scale Data Mining

In the meta-analysis of gene expression data, most methods attempt to identify genes that are consistently differentially-expressed across certain groups of samples (e.g. [Owen et al., 2003, Hong et al., 2006]) or to identify associations between genes through the analysis of general trends in expression patterns across the available data (e.g. [Campbell et al., 2007, Day et al., 2009]). If, however, the goal is to identify absolute patterns of the expression of groups of genes (as is the case in the biclustering analy-

sis discussed in the previous chapter), then it becomes especially useful to be able to make some inference regarding the biological significance of a particular gene expression measurement. The motivations for developing methods to enable such inference are described in greater detail below, and various approaches to transforming gene expression data with this objective are discussed, including the presentation of novel approaches.

4.1.1 Motivation

For the discovery of patterns in large datasets, some model of the structure of the desired patterns must be defined. A survey of various models of bicluster structure used by a range of algorithms is included in Madeira & Oliveira's 2004 review [Madeira and Oliveira, 2004]. Where the algorithms searching for biclusters with the given structures attempt to identify those patterns in data matrices of measured gene expression levels, but consider only localised structure within the data (as is the nature of the biclustering paradigm), a particular measurement of a given gene's expression is taken out of the context of the overall distribution of expression levels for that gene. As mentioned in Section (3.4.2), this may be problematic for the following reasons:

Firstly, the most comprehensive resources of genome-scale gene expression data contain data from gene expression microarrays. If cross-platform data analysis issues are to be avoided, it is particular Affymetrix microarray platforms that have by far the greatest coverage of different samples profiled in the public domain: the *HGU133plus2* and *MOE430v2* array platforms have data from approximately $46,000$ and $17,000$ samples (respectively) in the GEO repository at the time of writing (in 2010). It is well known that microarray platforms tend not to yield accurate absolute measurements of transcript abundance, owing to limitations of the array hybridization and imaging technologies [Draghici et al., 2006, Zilliox and Irizarry, 2007]. As a result, any meta-analyses of collections of this microarray-derived gene expression data require some way of transforming the measurements from different samples into a scale that enables appropriate comparison of a particular gene in one sample with that of any other sample. In addition, and possibly more importantly, the measurements obtained by microarray profiling are arbitrary intensity values, not necessarily corresponding directly to mRNA concentrations. In order to relate these (arbitrary) intensity values to biological significance of gene expression, analysis of microarray data tends to involve the identification of genes with high *differential* expression between particular groups of samples (usually a sample-type of interest and a control group). For meta-analysis of this data considering every possible combination of samples to retain the concept of differential expression, every potential pattern would have to be described in terms of the samples involved and the direction of all comparisons within those samples (i.e. which samples to compare with which others). This additional order of complexity in the search space of possible solutions to such a pattern-mining problem might explain why the majority of biclustering algorithms involve a preprocessing step to transform the data into a

scale that leads to a particular interpretation of any given value without requiring the specification of reference samples (as noted in [Madeira and Oliveira, 2004]).

In addition to the issues arising from considering large-scale meta-analysis of arbitrary measurement values such as those arising from gene expression microarrays, there remain further possible advantages to adopting a transformation of the data. For example, even if a large compendium of mRNA concentrations of every transcript across a comprehensive range of biological samples were to exist, there still remains the issue that an interpretation of the biological consequences of a particular concentration of mRNA would depend on knowledge relating to that gene, as the biological significance of a given concentration of mRNA differs from gene to gene. However, a comprehensive collection of gene expression data offers the potential to interpret any given gene expression measurement in the context of the distribution of that gene's expression measurements across all the samples in the dataset. This chapter concerns methods that attempt to utilise the whole range of biological contexts represented in a large gene expression data compendium and in doing so provide a means of transforming all the measurements in the collected dataset into a unified scale where a particular value implies a particular biological 'state' of gene expression, independent of the particular gene or sample that value concerns.

4.1.2 Discretisation Approaches

Discretisation of microarray data is a widely used preprocessing step for simplification of the computation challenge involved in performing biclustering (discussed in [Madeira and Oliveira, 2004]). In addition to simplification of the computational aspect of biclustering, discretisation of gene expression data is a means of achieving the objectives described above, in that a particular discrete value has a particular biological interpretation (i.e. high/low expression of the gene) regardless of which gene the value is measuring in which sample.

The availability of a discretised form of the underlying gene expression data enables a bicluster model to be proposed in such a way that it can be immediately and straightforwardly evaluated across any submatrix of the data. This discretisation is referred to in [Tanay et al., 2002] as the identification of a significant change in expression of a gene 'with respect to its normal level,' allowing the identification through the biclustering process of genes that are in a consistent state of extraordinary expression across the bicluster samples. A similar approach is taken in the Bimax algorithm, described as the identification of groups of genes in a bicluster that are displaying consistently significant 'change with respect to a control experiment' [Prelic et al., 2006], although the control experiment referred to here is a non-specific reference derived from the overall distribution of each gene's expression values. As these algorithms (particularly SAMBA) have often been demonstrated to be among the best-performing biclustering algorithms in

terms of various evaluation criteria ([Tanay et al., 2002, Prelic et al., 2006]) it would suggest that a meta-analysis approach based on discretisation of a large gene expression dataset has potential to be successful.

While the discretisation of gene expression datasets has useful applications (such as facilitating the bicluster pattern-mining approaches mentioned above), there is little discussion in the literature of the relative merits of different approaches to this task. Such a discussion of the relative merits of a number of reported approaches is given below, including the motivation for development of a novel approach and assessment of the results of applying these various methods.

Approaches to discretisation of gene expression data that are reported in the literature tend to belong to one of two categories. Those that fit a standard statistical distribution to each gene's expression values and apply thresholds corresponding to interpretation of the fitted distribution, and those that set a threshold at a fixed (relative or absolute) magnitude above each gene's minimum or average level.

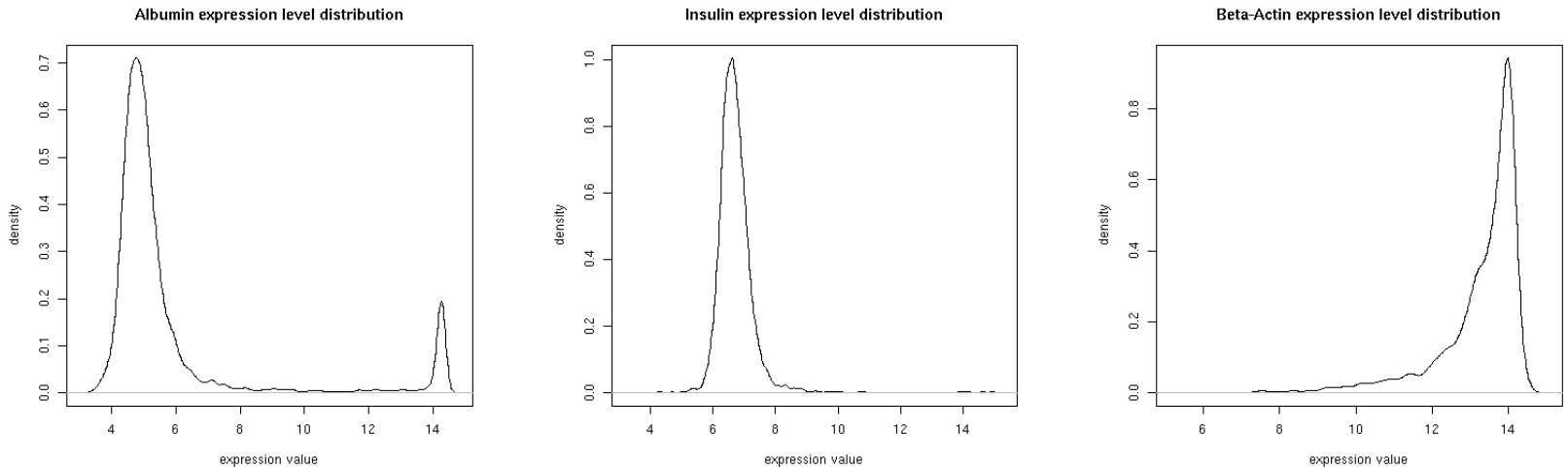
Any discretisation based on ranking can be seen as fitting an empirical distribution across that gene's expression values so that there is an equal probability associated with each of the measure values. Then applying a statistical threshold calling values with $P(val = `high') = p$ 'significantly upregulated' results in selecting the highest $(1-p)*100\%$ of the values. This is the default approach taken by the Bimax algorithm as implemented in *BicAT* [Barkow et al., 2006], using the top 10% of values. The approach taken by the SAMBA and xMotif algorithms involves assuming a normal distribution of each gene's expression values and classifying those outside a standard deviation of the mean as 'significantly changed' in the respective direction.

When examining the distributions of expression values across a large collection of samples for a number of genes, an immediate observation is that there is a high degree of variation in the shape of the distributions (as shown in Fig. 3.14). This variation in distribution clearly suggests that any particular standard distribution might not be an appropriate assumption to apply universally to all genes' expression values. In the case of the normal distribution assumption (as taken for the discretisation approaches in xMotif and SAMBA, and for the continuous z-score transformation utilised in ISA), the application of a Shapiro-Wilk normality test[1] to each gene's expression values across a large collection of data suggests that only 0.51% of the genes show a normal distribution across the dataset (to a confidence level of $P(values \sim N(\mu, \sigma)) \geq 0.5$), even when the number of samples in the dataset is very large.

[1]Shapiro-Wilk values averaged over 10 appropriately-sized random samplings of each gene's values, due to the fact that the Shapiro-Wilk test tends to underestimate the likelihood of data being normally distributed for large sample sizes

Similar issues apply to rank-based approaches, where different genes may have different proportions of their measured values corresponding to particular states of expression. To illustrate this concept, Fig 4.1 shows the distributions of values across a large dataset for three different genes. The first (albumin) shows a large proportion of low values but a still significant proportion with very high values. By contrast, the second (insulin) has only a few values with high expression, but these are still clearly distinct from the other, low values. Finally, the third example shows values for a gene ($\beta-$actin) that is constitutively expressed across the entire dataset. Such examples demonstrate that there is no standard approach to discretisation that will yield consistently accurate representation of biological states of expression when applied in the same way to expression values for different genes. As this is likely to be due to underlying differences in the biological mechanisms of activity of different genes, this point applies to the discretisation of gene expression data regardless of the technology used to generate the data or the normalization procedures applied prior to discretisation.

An alternative approach to this discretisation task that avoids the main issue described above is that of applying a fixed differential expression criterion to each value in turn for a given gene, using the remaining values as a collection of reference samples. This could potentially be achieved through a statistical approach to measuring differential expression or by applying a fold-change cut-off (e.g. classifying all values greater than $2*$ the median value as representing significantly high expression). The fact that the statistical approaches might run into problems estimating significance of a single measurement's difference to a large collection of reference values (as this is not what the widely-used methods for detecting differential expression were designed to do) might explain why this approach appears not to have been taken by any methods published in the literature. While the application of a fold-change threshold for classification is often successful as a means of detecting significant differential expression [Pearson, 2008] and avoids the issues described above for invalid assumptions on the distributions of each gene's expression values, it fails to take into account the fact that (as demonstrated in Fig. 3.14) the widely varying distributions of gene expression values result in a situation where a given fold-change threshold that appropriately separates distinct groups of expression for one gene may lie in the middle of a group of similar values for another gene. This is illustrated by the examples shown in Fig. 4.2, where the distribution of expression values for each of a set of genes are shown in terms of fold-change to the median. A consistent fold-change threshold would be indicated by a horizontal line drawn at the same height across all the plots. While subjective, suggested 'ideal' cut-off points have been drawn by hand in Fig. 4.2 with horizontal red lines, which illustrate that a threshold that is best for one gene may be inappropriate for another. This is further demonstrated in Fig 4.3, where panel (a) shows the scatter of measured expression values for each of the genes shown in Fig. 4.2, separated into 'high' and 'low' states of expression based on a median fold-change



(a) Density estimation for distribution of expression values of Albumin

(b) Density estimation for distribution of expression values of Insulin (Ins2)

(c) Density estimation for distribution of expression values of $\beta-$actin (Actb)

Figure 4.1: Distributions of expression values for three genes (from left: Alb, Ins2 and Actb) across a large collection of microarray data (described in Section (3.5))

threshold appropriate for Sox2. Panel (b) shows the same expression values separated into expression states based on a median fold-change threshold appropriate for alb. It is clear that the different shapes of expression value distribution for each gene mean that it is highly unlikely any universally-appropriate threshold will exist.

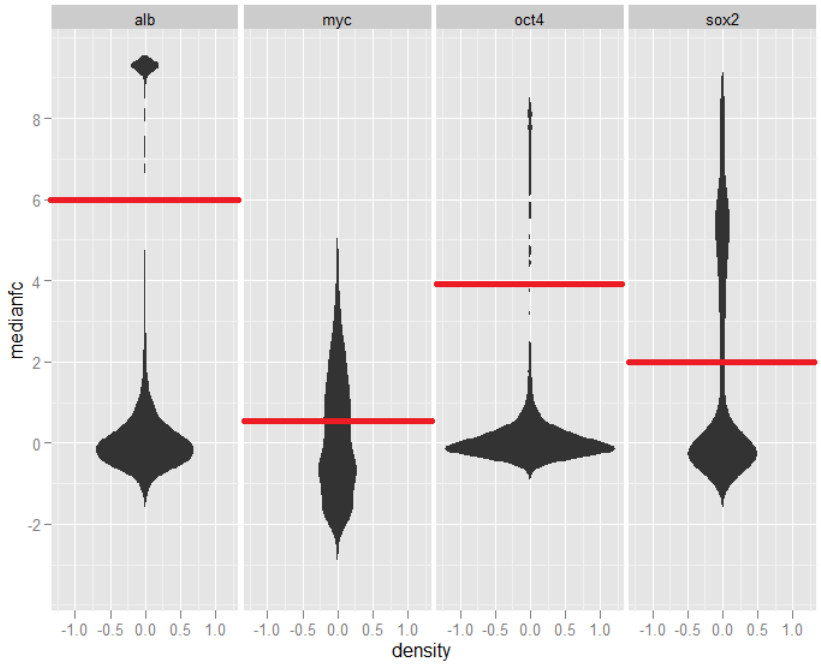
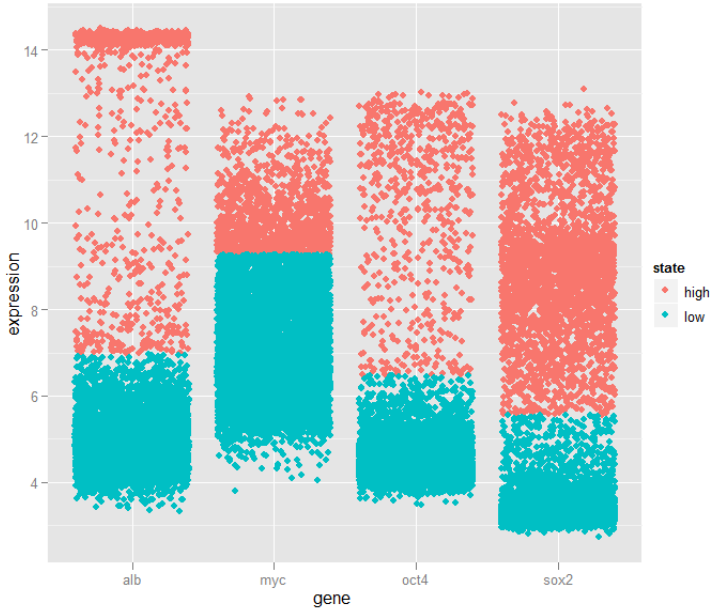


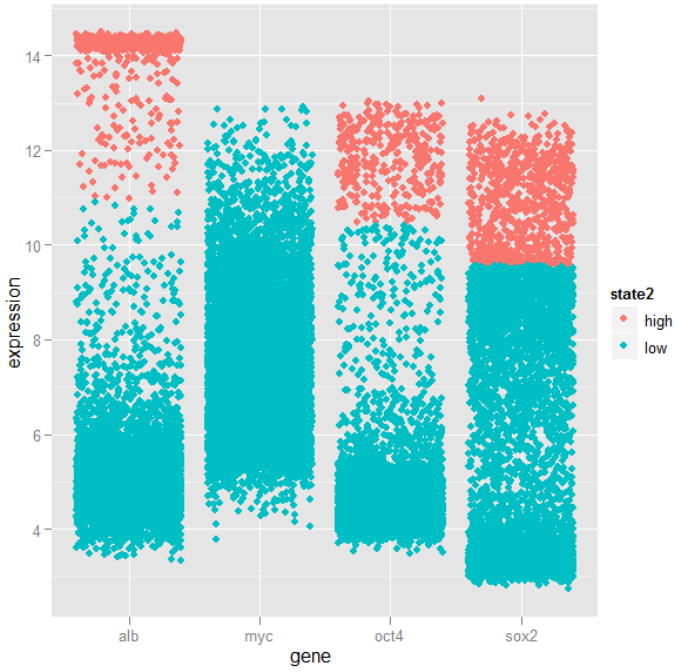
Figure 4.2: Distribution of expression values for a set of genes shown in terms of fold-change to median value (for each gene). Ideal classification thresholds were estimated for each gene, with position indicated by the horizontal red lines.

As a means of performing biological interpretation of gene expression data through discretisation, whilst avoiding the problems discussed regarding those commonly used approaches described above, a novel cluster-based discretisation method was developed to classify expression measurements for a given gene as in a significantly active transcriptional state or not, adapting to the particular characteristics of the distribution of expression measurements available for that particular gene. The key theoretical advantage of such an approach is that no structure is assumed on the distribution *a priori*, other than that for each gene there *may* be an interesting state of transcriptional activity indicated by a group of measurements of similar values, separated from the remaining measurements.

This cluster-based (binary) discretisation approach was described briefly in Section (3.4.4) as it was used for preprocessing of datasets in all BGA applications presented throughout Chapter 3. For quick binary discretisation, a statistic was developed to assess whether a gene has significant variation across a dataset, based on the range of values and the mean value in relation to the average range and mean of each gene across



(a) Low median fold-change threshold



(b) High median fold-change threshold

Figure 4.3: Classification of gene expression values on the basis of alternative median fold-change thresholds. Expression level measurement values for each gene are separated into low expression class (blue) and high expression class (red) on the basis of a median fold-change threshold. The same measurements are shown in both panels for the same genes, but each panel represents a different classification threshold, indicating that different thresholds are suitable for different genes.

the entire dataset. For genes with a relatively small range but high mean expression (i.e. a low value of s_i in Equation 3.1), it is determined that it is unlikely there is much variation with biological significance in terms of the expression of those genes as measured across the dataset. As the cluster statistic takes an essentially arbitrary value, an appropriate threshold (s_{min} for which $s_i < s_{min}$ implies gene i has little significant variation) should be chosen through empirical analysis of a carefully chosen[2] sample of genes from the dataset. For those genes classified by this method as having significant differential expression across the dataset, discretisation proceeds through binary classification according to k-means clustering (see [Hartigan, 1975]) with two cluster means (i.e. $k = 2$) initialised to the mean of each group of values either side of the greatest pairwise separation of any values.

To compare the performance of these various discretisation approaches in practical terms, evaluation of the results of applying each approach to the same datasets was required. An obvious approach to such evaluation involves the visual inspection of appropriateness of the assigned groupings, and for this purpose Figs. 4.4 - 4.6 show examples of gene distributions from the dataset described in Section (3.3), with classification boundaries from each of three discretisation approaches:

- Normal distribution:

$$val = \begin{cases} 1 & \text{if } val > \mu + \sigma \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

- Fold-change:

$$val = \begin{cases} 1 & \text{if } val > median(Vals) * 2 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

- Cluster-based discretisation: as performed by Algorithm (1) given in Section (3.4.4)

These figures illustrate that the adaptive approach taken by the novel cluster-based discretisation better fits the distributions of a variety of genes' expression values. However, such an evaluation is potentially subjective and certainly incomplete: the sampling of genes shown here is too small to represent the full range of $> 40,000$ such distributions across the dataset used in the above analysis.

For more complete and objective evaluation of the relative performance of each of the above discretisation approaches, simple measurements were devised to assess the

[2] chosen to represent the extremes of distribution patterns as well as those that are more commonly observed

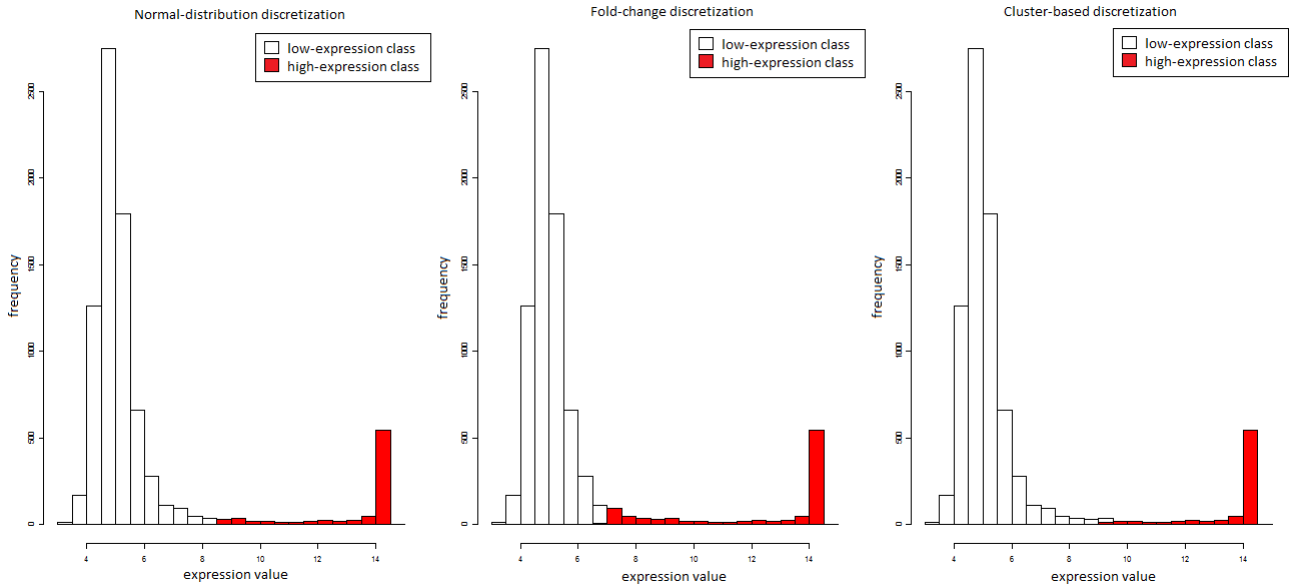


Figure 4.4: discretisation of gene expression levels for Alb across large gene expression dataset, from three methods. discretisation methods shown, from left to right, are: Normal-distribution, 2* median fold-change, and the novel cluster-based discretisation.

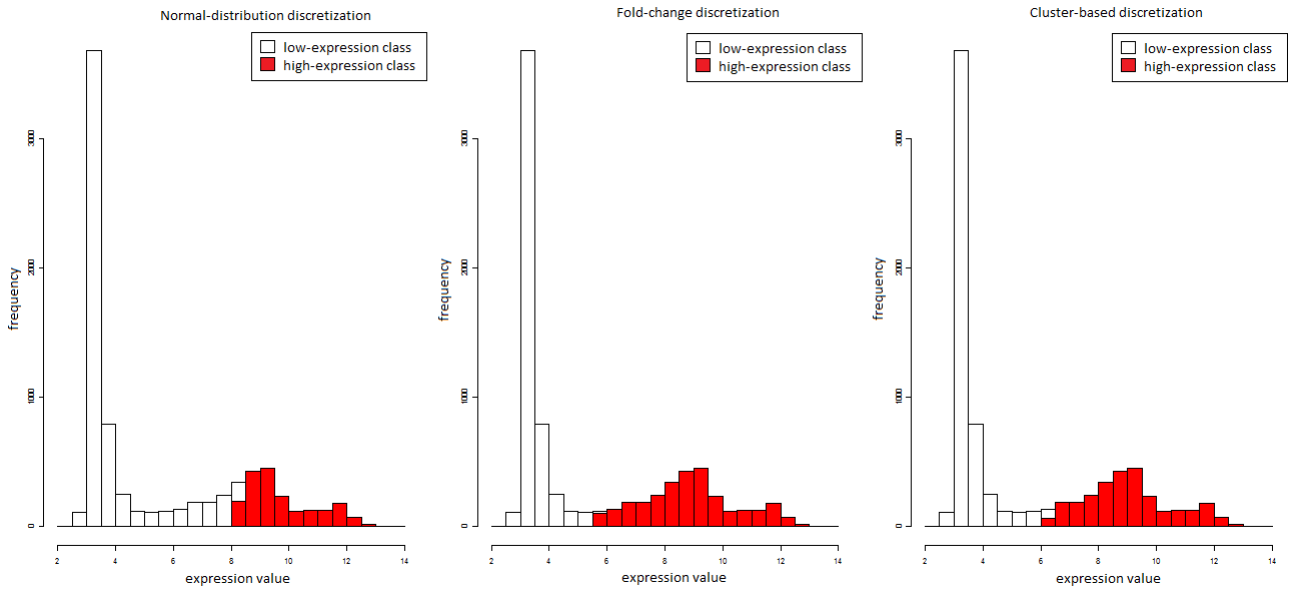


Figure 4.5: discretisation of gene expression levels for Sox2 across large gene expression dataset, from three methods. discretisation methods shown, from left to right, are: Normal-distribution, 2* median fold-change, and the novel cluster-based discretisation.

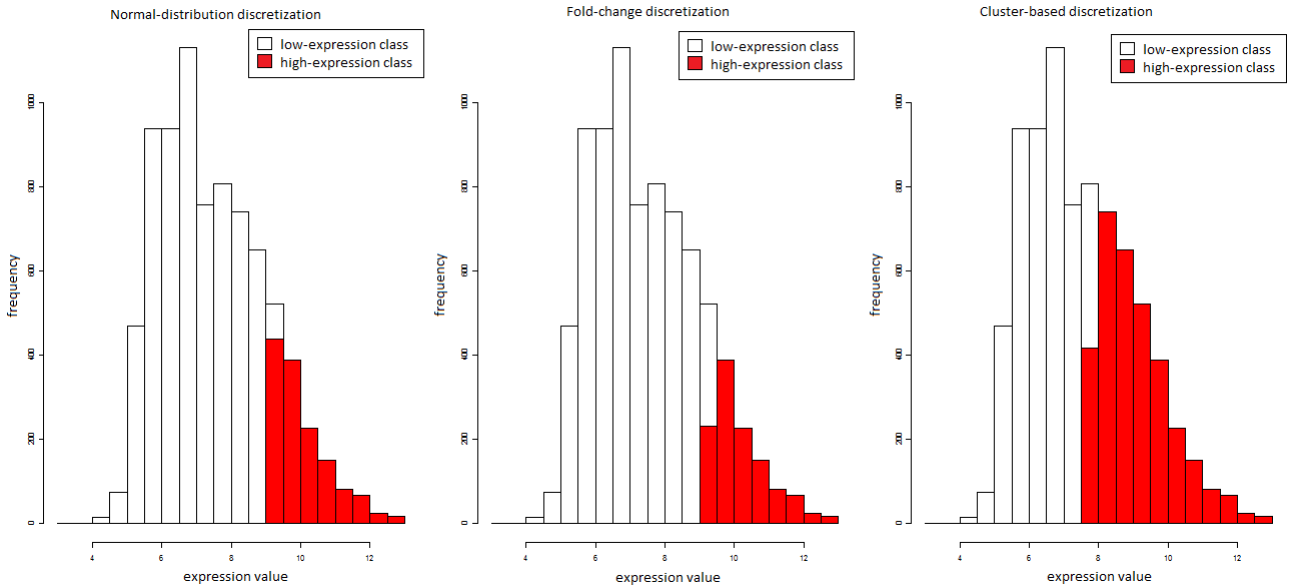


Figure 4.6: discretisation of gene expression levels for Myc across large gene expression dataset, from three methods. discretisation methods shown, from left to right, are: Normal-distribution, 2* median fold-change, and the novel cluster-based discretisation.

degree to which each method fits the underlying data. For this evaluation, the appropriateness of the partitioning performed by each method was assessed by calculating the ratio of the sum of within-cluster distances for the two clusters to the total sum of distances for each gene, as shown in Equation (4.3). In Equation (4.3) x represents the vector of expression values for the given gene, C_1 and C_2 are the sets of indices representing membership of low and high expression classes. The distributions of these inter-cluster distances for each of the above methods (and one additional method involving classifying the top 10% of each gene's values as representing high expression) are shown in Fig. 4.7, from 1000 genes chosen randomly from the set assigned 2 classes by all methods used.

$$d = \frac{\frac{\sum_{i \in C_1} \sum_{j \in C_1, j \neq i} (x_i - x_j)^2}{\frac{1}{2} * (|C_1|^2 - |C_1|)} + \frac{\sum_{i \in C_2} \sum_{j \in C_2, j \neq i} (x_i - x_j)^2}{\frac{1}{2} * (|C_2|^2 - |C_2|)}}{\frac{\sum_{i} \sum_{j \neq i} (x_i - x_j)^2}{\frac{1}{2} * ((|C_1| + |C_2|)^2 - (|C_1| + |C_2|))}} \tag{4.3}$$

This evaluation clearly demonstrates that the novel cluster-based discretisation approach introduced in Section (3.4.4) outperforms other discretisation approaches with reported application to gene expression data as a preprocessing step prior to bicluster analysis, as it reliably provides more tightly-grouped clusters than the alternative approaches whilst still defining only 2 classes of expression ('high' or 'low').

The main advantage afforded by all the above discretisation approaches is that a large gene expression dataset is transformed from numerical measurements into symbols with an immediate biological interpretation: one value implies that a gene is either expressed significantly above its 'background' level, another implies that it isn't. However, this simplicity has the potential to limit the success of any analysis method employing a discretisation approach, due to the fact that a threshold must be placed at some point in the distribution of expression values and, unless every gene (or at least the majority of genes) has clear separation between different groups of values, this may result in some values only marginally higher than those considered to be at a background expression level being classified as significantly highly expressed. An illustration of the practical problems that can result from this situation is presented below.

It was discovered in biclusters found in real datasets using the entropy-based and guidegene-dependent BGAs described in Section (3.6) that there was an enrichment in a range of biclusters for genes with particular distribution patterns. Across a set of biclusters discovered using the BGA with a range of different (and uncorrelated) guidegenes, bicluster genes appeared to be more likely to have a threshold towards the upper limit of the lower group of expression values (according to visual inspection)

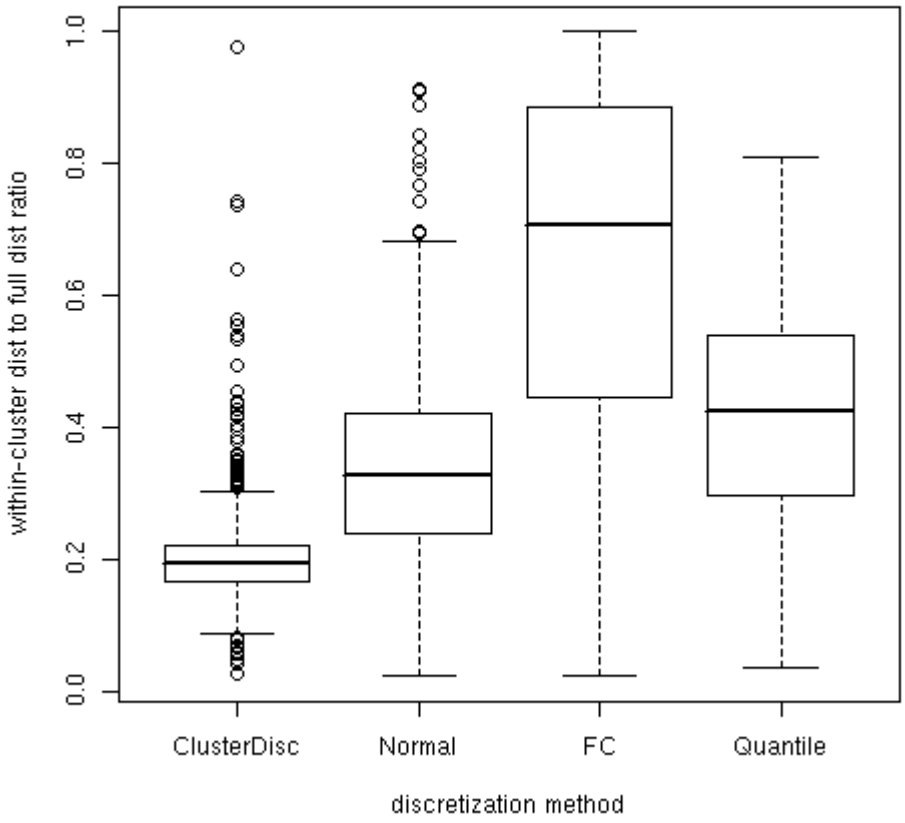


Figure 4.7: Distribution of scaled within-cluster distance to overall-distance ratios for a range of discretisation approaches. The distance ratios shown were calculated according to Equation (4.3). This measure represents the dissimilarity between the gene's measurement values that are classified as high, added to the dissimilarity between the gene's measurement values that are classified as low, and scaled by the overall dissimilarity between all of that gene's measurements. A lower score represents a better separation of highly-expressed and lowly-expressed measurements for the gene in question. Methods used, from left to right, are the novel cluster-based discretisation, discretisation based on the normal distribution as described in Equation (4.1), discretisation by median fold-change as described in Equation (4.2), and a universal threshold assigning the top 10% of a gene's measurements to the high-expression class and the rest to the low-expression class.

than by chance. In order to assess the generality of this observation, a quantitative measure was devised to provide an indication of a gene's presence in a bicluster on the basis of the distance from the mean of that gene's expression within the bicluster to the greatest value of the gene's expression when the discretisation process has assigned it to a 'low' class of expression, scaled by the distance between that gene's mean expression values for each of the 'high' and 'low' expression classes. The effectiveness of each discretisation method at facilitating effective bicluster discovery was evaluated by applying this measure (shown in Equation (4.4)) to a large set of biclusters discovered by the BGA (as described in Section (3.6.2 & 3.6.3)) and comparing the distribution of the resulting scores to a reference distribution obtained from applying the measure to a large number of randomly sampled genes for randomly chosen samples belonging to that gene's 'high' expression class set. These distributions are shown in Fig. 4.8, along with the distribution of scores for only the top ranking 200 genes in each bicluster's genelist. From these distribution plots it is clear that biclusters discovered in discretised data (preprocessed by a method shown in Section (4.1.2) to be more appropriate for preprocessing large gene expression datasets than any existing methods) using a probabilistic biclustering algorithm are more likely than would be expected by chance to involve genes which are included in the bicluster on the basis of consistently high expression and yet show poor separation from values classified as low expression, and that this observed effect applies across the top ranking genes (based on probabilistic bicluster-inclusion score) across each bicluster.

$$d = \frac{\dfrac{\sum\limits_{i \in Bicluster} x_i}{|Bicluster|} - \max\limits_{i \in LowClass}(x_i)}{\dfrac{\sum\limits_{i \in HighClass} x_i}{|HighClass|} - \dfrac{\sum\limits_{i \in LowClass} x_i}{|LowClass|}} \tag{4.4}$$

In the entropy-based framework for biclustering, this observed effect would be particularly problematic as it would result in genes being considered highly specifically expressed in a bicluster due to relatively insignificant expression level fluctuations either side of a discretisation threshold (akin to measurement noise). It is proposed that these genes might be confounding the attempt through application of the BGA to identify relevant transcriptional relationships involving particular genes of interest. This may explain why, although the enrichments of TFs' DNA-binding targets in biclusters shown in Section (3.6.3) was statistically significant, the actual number of known targets discovered in the bicluster lists was fairly low. At an average proportion of bicluster genes with DNA-binding evidence only 1% for Oct4 and Sox2, up to 10% for Esrrb and cMyc, a 1/10 to 1/100 chance of a bicluster gene being is probably too low to be very useful in guiding primary biological research.

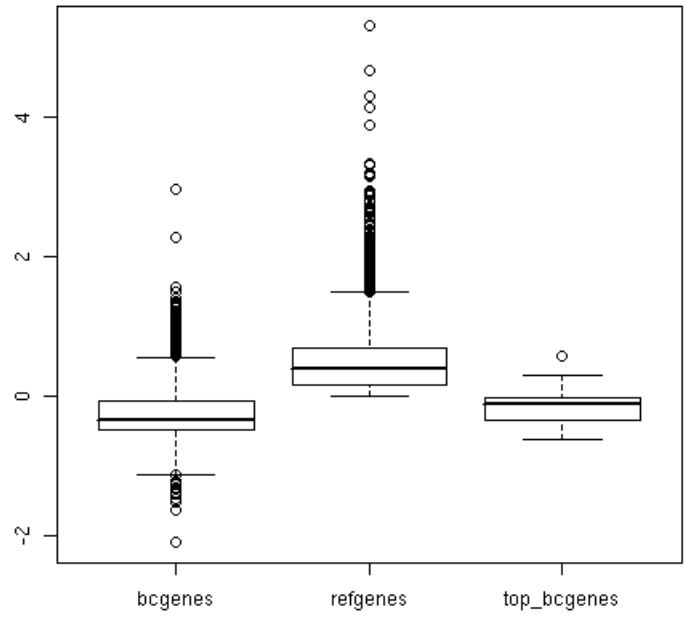


Figure 4.8: Distribution of scores for 'borderline' bicluster genes when compared with a randomly-sampled set of reference genes, with a lower value indicating a less clear distinction of a gene's expression between values assigned a 'low' expression class and values assigned a 'high' expression class (a necessary condition for membership of a bicluster.)

It therefore seems that it might be advantageous to produce a data transformation approach that can generate immediately biologically interpretable values on some continuous scale from underlying (potentially arbitrary) gene expression measurements. The remainder of this chapter deals with the development, evaluation and application of such continuous transformations for biological interpretability of gene expression measurement values.

4.1.3 Calibration with 'Spike-In' Datasets

To avoid the problems observed when applying pattern mining techniques to discretised gene expression data, it was deemed necessary to develop a continuous-scale data transformation approach that could provide biological interpretation of numerical expression values in isolation of any other information regarding the gene or sample involved.

One possible approach envisaged to achieve this transformation was to find a means of transforming median fold-change values (as in Equation (4.2)) into some standardised scale where a given fold-change measurement has a given biological interpretation, regardless of the context of that measurement. The proposed method of transformation involved using 'spike-in' reference datasets (e.g. [Choe et al., 2005]) to calibrate measured fold-changes into a scale representing the probability of a significant change in expression level. This could be achieved through estimation of false-positive rates of declaring significant biological variation at observed fold-changes. While spike-in datasets do not offer a direct estimation of biological significance, they do provide reference measurements from microarray platforms for known concentrations of mRNA. With a relatively simple assumption, such as that a doubling of the concentration of mRNA is always going to be biologically significant, this approach could however be applied to the task discussed here.

To assess the feasibility of such a false-positive calibration approach to the required data transformation, a predictor was created that could act as a transformation function to give estimated probability of significant change for any measured level of gene expression (in terms of fold-change to the median of all reference samples). In essence, such an approach aims to model the measurement errors of microarray platforms. Unfortunately, as there was not a spike-in dataset available for the microarray platform from which the majority of the data analysed in this work, a dataset [Choe et al., 2005] from a related platform (Affymetrix DrosGenome1 GeneChip) was used in lieu. Spike-in datasets on Affymetrix platforms are available for *Drosophila melanogaster* [Choe et al., 2005] and *Homo sapiens*[3] microarrays. The cyclic 'latin square' design of the human dataset did not provide a sufficient reference set of genes without differential expression between the dataset's samples. This rendered it impossible to derive the desired calibration from this dataset and the 'Golden Spike' Drosophila dataset of [Choe et al., 2005] was used. As the measurements from this dataset were on a different scale to those in the large meta-analysis dataset for which the transformations were intended, any results from analysis of this spike-in dataset would have to be transformed into the appropriate range. The spike-in dataset contains data from 'spiked-in' samples with mRNAs injected at known concentrations, each paired with a 'control' sample with all the mRNAs injected at a reference (lower) concentration.

[3] available online from http://www.affymetrix.com

From the measurements corresponding to each of these differentially injected ('spiked-in') mRNAs, measured fold-changes to median reference level were calculated for known fold-changes of mRNA concentration injected into the different samples. Fig. 4.9 shows the distribution of measured fold-changes for known changes in the spiked-in concentration. Interestingly, for samples with some higher levels of injected mRNA concentration the corresponding gene expression measurement on the microarray was lower than for samples with less mRNA injected. This observation suggests that there may be issues with attempting to use this data for estimation of false-positive rates arising from measured fold-changes: many negative measure fold-changes correspond to a false-positive rate of less than 1 in terms of representing real change in gene expression level (this observation is repeated in [Irizarry et al., 2006]).

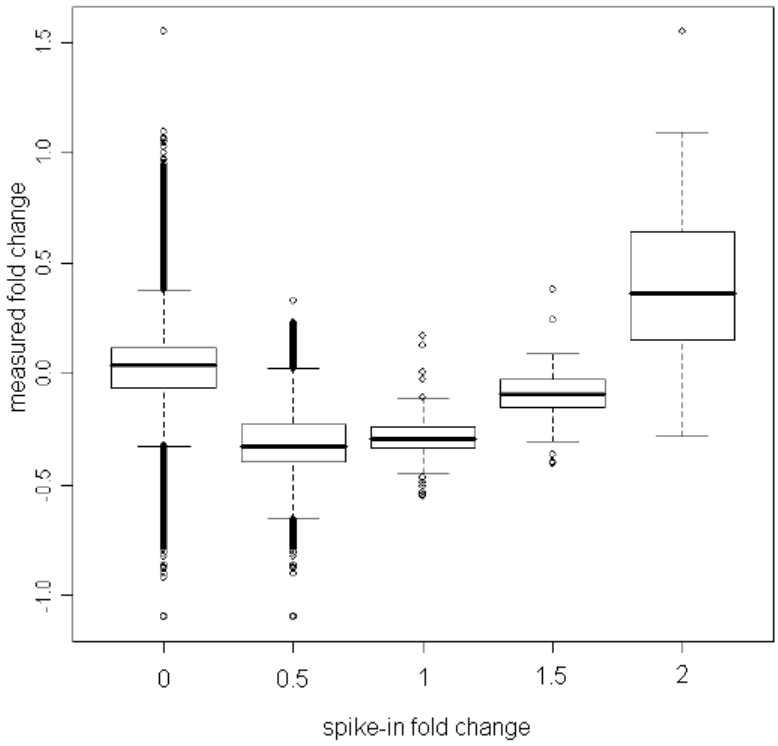


Figure 4.9: Distributions of measured fold-changes to median for different levels of differential mRNA spike-in. Value on the vertical axis represents fold-change of measured level of 'spiked-in' mRNA compared to median (non spiked-in) value for that gene. A box plot is given for each conentration at which mRNAs were spiked into the samples.

To continue with this approach, the measured fold-changes were transformed into the appropriate range for the meta-analysis dataset through quantile-alignment. An estimator of false-positive rates, corresponding to the probability that a given measurement doesn't represent at least a two-fold change in mRNA concentration was produced by creating a table of the measurements observed in the quantile-aligned spike-in dataset

and the probabilities that any given measured fold-change arises without there having been at least a two-fold increase in the injected mRNA in that sample (when compared to the median of others). Quantile-alignment was performed using the `glm` and `predict` functions available in the `glm` package within R. This false-positive rate estimator is shown in Fig.4.10, with the output of the predictor plotted against the input fold-change value.

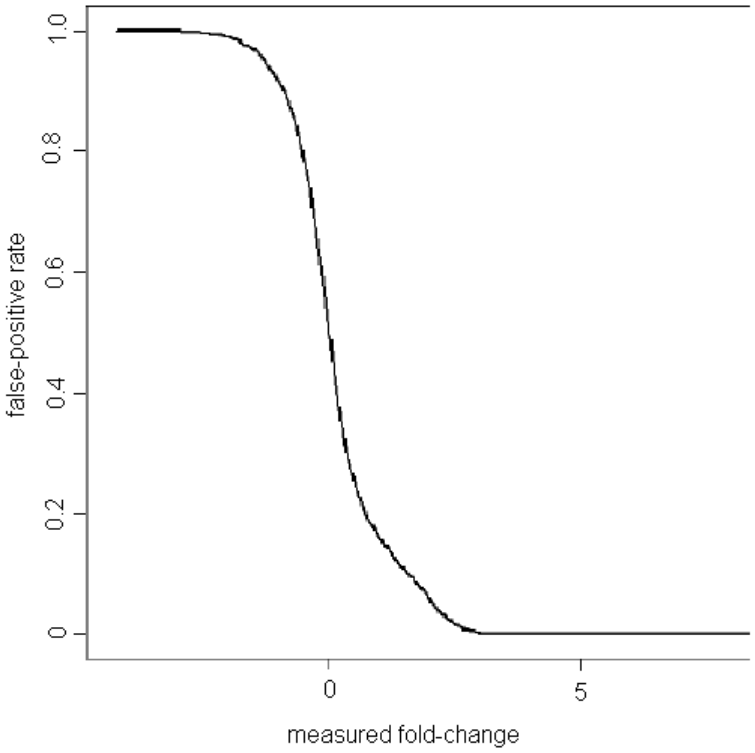


Figure 4.10: Estimates of false-positive rate for classifying given fold-changes as representing significant variation of expression level.

Finally, an estimator for true-positive rates of a given measured fold-change representing at least a 2-fold change in corresponding mRNA concentrations was created by fitting a sigmoid predictor to the $(1 - FP)$ values from the above table (for positive fold-change values only). A model of the form given in Equation (4.5) was fitted using the `nls` function in R, implementing a nonlinear least-squares model-fitting approach. The resulting predictor is shown as the red line in the plot in Fig. 4.11, which also shows the underlying $(1 - FP)$ values obtained from the quantile-aligned measurements.

$$P(x) = \frac{1}{1 + e^{-x}} \tag{4.5}$$

It is clear from the inspection of this predictor that this calibration approach would not be appropriate for the desired data transformation task, due to the fact that a

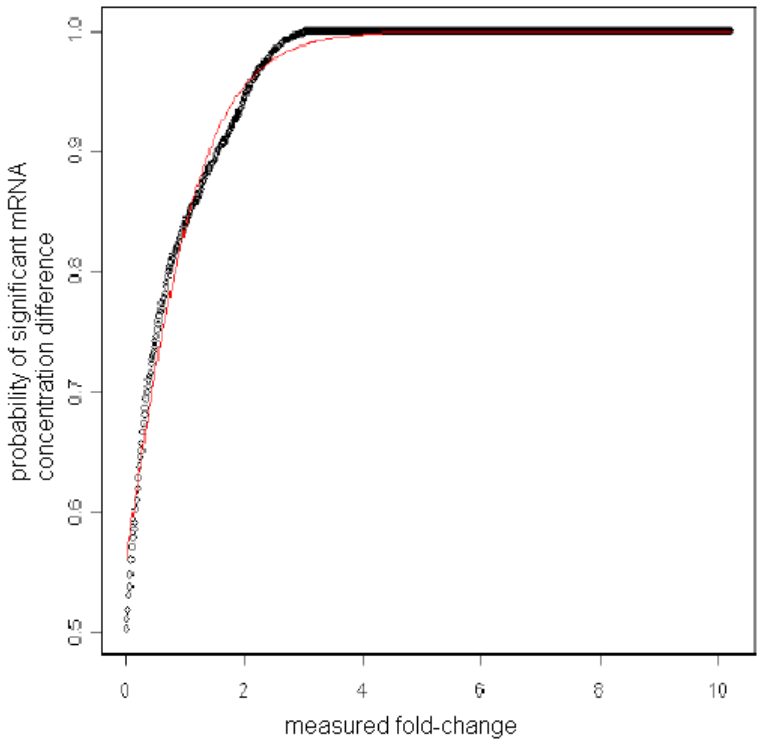


Figure 4.11: Predictor of probability of significant variation for given values of fold-change to median reference level

measurement at precisely the median value for that gene has an estimated probability of > 0.5 of being at a significantly higher expression level than the general reference 'background.' The plot in Fig. 4.11 shows that this calibration would result in a scale where nearly half the measurements for every gene are considered to be expressed at a significantly higher level than a general background level, with very little variation of estimated significance corresponding to very large (and presumably significant) differences in measured fold-changes: for example, an increase in fold-change above the reference level from 1x to 5x results in an increase in the estimate of significantly high expression of only approximately 0.8 to 1.0. In addition, this calibration was based on the assumption that a two-fold increase in mRNA concentration would always be significant, regardless of the gene involved. Owing to the obvious failing of this method, and potentially also of this assumption, an alternative approach was taken to attempt to utilise all the data available within the large meta-analysis dataset to model the biological significance of measure gene expression levels, and this alternative approach to the data transformation is the subject of the following section.

4.1.4 Expression-State Modelling

As the results of attempting to calibrate gene expression measurements using spike-in datasets (described above) were clearly inappropriate, an alternative approach to such calibration was sought. If the goal of such a calibration is to provide estimates for the probability that a given expression measurement (that is to say, both an intensity value and fold-change to some reference) corresponds to a significant change in expression level of that gene (compared to the reference level), this could be achieved through fitting an appropriate error model such as the Rocke-Durbin model for measurement error on microarray platforms [Rocke and Durbin, 2001]. This model, described in Equations (4.6-4.8), estimates a distribution of measured values, Y, likely to arise from a single underlying value, x.

$$y = \alpha + xe^{\eta} + \epsilon \tag{4.6}$$

$$\eta \sim N(0, \sigma_\eta) \tag{4.7}$$

$$\epsilon \sim N(0, \sigma_\epsilon) \tag{4.8}$$

By appropriately fitting the model parameters to the observed data, a probability can be estimated that any two measured values might arise from the same underlying expression level, and thus a predictor of the significance of any given difference in measured values can be obtained. However, this 'significance' refers only to the probability that there is a real difference in the expression levels represented by the given measurements. While an assumption such as that a 2-fold change in underlying expression level is biologically significant could be incorporated into this framework so that estimates for $P(x > 2\bar{X}|y, \bar{Y})$[4] could represent a probability of biologically significant variation, such an approach is still dependent on an essentially arbitrary and universally applied assumption of significance of differential expression. In order to get around this issue, more involved models of the biological significance of particular expression levels of each gene would have to be constructed individually.

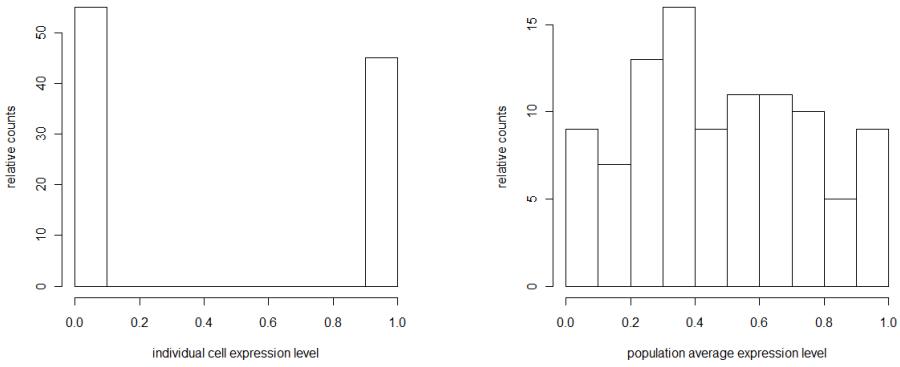
Motivation

Given that the meta-analysis proposed in this work is intended to involve as comprehensive a collection of gene expression data as is possible, this ought to imply that as the datasets involved cover more samples from a given organism, so the set of all (comparable) measurements of each gene tend towards a complete distribution representing the full range of expression levels that gene may take. Providing that this assumption holds, it would then be possible to analyse these distributions of expression values for

[4] where x and y are an expression value and measurement, respectively, and $\bar{X}$ and $\bar{Y}$ are the averages across the distribution of expected underlying expression values and the observed distribution of measurements

each gene to identify natural (or in rare cases forced) 'states' of expression, represented by groups of samples with similar values.

One possible confounding factor to such an analysis arises from the fact that the gene expression data come from (sometimes heterogeneous) mixtures of cells. As a result, even if a hypothetical gene is only ever expressed at one of two levels (0 or 1, say), owing to a complete range of possible proportions of cells expressing the gene at either level in every measured sample, the observed measurements could lie anywhere in the range $(0, 1)$. This principle is illustrated in Fig. 4.12, where distributions are shown for the expression levels of the hypothetical gene across a set of individual cells and across a set of mixtures of cells (it is assumed that all measurements are amplified/normalized to the same scale).



(a) Individual cells (b) Mixtures of cells

Figure 4.12: Hypothetical distributions of measurements from individual cells and from mixtures of cells

However, unless the mixing of cells in the input samples is uniformly distributed (over the range of expression levels the gene in question may take) across all the samples in the (assumed to be comprehensive) dataset, and such a situation would imply no discernable link between the expression level and the biological context, the different states of expression of that gene will emerge as components of the overall distribution of measured expression values. Therefore, by modelling the underlying components of the distribution of measured expression levels for each gene, it is proposed that the biological significance of any given measurement could be inferred from the estimated contribution of different expression 'states' of that gene constituting the given sample's measurement.

Of course, some genes may have biological consequences based on a finely-controlled quantitative level of the abundance of transcript, but the proposed modelling approach

doesn't preclude such cases being considered appropriately: as the intended output of the modelling proposed here is a set of values on a standard numerical scale indicating the biological significance of expression, those corresponding values will vary accordingly across the spectrum of observed values while conforming to any structure implied by the distribution of these measurements over the whole data.

It may also be worth noting that such an approach involving component-based modelling of gene expression distributions over large collections of data has been applied to other tasks: the estimation of gene expression variance in improved statistical measures of differential expression [Kim et al., 2010b] and the discretisation of large-scale microarray datasets as a precursor to sample classification [Zilliox and Irizarry, 2007]. However, there are no known reports in the literature of attempts to model the biological significance of gene expression measurements given a compendium of data, aside from that implicitly modelled in discretisation approaches, so the work presented below represents exciting novel developments in the theory and potential practical applications of the analysis of large-scale gene expression datasets.

Approach and Implementation

As described above, the goal of the intended modelling approach is to identify structural components within the distributions of measurements for a particular gene across the whole collection of samples profiled in the dataset. Once these structural components have been discovered (if they exist), a classifier can be constructed to compute probabilities that any observed measurement arose from each underlying distribution: that is, the likely contribution of each component can be estimated for any value. From this general modelling framework, a biological interpretation can be afforded by combining the component classification probabilities into a linear scale representing the probability of a high-active transcriptional state of the given gene (with intermediate values quite possibly representing intermediary biological states of expression). If preferred, the proposed framework is flexible enough to allow the output to be expressed in terms of classification probabilities for any number of distinct biological states of expression that appear to emerge from the pattern of distribution of measurements for the given gene across the dataset.

An implementation of the data transformation approach based on the proposed modelling framework requires the performance of a number of constituent tasks, described below in terms of the implementations used in this work.

Model Construction and Fitting To model any component-based structure within the distribution of expression values for a gene, a Gaussian mixture model (GMM) approach was taken. GMMs provide a flexible framework for modelling data distributions involving different component distributions: this is especially useful when applied to

multi-modal distributions [Fraley and Raftery, 2002] such as the example in Fig. 4.13. A GMM models a given data distribution as a weighted sum of component Gaussian distributions, as in Equations (4.9-4.10). The examples here concern only univariate data, as is the case with the data to be modelled, although the GMM framework is extendable to data of multiple dimensions. According to the definition in Equations (4.9-4.10), the GMM is parameterized by a set of weights, means and standard deviations for each distribution, and implicitly parameterized by the number of mixture components specified.

$$p(x|\theta) = \sum_{i=1}^{m} w_i g(x|\mu_i, \sigma_i) \tag{4.9}$$

$$g(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.10}$$

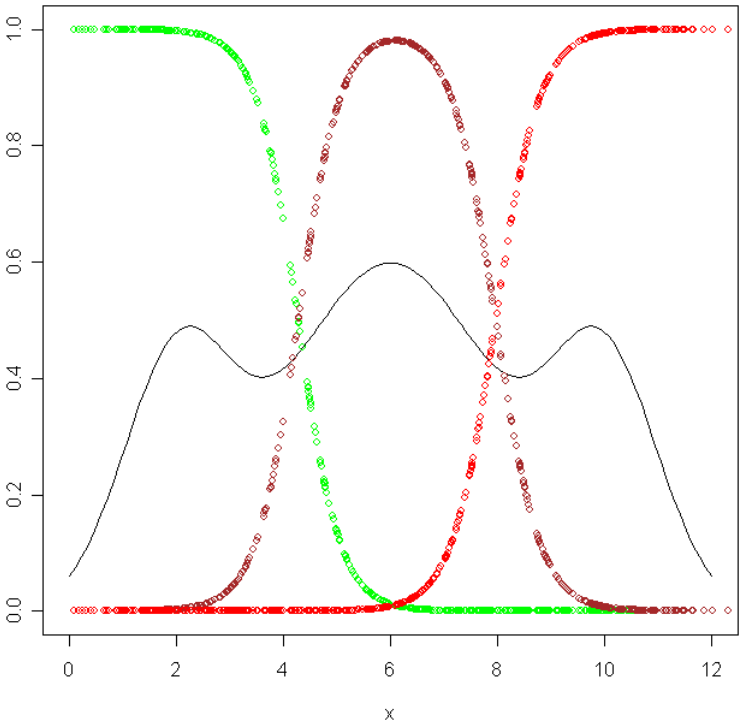
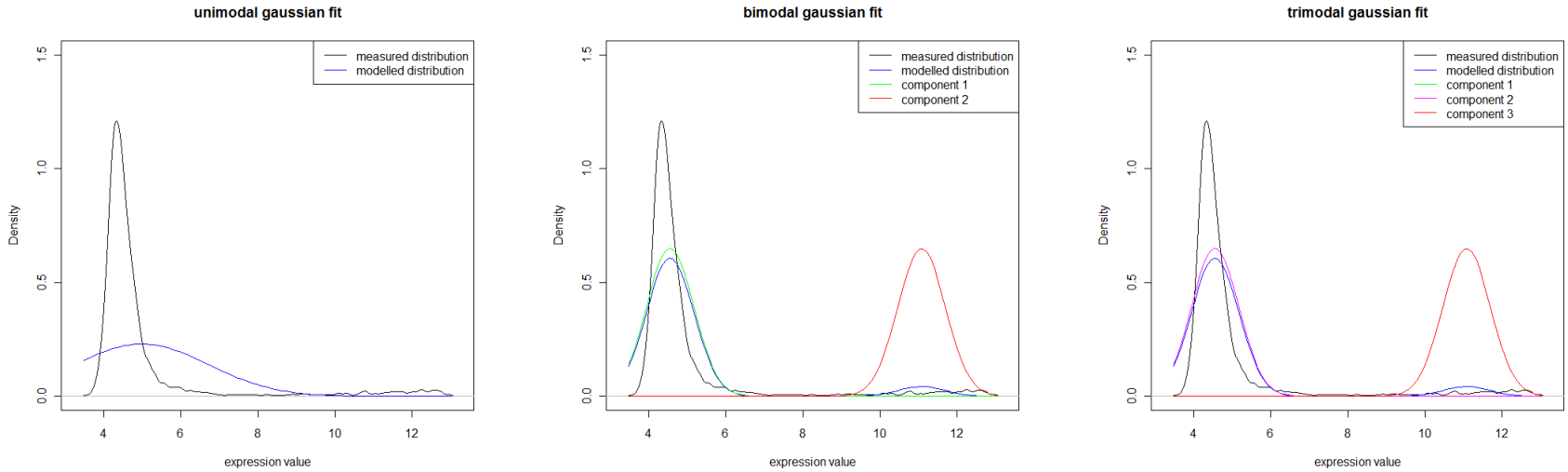


Figure 4.13: Illustration of Gaussian components (red, brown and green) from a multi-modal distribution (black line)

In this modelling of the distribution of gene expression values, the number of components in the distribution is not known in advance, in fact the estimation of this property is often one of the key steps in specification of the appropriate model (and is indirectly one of the key outcomes of the modelling process). The Bayes Information Criterion (BIC) [Schwarz, 1978] provides a means of assessing different parameterizations of a model, through calculating maximised loglikelihood whilst penalising increasing num-

bers of parameters in the model. Fitting a number of GMMs with different specified numbers of mixture components, then selecting the model with the best BIC score provides a natural means of selecting the most appropriate model for the observed data. The `mclust` package in R provides a function `Mclust` that implements the above procedure for fitting a number of GMMs to the data given (with parameter estimation for model fitting performed by expectation-maximization (EM) algorithm) and selecting the model that best fits the data according to the BIC. For further details of the procedures implemented in `Mclust`, see [Fraley and Raftery, 1999]. An illustration of this GMM-fitting procedure is given in Fig 4.14, in which a (bimodal) gene expression value distribution is modelled by GMMs with 1,2 and 3 components. The overall data distribution is shown in black, the component distributions in green and the overall model distribution in red. Fig 4.15 shows the BIC scores for each of the models (and up to 5 Gaussian components), illustrating that the model with 2 Gaussian components did indeed best fit the data according to this criterion.

A potential problem that arose from this application of `Mclust` to the modelling task at hand is illustrated in Figs. 4.16 & 4.17. These show similar plots to those in Figs. 4.14 & 4.15, although the (clearly bimodal) distribution is best fitted by a higher-order model (one with more components than appear to be present in the underlying data). In addition, two of the components of the model clearly map to the same component in the data, presumably due to the fact that this component in the data is not Gaussian in shape. However, this 'over-specified' model does fit the underlying data better, so it was deemed appropriate to use `Mclust` to fit a GMM whether over-specified or not, and use the individually fitted components non-naively to produce the desired scale of biological significance.



(a) 1 component (b) 2 components (c) 3 components

Figure 4.14: GMMs with different number of mixture components fitted to data

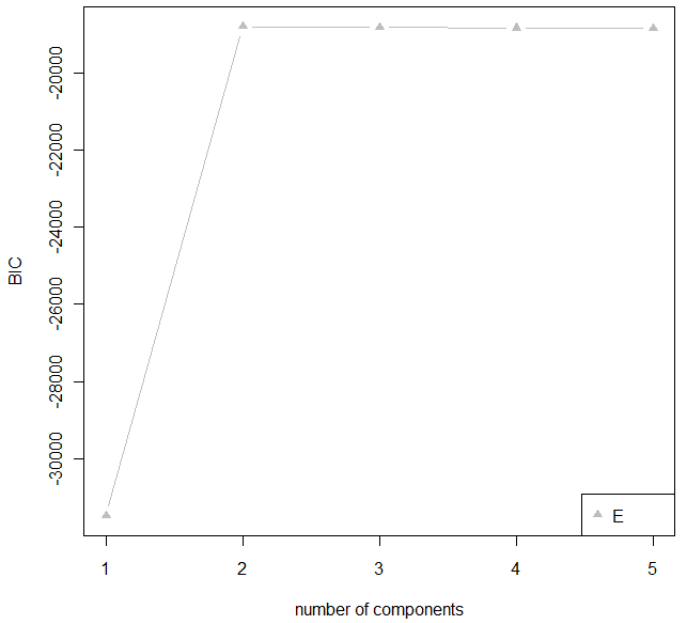


Figure 4.15: Bayes Information Criterion scores for models fitted in Fig. 4.14

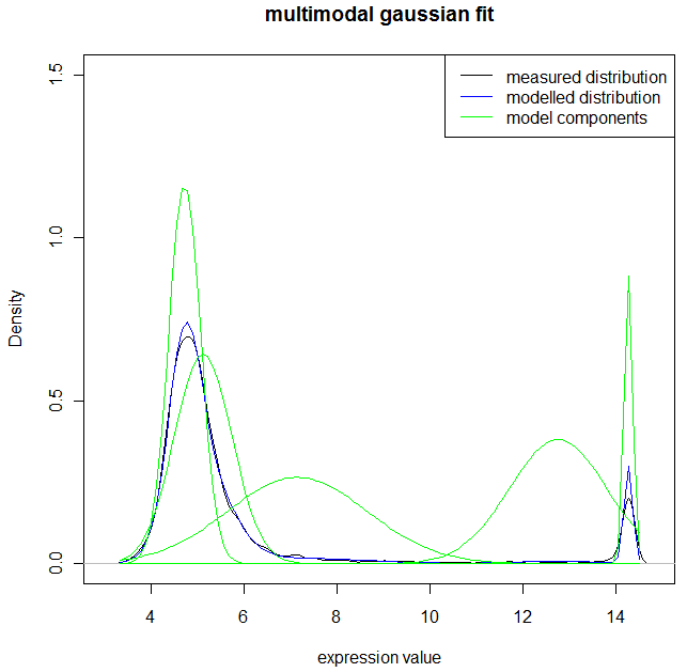


Figure 4.16: Distribution plots for a GMM (blue) with five individual Gaussian components (shown in green), fitted to a clearly bimodal data distribution (shown in black).

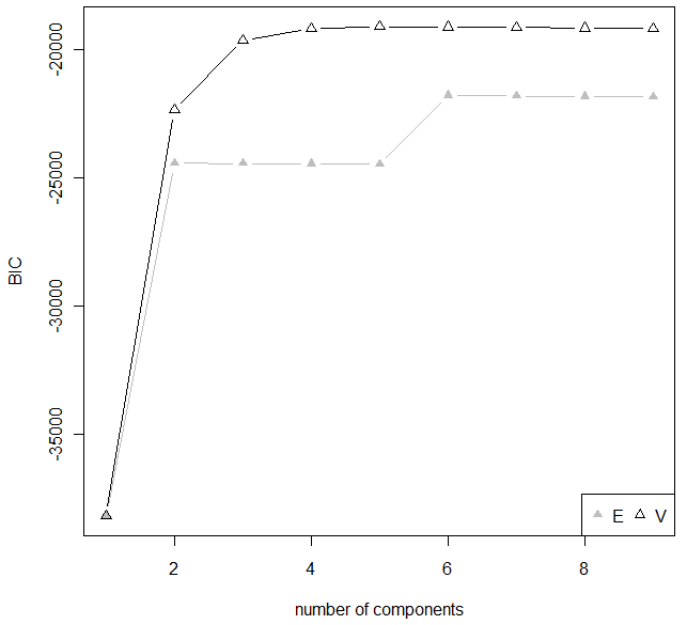


Figure 4.17: Bayes Information Criterion scores for models fitted in Fig. 4.16

Construction of Biological Significance Predictor In order to use the output from `Mclust` GMM-fitting to produce an estimator for biological significance of any measurement of a given gene, firstly a heuristic was developed to merge highly overlapping Gaussian components in the model. This merging heuristic proceeds on the basis that if a number of components each contribute a significant proportion of the overall density to a number of values (i.e. the Gaussians overlap) but have little or no such overlap with any other components, it is likely that these overlapping model components represent different features of a single structural component in the data and should therefore be merged. The result of merging becomes apparent in the creation of the estimator for biological significance, where classification confidences from the e-step of the EM algorithm are used to generate biological significance values according to the probabilities that a given measurement belongs to each of the biological states of expression represented by the structural components within that gene's expression data distribution. In this process, the confidences of classification for each of the merged components are added together to give a probability that the measurement belonged to the expression state corresponding to the structural feature represented by the combined modelling component.

Another issue remains in the development of a universally interpretable scale of biological significance from a series of fitted GMMs, due to the fact that these GMMs may have different numbers of components, owing to differences in the characteristic expression patterns of the genes they represent. To provide a means of unifying the classification confidences for each of the components of these different models, it was noted that there will always exist such confidences for values to be classified as belonging to the lowest and highest components of any model with more than a single component. For models with a single component, the cumulative probabilities of each measured value provide a confidence score for classification to the high state of expression, and can be subtracted from 1 to give a classification confidence score for a low state of gene expression. If the unified scale concerns the confidences of a gene being in its highest or lowest observed expression state, values on such a unified scale can be obtained regardless of the number of components in the respective fitted GMM by producing a monotonically decreasing classification score for values belonging to the lowest expression state through subtracting all higher components from the classification confidence of the value arising from the lowest component in the model (with a minimum classification score of 0 and maximum 1). Similarly, a monotonically increasing classification score can be obtained for values belonging to the highest expression state, regardless of the number of components in the fitted model (assuming that number is at least 2). This approach additionally means that, assuming monotonicity of the classification scores is appropriately ensured, it is not so great a problem to have overlapping model components (such as if the merging heuristic fails) due to the fact that from the point that the most extreme (highest or lowest) model components contribute most to

the overall distribution, this and any other components will be combined to produce the appropriate score. An illustration of this behaviour is shown in Fig. 4.18, with a hypothetical gene expression value distribution plotted, including the fitted model components (brown) and the resulting classification scores for low expression (green) and high expression (red). A straightforward linear combination of these two component scores, given in Equation (4.11), results in a unified scale for biological significance of expression level, illustrated in blue in Fig. 4.19. It should be noted that this scale does not enforce bimodality (all values being either 'high' or 'low') due to the intermediary components reducing the confidence of classification to the nearer of the extreme components. Therefore, intermediate states of expression will appear as values on this unified scale of around 0.5 (with confidence of not being in either a high or a low expression state).

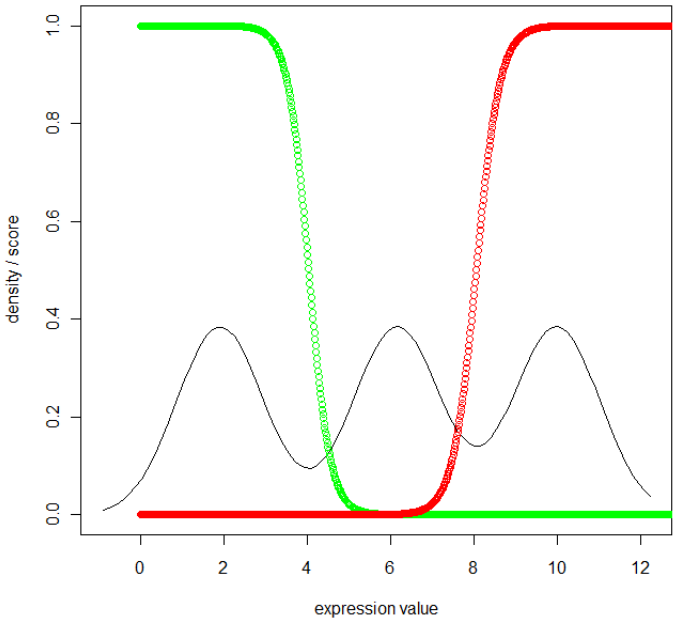


Figure 4.18: Classification scores GMM from a multi-modal distribution (black line)

$$score_{overall} = \frac{1 + score_{hi} - score_{lo}}{2} \tag{4.11}$$

The above findings are based on the assumption that monotonicity of the respective high- and low-class classification scores is ensured appropriately. This is not necessarily straightforward, but a successful approach was implemented finding the points (if such points exist) at which classification scores for the most extreme components revers (e.g. lower expression values have *lower* classification scores for the lowest component) and adding to those incorrect classification scores the classification scores for all other classes, weighted by mixture proportions derived from a cumulative normal distribu-

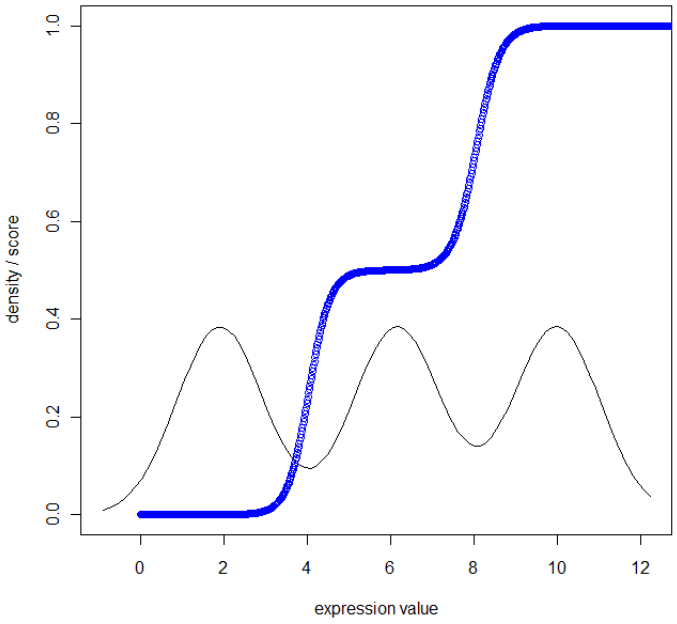


Figure 4.19: Unified expression state values for hypothetical gene with a multi-modal distribution (black line)

tion. This ensures that the additional intermediary components do not contribute too greatly to the classification score for the extreme class when the value may well have a significant change of belonging to an intermediate expression state.

The above discussion concerns only the distributions of the measured expression values. However, as mentioned at the beginning of this section, it is possible to fit measurement error models to obtain (from the measured values) estimates of underlying expression level.

Error Model Incorporation As described at the beginning of this section, the Rocke-Durbin error model provides a method of estimating the range of measurements (and their corresponding probabilities) that might arise from a given underlying value. This incorporates a general background level (α in Equation (4.6)), an intensity-dependent error term proportional to the logarithm of the underlying value (η in Equation (4.6)) and a constant measurement error term (ϵ in Equation (4.6)). For further details regarding the motivations for this form of error model, see [Rocke and Durbin, 2001]. Appropriate estimation of the error model parameters therefore results in a method to estimate, for any given measurement, the distribution of underlying values likely to give rise to such a measurement. An expression corresponding to this method is given in Equations (4.12-4.14).

$$x = \frac{y - \alpha - \epsilon}{e^{\eta}} \tag{4.12}$$

$$\eta \sim N(0, \sigma_\eta) \tag{4.13}$$

$$y - \alpha - \epsilon \sim N(y - \alpha, \sigma_\epsilon) \tag{4.14}$$

Derivation of an analytical expression for the resulting distribution of x in terms of y, σ_η, α and σ_ϵ is not clear, however values from this distribution can be generated by computer simulation.

This error model can be incorporated into the data transformation by providing, for every measurement, ranges of possible underlying values that could be mapped onto the biological significance scale for that gene. The calculation of the expected biological significance value for the underlying expression value additionally incorporates a notion of the expected measurement error involved in each value. Therefore, while the transformed values for a given gene will always lie in the range $(0, 1)$, the incorporation of the error model would mean that both genes with generally low expression levels and those with a low expression range would have expected variation in the measurements covering a significant part of their overall distribution. This in turn would mean that less confidence could be assigned to the prediction of a value being in a particularly high or low expression state. This expected effect was confirmed by measuring the range in expression state scores across the dataset for each gene in turn and calculating the correlation to the range in expression values across the dataset for each gene and the average expression values across the dataset for each gene. Pearson correlation tests (applied using `cor.test` in R) revealed high statistically significant positive correlation with coefficients $\rho = 0.58$ and $\rho = 0.22$, respectively (for both correlation tests $p < 2.2 * 10^{-16}$). As the Pearson correlation coefficient of the association between the ranges of transformed expression state scores and the untransformed gene expression value ranges was higher than that between the ranges of transformed expression state scores and the untransformed gene expression value means, the range in expression appears to be the dominant effect for reduction of confidence in expression state assignments as a result of the incorporation of the Rocke-Durbin error model.

Additionally, the availability of such an estimator as that provided by the Rocke-Durbin error model (Equation (4.6)) results in the ability to test the distributions of each gene's expression values against a 'constant expression reference distribution' representing the distribution of measurements that would be expected to arise from repeated measurement of the same value. If the observed distribution is sufficiently close to the expected distribution for constant expression (e.g. according to a Kolmogorov-Smirnov test (KS-test) against the reference distribution) the gene can be classified as invariant across the whole dataset.

The above discussion concerns a Rocke-Durbin error model with appropriate parameters. These parameters must be estimated from the data, and suggested methods for performance of this parameter estimation are described in [Rocke and Durbin, 2001]. The implementation used here revolves around the identification of groups of 'expected replicate samples' according to overall similarity of the samples' profiles across all genes. For each replicate group identified (with a very high degree of overall similarity), estimates for the model parameters α, σ_ϵ and σ_η are calculated and then combined in an average weighted by the relative numbers of samples in each corresponding replicate group. For each replicate group, sets of 'low-expression genes' and 'high-expression genes' are obtained (as suggested in [Rocke and Durbin, 2001]). The α parameter is estimated as the mean value of the low-expression genes, σ_ϵ is estimated as the square root of the average variance of each of the low-expression genes, and σ_η is estimated as the square root of the average variance of each of the high-expression genes.

It should be noted that, while the precise form and parameter estimation methods for the error model are specific to data from gene expression microarrays, the whole data transformation as presented in this section would be applicable to gene expression data from any source, provided an appropriate error model can be specified. This data transformation process, as described through Section (4.1.4) is called the Gene Expression State Transform (GESTr). There follows an illustration of the results of applying the GESTr to real gene expression datasets.

Discussion

As the GESTr method attempts to perform a task not previously reported in the literature, a general and straightforward comparitive evaluation is not obvious. Attempts to evaluate the method more fully in terms of a demonstration that expected patterns within the data are not disrupted and, critically, the fact that the existence of method enables improved biclustering performance on large-scale meta-anlaysis (for which the transformation was developed in the first place) are left to the following section of this chapter. As a first illustration of the success of the transformation in terms of modelling the observed variation of each gene's expression, a series of plots are presented in Fig. 4.20 to show that the transformation reliably adapts to structures within each gene's expression level distribution, despite there being considerable differences between these distribution patterns.

The notion of a biologically significant scale of expression level is presented here as a concept based on the observed pattern of distribution of each gene's expression levels across a data compendium that is assumed to provide a comprehensive (and ideally representative) sampling of all possible (at least physiological) biological contexts. Even when the dataset is insufficient to make such an assumption, or if the modelling assumptions were to turn out to be invalid from a biological perspective, the methods

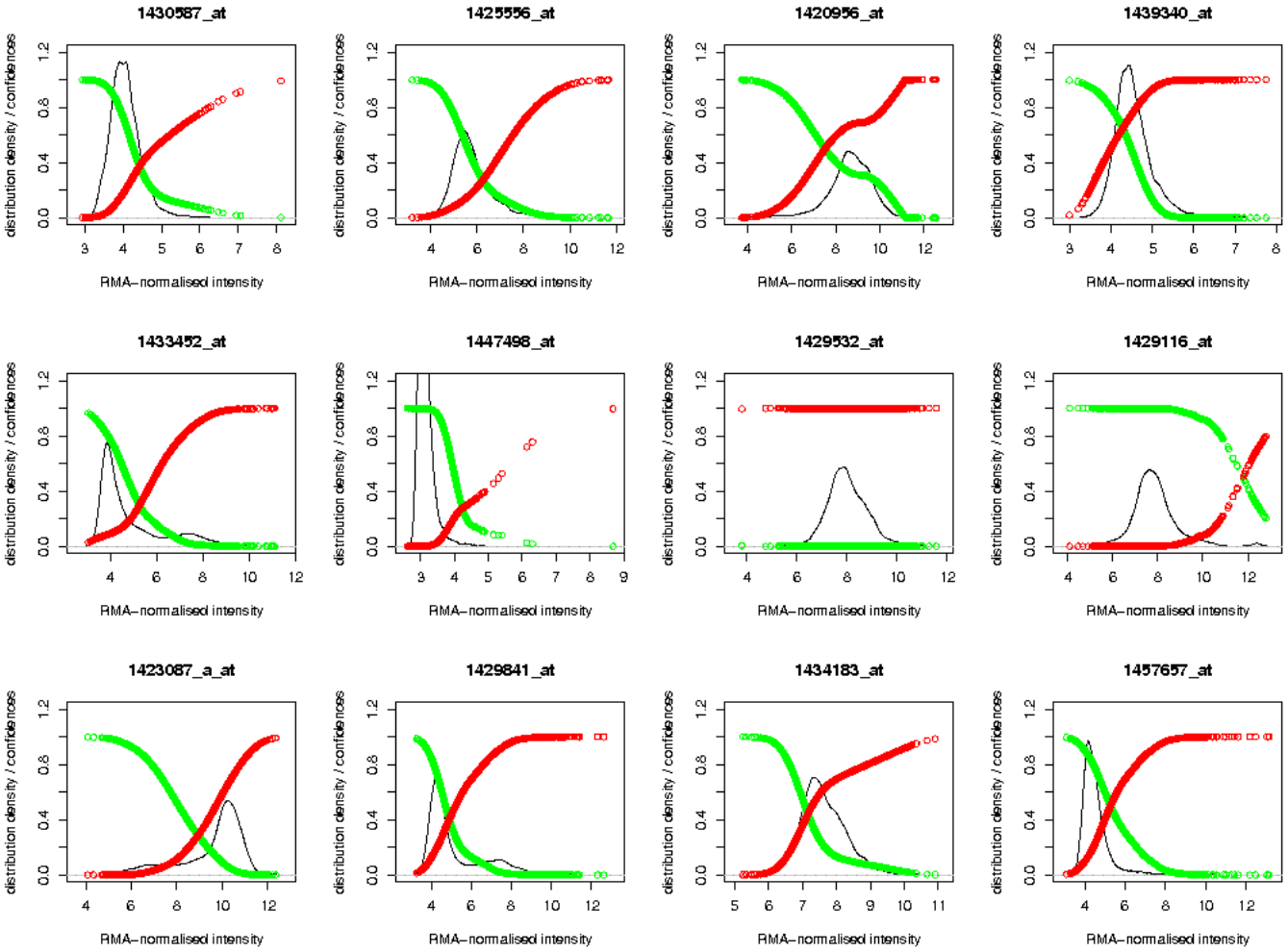


Figure 4.20: GESTr outputs for panel of randomly chosen genes

presented here still provide a means of transforming potentially arbitrary gene expression measurements into a unified scale that takes into account the precise patterns of variation of expression observed for each gene across all the available data. This provides a means of accurately estimating a confidence of 'relatively high' or 'relatively low' (etc.) levels of expression of a gene according to a single measurement, as compared to a full reference distribution. The utility of such a transformation is potentially widespread, but in the least case it provides a method for facilitating large-scale pattern mining across gene expression datasets, such as biclustering methods adapted from the MNC BGA presented in Chapter 3.

4.2 Evaluation of Expression-State Modelling

Following the above demonstration that the novel gene expression state modelling transformation (GESTr) presented above effectively produces a unified, continuous scale representing a biologically significant expression level, accounting for differences in the distributions of expression levels of different genes, a more objective and comprehensive evaluation of this GESTr would be helpful for demonstration of the validity of this approach and, by extension, any methods that utilise it. The GESTr introduced in Section (4.1.4) is the first reported approach for transformation of gene expression measurements into a unified scale for universal interpretation of the biological significance of any measurement. As no standard evaluation criteria exist for this task, novel evaluation methods had to be devised. Two such evaluations are proposed here: one based on demonstrating that expected features in the data are not distorted (i.e. the transformation doesn't introduce non-biological variation), and another based on demonstrating the improved performance of a large-scale pattern mining approach afforded by the incorporation of the transformation as a data preprocessing step.

4.2.1 Redundant Probeset Variation

On the Affymetrix MOE430v2 microarray platform, for which a large amount of mouse gene expression data is publicly available, a number of genes are represented on the array with multiple probesets. These sets of multiple, redundant probesets provide an opportunity to measure non-biological variation of gene expression measurements, under the assumption that in all samples measured, multiple probesets mapping to the same gene ought to result in a similar measurement of gene expression level. This in turn provides a means of evaluating the treatment of non-biological variation by the transformation, at least in comparison to the raw data measurements. If the transformation does indeed map gene expression measurements to a biological state of expression, and there is variation in the responsive range of each of a set of redundant probesets, it might be expected that the GESTr would reduce the overall differences within each set of redundant probesets. At least a demonstration that the GESTr doesn't increase the

within-replicate set variation over that observed in the raw (un-transformed) expression measurements would be important to illustrate that the transformation doesn't introduce non-biological variation in its transformation of each gene's expression measurements into a unified scale for assisting interpretation of the measurements from different genes.

An important caveat to consider in this evaluation is the fact that some probesets annotated as measuring the same gene may in fact detect different transcripts [Stalteri and Harrison, 2007]. As variation between the levels of expression measured for the same gene by different probesets might reflect the abundances of different transcripts, as opposed to measurement errors, the GESTr might accurately map gene expression measurements into a biological state of expression without reducing variation between 'replicate' probesets.

Sets of redundant probesets (defined here as multiple probesets annotated as mapping to the same gene) on the Affymetrix MOE430v2 microarray platform were identified. The list of redundant sets included 12,607 probesets mapping to 5,182 unique genes.

For each of these redundant sets of probesets, the pairwise Euclidean distances across the whole meta-analysis dataset described in Section (3.3) (for all possible pairs of probesets mapping to the same gene) for both the untransformed data and the dataset resulting from application of the GESTr to the data. As the transformed and untransformed values clearly vary on different scales, a reference set of measurements for each dataset was provided by measuring the pairwise Euclidean distances across the respective datasets of $10,000$ randomly-chosen probeset pairs (that do not map to the same gene), so that performance could be assessed (in terms of non-biological variation introduced) on the basis of the reduction in observed distances from the respective reference set of probeset pairs to that of the set of redundant probeset pairs. The distributions of such distances are shown in Fig. 4.21 for both the untransformed values and for the corresponding values after application of the GESTr.

As the plots are very similar, it is clear that, while the GESTr doesn't appear to increase the non-biological variation observed in the dataset, it doesn't *reduce* it from the untransformed measurements. It was critical to demonstrate that the GESTr doesn't introduce non-biological variation into the unified scale of gene expression values, and as such this objective is achieved. However, the observed results may disappoint somewhat on the grounds that the transformation might have been expected to reduce the non-biological variation observed in the dataset. A possible explanation for this observation is suggested by investigating the respective distributions of values for the untransformed and transformed datasets (across redundant probesets only), as illustrated by the histograms shown in Fig. 4.22. These histograms show that the majority

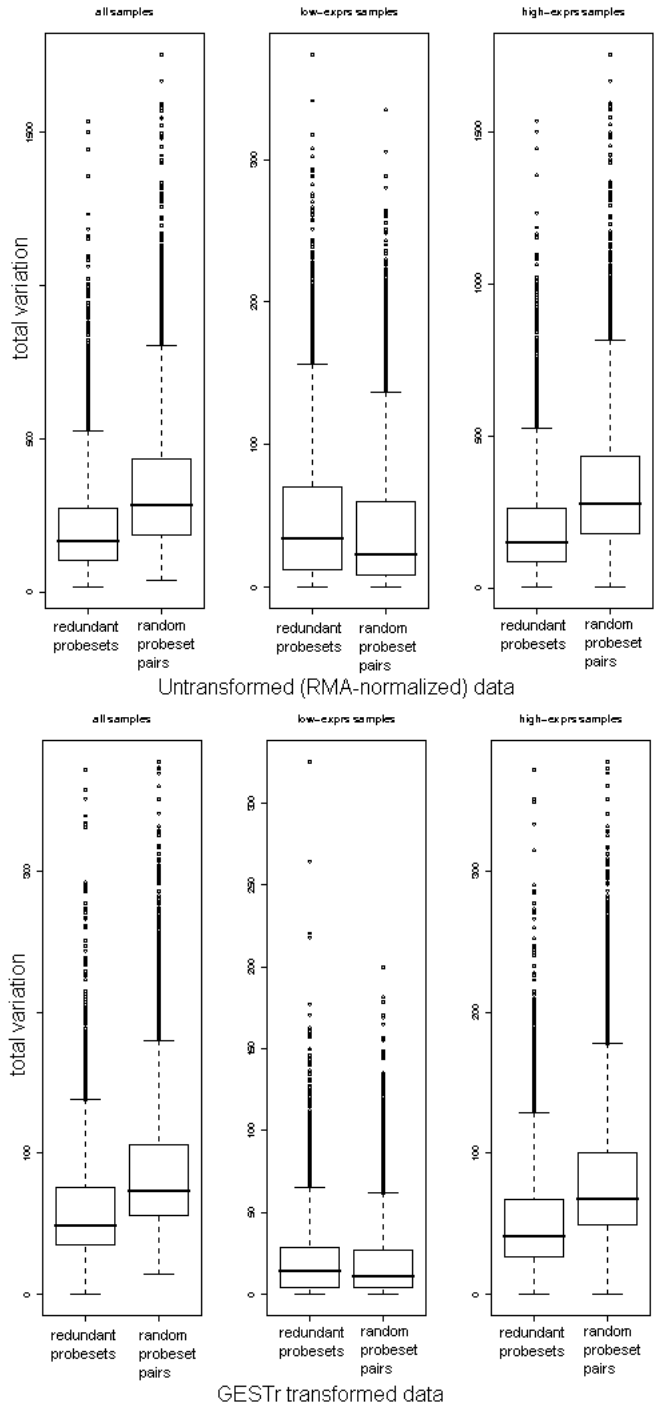
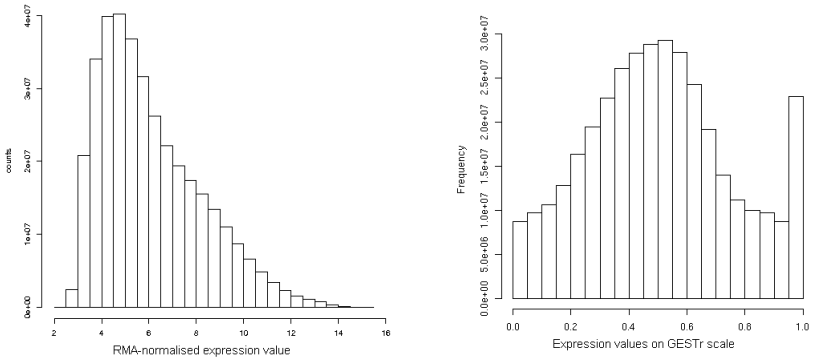


Figure 4.21: Distributions of Euclidean distances for within redundant probesets and reference probeset pairs for untransformed data (top 3 panels) and GESTr-transformed data (bottom 3 panels). For each row of panels, redundant probeset variation is measured across all samples in the dataset (leftmost panel), only those samples in the dataset with low values of expression for all the probesets (middle panel) and only those samples in the dataset with high expression values for any of the probesets (rightmost panel). Within each panel, distances between pairs of redundant probeset are shown in the left-hand box and distances between random probeset pairs are shown in the right-hand box. The greater the difference between the two distributions shown in any panel, the less variation is observed across probesets mapping to the same gene.

of the untransformed values lie towards the lower end of the range of values, whereas a greater proportion of the transformed values lie at the upper end of the range of values. This relative preponderance of values at the upper end of the numerical range for the transformed dataset would have the expected consequence that the differences between similar values would be greater for the transformed datset. To take this into account, a relative measure of distance between replicate measurements (those from redundant probesets from the same sample) was used to create a fairer basis for comparison. This scaled distance metric is given in Equation (4.15), and results of its application to the redundant probesets for each of the untransformed and transformed datasets are shown in Fig. 4.23.

$$dist(x,y) = \sqrt{\frac{(x-y)^2}{\frac{x+y}{2}}} \tag{4.15}$$



(a) Raw expression values

(b) Transformed expression-state values

Figure 4.22: Distributions of values across each of the raw and transformed datasets. A far greater proportion of untransformed expression values lie towards the lower end of the measured range than for the transformed values, thus the sum of the distances between pairs of values will tend to be a lesser proportion of the measured range for the untransformed values than it will be for the GESTr-transformed values.

While these results go some way towards demonstrating the biological relevance of the novel transformation at least in terms of a reduction of the non-biological variation observed across probesets with varying dynamic ranges, its motivation for existence was based on the idea of facilitating large-scale gene expression data mining. Therefore, another potentially useful evaluation of the novel GESTr would be a demonstration that it does indeed facilitate such large-scale pattern mining of gene expression datasets as intended. An evaluation of this desired property of the GESTr approach follows.

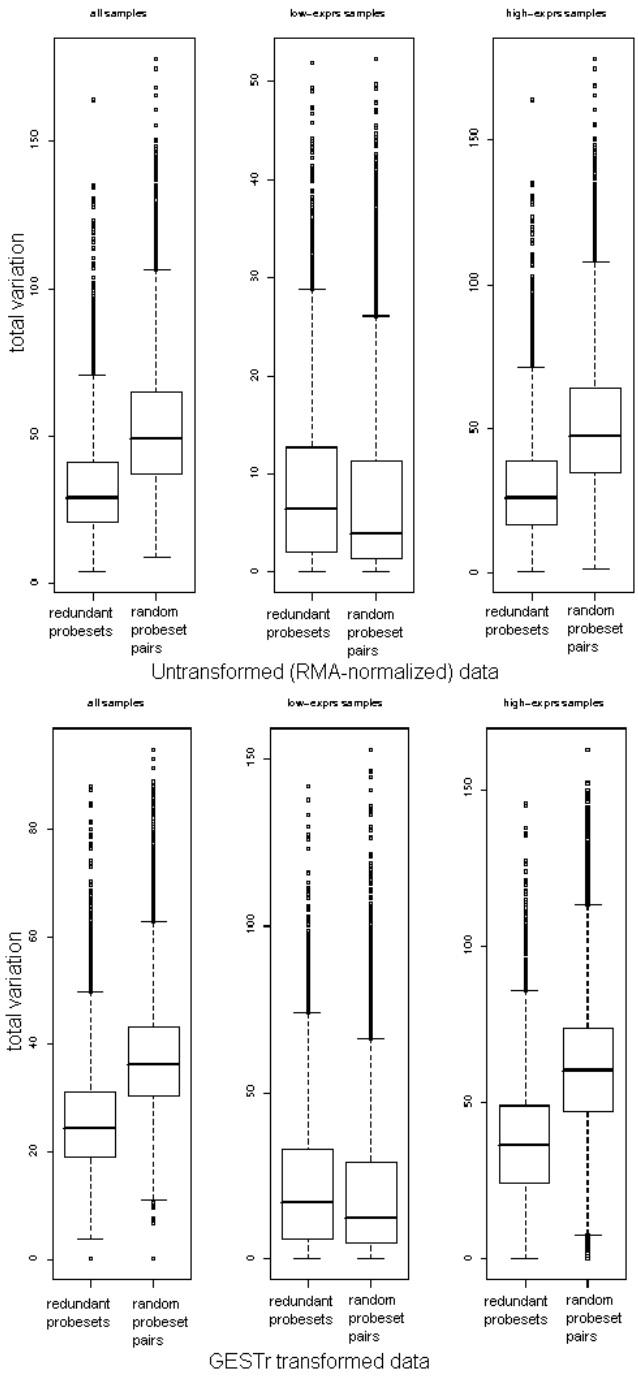


Figure 4.23: Distributions of scaled Euclidean distances for within redundant probesets and reference probeset pairs for untransformed data (top 3 panels) and GESTr-transformed data (bottom 3 panels). For each row of panels, redundant probeset variation is measured across all samples in the dataset (leftmost panel), only those samples in the dataset with low values of expression for all the probesets (middle panel) and only those samples in the dataset with high expression values for any of the probesets (rightmost panel). Within each panel, distances between pairs of redundant probeset are shown in the left-hand box and distances between random probeset pairs are shown in the right-hand box. The greater the difference between the two distributions shown in any panel, the less variation is observed across probesets mapping to the same gene.

4.2.2 Improving Biclustering Performance

As the primary purpose for developing a method of transformation of gene expression measurements to a unified, continuous scale representing relative transcriptional state for each gene was to facilitate large-scale pattern mining, it seemed that a demonstration of this ability afforded by the GESTr would be appropriate. The idea behind developing such a transformation for this purpose was to avoid situations where discretisation of gene expression levels, performed to facilitate comparability of expression variation of different genes, results in probabilistic pattern mining approaches being misled by small measurement variations (such as those due to noise) resulting in insignificant observations being classed as rare and significant due to the positioning of an absolute threshold. Therefore, it would be pertinent to provide a validation scenario that demonstrates the ability of the GESTr to overcome this problem, and that provides an indication of the corresponding improvement in results afforded by the adoption of this novel transformation as a preprocessing step applied before data analysis. Such an evaluation procedure was developed based on using 'semi-artificial' datasets: real datasets with carefully-chosen permutations applied to result in known implanted structures with desired properties.

By using permutations of real datasets, the overall distributions of values for each gene would precisely reflect those observed in real data, but particular bicluster structure could be imposed on chosen subsets of the data matrix. By enforcing certain genes to be as consistently high across all the chosen bicluster samples as their respective distributions of values allow, and ensuring that all other genes showed random variation across the whole dataset (including the bicluster samples), individual datasets could be constructed with known biclusters through which effective discovery rates could be evaluated. In addition, various bicluster properties could be enforced (if not completely controlled), and the effect of variation of each of these properties on the impact on implanted bicluster recovery rates of the application of different preprocessing methods could be investigated. In addition to varying the numbers of genes and samples in the biclusters, one property that was controlled was the average specificity of expression of bicluster genes in the bicluster samples: that is, for each gene chosen to be in the bicluster, the number of bicluster samples with high expression of that gene expressed as a proportion of all the samples in the dataset with high expression of that gene. This was determined by evaluating the difference in the average expression value across the bicluster samples to the average expression value across the non-bicluster samples. For each artificial dataset to be created, a permutation matrix was created by selecting a specified number of samples to be in the bicluster, then evaluating the bicluster specificity scores for each gene as described above (after re-ordering of samples so that the top-ranking expression levels corresponded to the bicluster samples) and selecting those genes with specificity scores closest matching the desired level specified. Those genes

not selected to be in the bicluster each had random permutations applied to remove the enforced structure where not appropriate.

With 'semi-artificial' datasets generated with desired properties (as described above), it was possible to use simple bicluster discovery algorithms to return scored (and ranked) genelists for the specified bicluster samples (known for each dataset), using datasets generated by applying permutation matrices to each of: the untransformed data, the dataset as discretised by the cluster-based discretisation method described in Section (4.1.2), and data transformed by the GESTr method. The simple bicluster discovery algorithms identify genes with consistently high expression across the specified samples (defined by specifying median value thresholds in the continuous data) and ranking those genes that pass the consistency filter according to specificity of expression to the bicluster samples (as defined above for the continuous data, and as the proportion of all the gene's 'high' values that are in the bicluster for the discrete data). Given the known sets of 'truth' genes that were implanted in the bicluster structure for each dataset, a Receiver Operating Characteristic (ROC) curve could be calculated for the ranked genelist discovered for each implanted bicluster in each type of dataset (preprocessed or not). A ROC measures the ability of a classification (ranking) system to reliably rank known positive cases above known negative cases. For a full discussion of ROC applications in evaluation of the performance of classification algorithms, see [Fawcett, 2006]. A useful property of the ROC is that the area under the ROC curve (the ROC-AUC) generated by plotting the number of false-positives (i.e. classification errors) against the number of classifications made (i.e. the length of the ranked list) evaluated for each possible length of ranked list as output from the classification algorithm, provides a numerical measure of success of the classification algorithm with a value of 1 representing a perfect classification (where all 'truth' inputs are ranked above all non-truth inputs) and a value of 0.5 representing a wholly uninformative classification (i.e. the expectation of the performance of a completely random ordering). ROC-AUC measures were computed for the ranked genelists produced by evaluating the implanted bicluster samples in each semi-artificial dataset, for each data type. A plot comparing these ROC-AUC scores for GESTr processed data to those for discretised data is given in Fig. 4.24, showing the dependency of the results from each data type on the number of samples in a bicluster.

The results shown in Fig. 4.24 clearly demonstrate that the GESTr method results in improved discovery of implanted biclusters when compared with a discretisation-based method, with this improvement becoming more pronounced as the number of samples in the biclusters decreases. The significance of this discrepancy becomes apparent when considering the distribution of numbers of samples forming bicluster-like groups in real

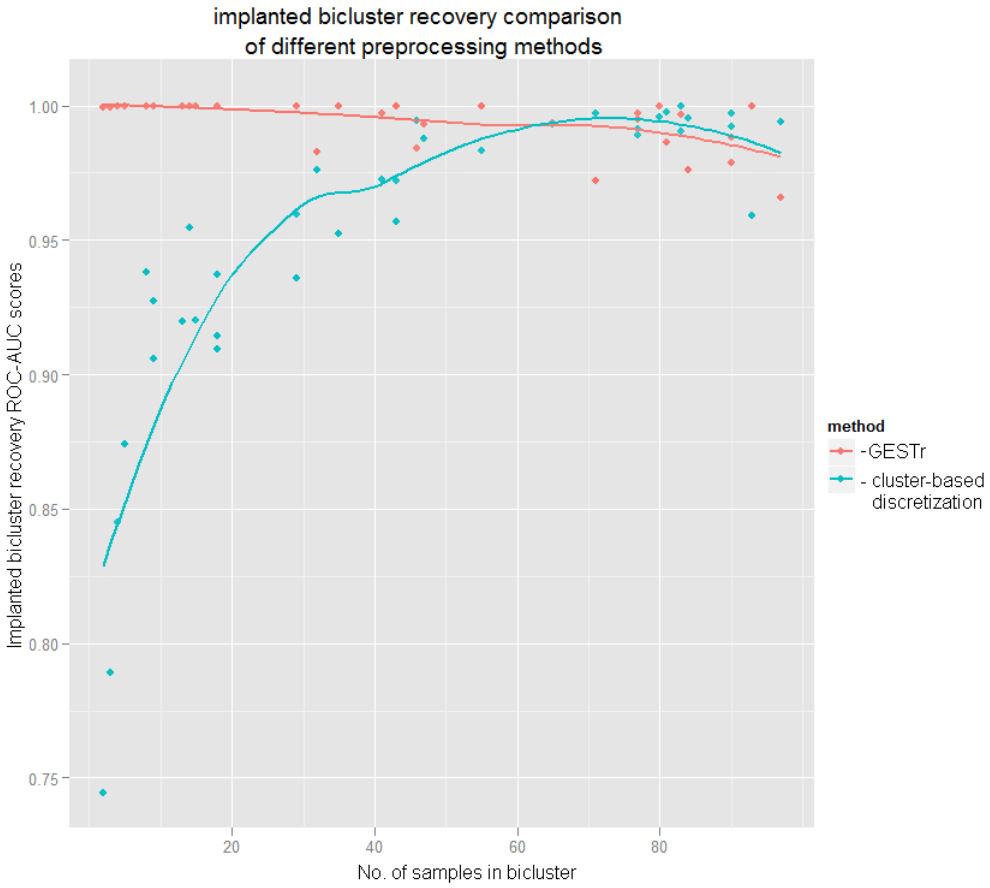


Figure 4.24: ROC-AUC recovery scores from semi-artificial datasets with implanted biclusters of varying size. Note contrast between scores for GESTr processed data (shown in orange) compared to scores from data processed with a cluster-based discretisation approach (shown in blue)

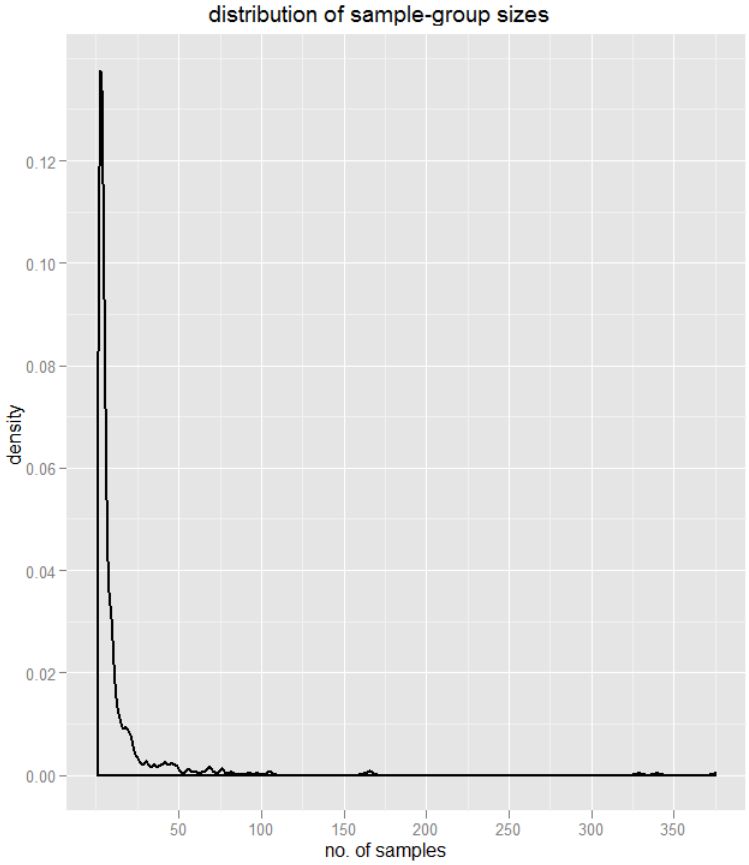


Figure 4.25: Density plot showing distribution of numbers of samples in bicluster across a set of groups of samples from a large collection of gene expression data, where each group comprises a significantly greater number of samples within a specified dissimilarity threshold than would be expected by chance (for a range of such thresholds). These sample groups approximate 'natural' biclusters within the dataset.

data[5], as illustrated in Fig. 4.25. This shows that the vast majority of biclusters likely to occur in a dataset will involve sample numbers where the GESTr method results in the greatest improvement over discretisation.

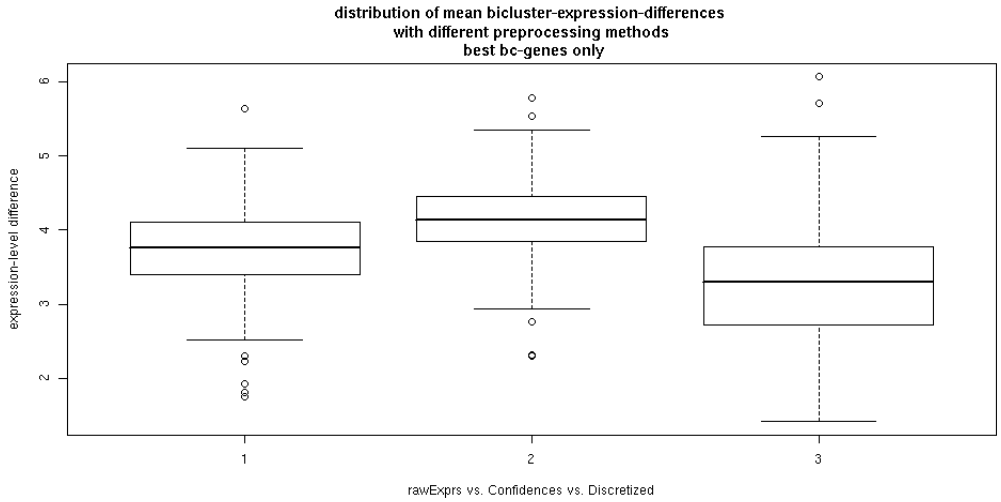
As an additional means of evaluating the effect of incorporating the GESTr as a preprocessing step to large-scale gene expression data mining, desirable bicluster properties were measured and compared for biclusters across the same sets of samples evaluated on different versions of the same dataset (i.e. with different preprocessing methods, or none at all, applied). For this evaluation, the same simple bicluster discovery algorithms mentioned above were used to evaluated genes corresponding to bicluster expression patterns across groups of highly similar samples. Highly-similar sample groups were used to represent cell-type- or tissue-specific patterns in the data. Details for the con-

[5]These bicluster-like sample groups were identified using the similar-sample modelling heuristic approach taken to identify potential biclusters, described in detail in Section (5.1.4)

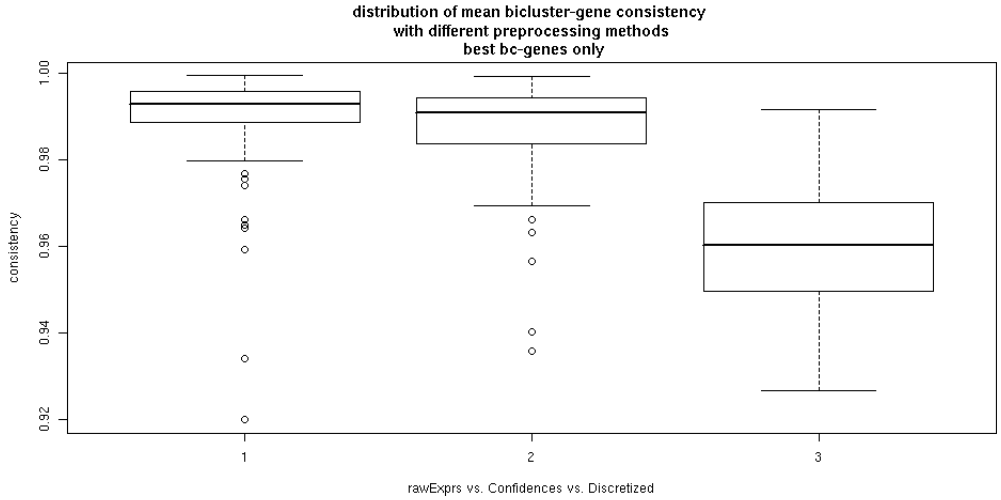
struction of these sample groups are given in the following chapter (Section 5.1.4). It is sufficient for the purposes of the evaluation presented here to know that the sample groups each consist of a set of highly similar samples (according to overall Euclidean distance). The same set of sample groups was applied for bicluster evaluation with each data type to allow fair comparison of the discovered bicluster properties. This approach based on specified sample groups avoids confounding the comparison due to differences in the effect of the data transformation on the precise mechanism of any biclustering algorithm's implementation. For each type of input dataset, distributions across the biclusters of the mean difference in level of expression of each bicluster gene between bicluster and non-bicluster samples and the mean consistency of high expression of bicluster genes across bicluster samples are shown in Fig. 4.26.

The plot shown in Fig. 4.26 (a) indicates that processing data with the GESTr method results in prioritization of bicluster genes with a clearer pattern of expression level difference between the bicluster samples and non-bicluster samples (which is related to the specificity of the genes to that bicluster) than compared to data processed with the cluster-based discretisation method described in Section (3.4.4), and by this metric the GESTr method appears to improve biclustering performance over application of the same gene prioritization approach to raw expression data. Additionally, Fig. 4.26 (b) shows that application of the GESTr method results in prioritization of genes with better consistency of expression level across the respective bicluster than when compared with results on data processed using the cluster-based discretisation approach, and results in consistency close to that of untransformed expression data.

The results of the evaluations presented above demonstrate that the GESTr method of transforming gene expression measurements from a large, heterogeneous collection of data facilitates complex large-scale pattern mining (such as with biclustering methods) in a superior manner to a successful discretisation-based transformation method. As this was the principal motivation behind the development of the GESTr method, these results indicate that it was successful at performing its desired task, and thus the development of this novel data transformation method has been justified at least on account of providing the opportunity for improved large-scale pattern-mining in gene expression data. While the development of the GESTr method enabled the continuation of the study of large-scale meta-analysis of gene expression data through biclustering and ultimately led to the development of effective meta-analysis approaches based on biclustering (as shown in the following chapter), the benefits of applying this novel gene expression data transformation method may extend beyond this pattern-mining context.



(a) Distribution of magnitude of expression differences for top bicluster genes between bicluster samples and non-bicluster samples



(b) Distribution of consistency of expression level (as a proportion of maximum) of top bicluster genes across bicluster samples

Figure 4.26: Comparison of properties of biclusters (gene prioritisations for specified groups of samples) identified in data processed by no method (1), GESTr (2) and cluster-based discretisation (3).

4.3 Interpretation of Microarray Datasets

As the GESTr presents a means of inferring a biologically-relevant state of transcription of a gene in a sample on the basis of any single gene expression measurement (provided a comprehensive reference dataset of comparable measurements is available), this may have a range of potential uses outside application to large-scale pattern mining such as the biclustering case given above. One particular such use is related to the idea presented in [Kim et al., 2010b], where estimates of gene expression variation across all available data from a given microarray platform are used to 'regularise' t-tests for differential expression across small datasets for which estimates of sample variance (required for the t-test) are liable to be inaccurate. As an extension of this idea, if it is possible to map expression values from a novel dataset to equivalent values in the reference dataset that have been supplied as input to the GESTr, it would also be possible to infer biological significance of the observed expression measurements from the novel dataset.

As the novel datasets intended for this reference-mapped GESTr interpretation will typically involve relatively few samples from a restricted biological domain, it may well be that typical differential-expression analyses across this restricted domain miss biological information regarding each gene's expression range and distribution pattern across a broader spectrum of biological contexts. This information would be provided by mapping the novel dataset's measurements to the GESTr scale, on the basis of the models fitted to the reference dataset.

In order to utilise the gene expression models provided by application of the GESTr to a reference dataset for interpretation of a novel dataset, first the novel dataset's values must be brought into the same scale as those of the reference dataset. Due to this critical step, it is highly recommended that this mapping only be performed using a reference dataset from the same gene expression measurement platform as that used for the novel dataset, unless normalization methods capable of providing universally comparable values exist for all platforms involved. In the normalization of the large gene expression dataset used for the redundant probeset evaluation of the GESTr presented in the previous section, normalization parameter sets for the RMA algorithm [Irizarry et al., 2003] were generated (through use of the `RefPlus` package [Harbron et al., 2007] in R that implements the RefRMA method [Katz et al., 2006] of normalization of very large microrarray datasets). As described in Section (3.3), these normalization parameter sets could be used to apply an identical normalization process to a novel dataset obtained from the same platform (Affymetrix MOE430v2 microarray), thereby mapping the novel dataset's measurements onto the same scale as that used to apply the GESTr. Following this mapping, the expression values from the novel dataset can be used to obtain appropriate transformed values for the novel dataset's

measurements through interpolation of the GESTr models (presumed to be available from the relationship of the reference dataset's GESTr-transformed values with the corresponding untransformed values). An example of this process is illustrated in Fig. 4.27, which shows the distribution of GESTr-transformed values across the reference dataset's expression distribution for one gene (Pou5f1) and points corresponding to five measurements from a novel dataset[6] not included in the reference dataset, interpolated onto the GESTr output value scale with the vertical lines shown linking the expression values on the x-axis to the GESTr scores on the y-axis.

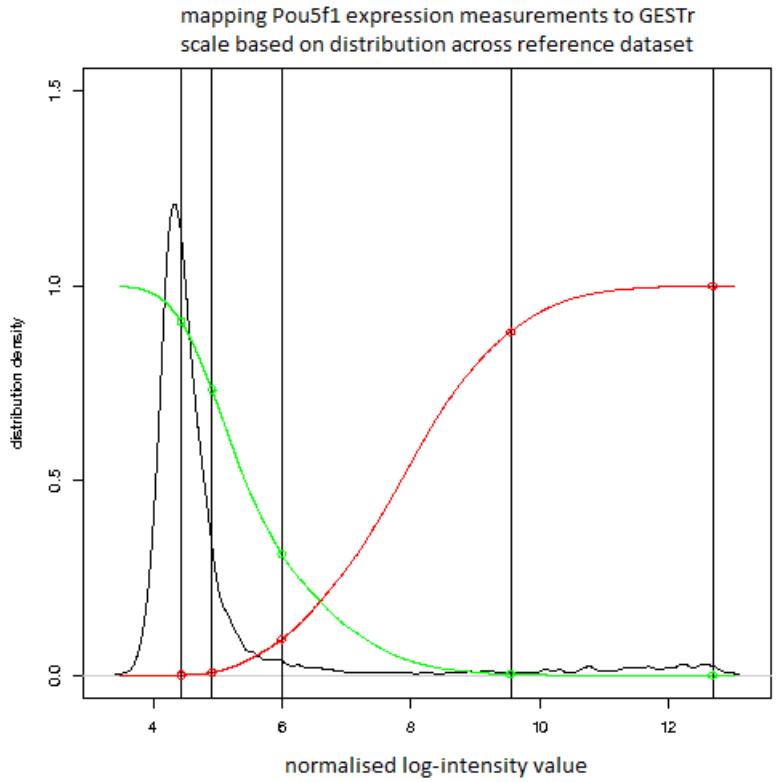


Figure 4.27: Interpolation of GESTr scores from mapped expression measurements from a novel dataset. Black vertical lines indicate expression values of Pou5f1 measured in a time series experiment described in [Hall et al., 2009], intersecting with GESTr scores for confidence of low expression state (green curve) and confidence of high expression state (red curve) as derived from the distribution of Pou5f1 expression across a large reference compendium of gene expression data (black curve).

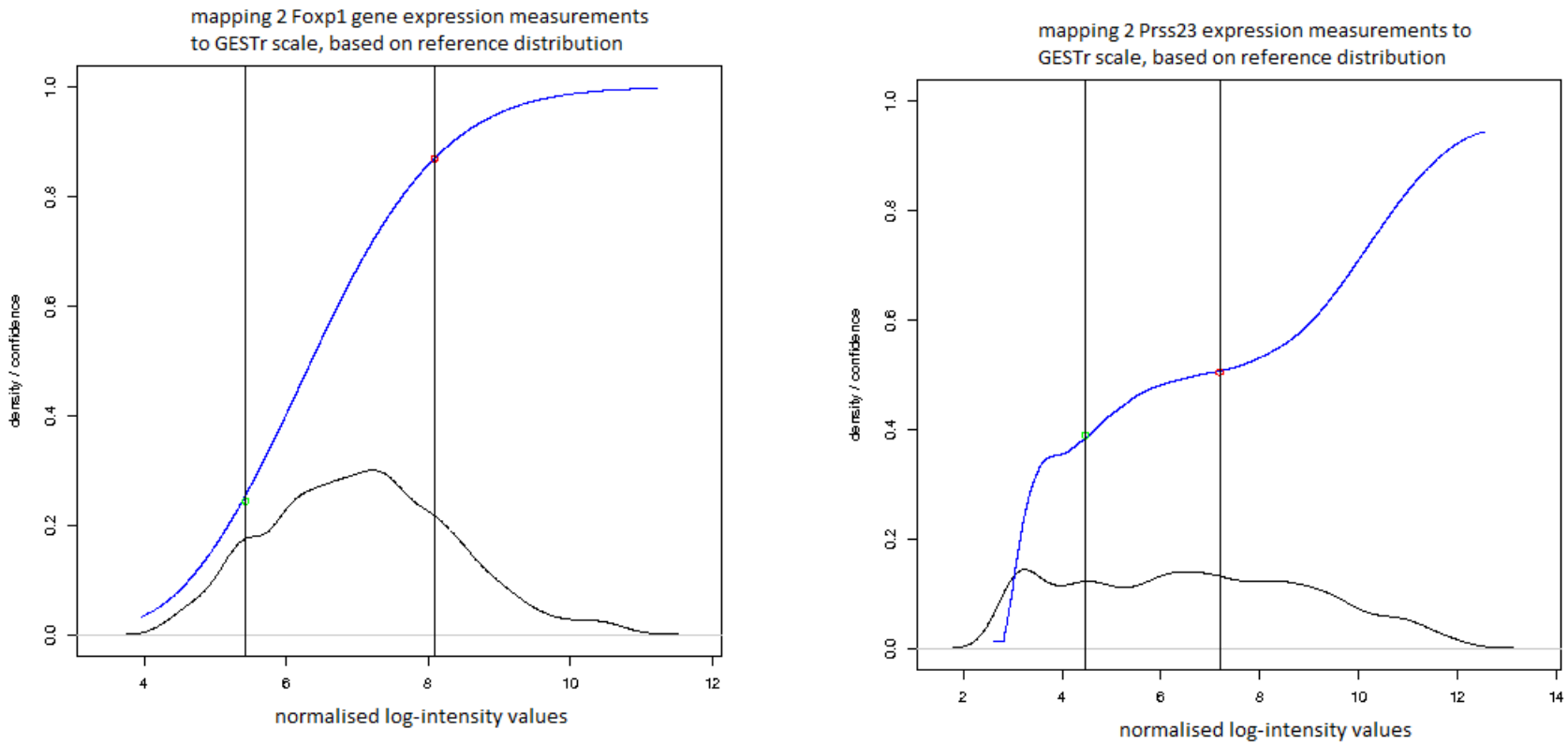
As an illustration of a way in which this information provided by the interpretation of a novel dataset in the context of a broader reference of expression patterns (through the application of the reference-mapped GESTr as described above) could be useful, an example is given from the novel dataset used in the previous illustration (the Pou5f1 knock-down time-series from [Hall et al., 2009]) obtained from Affymetrix MOE430v2

[6] an experiment profiling gene expression during a time series knock-down of Pou5f1 expression in mouse ES cells, described in [Hall et al., 2009]

microarrays, in which two genes were identified with similar fold-changes and differential expression levels (and statistical scores of differential expression) across two groups of samples of interest in this novel dataset. The respective low and high expression values for each gene are shown in Fig. 4.28 in the context of the respective gene expression distributions across the reference dataset. It is clear from this illustration that the differential expression observed in the novel dataset for the gene shown on the left (Foxp1) spans a greater proportional range of that gene's observed expression levels, whereas the gene shown on the right (Prss23) varies across the novel dataset a smaller proportion of its overall expression range. The information gained from mapping the novel observations to the reference dataset and applying the GESTr shows that the expression distribution pattern for one of the genes Foxp1) implies that the observed levels in the novel dataset represent significant differences in terms of the range of expression levels of that gene across a broad range of biological contexts, whereas the other gene appears to show only variation within one component of its general distribution of expression levels and so this observed variation is less likely to be as significant from a biological point of view. Interestingly, Foxp1 was identified as a novel candidate pluripotency gene in a recent screen performed in mouse ES cells [Ding et al., 2009] and was shown to be a DNA-binding of Oct4 in [Loh et al., 2006].

This demonstration is not intended as proof that the suggested application of the GESTr is valid, rather to illustrate the motivation for this application and to show that it offers an interesting, novel means of interpreting individual gene expression datasets in relation to observed patterns of gene expression distribution across a broader biological context. The above illustration serves to show that there is potential merit in this interpretation through supplying additional information that can be used to distinguish biological differences between sets of observed expression measurements that would otherwise be unavailable.

As previously mentioned, due to the fact that this novel approach to interpretation of gene expression datasets appears to be the first of its kind reported (or even suggested), a means of critical evaluation is not obvious. Some means toward an evaluation of the GESTr-based approach to interpretation of gene expression data can be provided through demonstration of a successful practical application of the method. In utilising this transformation of a novel dataset to a unified scale of biologically significant expression levels, novel approaches to the identification of differentially-expressed genes can be developed to identify genes with expression differences across sample groups in a novel dataset that span a significant range of the gene's expression distribution across the whole reference dataset.



(a) Gene1 (Foxp1)

(b) Gene2 (Prss23)

Figure 4.28: Mapped, GESTr-transformed expression scores for measurements for two genes similarly differentially expressed in a novel dataset. Black vertical lines indicate expression values of the gene in question (left Foxp1, right Prss23) measured in a time series experiment described in [Hall et al., 2009], intersecting with GESTr scores for expression state (blue curve) as derived from the distribution of the gene's expression across a large reference compendium of gene expression data (black curve).

4.3.1 Identification of Differentially-Expressed Genes

As mentioned in Sections (3.4.2 & 4.1.1), typical transcriptome-profiling experiments involving genome scale gene expression measurement technologies (such as microarrays) tend to involve the identification of differentially expressed genes (DEGs) between groups of (replicate) samples. While a commonly used pragmatic approach of identifying genes with the greatest relative (fold-) change in measured expression level between the groups is often successful [Pearson, 2008], a large number of methods exist for the identification of genes with statistically significant differential expression between groups of interest among the experimental samples (e.g. LIMMA [Smyth, 2004], SAM [Tusher et al., 2001], GEOMixBayes [Kim et al., 2010b], etc.). However, by applying the GESTr models (calculated from a comprehensive reference dataset) to a novel dataset, it may be possible to attempt to identify genes with biologically significant differential expression between the sample groups of interest.

An immediate (and simple) approach would be to identify genes with the largest change (either absolute or in a particular direction) between the GESTr-transformed values of the sample groups of interest. This approach is similar to extending the idea of fold-change, where one of the sample groups is implicitly used to establish a reference level of expression of each gene, to utilising the distribution of expression across a broad range of biological contexts to establish a reference pattern of expression of each gene.

It was proposed that a method be developed to assess the measure of biologically significant differential expression provided by the GESTr-transformed data on a statistical basis. Statistical measures of differential expression take into account, as well as the magnitude of the observed change between the sample groups, the variation of the gene's measurements within each sample group. Both the level of change and the consistency are important, particularly as the analysis tends to be performed on very small sets of replicates [Kim et al., 2010b]. While a large number of methods exist for estimating the statistical significance of differential expression of genes, attempts to use these on data after GESTr-transformation proved unsuccessful due to the preponderance of zero-values for fold-change reference and estimates of zero variance, which result in t-tests (and any variants of this approach, which make up the majority of statistical tests for differential expression [Dudoit et al., 2002]) evaluate to infinity.

To avoid these problems, an approach based on that taken by SAM [Tusher et al., 2001] was taken. In SAM, 'balanced' permutations of the dataset are created so that an even distribution of samples from each experimental group is maintained in each permuted group, and the 'expected' variation of each gene is calculated based on the values obtained from analysing these balanced permutation groups (where no differential expression ought to be significant). Based on the observed differences between each set of balanced permutations, an estimate for the proportion (and absolute number)

of false-positives can be produced for any setting of an otherwise arbitrary differential expression threshold. An implementation of this SAM approach was created for application to GESTr-transformed data, here referred to as 'TranSAM' (transformed significance analysis of microarrays). Balanced permutation groups are generated in this implementation using the `combinations` function from the R package `gtools` to find all combinations of $\lfloor(\frac{n}{2})$ samples from each group (where n is the size of the smaller group) and these combinations are combined to form all unique sets of partitionings of the experimental groups' samples into two groups with an even number of samples from each experimental group in each of the resulting partition groups.

As a further demonstration of the utility afforded by the GESTr in conjunction with the TranSAM method for identification of genes with biologically significant differential expression, a comparison is shown in Fig. 4.29 of the numbers of Pou5f1 DNA-binding targets (as used in the enrichment analysis above) found in top-ranking genelists of various size obtained from each of a standard SAM analysis on the RMA-normalized data from the Pou5f1 knock-down time-series dataset used in the analysis presented earlier in this section and an analysis using TranSAM on the GESTr-transformed values from the same dataset (obtained as described above). This plot (similar to a ROC) shows that the TranSAM approach involving GESTr-transformed data is better (in this example) at predicting DNA-binding from analysis of expression in a single dataset. Again, while this is by no means a conclusive evaluation of the novel data transformation approach presented in this chapter[7], it suggests both a utility over and above that which the GESTr transformation was designed to provide, and that these novel approaches to interpretation and analysis of gene expression datasets may prove useful for future biological research.

4.4 Discussion

For pattern mining in large-scale gene expression datasets, where numerical measurement values for different genes may not be directly comparable in a biological context (as discussed in Section (3.3.2)), it simplifies the complexity of the pattern mining task to have a representation of the data values in some unified scale. With a unified scale of values, a particular defined pattern of interest can immediately be assessed in any submatrix of the whole dataset, regardless of which genes' measurements are involved. When the pattern mining task involved is sufficiently complex, as is the case with biclustering, the simplicity of assessment of a particular pattern in any given submatrix of the dataset enables construction of more efficient algorithms for discovery of such patterns that, in turn, enable application of computationally complex analysis tasks to especially large collections of data. Owing to this, a number of biclustering algorithms

[7]Further evaluation was left for future work, on the grounds that work on large-scale meta-analysis of gene expression data could progress as a result of the development of the GESTr transformation.

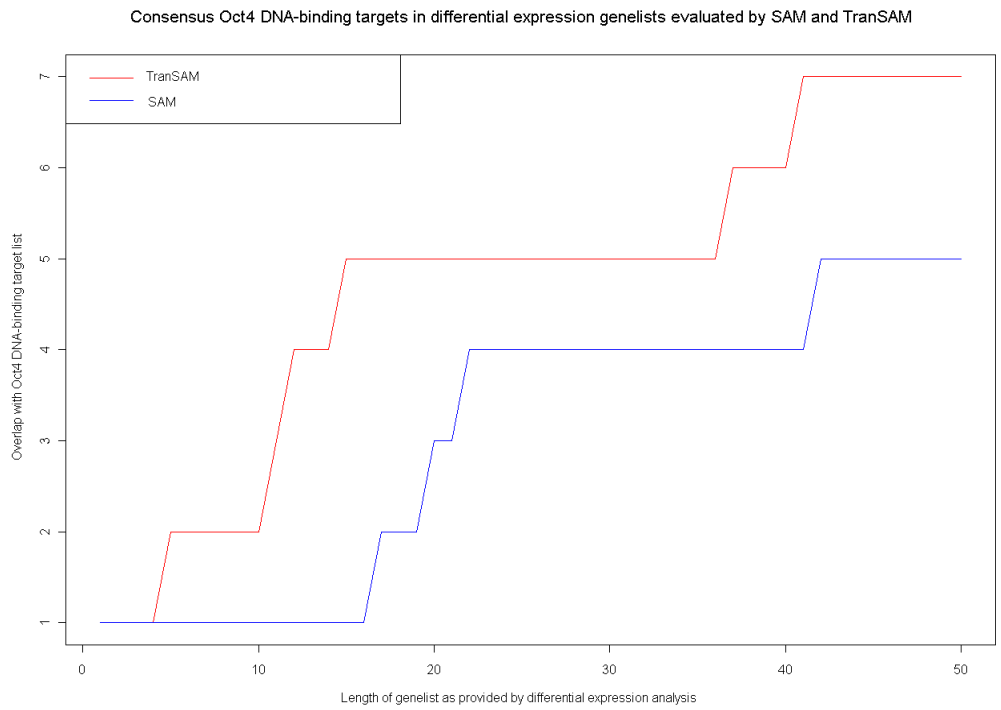


Figure 4.29: Comparison of prediction rates of DNA-binding from lists of genes with significant differential expression in a single Oct4 knock-down dataset as identified by SAM (blue) and the novel TranSAM method (red).

incorporate a data preprocessing step to bring all measurements into a comparable scale.

As common approaches to this preprocessing revolve around the discretisation of the data from measured values to symbols representing some state of expression (e.g. 'high','low' or 'insignificant' in relation to background expression for the gene in question), and a novel discretisation approach had been developed for the implementation of biclustering algorithms presented in the previous chapter, a comparison of a number of alternative approaches to discretisation of large collections of gene expression data was performed to provide an evaluation of the relative merits of different approaches. Examples given in Section (4.1.2) illustrate the advantages of the novel cluster-based discretisation method that was used for the biclustering analysis of the previous chapter and described in Sections (3.4.2 & 4.1.2). However, owing to trends discovered in the output of biclustering performed using discretised data, where genes were classified as significant in a bicluster due to small measurement variations either side of a discretisation threshold, it was considered appropriate to evaluate the effect of adopting a continuous (rather than discrete) unified scale of expression measurement.

A novel data transformation, the Gene Expression-State modelling Transformation (GESTr) was developed, based on inference of biological states of expression from mod-

els of the distribution of each gene's expression values across a large reference dataset (assumed to represent a comprehensive range of biological contexts for the organism in question). This GESTr method is described in Section (4.1.4) and demonstrations of the results produced by the method are provided towards the end of that section. The supposed biological significance of this method was evaluated using redundant sets of microarray probesets as a means of measuring non-biological variation across a large dataset. The results given in Section (4.2.1) show that the transformation provides a unified scale for interpretation of gene expression measurements without introducing non-biological variation.

As the data transformation methods described in this chapter were developed with the aim of facilitating large-scale gene expression data mining through biclustering, the effects of applying a number of the methods described in this chapter (including both the novel cluster-based discretisation algorithm and the GESTr method) were evaluated in terms of the results of applying equivalent bicluster evaluation algorithms to recover known gene expression patterns in 'semi-artificial' datasets and to discover biologically significant groupings of (cell-type- and tissue-specific) genes. The results of these evaluations, shown in Section (4.2.2), demonstrate that the novel GESTr method improves the performance of biclustering, allowing the identification of transcriptional relationships with greater biological significance.

Additionally, the potential utility of the novel GESTr method outside the context of large-scale gene expression data mining was demonstrated through its application to provide a novel means of interpreting gene expression datasets. A novel approach to interpretation of individual gene expression datasets was developed, based on using the GESTr (as performed on a large reference collection of gene expression data) to relate measurement observations from a novel dataset to the general gene expression patterns observed across a broad range of biological contexts. An illustration of the conceptual value of the information gained through this interpretation of data in terms of a broader biological reference, and in particular, in terms of the modelled biological significance of each gene's measured expression levels as calculated in the GESTr method, is given in Section (4.3). A further utility of this interpretation to discover genes with biologically significant differential expression between sample groups of interest in individual gene expression datasets was revealed through the development of tools (e.g. 'TranSAM' described in Section (4.3.1)) to identify such genes, given a novel dataset that has been mapped onto the GESTr output scale. An example where such a method has been used to predict DNA-binding of a TF from a knock-down time-series microarray profiling experiment to greater effect than a standard statistical method for detecting differentially-expressed genes applied to the same dataset is given in Section (4.3.1). Further applications of the output of GESTr transformation of gene expression compendia are currently being explored in a number of ongoing research projects.

In summary, novel data transformations have been developed through the course of this work, to provide a unified scale representing biologically significant levels of expression from gene expression measurements where the significance of a given measurement may vary from gene to gene. The GESTr method of performing such a transformation provides a novel means of interpreting gene expression measurements in a wider biological context, in addition to facilitating improved large-scale pattern mining of gene expression data. This method has potential implications for providing improved means of interpreting gene expression datasets which, it is hoped, will be revealed through successful application to future problems in biological research. However, the fact that the GESTr method enabled improved bicluster discovery in large collections of gene expression data resulted in the possibility to develop and apply novel meta-analysis methods that could answer particular biological questions regarding transcriptional relationships involving genes of interest in specific biological contexts. The development of such a meta-analysis method is described in the next chapter.

Chapter 5

A Probabilistic Approach To Localised Co-Dependency Analysis

With the availability of the GESTr method, capable of transforming whole-genome scale measurements of transcription across large collections of biological samples, comes a simplification of gene expression data mining tasks as discussed (specifically for biclustering) in Section (4.1.1). Although the biclustering algorithms presented in Chapter 3 demonstrated an ability to find biologically significant localised expression patterns in very large datasets (with the GDepBGA of Section (3.6.3) being particularly successful) there still seemed to be some limitations of these approaches in terms of the desired output for application to real problems in biological research.

As it was noted in Section (4.1.2), biclusters from the BGA had a tendency to include genes that did not appear to have the significant gene expression patterns reflected in the non-discretised data that would have been expected due to their inclusion in the bicluster. Owing to the fact that it was established that this observed problem appeared to be due to discretisation thresholds positioned in-between similar measurement values, and inspection of gene expression distributions suggest that this problem would remain for any discretisation-based approach, a novel transformation method was developed (as described in Section (4.1.4)) to provide a mapping from measurements to a unified, continuous scale based on patterns in each gene's distribution of expression measurements across a large number of samples spanning a broad range of biological contexts.

The biclustering algorithms presented in Chapter 3 had to be adapted to utilise input data mapped to this continuous, unified scale. This chapter describes these necessary modifications, along with subsequent alterations and improvements that were involved in adapting the biclustering framework into a meta-analysis approach suitable for application to real questions in biological research. This meta-analysis approach is termed HBLCA, Heuristic Biclustering for Localised Co-Dependency Analysis. One crucial stage in development of the HBLCA approach was the formulation of a probabilistic framework for gene expression co-dependency analysis, which is described in Section (5.1.2)

Evaluations of the success of the HBLCA approach are presented in Section (5.2), comparing performance on meta-analysis tasks with that of alternative approaches to large-scale meta-analysis of gene expression data. A description of the way in which this novel approach can be used in conjunction with data from genome-wide DNA-binding studies to provide further insight into transcriptional mechanisms is given in Section (5.3.1).

5.1 Description of Co-Dependency Meta-Analysis Approach

To take advantage of the novel GESTr transformation (described in Section (4.1.4)) of gene expression measurements into a unified, continuous scale representing the likely biological significance of the measured level of transcription for each gene, the biclustering methods developed for large-scale meta-analysis of discretised gene expression data would have to be altered. In the very least, the probabilistic framework for biclustering upon which the BGA presented in Section (3.6.3) was based would have to be modified to take into account distributions of continuous (as opposed to binary) values. This section provides details of the necessary modifications to the GDepBGA, including the adoption of a probabilistic framework for localised gene expression co-dependency analysis, followed by the description of a heuristic employed to bypass problems affecting the resulting biclustering GA, giving rise to the HBLCA method that provides practically useful results from the application of biclustering to large-scale meta-analysis of gene expression data.

5.1.1 Biclustering with Expression-State Confidences

The notion of probabilistic assessment of bicluster desirability through modelling the entropy of a particular bicluster pattern in a given dataset was presented in Section (3.6.2), and results of application suggest that the adoption of this probabilistic approach resulted in a more successful biclustering algorithm than an equivalent algorithm using a naive model of bicluster desirability. Given that the GESTr output is intended to represent biological significance of a measured expression level in terms of a confidence of the respective gene being in a high or low state of expression, this measure of confidence of expression across the bicluster can be used to introduce a quantitative element to the scoring of consistency across a bicluster.

Using the GESTr output, the desirability of a bicluster can be assessed in terms of the confidence of consistently high/low expression of each gene across the bicluster and the unlikelihood of the observation of consistent expression of that gene across a randomly-chosen set of samples of the same size as the bicluster. The GESTr output values provide a straightforward way to estimate the confidence of high/low expression of any given gene in any sample in the dataset, so all that need be considered regarding the confidence of consistent expression of a gene across a bicluster is the method employed to summarise those values across the bicluster. This could be the mean, median or some other function of the individual values. However, modelling the probability of a gene's consistency of expression to a given (summary of) confidence across a given number of randomly-chosen samples is less straightforward and is in fact dependent on the summary function used.

Fundamentally, in order to model this probability of consistency of a gene's expression across any number of random samples, it would be ideal to obtain an expression (or at least an estimate) for the cumulative distribution function for every number of samples that could be combined. That is, if there are values x_i sampled from biclusters with numbers of samples ranging from 2 to n, and the CDF of the distribution of values across all the samples in the dataset is $Y = P(X < x)$, then the probability of that gene's expression pattern in any observed bicluster could be estimated if $Y^2, Y^3, ..., Y^n$ were all available. A number of approaches to modelling this probability of observed bicluster-expression for any gene were investigated for applicabilty within a biclustering algorithm.

Resampling

To model the probabilities of observing any particular set of bicluster values for each gene, it would be possible to use a resampling-based approach to generate estimates for the distribution of the given summary statistic over large numbers of random samplings from that gene's values in the dataset with each possible bicluster-size as the number of values taken in each sampling. As a result, sample distributions of the probabilities of a range of values for each gene, for each bicluster size could be estimated. Given a large enough sample-size (i.e. large biclusters), it can be assumed from the central limit theorem (CLT) that the distribution of the means of any set of values will approximately follow a normal distribution [Tijms, 2004]. Therefore, if the summary statistic used is the mean of the gene's observed bicluster values, the distributions of observation probabilities of each gene for each size of bicluster could therefore be represented with a value for each of the mean and standard deviation of the characterised (normal) distribution of summarised random-bicluster scores.

This approach provides a means for fast evaluation of individual bicluster scores through evaluation of normal distributions, but requires a very large number of simulation runs as a preprocessing step to estimate the distribution parameters for each gene's summarised GESTr values over differently-sized subsets of the data. This preprocessing step may be prohibitively time-consuming: for example, to evaluate these distributions over all possible bicluster sizes up to, say, half of a $10,000$-sample dataset involving $10,000$ genes with $10,000$ resampling samples to estimate each value would involve 10^{12} generations of random samplings and computation of means. In addition, the validity of the CLT-driven assumption of normality of distribution of means may not be appropriate for estimating the probabilities of observations when the number of bicluster samples is small. Owing to these possible limitations, alternative approaches to the estimation of the desired probabilities were sought.

Fast Hypergeometric Approach

Another possible approach to modelling the probability of a gene's expression pattern across a bicluster is to evaluate the probability of observing a random combination of n samples, where n is the bicluster size, taken from the dataset that would have a summarised consistency score at least as great as that observed in the bicluster. A fast means of estimating the probability of observing at least $(\frac{n}{2}\rfloor+1)$ values[1] from a random sampling of n values from a given gene's distribution can be obtained through evaluation of the appropriate hypergeometric distribution. As the hypergeometric distribution can be used to evaluate probabilities for sampling at least x values with a particular property out of a sample of k values (taken without replacement, as is the case here) from a total population of t values, of which m display the given property and $(t-m)$ don't, it is possible to estimate the probability of observing at least $n/2$ values above the median summary score for the bicluster out of a random sampling of n values from the distribution of all the given gene's GESTr-transformed values across the dataset (from which the number of values above the bicluster's median summary score for expression of that gene can be obtained trivially), where n is the size of the bicluster. This provides a fast and accurate means of estimating the bicluster observation probabilities for each gene, providing that the summary statistic used is the median. For fast estimation of the desired probabilities of bicluster observations for summary statistics other than the median, an alternative approach would have to be taken.

Fast Approximate Approach

As the distribution of observed GESTr-transformed values for any given gene is discrete, it is relatively straightforward to produce for each gene a sample cumulative distribution function (CDF) for the probabilities of observing greater than any given value. A simple heuristic approach to estimating a gene's bicluster observation probabilities might be to estimate an overall, indicative probability of observing greater than each measured value. This presents a situation similar to that observed for the discrete data bicluster entropy estimates such as that given in Equation (3.10), where each gene's contribution to bicluster information is estimated as a product of the number of samples in the bicluster and the negative logarithm of the indicative per-sample probability estimate of an observation relevant to the bicluster regarding the gene in question. Using the mean of the CDF-derived probabilities of each observed measurement in the bicluster as an indicative estimate of per-sample probability of significance of expression for the gene in question as observed across the bicluster, the simple bicluster information content estimator becomes the expression given in Equations (5.1-5.2).

[1] in this expression, $\frac{n}{2}\rfloor$ denotes the integer part of the fraction $\frac{n}{2}$

$$f(x) = -n \sum_{g \in genes} \log(\text{mean}_{s \in samples}(F_g(D_{g,s}))) \quad (5.1)$$

$$F_g = P(D_g > D_{g,s}) \quad (5.2)$$

In this fast method of estimating an approximate probabilistic score for information content of a bicluster, the mean summary function may be replaced with any other summary to estimate the probability of the observation of expression across the bicluster for each gene.

In summary, a number of different methods for evaluating the significance of a given bicluster in terms of the probabilities of observing such consistently significant expression scores (for the genes in the bicluster) from the GESTr-transformation of the dataset across a random subset of samples of the same size. One of these approaches, based on resampling of the summary-statistic of the given gene across different possible biclusters provides an accurate estimate of the above probabilities but requires a (possibly prohibitively) time-consuming pre-computation step and may not in fact be accurate for small-sized biclusters. An estimate for the exact probabilities calculated for the median summary score distribution was provided by utilisation of analysis of the hypergeometric sampling distribution to evaluate the probability of obtaining a random sample of values at least as good as the bicluster (according to a median summary statistic). Finally, a quick approximate estimator was developed to provide a means of estimating the desired probabilities when a summary statistic other than the median is desired, when speed of evaluation is critical and no limits to bicluster size are desired. The precise method of probabilistic bicluster evaluation using GESTr-transformed expression values will therefore depend on the context of bicluster evaluation and the desired properties entailed with that context.

5.1.2 Probabilistic Modelling of Gene Expression Co-Dependency

In addition to the entropy and confidence of expression based measure of bicluster desirability described above that utilise GESTr output values, the other component of the GDepBGA's bicluster evaluation function (Equations (3.14-3.17)), that of estimating the expected degree of co-dependency of expression of each bicluster gene with a guide gene of interest, would have to be modified to utilise GESTr output values. As discussed in Section (3.6.3), it is impossible to distinguish between co-dependency and dependency of expression using only (steady-state) gene expression measurements. Given the essentially arbitrary nature of the guide gene-dependency evaluation approach given in Section (3.6.3), it would be advantageous to have a means of providing estimates of the probability of co-dependency of expression, given available GESTr-transformed gene expression data. Such a means of evaluation would provide a clear interpretation

of the resulting scores for each genes. Furthermore, with a fully probabilistic means of evaluating a bicluster it may be possible to produce a (p-value) significance estimate for each gene in a bicluster based on the probability of observing such an expression pattern as that observed with the bicluster in question if the gene were not expected to share co-dependence of expression with the guide gene.

A measure of localised gene expression co-dependency as introduced in Section (3.6.3) an attempt to dissect the multitude of gene expression patterns within even a small subset of a large dataset, in order to distinguish trends in expression level across certain samples that associate any gene's expression level to one particular gene (or set of genes) of interest from other transcriptional effects. The ability to distinguish such patterns will clearly be dependent on the degree of change of the rest of the transcriptome across the samples in which the given pattern is observed, therefore it would be helpful to be able to estimate the expected variation in expression level of a gene between any two samples based on the overall difference in the general transcriptional program between those samples. With a model to estimate the expected variation of a gene's expression between any two samples, probabilities for co-dependency of expression of any pair of genes can be estimated through some measure of coincident variation in expression of those genes that is unlikely to be observed as a consequence of the general differences between the samples involved. Following the above ideas, the first part of this section describes a model for estimating the expected variation of expression of a gene between a pair of samples as a function of the overall Euclidean distance between the expression profiles of those samples, and the second part of this section presents methods for utilising these estimates of probability of observed variation being explained by general differences between the samples involved to produce probability estimates of context-specific co-dependency of gene expression based on bicluster-related observations.

Modelling Expected Gene Expression Variation

In order to estimate the expected variation in expression level (as a GESTr-transformed value) between any pair of samples, it may be possible to measure the differences in all genes' expression levels between sampled pairs of samples and to construct a model of the corresponding distributions of expression differences for each pair of samples, in such a way that the parameters of this model can be predicted from the overall Euclidean distance between the samples. If such a modelling process could be performed successfully, it would then be possible to estimate probabilities of observing a given gene expression difference between two particular samples in a randomly-chosen gene (and additionally the expected number of genes with at least that difference in expression level) through analysis of the expected gene expression difference distribution model using the (pairwise) Euclidean distance between the samples as a hyperparameter by which the parameters of the underlying probability model are specified. To set about this task of identifying a possible model with which to estimate gene expression

difference distributions for arbitrary sample pairs, a large number of such expression difference distributions were inspected (via plots using the `density` probability density estimation function provided in `R`).

From inspection of a large selection of such distributions and comparison to a number of standard statistical distributions, it appeared that the general expected gene expression difference distribution might be effectively modelled by a Laplace (double-exponential) distribution. The probability density function of the Laplace distribution (for zero-centred distributions) is shown in Equation (5.3). The shape of the Laplace distribution and its correspondence to the shape of differences for gene expression difference between two samples can be seen in Fig. 5.1

$$f(x) = \frac{e^{-|\frac{x}{\beta}|}}{2\beta} \tag{5.3}$$

To test the applicability of the Laplace distribution as a model of the gene expression difference distributions, such distributions could be fitted (through estimation of the parameter β in Equation (5.3)) to a number of sample distributions and the goodness of fit of these fitted distributions to the target sample distributions evaluated. As a means of estimating appropriate values of the parameter β for the Laplace distribution, there exists a closed-form maximum-likelihood (ML) estimator which is given in Equation (5.4). An alternative approach that could be taken if the MLE were to prove inaccurate would be to use the nonlinear least-squares estimate as calculated using the `nls` function in R.

$$\hat{\beta} = \frac{\sum_{i=1}^{N} |X_i - \bar{X}|}{N} \tag{5.4}$$

An example of a fitted Laplace distribution using the MLE for β is given in Fig. 5.1, with plots showing the underlying (observed) distribution in black and the fitted Laplace distribution in blue. As these plots appear to indicate that the MLE $\hat{\beta}$ is a consistent over-estimate for the underlying distributions, alternative estimation of the β parameter through nonlinear least-squares optimisation was attempted using the MLE $\hat{\beta}$ as an initial value. As this approach repeatedly failed to improve upon the ML estimate, it can be assumed that better estimates of this β are not readily obtainable. A particularly relevant observation regarding the Laplace model fits (such as that shown in Fig. 5.1) is that the fitted Laplace distributions generally seem to underestimate the density at the tails of the distribution, especially when the overall distance between the samples (and hence the parameter β) increase: this would entail underestimates of probabilities and therefore overestimates of the significance of all observations of large expression value differences.

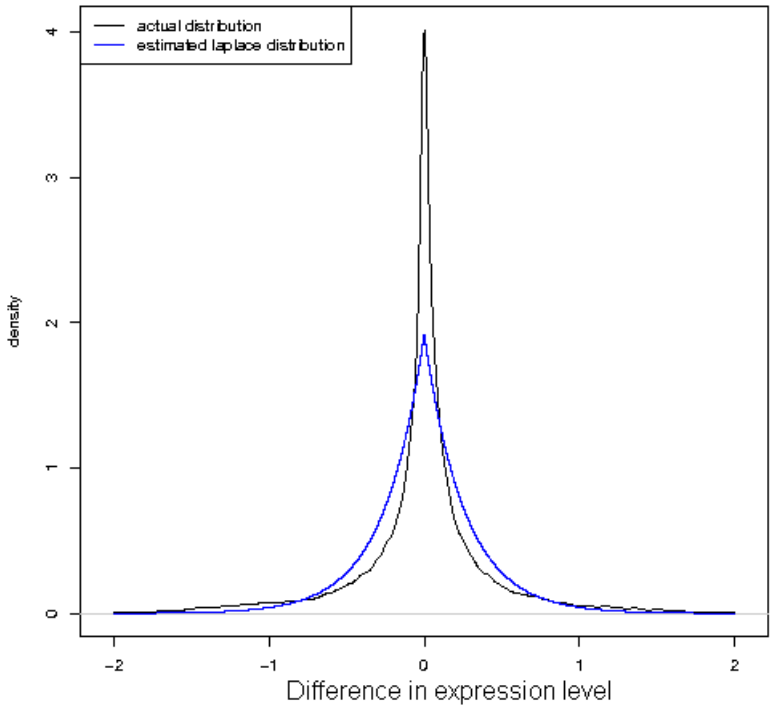


Figure 5.1: Fitted Laplace distribution to underlying gene expression difference distribution. Observed distribution of expression differences for all genes between two selected samples is plotted in black. Laplace distribution fitted with MLE to the observed values is plotted in blue.

Given that these consistent errors occur in the most critical parts of the distributions, and also the fundamental difference between the support of the observed distribution (on the scale of the GESTr-value input differences $-2 \leq x \leq 2$) and the models using the Laplace distribution (for which $-\infty < x < \infty$), an alteration to the model from the Laplace distribution was sought that could account for these critical differences and hopefully obtain a better fit to the observed values.

To create a distribution more appropriate than the Laplace distribution for modelling the desired gene expression difference distributions, alterations would be required to increase the values in the tails of the distribution (that appear to be more linear in the observed distributions than the Laplace distribution, particularly as distance between the samples increases). A method was proposed for this alteration in the form of adding a parameter-dependent linear decay term into the function that would be proportional to a parameter based on the distance between the samples and inversely proportional to the magnitude of the difference value. Incorporating appropriate bounds to the support of the distribution would be trivial, although scaling of the resulting density function would be required to maintain the essential property for any probability distribution that the integral of the density function evaluated across its whole range of inputs must equal 1. The resulting function on which the PDF would be based is given in Equation (5.5).

$$f(x|\alpha,\beta) = \begin{cases} 0 & \text{if } x < -2 \\ \frac{e^{\frac{x}{\beta}}}{\beta(2+4\alpha)} + \alpha(x+2) & \text{if } -2 \leq x < 0 \\ \frac{e^{\frac{-x}{\beta}}}{\beta(2+4\alpha)} + \alpha(2-x) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases} \tag{5.5}$$

Evaluating this as an integral gives a CDF (which is required for the evaluation of the resulting distributions) as shown in Equation (5.6). In order to scale the CDF appropriately, the value used is F^* as given in Equation (5.7).

$$F(x|\alpha,\beta) = \begin{cases} \frac{e^{\frac{x}{\beta}}}{2+4\alpha} & \text{if } x < -2 \\ \frac{e^{\frac{x}{\beta}}}{2+4\alpha} + \alpha(\frac{x^2}{2} + 2x + 2) & \text{if } -2 \leq x < 0 \\ (1 - \frac{e^{\frac{-x}{\beta}}}{2+4\alpha}) + \alpha(2x - \frac{x^2}{2}) & \text{if } 0 \leq x \leq 2 \\ (1 - \frac{e^{\frac{-x}{\beta}}}{2+4\alpha}) & \text{if } x > 2 \end{cases} \tag{5.6}$$

$$F^*(x|\alpha,\beta) = \frac{F(x|\alpha,\beta)}{F(2|\alpha,\beta) - F(-2|\alpha,\beta)} \tag{5.7}$$

With this adjusted distribution, attempts were made to model the observed gene expression value difference distributions by fitting α and β parameters to each observed distribution through simultaneous nonlinear least-squares optimisation of initial estimates $\beta = \hat{\beta}$ and $\alpha = \frac{dist^2}{2}$. An examples of such a fitted model distribution with

underlying observed distribution is shown in Fig. 5.2, similar to the equivalent shown for the Laplace distribution in Fig. 5.1.

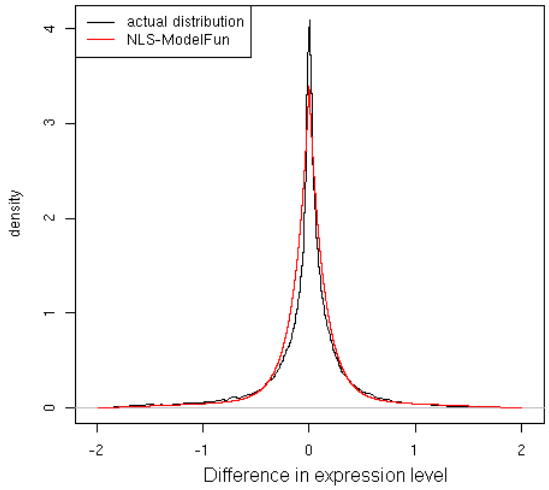


Figure 5.2: Fitted adapted model distribution to underlying gene expression difference distribution. LiTAL model fitted to observed gene expression differences by NLS regression and plotted in red. This clearly fits the observed distribution (black) better than the Laplace distribution shown in Fig. 5.1.

In order to provide an objective means of testing whether these alterations improved the modelling accuracy or not, χ^2 goodness of fit tests were employed to assess comparative goodness of fit. A χ^2 goodness of fit test was performed for each model fit to the appropriate underlying distribution by generating empirical discrete distributions for each of the observed difference distributions, the fitted Laplace distributions and the fitted linear-adjusted model distributions with a fixed number of discrete 'bins,' then computing a χ^2 statistic for each fit as the sum of the squared errors in bin totals as predicted by the model (and compared to the observed distribution) divided by the predicted totals for each bin. Owing to the large number of observations (approximately $45,000$), goodness of fit tests will tend not to provide any useful insight into significance testing for model fitting, but can be used effectively for comparative purposes [Ajiferuke et al., 2006]. According to such χ^2 goodness of fit comparisons, 66% of approximately 5000 randomly selected sample-pair gene expression value difference distributions were modelled better using the 'linear-tail adjusted Laplace' (LiTAL) distribution than the standard Laplace distribution. In addition to this, the observation that errors in the (critical) tail regions of the model distributions tended to be less for the LiTAL models than the Laplace models further motivates the use of the LiTAL distribution model.

Given the motivation for adopting the LiTAL distribution to attempt to model expected gene expression value difference distributions, a crucial step towards practical application of these models lay in the formulation of predictors for the individual α and β parameter values based on the observed sample distance. As a first step to carrying out this task, fitted parameter values were obtained for a large number of sample pairs spanning the range of possible distances, and the values for each of the parameters were plotted against the sample distances from the observations used to fit each of the parameters. These plots are shown in Fig. 5.3.

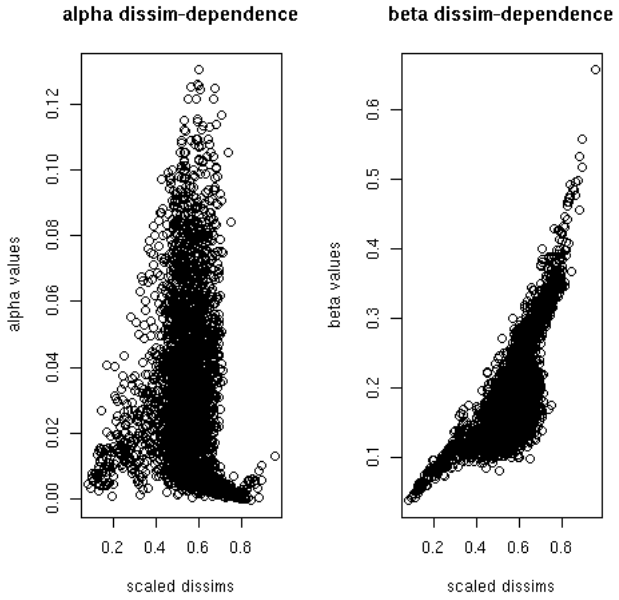


Figure 5.3: Dependency of each LiTAL model parameter on sample distance: fitted values of each LiTAL model parameter (α in left-hand panel, β in right-hand panel) plotted against the dissimilarity of each sample-pair analysed to obtain underlying distribution of gene expression differences to which LiTAL model was fitted.

While the β parameter appears to be clearly dependent on the sample distance, the relationship is less obvious for the α parameter. However, a similar plot is shown in Fig. 5.4 showing only the parameter values for those models with (an arbitrarily defined threshold of) good fit to the underlying data, according to the χ^2 values. From these models with good fit, an approximately linear relationship of both parameters to the sample distance emerges.

A series of linear (polynomial) models were fitted to each of the parameters' distributions as a function of the sample distance, using the `lm` function in R. ANOVA was used to compare the models to identify polynomial terms that failed to contribute significantly to the accuracy of the predictor model (given that an additional term was added). The resulting models are given in Equations (5.8-5.9) where a,b,c,d and g

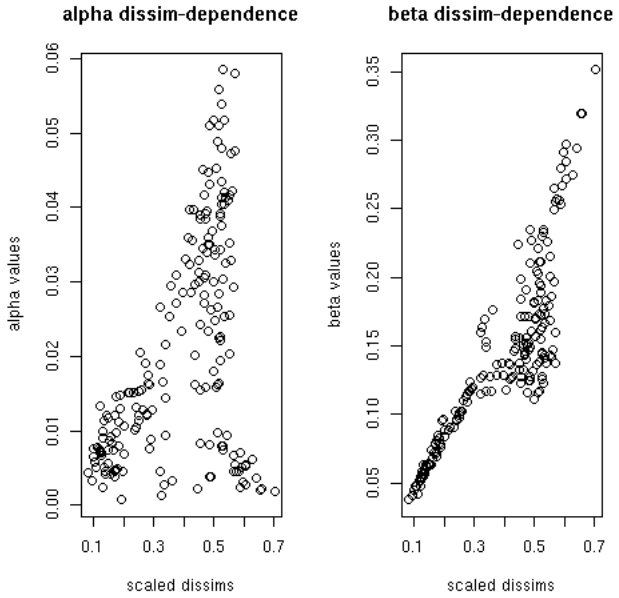


Figure 5.4: Dependency of each LiTAL model parameter on sample distance. After filtering out poorly-fitted distributions a clear linear relationship emerges between fitted values of each LiTAL model parameter and the dissimilarity of the samples analysed to obtain the gene expression difference distributions.

represent parameters 'learned' from the respective parameter distributions.

$$\alpha = a * dist \tag{5.8}$$

$$\beta = b + c * dist + d * dist^2 + g * dist^3 \tag{5.9}$$

A final remaining test was to ensure that the resulting 'typical' distributions were good predictors of the underlying values even when using the sample distance as a hyperparameter to determing the models' parameter values, rather than fitting the parameters directly to each distribution. A comparison of χ^2 values for model fits of these predicted LiTAL models based on the sample distance hyperparameter and equivalent predicted Laplace models (using a similarly learned hyperparameter-dependent distribution of the β parameter of the Laplace distribution) is shown in Fig. 5.5, with the χ^2 values plotted against sample distances for each of the tested model fits for the LiTAL models (red) and Laplace models (blue). This shows a clear improvement in the LiTAL model fits compared to those of the predicted Laplace models, as a lower χ^2 value corresponds to a better fit. It was also observed that the goodness of fit tended to decrease (χ^2 values increase) with increasing sample distance as shown in Fig. 5.6, a point that will be referred to in the description of guide gene co-dependency probability estimates that follows.

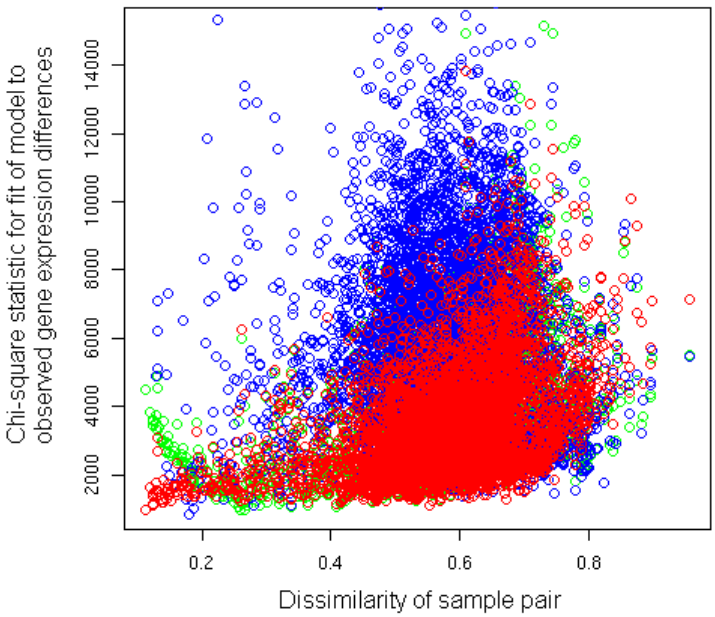


Figure 5.5: χ^2 goodness of fit scores for different models of 'typical' gene expression difference distributions. Goodness of fit scores for Laplace models fitted to gene expression difference distributions are shown in blue. Scores for the LiTAL model fitted through hyperparameter-based estimation of model parameters is shown in red. Scores for LiTAL model fitted through NLS regression shown in green. Lower values represent a better fit to the observed gene expression differences, indicating that the fitted LiTAL models more accurately describe the observed differences than the Laplace model, even when model parameters estimated from hyperparameters fitted according to Equations (5.8 & 5.9).

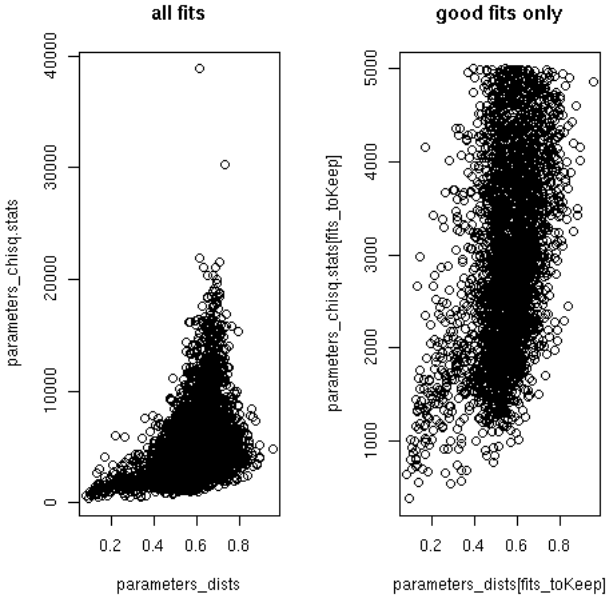


Figure 5.6: Dependency of goodness of fit of predicted LiTAL model on sample distance

It should also be noted that this process of fitting expected gene expression difference models may be dependent on the dataset. It is assumed that the general form of models used (but not the precise sample distance dependency functions for prediction of individual parameter values) will be consistent across other gene expression datasets[2], but as this is not known (especially for data from other types of gene expression measurement technologies) it is suggested that these steps are implemented with caution and with regard to alterations that may be required in the models described above when this process is applied to other datasets.

Estimating Probabilities of Gene Expression Co-Dependency

With a model to estimate expected distributions of the differences in gene expression level between any pair of samples, it may be possible to use such a model in conjunction with observed gene expression values across bicluster samples and appropriate comparison samples to provide an estimate of the probability of co-dependency of expression of two genes in the biological context represented by the bicluster samples. As with the case for evaluation of guide gene dependency using binary gene expression data (as presented in Section (3.6.3)), a crucial step in estimating context-specific co-dependency of expression involves the identification of samples with similar transcriptional programs to the samples in the bicluster but with contrasting levels of expression of the guide gene of interest. The aim of using such samples for comparison is to identify genes with expression patterns that appear to be dependent on some feature represented by the expression pattern of the guide gene of interest and not dependent on any other feature in the data, and to provide a measure of the confidence of each given observation of expression co-dependency based on the degree of correspondence of the respective expression profiles and the specificity implied by the appropriateness of the samples available for comparison with the bicluster.

The model for expected gene expression differences provides a means to assess the significance of any potential sample for use as a comparison with a set of bicluster samples, simply through calculating (from the expected distribution of differences) the probability of observing by chance a contrast equal to or greater than the difference in expression value of the guide gene between the bicluster samples and the given potential comparison sample. This probability estimate provides the ability to distinguish the guide gene's expression pattern from expected transcriptional differences between the samples in question. An additional point to consider is the fact that, with large datasets, a potentially large number of samples may be evaluated (or be available for evaluation) for comparison in this way, which introduces the chance that some insignificant comparisons would be evaluated as significant purely by chance, because a large number of such comparisons might be evaluated. It is therefore important to account

[2]especially as the GESTr transformation ought to make the values from different sources similar to those presented here

for such errors by considering the Family-Wise Error Rate (FWER), that is, the chance of *any* insignificant case being evaluated as significant, as opposed to the chance of a particular individual insignificant case being evaluated as significant. There is a range of multiple testing correction techniques available to deal with this situation, and adjustment of significance p-values can be performed by any of a number of these methods with the `p.adjust` function in R.

With an estimate for the probability of a given pair of samples involving a (significant) change in expression level of a guide gene of interest not simply as expected due to the overall differences between the samples, the probability of co-dependence of expression observed in that individual comparison could be estimated by multiplying this probability of significance of observation with an estimate of the probability that the expression levels are changing in a related manner. This would overall give an estimate for the probability that the expression level of a given gene is changing in a similar and significant manner with that of the guide gene, between the two selected samples. As there is a large number of genes that may be evaluated for significance in this way, multiple testing correction will again need to be applied to control FWER.

A clear way to estimate the probability that the expression levels of two genes are changing in a related manner between two samples purely by chance (due to general transcriptional differences between the samples) would be to evaluate the difference in the CDF for expected expression level differences between the two samples between bounds defined by the observed expression level differences for the two genes in question. In this way, the expected proportion of all genes in the dataset to have more similar expression level differences purely by chance is represented by the integral of the expected difference density function for the samples in question, evaluated between the two observed differences. These two properties described are equivalent. Such an approach means that genes with expression differences greater than that of the guide gene will be considered more significant (that is, less likely to occur by chance) than those with expression differences less than the guide gene, according to the estimated distribution's expected proportion of genes with lower expression level differences by chance. This property is illustrated in Fig. 5.7, showing two such integrals (area under probability density curve) for a hypothetical distribution of gene expression differences: the difference observed for the guide gene is indicated with a black vertical line, and estimates for probability of similarity of expression pattern occuring by chance for two different genes, one with higher difference in expression than the guide gene (area shown in red) and one with an equally lower difference in expression than the guide gene (area shown in green). Additionally, this formulation of probability estimate provides the desired property that a guide gene's estimated probability of similar change of expression with itself will evaluate to 1 (as should obviously be the case).

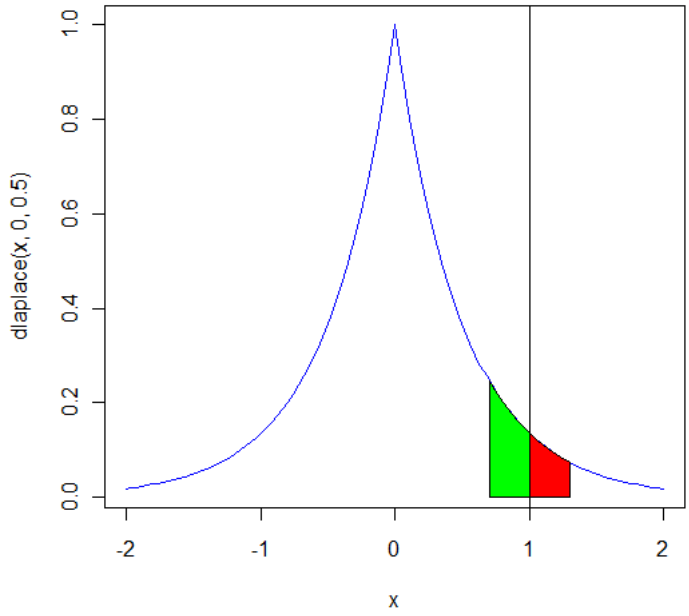


Figure 5.7: Estimates for probability that similar expression differences occur by chance, with a hypothetical distribution of gene expression differences between two samples, a hypothetical observed change in guide gene expression (black vertical line) and areas indicating the probability estimates for each of two hypothetical genes with lower and higher expression difference (green and red areas, respectively) than the guide gene

In the event that the expression of a gene under investigation varies along with other known transcriptional regulators in most situations in which the gene of interest varies, and the desired outcome of the investigation is to identify genes with exclusive expression co-dependency relationships with the gene of interest and not the other known regulators, it may be desirable to allow specification of 'contrast genes' in the analysis for which patterns discovered by the meta-analysis approach are specifically penalised if co-dependency is observed. A straightforward implementation in terms of expression co-dependency estimation involves estimation of co-dependency likelihood for all potential bicluster genes with the gene of interest and each of the specified contrast genes, with final co-dependency estimates provided by Equation (5.10). In Equation (5.10) G is the set of guide genes, C is the set of contrast genes, t is the potential target gene being evaluated and $P(dep(a,b))$ denotes the probability that genes a and b are observed to be co-dependently expressed in the given comparison.

$$P(dep(t,G)\ \&\ \overline{dep(t,C)}) = \left(\prod_{g_i \in G} P(dep(t,g_i))\right) * \left(\prod_{c_i \in C} 1 - P(dep(t,c_i))\right) \quad (5.10)$$

The manner in which the available estimates of similar and significant change in expression level with the guide gene between a range of pairs of samples might be utilised to produce an estimate for probability of co-dependency of expression in the biological context represented by the bicluster (on the strength of the available gene expression data), is not obvious. At least, there may be a range of possible approaches with different motivations and characteristics that might suit generation of output most useful for different biological questions. The examples given below by no means constitute an exhaustive list of options, but provide possible implementations with certain advantages (and drawbacks).

One such approach to generating a summary guide gene dependency probability is to estimate the probability that over a set of relevant comparisons, the whole set of observations all indicate similar and significant change in expression level of the guide gene and the gene in question. In such a way, the cumulative product of the estimates of probability of a significant comparison for the most significant comparisons may be evaluated to find the largest set of observations that is considered entirely significant (to a given significance threshold). Subsequently, the product of all the estimated probabilities of similar and significant change may be evaluated to obtain an estimate for the probability that the full set of observations entail significant and similar expression changes between the guide gene and the gene in questions, and therefore indicate likely co-dependence of expression. This approach has the advantage of being less dependent on the number of significant comparisons available to a bicluster for a given guide gene than other potential significance combination approaches (such as

taking all comparisons significant to a given threshold and applying a χ^2 combination to the set of resulting p-values), but tends to result in the undesirable situation in which good sets of comparisons (i.e. large ones) are penalised due to the fact that a gene's observations are less likely to be consistently similar to those of the guide gene across a large number of observations than across a smaller number of observations. This effect may be balanced by including a term to represent the probability of the collection of observations of similarity of expression arising by chance, resulting in an overall estimate for a gene i over a set of comparison samples $j \in$ ComparisonSamples as given in Equations (5.11-5.18).

$$score_i = e^{\frac{\sum_j log(1 - x_{ij})w_j}{\sum w_j}} \tag{5.11}$$

$$w_j = \frac{log(v_j^\dagger)}{\sum_{\forall k} log(v_k^\dagger)} \tag{5.12}$$

$$\mathbf{v} = \{\text{Csig}(ggVar_j, d_j)).\forall j \in \text{ComparisonSamples}\} \tag{5.13}$$

$$\mathbf{v}^* = \{..., v_i, v_j, ...\}\forall i, j.v_i \leq v_j \tag{5.14}$$

$$\mathbf{v}^\dagger = \{v_1^*, ..., v_n^*\}. \prod_{k=1}^{n} v_i^* \leq \theta \tag{5.15}$$

$$x_{ij} = \text{Cpval}(gVar_{ij}, ggVar_j, d_j) \tag{5.16}$$

$$\text{Csig}(v, d) = P(X \geq v|d) \ . \ X \sim \text{LiTAL}(\alpha(d), \beta(d)) \tag{5.17}$$

$$\text{Cpval}(v_1, v_2, d) = |Csig(v_1, d) - Csig(v_2, d)| \tag{5.18}$$

An alternative approach is based on limiting the maximum allowed distance between comparison samples and the bicluster samples, to take into consideration both the fact that (as shown in Fig. 5.6) the accuracy of the models of the expected distribution of gene expression differences between two samples decreases as the distance between the samples increases, and the general chance of observing gene expression patterns due to differences in biological context becomes greater as the distance between samples increases. With a restriction on the maximum distance allowed between a comparison sample and the bicluster samples, the number of potential samples for comparison may be significantly reduced. While in general a smaller maximum distance implies that closer (and therefore more appropriate, for a given significance level) comparisons will be used to assess gene expression codependencies, there may be a trade-off between the availability of similar samples for comparison and the significance of difference in guide gene expression between the bicluster and comparison samples. As such, it would be wise to examine the precise expression patterns used as evidence for expression co-dependency estimates for each bicluster, as these may provide insight into what constitutes appropriate (and inappropriate) settings of maximum sample dis-

tance and minimum significance level thresholds for each (set of) bicluster(s). As it is desirable to have as many highly significant observations as possible and especially undesirable to have observations of clearly dissimilar expression patterns across significant comparisons, the individual observation probabilities can be combined through a product weighted by each comparison's significance (the exponential of an appropriately weighted sum of logarithms) into an overall estimate of probability of co-dependent gene expression, given the observations available. Due to the fact that this weighted product is more likely to contain low values (and thus evaluate to a low value overall) if there are more significant comparisons available, this resulting estimate is scaled by the overall significance of the observations, as calculated through a χ^2 combination of the significance p-values (where the p-values are given by $1 - P$, P being the probability estimate) for each observation (demonstrated in Equation (5.20)). The overall estimate for the probability of guide gene dependence for a gene i over a set of comparison samples $j \in Csamples$ is given in Equations (5.21-5.25).

$$overall_p = F(-2 * \sum \log \mathbf{p} | \mathbf{p}) \tag{5.19}$$

$$F(x|\mathbf{p}) = \frac{1}{\Gamma(\frac{2*length(\mathbf{p})}{2})} \gamma(\frac{2 * length(\mathbf{p})}{2}, \frac{x}{2}) \tag{5.20}$$

$$score_i = (1 - (1 - z_i)(1 - y_i))(overall_p | \mathbf{p} = \text{Csig}(ggVar_j, d_j) \forall j) \tag{5.21}$$

$$z_i = e^{\frac{\sum_j log(1 - x_{ij}) w_j}{\sum w_j}} \tag{5.22}$$

$$w_j = \frac{log(\text{Csig}(ggVar_j, d_j))}{\sum_{\forall k} log(\text{Csig}(ggVar_k, d_k))} \tag{5.23}$$

$$y_i = |genes| \prod_j x_{ij} \tag{5.24}$$

$$x_{ij} = \text{Cpval}(gVar_{ij}, ggVar_j, d_j) \tag{5.25}$$

With the combined effect of the methods presented in this section, there exist methods to estimate the probability of co-dependency of expression between any given gene and a guide gene of interest, as evidenced by the gene expression patterns across a set of bicluster samples and appropriate comparison samples in the dataset. Used together with the estimates of entropy-based bicluster desirability, biclusters and genes can be evaluated for significance and those observations most appropriate for providing an answer to particular biological questions through analysis of bicluster-based patterns in gene expression data can be identified using an appropriate search algorithm.

5.1.3 Practical Limitations of Genetic Algorithm Approach

Practical application of a GA approach to biclustering, for identification of those potential biclusters in a large dataset that involve gene expression patterns that best provide answers to questions regarding transcriptional relationships involving a gene of interest in particular biological contexts, requires a very fast method of evaluating the score of a bicluster as the GA approach involves a very large number of such evaluations (the entire population of candidate solutions must be evaluated in every generation). Unfortunately, as the bicluster evaluation procedures described above involve evaluating models of expected distributions of differences for every comparison, for every gene, for every bicluster, even if the number of comparisons allowed is restricted (to increase evaluation speed) the typical evaluation time for an individual bicluster is of the order of a few minutes (fully utilising a single 2GHz core of a server's 8-core processor). Even for a relatively small GA run with 100 chromosomes in a population evolved over 100 generations, this would take weeks to obtain a set of biclusters for a single query. Using application to the dataset described in Section (3.3) as an example, for a target execution time of 10hrs and with the additional computation of the GA taken into consideration, the bicluster evaluation step must take no more than 3s. Assuming the sort of time frame described above is impractical for the desired application of the biclustering algorithms, a simpler approach to evaluation of guide gene expression dependencies would be required for a BGA to utilise GESTr-transformed values in performing large scale meta-analysis of gene expression data.

Reverting from the probabilistic evaluation of guide gene co-dependency analysis described in Section (5.1.2) to an approach more similar to the essentially arbitrary guide gene dependency scoring presented in Section (3.6.3), patterns of expression co-dependency represented in a bicluster can be assessed on the basis of the significance of available samples to use for comparison with the bicluster samples and the significance of each gene's observed expression variation between the bicluster samples and the comparison samples. However, in taking arbitrary bicluster scoring approaches, a number of problematic situations arose when considering the complex and subtle pattern mining task involved in implementing localised gene expression co-dependency analysis. Owing to time constraints for the completion of this work, it was necessary to produce a means for obtaining results from applying localised co-dependency analysis to large collections of gene expression data that were reliable enough to use to generate suitable biological hypotheses regarding research questions of interest. As such a point had not been reached with the GAs, and a route to obtaining such results with GAs was not immediately apparent, any further study and development of this family of methods had to be left for future work and an alternative approach to obtaining satisfactorily reliable results had to be implemented.

5.1.4 Sample-Grouping Heuristic

If the goal of bicluster-based meta-analysis of a large gene expression dataset is to identify expression patterns across particular (specific) biological contexts, it may simplify the bicluster search to restrict the search space to involve only potential biclusters with samples that are generally similar (for some definition of similarity). Such an approach would provide the additional advantage that interpretation of biological relevance of each bicluster to any particular context would be straightforward through inspection of the bicluster sample annotations (which would presumably reflect a consistent biological context as enforced through the pre-specified overall transcriptional similarities). If a set existed containing all subsets of the samples in the dataset that each represented a homogenous group of biological samples, bicluster search would involve screening of each of these groups to identify those with consistently high expression of the guide gene of interest and a set of associated samples available to enable significant (and relevant) comparisons of expression level. Following screening of sample groups to identify those that would make suitable biclusters, the guide gene-dependent expression patterns based around the maximal biclusters specified by each group's particular samples could be evaluated to obtain a full set of relevant bicluster-based expression patterns involving homogenous biclusters, and should involve sufficiently few bicluster evaluations to enable utilisation of the fully probabilistic approach to assessing both bicluster entropy-based significance (as described in Section (5.1.1)) and guide gene co-dependency of expression (as described in Section (5.1.2)). However, such an approach requires a set of potential bicluster sample groups and a means of assessing which of these are likely to result in a good bicluster for identification of context specific expression patterns involving a particular gene of interest.

The task of generating a set of groups of similar samples is similar to producing a full hierarchical clustering of the samples in the dataset (based on pairwise Euclidean distances) and selecting those groups of samples that appear 'appropriately connected' through similarity in the corresponding tree. This is therefore akin to finding clusters of high similarity within the matrix of all pairwise sample distances. However, determining what constitutes an appropriate level of similarity between samples may not be trivial, not least because it may be desired to have a number of different sample groups involving a particular sample but representing different degrees of specificity (e.g. a group involving replicates, a group involving all samples of that cell type and a group involving samples from tissues containing that cell type, etc.). One approach devised was to construct a probability model to estimate the expected number of samples sharing pairwise distances of no more than a particular threshold for a range of such thresholds, and to use these estimates to identify groups at each similarity threshold with significantly more samples similar (to the specified degree of similarity) than expected by chance according to the probability model. Thus, the task presented is

first to develop an appropriate probability model for the expected numbers of samples with a particular degree of similarity to a randomly chosen sample, for any similarity threshold.

For a given set of samples to constitute a group of similar samples as defined above, all pairwise distances between each of the samples must be below the specified similarity threshold. That is, the submatrix of the full matrix of all pairwise distances between the samples of the dataset for those samples in the given set must contain no values greater than the similarity threshold. In order to evaluate the probability of such observations occurring by chance, the conditional probabilities of two given samples being similar to each other (according to a given definition of similarity) taking into account the set of all those shared samples to which it is known both of the given samples are similar would have to be considered. It is unlikely that knowing that two samples are both similar to a third sample would not influence the probability that those two samples are similar. Measurements of observed numbers of similar samples shared between each pair of a large number of randomly chosen sample pairs and the corresponding distance between the samples of each sample pair were generated and linear models to predict distance based on number of similar samples shared were fitted to these measurements. Significance analysis of the fitted models through ANOVA suggested that the number of similar samples could indeed be used to predict inter-sample pairwise distances (data not shown), although the impact on accuracy of such a predictor of each shared sample considered diminished with each successive shared sample after the first. This implies that knowing that a few (or at least one) samples are shared between two randomly chosen samples strongly influences the probability that those two samples will be similar. The full conditional probability of similarity to a given threshold of a random pair of samples given a particular number of shared samples known to be similar to both will be different for increasing numbers of shared samples, but the significance tests of fitted linear models suggest that these will tend to converge as the numbers of shared samples increase. Given the suspected convergence of such conditional probabilities, the number of conditional probabilities necessary to evaluate in order to predict overall probability of a given observation of similarity can be limited without significantly affecting accuracy of the resulting estimations.

Despite the possible considerations for producing a probability model for such sample similarities based on available data, it was relatively straightforward to estimate (through a sampling approach) the cumulative probabilities of observing each of a given number of similar samples by recording numbers of observed samples in a similar-sample group involving each of a large number of randomly selected samples (with replacement), using a fast method to find 'dense' regions of similarity above the set threshold within the submatrix of pairwise sample distance involving all those samples similar (to the specified level) to the randomly chosen sample. With such cumulative probabilities

of observing a given number of samples in a group (with each similar to all of the others) estimated for a range of numbers of samples, for a range of thresholds, the significance of observing any group of samples in the dataset for which each is similar to all the others (i.e. the sample groups intended to be identified) to a given similarity threshold can be trivially obtained through looking up the estimated cumulative probability of observing by chance at least as large a group of interconnectedly similar samples as the one being evaluated. Due to the fact that these probabilities are estimated entirely from properties of the data, the software produced to estimated such probabilities is equally applicable to any dataset without any necessary alteration.

With significance tables available (through the above sampling-based method) for any group of similar samples at a range of pre-specified similarity thresholds, all potential groups of samples can be screened for statistical significance by finding for each sample in the dataset the group of samples with fully interconnected similarity according to each similarity threshold, then comparing the corresponding p-value (obtained from the pre-computed significance tables) to a set significance level. Those groups with sufficiently low significance p-values can be added to a set of significantly similar sample groups. To demonstrate the results of application of this approach to obtain a set of sample groups from the large collection of gene expression data from microarrays described in Section (3.3), Fig. 5.8 shows a sample tree (hierarchical clustering dendrogram) of the samples from a number of different sample groups at different levels of similarity. Highlighted in the left-most branch of the tree is an example of the hierarchical structure of the sample groups: three statistically significant sample groups (shown in red, blue and green) are also part of a larger statistically significant sample group (shown in purple). This creation of statistically significantly similar sample groups, some of which being subsets of larger groups, extends across the entire sample tree for all samples in the dataset. The full sample tree for the dataset is shown in Fig. 5.9, illustrating the fact that there are too many possible combinations for an exhaustive search. The sample grouping algorithm presented here is a means of defining a subset of distinct groups from this full tree that represent the similarity structure of the dataset. A concise description of this approach is provided in Algorithm (4), which features the 'pingpong' algorithm for finding submatrices of `1`s within binary matrices as described in [Oyanagi et al., 2001].

Input: Distance matrix for all samples, D, similarity thresholds, θ, significance table for sample group sizes at each threshold, P

Output: Set of groups of similar samples, S, set of significance p-values for each sample group, Sp

Initialise $sPool$ = Samples ;
while $|sPool| > 0$ **do**
  Set $s = sPool_1$;
  foreach $\theta_i \in \theta$ **do**
    Find similar samples $sSim = \{\text{Samples}_j\} \forall j.D_{sj} \leq \theta_i$;
    Find consistently similar samples $cs = \text{pingpong}(D_{sSim,sSim} < \theta_i)$;
    Look up pvalue $pv = P_{\theta_i,|cs|}$;
    if $pv \leq 0.5$ **then**
      Add cs to S ;
      Add pv to Sp ;
    end
  end
  Remove s from $sPool$;
end
Remove redundant groups from S ;

Algorithm 4: Creation of a set of groups of similar samples

Once such a set of sample groups is created, guide gene-dependent biclustering can be performed by evaluating each group of samples for its potential as an informative bicluster regarding the guide gene in question and then performing full bicluster analysis on the subset of the set of all groups corresponding to those samples predicted to result in informative biclusters. In order to filter the set of all 'potential bicluster' groupings of samples to find those that would be likely to give rise to informative biclusters, both consistency of high expression of the guide gene across the bicluster and availability of samples to provide appropriate comparison with the bicluster samples should be considered (as mentioned earlier in this section).

In principle, a bicluster will be particularly informative regarding expression patterns involving a given gene if the bicluster represents a group of samples from a particular biological context with consistently high expression of the gene of interest and an associated group of samples that are generally similar to the bicluster samples (from a global transcriptional perspective) but with significantly lower expression of the gene of interest than in the bicluster samples. Full evaluation of biclusters can be performed as described in Sections (5.1.1 & 5.1.2) although it would be useful to perform a quick pre-evaluation screening to filter out those potential biclusters unlikely to provide any useful results from evaluation. Therefore, if a potential bicluster (i.e. a sample group from the set created as described above) has both consistently high expression of the

Figure 5.8: Dendrogram of samples from a set of groups of significantly similar samples, as calculated by method described above

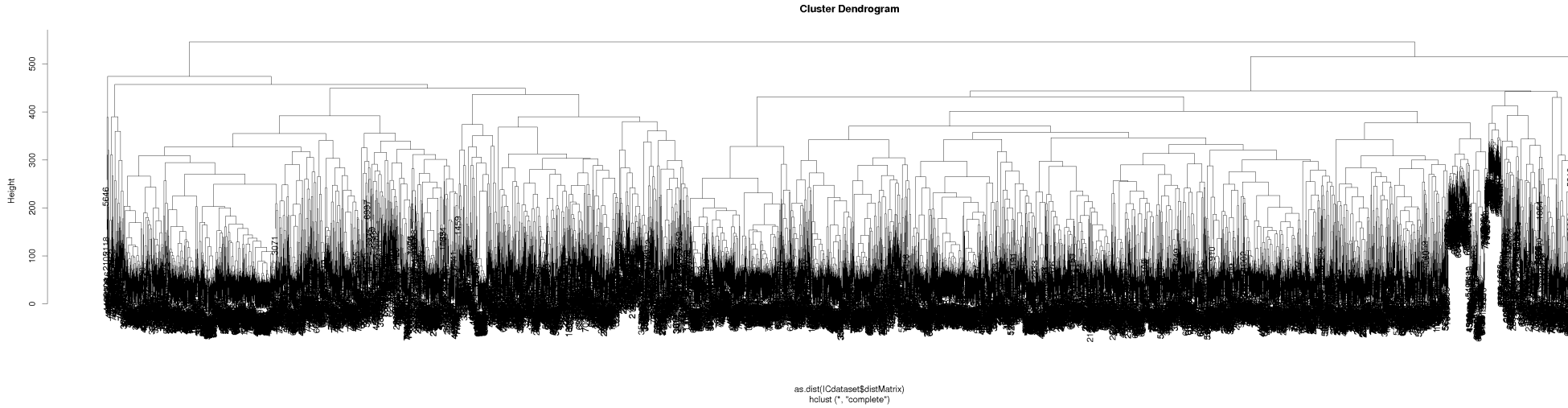


Figure 5.9: Full dendrogram of all samples in large dataset arranged by pairwise Euclidean distances

guide gene and appropriately similar samples with significant contrast in the guide gene expression when compared to the bicluster, it should be evaluated as an informative bicluster.

For filtering sample groups on consistency of guide gene expression, a criterion based on mean or median (or some other specified quantile) of GESTr-transformed expression values for the guide gene across the bicluster samples being above a set threshold is appropriate, provided the threshold represents a minimum required confidence of consistently high expression across the bicluster. The use of the GESTr-transformed values ensures a clear interpretation of any threshold applied.

The filtering of informative biclusters additionally requires evaluation of the availability of appropriate comparison samples. The modelling of gene expression variation developed for probabilistic modelling of co-dependency of expression, as presented in Section (5.1.2), could be used to estimate significance of variation of a guide gene between bicluster samples and potential comparison samples (in fact this is as described for part of the probabilistic guide gene co-dependency evaluation of biclusters presented in Section (5.1.2)). If the maximum allowed overall Euclidean distance between bicluster samples and comparison samples is restricted (as in the second bicluster guide gene dependency evaluation method described in Section (5.1.2)), sample groups can be filtered for availability of samples for appropriate guide gene expression comparison on the basis of the existance of samples within the specified maximum allowed distance to the 'median profile' of the bicluster that have statistically significant (to some significance level threshold) variation of expression level of the guide gene, given overall similarity. In fact, this approach would also work in the case where no maximum allowed distance was specified.

While the filtering methods above require specification of absolute thresholds, as both the confidence of gene expression level and the significance of guide gene variation involve values with a clear interpretation, these thresholds can be set to filter sample groups on the basis of known criteria. It is accepted that absolute thresholding is not completely desirable, but as the bicluster passing the filters will be evaluated and the properties on which the sample groups were filtered contribute to bicluster (and bicluster-gene) scores, setting thresholds to a minimum acceptable level will speed up the biclustering process without sacrificing accuracy of results. Using the above described approach to obtain a set of groups of (statistically) significantly similar samples, then filtering this set of groups to find the subset likely to give rise to informative biclusters for a particular guide gene (or set of genes) and evaluating the bicluster arising from each group of samples in the subset, a fast heuristic method is provided for identification of statistically significant local gene expression patterns involving a gene of interest, and estimating the significance of each such pattern for every gene identified

as following a similar expression pattern to that of the guide gene(s) in each localised context. This method is termed Heuristic Biclustering for Localised Co-Dependency Analysis, or HBLCA.

5.1.5 Integration of Biclusters

For the originally proposed task of utilising relevant gene expression data from as wide a range of sources as possible, the HBLCA framework described above identifies a number of supposed significant patterns that may each involve only small numbers of samples. While this approach utilises the full range of the data in a sense by identifying localised patterns across different subset of the whole dataset in the context of the overall gene expression distributions (as determined by the GESTr process), the ability to identify patterns which appear to be consistent across a number of the biclusters obtained for a single guide gene would give the potential for improved reliability and generalisability of inferred expression patterns (as discussed in Section (2.3.1) regarding the benefits afforded by meta-analysis of multiple datasets). An approach developed for the integration of results from multiple biclusters is described in the remainder of Section (5.1.5).

Bayesian Integration of Probabilistic Biclusters

An approach was desired for integration of results across a set of biclusters that would provide useful output, prioritising genes that score particularly highly in some biclusters (especially if scoring highly in many biclusters). The goal of applying such a method would be to identify particularly reliable observations, incorporating all evidence from a set of biclusters. If each bicluster provides a list of estimates of probability for each gene that the expression data available appears to be consistent with a co-dependency of expression of that gene and the guide gene in the biological context represented by the bicluster and a set of biclusters is available, assumed to represent a consistent biological context (for some desired degree of specificity), a bayesian integration approach can be taken to obtain estimates for each gene for the probability that gene is co-dependently expressed with the guide gene, in the biological context represented by the given set of biclusters, given the available evidence (i.e. the set of biclusters).

In a bayesian framework for integrating probability estimates for a particular event from multiple evidence source, given prior probability estimates for the event occuring and for the event not occuring ($P(E)$ and $P(F)$, respectively) the posterior probability estimate for that event occurring given the available evidence is defined in Equation (5.19), where $O = \{O_1, ..., O_n\}$ is the set of available evidence sources (that is, the data).

$$P(E|O) = \frac{\prod_{i=1}^{n} P(O_i|E)P(E)}{\prod_{i=1}^{n} P(O_i|E)P(E) + \prod_{i=1}^{n} P(O_i|F)P(F)} \tag{5.26}$$

In the case at hand, the event considered is that the gene in question is codependently expressed with the guide gene in the biological context represented by the set of biclusters, which are the separate evidence sources. A prior probability was estimated through the proportion of genes with expression values (across all the samples of the dataset) that have a Pearson correlation coefficient with those of the guide gene of at least some set threshold (e.g. $\rho = 0.6$). The correlation of each gene's expression profile with that of the guide gene could be used to obtain a different prior probability for each gene, although that would require development of a probability model for estimation of likelihood of gene expression co-dependency based on correlation coefficient values, which is not attempted here.

One important point to note regarding the adoption of the above approach is that the Bayesian integration model specifically includes estimates for the likelihood of each given bicluster pattern if the gene in question is not co-dependently expressed with the guide gene. This results in the estimates that a given gene is not co-dependently expressed with the guide gene because of a poor (or absent) estimate in a single bicluster proving crucial. If there is an assumption that the biclusters all represent the same context and that if an expression pattern is absent from any one bicluster then it should not be considered reliable, then this case should be treated differently to one in which it is assumed that the biclusters may reflect different underlying expression patterns or that some may be unreliable, meaning that genes will be considered especially reliably co-dependently expressed with the guide gene if they have high co-dependency probability estimates in a number of the available biclusters even if completely absent from the patterns represented by some of the other biclusters in the set being integrated. Due to the possible range of desired treatments of such cases, a minimum allowed likelihood for a single bicluster observation given a gene being codependently expressed with the guide gene and a corresponding maximum allowed likelihood for a bicluster observation given the gene not being codependently expressed with the guide gene should be specified in the particular case of integration.

An additional consideration of the integration of results across a set of biclusters is the possibility that there may exist subsets of biclusters within that set, with each subset representing a different underlying gene expression pattern. Therefore, it would be useful to have a means of identifying such structure within a given set of biclusters. In order to achieve this, a method was developed to identify clusters in a matrix of bicluster observation scores for each gene. Firstly, a matrix of likelihood scores for each gene for

each bicluster is created, setting scores to 0 for genes not in a bicluster. Principal Component Analysis (PCA) is used to reduce the dimensionality of this bicluster-attribute matrix from a possible $O(10,000)$ attributes down to a number more appropriate for clustering analysis to find general structure - the fewest number of characteristic 'principal components' that together explain at least a set proportion of the overall variation observed in the matrix (say 95%). Using the `prcomp` function (in the `stats` package) in R, PCA is performed using singular value decomposition of the input matrix and proportion of variation of the input matrix explained by the given number (say n) of principal components can be calculated by dividing the sum of the largest n eigenvalues by the sum of all the eigenvalues obtained through eigenvalue decomposition of the matrix (the eigenvalues corresponding to each of the principal components as eigenvectors of the matrix are provided in the output of `prcomp`). A series of k-means clusterings ([Hartigan, 1975]) of the resulting reduced-dimensionality bicluster-attribute matrix were performed, with the *gap statistic* [Tibshirani et al., 2001] calculated for each clustering performed and the optimal number of clusters selected as that resulting in the best gap statistic score of clusters calculated through k-means clustering. These steps are performed in the `R` functions `kmeansGap` and `gapStat`, available from the package `SLmisc`.

Using the method described in the previous paragraph, a set of biclusters may be separated into a number of subsets reflecting structure within the collection of gene expression patterns represented by the given set of biclusters. Following this decomposition, each subset of biclusters may be integrated individually to identify reliable gene expression co-dependency patterns (involving the guide gene) across each context (as reflected in the underlying structure of the collection of bicluster expression patterns). Additionally, it may be interesting to identify (if present) any genes with consistent guide gene expression co-dependency patterns across all biological contexts represented by the set of all biclusters, in which case integration of probability estimates can be performed across the full set of biclusters available for the guide gene, before being performed across each subset identified through the decomposition method described above.

It should also be noted that the gene association list clustering approach described here may be useful for applications other than that presented here, such as the decomposition of a heterogeneous genelist into groups with similar bicluster patterns (represented by the results of bicluster genelist integration), as is utilised in Sections (5.4 & 5.6).

The method described in this section provides a means to obtain estimates of co-dependency of gene expression for each gene represented in a dataset with a guide gene (or set of guide genes) of interest, incorporating evidence from multiple biclusters: first

from a (global) set of all available biclusters for the guide gene and subsequently from individual subsets of these biclusters, each representing a different component of the expression co-dependency patterns observed across the full set of available biological contexts. This reflects a natural decomposition of the full range of biological contexts into those displaying similar expression patterns involving the guide gene.

5.1.6 Implementation of HBLCA Algorithm

The HBLCA approach provides a means to identify genes with co-dependent expression patterns across individual biological contexts, through meta-analysis of large collections of gene expression data. An implementation of the HBLCA approach, combining the components described through Section (5.1), is outlined below:

1. Apply GESTr method (of Section (4.1.4)) to generate a matrix of universal gene expression state confidence values for all available data
2. Create a set of similar sample groups through estimation of significance values for a range of possible sizes of similar sample groups at each of a number of specified similarity thresholds, as described in Section (5.1.4)
3. Optionally, select a subset of the sample groups to perform biclustering across, if only certain biological contexts are to be considered in the meta-analysis
4. Obtain a filtered subset of sample groups evaluated as likely to give rise to informative biclusters for a given query gene (or set of genes), as described in Section (5.1.4)
5. Evaluate the bicluster formed by each sample group in the filtered subset, combining the component probabilistic bicluster evaluation methods described in Sections (5.1.1 & 5.1.2)
6. Create a bicluster-attribute matrix for all biclusters evaluated and perform clustering as described in Section (5.1.5) to identify grouping of observed expression patterns within the set of returned biclusters
7. Perform bayesian integration of each gene's probability scores evaluated for each bicluster, as described in Section (5.1.5), firstly for the set of all biclusters then for each individual group of biclusters identified in the previous step

A suite of `R` functions was created to provide a usable implementation of the HBLCA approach. With the development of the HBLCA procedure to identify and evaluate the evidence for gene expression co-dependency patterns provided by a large collection of gene expression data, some demonstration of the success of this approach is important in order to make the case for its application to problems in biological research.

5.1.7 Visualisation of Output from HBLCA Algorithm

The HBLCA approach presented above provides a means of identifying patterns of gene expression co-dependency across samples within a consistent biological context, involving a given gene (or set of genes) of interest. Information represented in each bicluster output object (through the calculations involved in the biclustering process) includes similarity of samples used for comparison of expression profiles, expression level contrast of the gene(s) of interest, expression level contrast of identified genes with relevant expression patterns, and a likelihood of each codependent expression observation. In order to assess the validity of a particular result obtained using this novel HBLCA approach, it would be helpful to have a means of scrutinising simultaneously each aspect of the information represented by a particular bicluster output object. As many biclusters with relevant expression patterns may be discovered for a single meta-analysis query, it would be especially useful to be able to visualise these many aspects of the meta-analysis output so that patterns identified through the meta-analysis (or patterns involving any given set of genes) can be assessed through inspection across a set of biclusters and in such a way, the magnitude and consistency of predicted expression co-dependency patterns can be evaluated.

To achieve this aim, a bicluster plotting tool was developed. This tool utilises plotting capabilities of `R` to provide a means of simultaneous visualization of all the aspects of useful information captured by the HBLCA approach, as described above. An example plot produced by this visualization tool is shown in Fig. 5.10, with the various components of the plot described as follows:

1. Expression profile of guide gene across bicluster samples and relevant comparison samples, shown with dotted red line. Expression levels are by default shown in the unified GESTr-transformed scale from 1 (high expression state) to 0 (low expression state), but may also be shown in terms of the untransformed expression measurement values.

2. Expression profiles of genes under investigation (also shown across bicluster samples and relevant comparison samples), shown with non-dotted lines of colour varying from red to blue.

3. Assessment of likelihood of each expression profile representing codependent expression with the guide gene, encoded in the colour of each line along a (purple) spectrum from red (likelihood=1) to blue (likelihood=0).

4. Indication of which samples shown are in the bicluster and which are the comparison samples. Bicluster samples have a black bar shown underneath.

5. Similarity of comparison samples to bicluster samples. For each comparison sample, Euclidean distances are calculated between that sample and each of the bi-

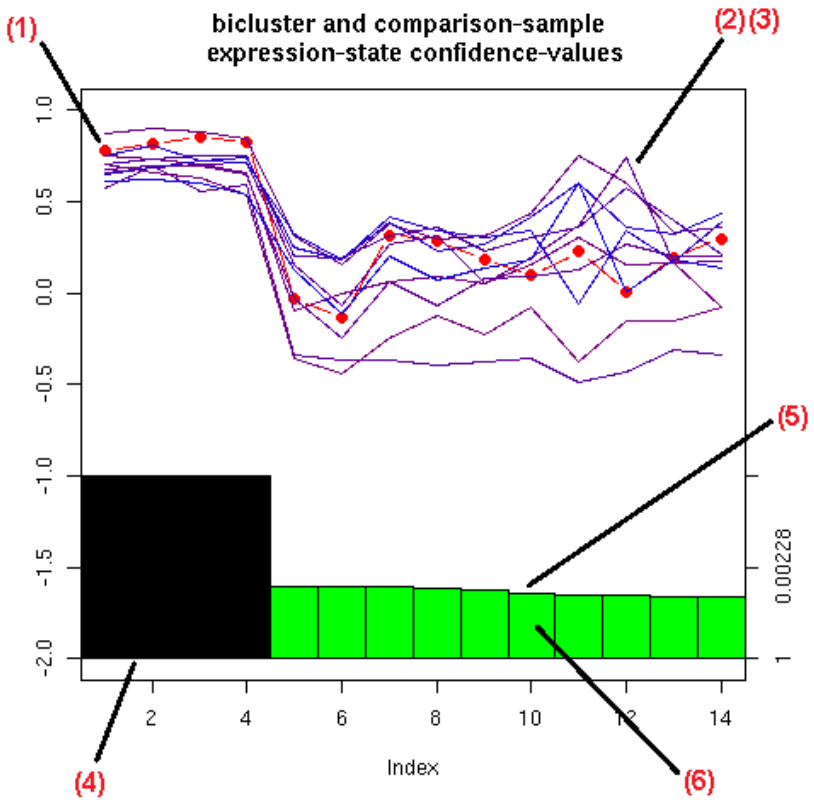


Figure 5.10: Illustration of output of HBLCA through a novel bicluster visualization tool. Numbers refer to points in enumerated description of the visualization tool output provided in Section (5.1.7)

cluster samples. These distances are averaged and the average distance is transformed to the negative logarithm of the probability of observing at least as similar samples purely by chance, evaluated from the cumulative normal distribution.

The normal quantile plot (data not shown) illustrated that this is a fair approximation for the probability estimate. The resulting similarity-significance values are shown in the bicluster plot by the height of the bars underneath each non-bicluster sample included in the plot. The similarity significance p-values can be read from the axis on the right hand side of the plot, as the height of the bars are given in terms of the similarity significance of the bicluster samples to one another.

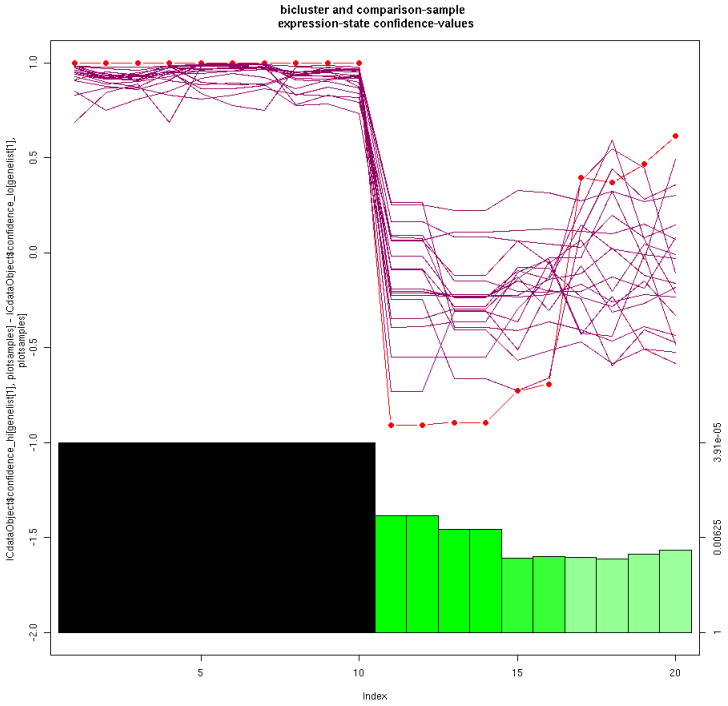
6. The significance of each comparison, as evaluated through the method described in Section (5.1.2), is indicated by the colour of the bars underneath the comparison samples, with a more intense green representing a more significant comparison (relative to the best comparison shown), fading to whiter bars for less significant comparisons. This information can also be inferred through inspection of the magnitude of expression level change of the guide gene between the bicluster

samples and each comparison sample in turn, considering the similarity significance of each comparison sample to the bicluster samples shown as described above.

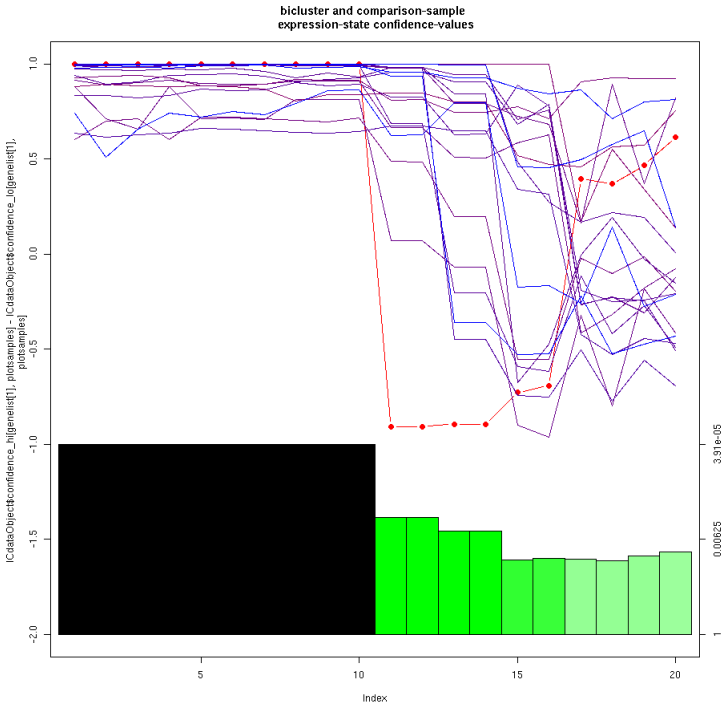
This visualization tool provides the opportunity to apply any (arbitrary) *post hoc* filtering of the biclusters generated through the HBLCA approach so that only those with desired properties are used for further analysis. In this way it also provides a means for validation of the bicluster patterns discovered by the HBLCA algorithm, through inspection of the expression profiles and the comparison distances and significances shown in the plots. With any high-throughput statistical data mining approach, it is important to check that the patterns discovered do indeed have the desired properties.

Additionally, the visualization tool (in conjunction with the implementation of the HBLCA algorithm) enables intuitive assessment of context-specific expression patterns involving any given gene list and a guide gene of interest, in any specified set of samples. Following evaluation of the bicluster properties based on gene expression co-dependency patterns with the guide gene in the samples of interest, the bicluster expression patterns representing the input genelist can be plotted by the visualization tool and shown in relation to the expression of the guide gene and the overall similarity of the samples identified by the meta-analysis algorithm as best illustrating a contrast in level of expression of the guide gene within the appropriate biological context. As an example of this application, Fig. 5.11 shows two bicluster plots produced by the visualization tool, illustrating differences in Oct4-dependent expression in iPS cells between two genelists: one (top) identified through subset-integrated biclustering performed using the novel meta-analysis approach (from results presented in more detail in Section (5.7)), and the other (bottom) identified as genes with greatest Pearson correlation coefficients with Oct4 across a large collection of gene expression data.

Visualization of biclusters in an area of research that has attracted some attention (e.g. [Grothaus et al., 2006, Cheng et al., 2007, Santamaria et al., 2008]) due to the fact that visual approaches to analysis can complement statistical approaches applied to large datasets. However, the guide gene-dependent bicluster model utilised by the novel meta-analysis approach presented and used in this work incorporates a number of features not considered by other biclustering approaches, and therefore a novel visualization tool was required in order to show this additional information regarding global similarities between bicluster samples and chosen 'contrast samples,' and significance estimates for each given comparison for the 'guide' gene of interest. This visualization tool is useful for confirmation of expression patterns of interest identified through HBLCA, application of arbitrary filtering of biclusters for meta-analysis and investigation of context-specific expression codependencies involving genes of interest.



(a) iPS-specific Oct4 codependent genes



(b) Oct4 correlated genes

Figure 5.11: Illustration of the differences in Oct4-dependent expression patterns in iPS cells only between a set of genes identified as iPS-specifically Oct4 codependent and a set of genes whose expression is generally correlated with that of Oct4. Visualisation tool clearly shows genes in top panel more closely follow the level of Oct4 expression (dotted red line) in this comparison, involving iPS cells, than the more general set of Oct4-correlated genes shown across the same samples in the bottom panel.

5.2 Evaluation of Similarity-Biclustering Approach

As discussed in Section (2.3.6) any method developed for identification of transcriptional relationships between genes through the analysis of gene expression data must be shown to identify known relationships with biological relevance in order for novel findings through application of the method to real problems in biological research to be trusted. To this end, example tasks were performed through meta-analysis of a gene expression dataset using the HBLCA approach presented in Section (5.1), in such a way that the results could be evaluated for the ability of the method to recover known transcriptional relationships. The results from application of the HBLCA approach are compared with those of existing meta-analysis approaches to give an indication of the significance of each observed result.

5.2.1 Prediction of Differentially-Expressed Genes

A significant part of the motivation for performing meta-analysis of large collections of gene expression data comes from the desire to make more generally applicable inferences regarding transcriptional mechanisms, as discussed in Section (2.3.1). A widely adopted application of microarray technology has been to perform transcriptional profiling of a biological sample type of interest with and without some forced alteration of the expression level of a gene of interest. From the resulting data, genes most responsive to forced change in expression level of the gene of interest can be identified. This has proven to be an effective means of identifying transcriptional targets of genes of interest (e.g. [Hall et al., 2009, Levy and Hill, 2005, Loh et al., 2006, Ivanova et al., 2006]). Due to this application, data from a number of controlled transcriptional profiling experiments involving genetic or chemical alteration of the expression level of a TF were available to assess the generalisability of inferences of gene expression relationships performed by the HBLCA approach. Lists of genes differentially expressed along with a gene of interest in controlled, targeted differential expression transcriptional profiling experiments that were not included in the meta-analysis dataset were obtained. These genelists were obtained through application of statistical tests for differential expression using the `affy` and `limma` packages within Bioconductor, as described in [Smyth, 2004]. It would be expected that a list of genes with reliable transcriptional relationships involving a particular gene of interest would include genes differentially expressed in experiments involving targeted alteration of the expression of that gene of interest.

Comparative enrichments of such lists of genes differentially expressed in individual experiments were evaluated for genelists resulting from meta-analysis performed by the HBLCA approach presented earlier in this chapter and for genelists from an established correlation-based method. This simple correlation-based approach to meta-analysis of gene expression data involves obtaining a ranked list of correlated genes

based on Pearson correlation coefficients for each gene's RMA-normalized gene expression values with those of the gene of interest, across the dataset used for HBLCA. The effectiveness of this approach was demonstrated in [Day et al., 2009], where it was used to infer transcriptional associations between genes that were shown to be involved together in the same disease processes. Furthermore, applications of hierarchical clustering to gene expression data tend to use Pearson correlation as the measure of similarity [Eisen et al., 1998], making this one of the most widely-used means of inferring associations between genes from expression data. Plots of these enrichments for each meta-analysis approach for each of a number of genes for which targeted differential expression studies were available are shown in Figs. 5.12 & 5.13. The respective targeted expression studies used to obtain 'target' gene lists were as follows:

- Oct4: Hall [Hall et al., 2009], GSE8617 [Sharov et al., 2008], GSE4679 [Ivanova et al., 2006]
- Nanog: GSE4679 [Ivanova et al., 2006], Chambers (personal communication), GSE8617 [Sharov et al., 2008]
- Sox2: GSE4679 [Ivanova et al., 2006], GSE5895 [Masui et al., 2007]
- Ppara: GSE6864 [van den Bosch et al., 2007]
- Srf: GSE7412 [Fleige et al., 2007]
- Klf9: GSE6443 [Simmen et al., 2007]

The plots shown in Figs. 5.12 & 5.13 indicate that the HBLCA approach identifies genes with reliable expression patterns involving the gene of interest used in the analysis, at least as well as an effective established method. Although, it should be noted that using lists of genes obtained from analysis of gene expression datasets from individual targeted experiments may be a distorted proxy for representing the desired transcriptional relationships, owing to the dependency of each experiment's results on the particular cells involved (which may differ significantly from the general biological context in which gene expression relationships are sought to be identified) and the precise conditions in which that experiment was performed, or due to experimental measurement error or stochastic 'transcriptional noise' [Li et al., 2008]. The identification of genes with differential expression in individual targeted experiments will therefore not represent a perfect validation scenario for methods for predicting transcriptional relationships. However, the results presented in this section do indicate that the HBLCA approach identifies gene expression patterns that are likely to be observed in novel experimental datasets, at least as well as and often considerably better than an effective existing meta-analysis approach.

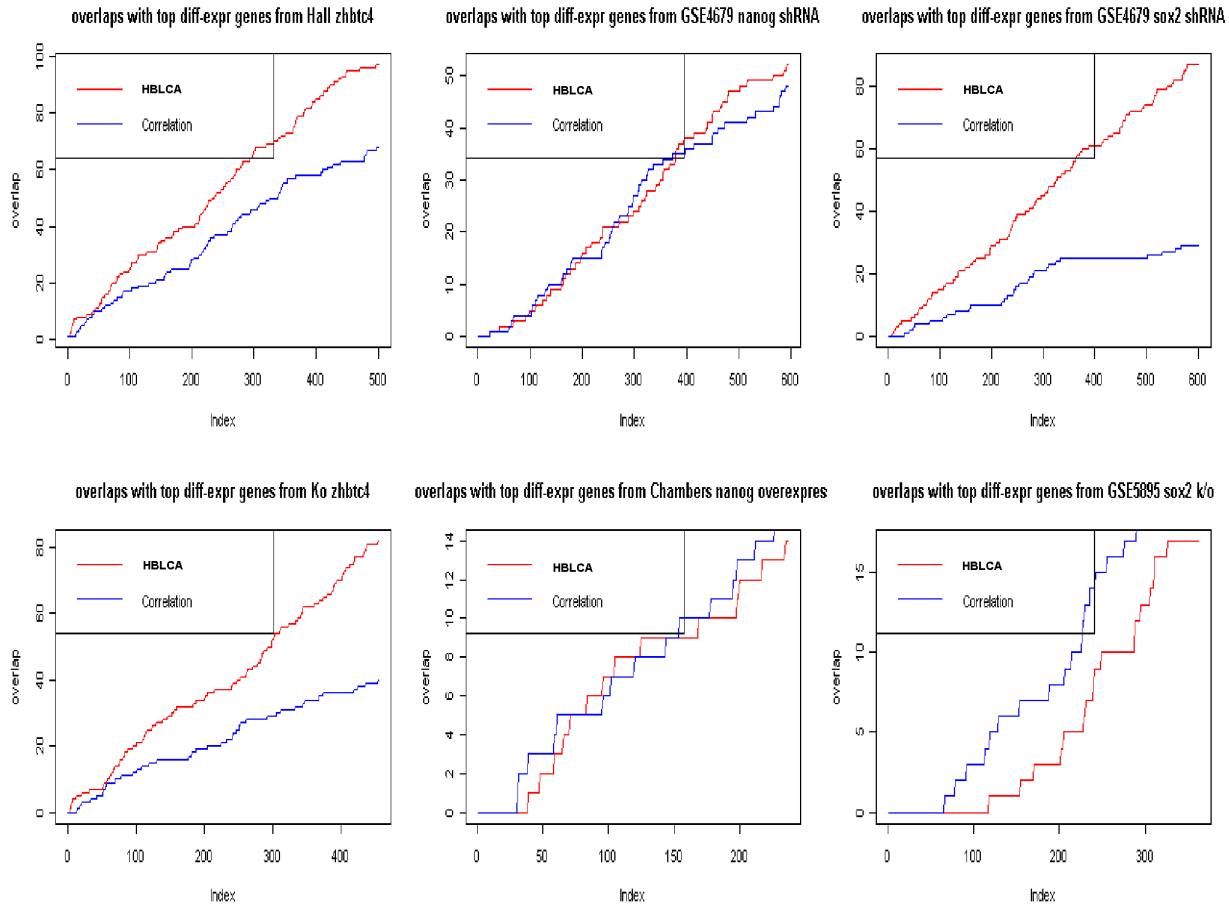


Figure 5.12: Comparative rates of discovery of differential expression targets identified by HBLCA (red) and correlation-based meta-analysis (blue). Numbers of targets in top ranking genelists obtained from novel biclustering approach are plotted in red and those obtained from Pearson correlation of gene expression profiles across meta-analysis dataset are plotted in blue. Horizontal axis indicates length of genelist.

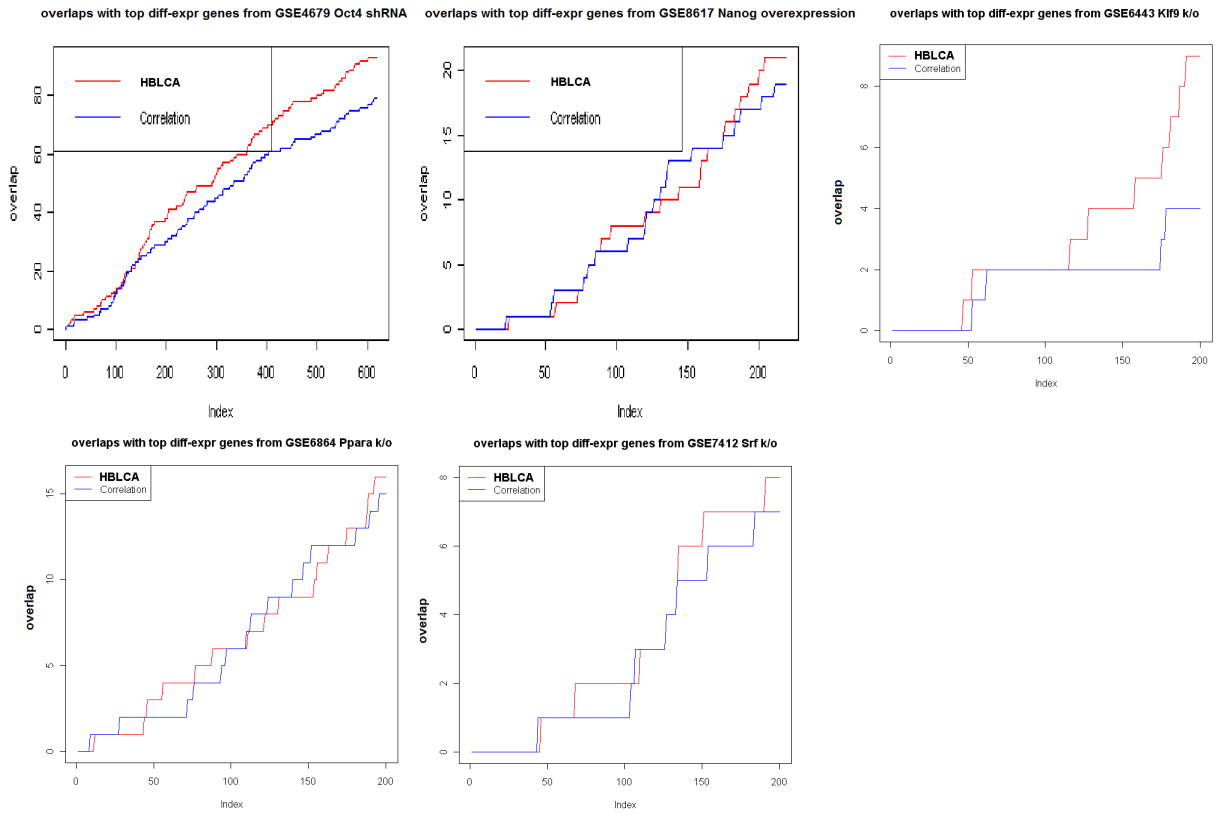


Figure 5.13: Comparative rates of discovery of differential expression targets identified by HBLCA (red) and correlation-based meta-analysis (blue). Numbers of targets in top ranking genelists obtained from novel biclustering approach are plotted in red and those obtained from Pearson correlation of gene expression profiles across meta-analysis dataset are plotted in blue. Horizontal axis indicates length of genelist.

5.2.2 Enrichment of Targets Identified From ChIP Studies

As with data from high-throughput gene expression studies (such as those involving microarrays), data from high-throughput DNA-binding studies (such as those based on chromatin immunoprecipitation (ChIP) assays) are being released into the public domain following publication of the original findings from the study in question. The publication of data from high throughput ChIP assays from an increasing range of transcription factors (TFs) provides lists of genes with known DNA-binding by those assayed TFs. Lists of genes with known DNA-binding identified through ChIP assays therefore provide an opportunity to assess the effectiveness of a gene expression data meta-analysis approach for predicting transcriptional interactions. This section presents observations from analysis of genome-wide ChIP data with relevance to the application of ChIP data to evaluate gene expression data meta-analysis approaches, followed by the results of comparative evaluation of the HBLCA method against established techniques for predicting transcriptional relationships between genes, in terms of the identification of TF DNA-binding targets.

DNA-Binding Data From Genome-Wide ChIP Studies

With increasing numbers of ChIP datasets released into the public domain comes firstly an increased chance that DNA-binding data is available for a particular TF of interest and secondly, the opportunity to perform meta-analysis by investigating together data from multiple platforms, laboratories and TFs.

In performing such meta-analyses it is worth bearing in mind the differences in experimental techniques for generating high-throughput ChIP data. The chromatin immunoprecipitation (ChIP) technique involves cross-linking DNA (chromatin) to all proteins bound to it, followed by fragmentation of the DNA, use of an antibody to the protein of the TF of interest to 'pull down' those fragments of DNA bound by the TF, then reversal of the DNA-protein cross-linkages and purification of the DNA fragments. The results of this procedure will be a collection of DNA fragments thought to be bound by the TF in question, so somehow the original genomic location of these fragments must be identified and (optionally) associated to 'target' genes. It is in this step that the variety of high-throughput technologies tend to differ the most: two main families exist, one based on microarray technology to identify target genes through hybridisation of the purified, pulled-down DNA fragments to probes with complementary sequences to predicted promoter regions for all (or some subset) of the genes in the relevant organism's genome; the other is based on sequencing each of the individual DNA fragments followed by computational alignment to a reference genome and identification of genomic features (in particular, genes) associated to each of the unambiguously aligned sequences. Obviously, the so-called ChIP-chip technologies may only identify binding to particular, pre-specified regions of the genome, whereas those based on sequencing (ChIP-PET,

ChIP-seq) are not restrained in this manner. Therefore, it would be expected that ChIP-seq studies of DNA-binding of a given TF would identify many potential targets not identified by an equivalent ChIP-chip study, as is the case with example studies for Oct4 [Chen et al., 2008, Kim et al., 2008, Marson et al., 2008, Sharov et al., 2008] shown in Fig. 5.14 (data from S. Morfopoulou, personal communication).

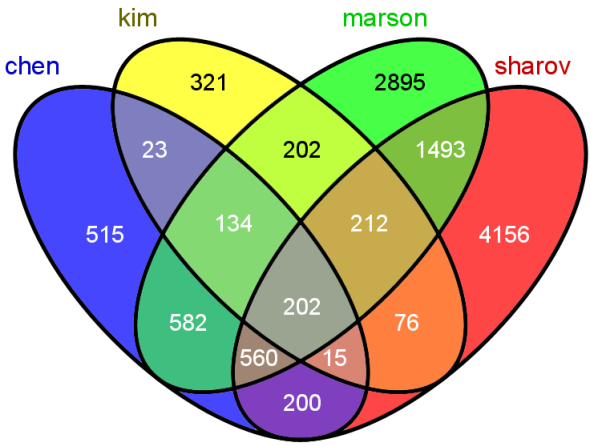


Figure 5.14: Overlaps between Oct4 ChIP studies: 'kim' targets from ChIP-chip study [Kim et al., 2008], 'chen' 'marson' and 'sharov' targets from sequencing-based ChIP studies [Chen et al., 2008, Marson et al., 2008, Sharov et al., 2008]

An additional consideration when performing meta-analysis of ChIP studies is that the methods used to generate results from the raw measurements obtained in the experiment may differ from study to study (e.g. in [Chen et al., 2008, Sharov et al., 2008]). A number of factors may influence the output of such data analysis (e.g. measurement intensity thresholds for calling positive binding, genome build used to determine alignments, models of the expected measurement observations corresponding to certain underlying DNA-binding, methods of association of sequences to genes, etc.), and may obstruct fair comparison or effective integration of data from multiple studies. Bearing this in mind, for effective meta-analysis of data from multiple ChIP studies such approaches should be standardised (performed in the same way) as far as is possible from the data available. Some differences will tend to remain due to the methods of obtaining sequences of bound DNA fragments from the particular experimental platform, but if these sequences are available then alignment to the genome and association to genes can be performed in a standardised manner to remove some of the (possibly systematic) discrepancies in the results from different studies. However, even when this is done, an interesting result is illustrated in Fig. 5.15 emerging from comparative analysis of DNA-binding data from different studies. Data from a number of ChIP studies from different laboratories, using different technologies to predict DNA-binding

of a number of key pluripotency TFs (in mouse ES cells), were assembled and analysed using a standardised method to obtain lists of predicted targets corresponding to each TF from each study (this standardised analysis was performed by S. Morfopoulou). From these lists, a target inclusion matrix was created with each study represented by a (column) vector of binary values for each gene: `1` representing the gene was identified as a predicted target of the appropriate TF in the study, `0` representing the gene wasn't identified as such a predicted target. For a closer comparison, only genes represented on the ChIP-chip promoter array were included in the target inclusion matrix. Hierarchical clustering of this matrix was performed using Euclidean distance as the distance metric, resulting in the dendrogram shown in Fig. 5.15.

It is particularly interesting to note that the target lists from the ChIP studies for Oct4 and Nanog as published in [Sharov et al., 2008] are more similar to one another than the respective target lists for Oct4 and Nanog from other studies [Kim et al., 2008, Chen et al., 2008]. This is especially interesting because the Sharov and Chen studies ([Sharov et al., 2008, Chen et al., 2008] respectively) use ChIP-seq technology but the Kim study [Kim et al., 2008] uses ChIP-chip technology: it would probably be expected therefore that the Kim studies would be further from the other two (for each TF), but that is clearly not the case. Additionally, it is interesting to note that there is a relatively high degree of similarity between the target lists from Oct4, Nanog and (the one study with) Sox2, while those for cMyc and Klf4 are clearly distinct (despite coming from different platforms). Precisely what these discrepancies represent (and their likely causes) is left for further study as it is not so relevant to the main theme of this work, but the results shown in Fig. 5.15 and discussed here demonstrate some of the motivation for meta-analysis of data from multiple ChIP studies and for integration with reliable expression data observations to predict likely transcriptional targets of particular TFs of interest. Further motivation for such integrative analysis is provided, with analysis of gene expression patterns of predicted DNA-binding target lists from ChIP studies and the corresponding TFs.

Gene Expression Dependencies of ChIP Targets on Associated Transcription Factors

It has been reported in the literature (e.g. [Li et al., 2008, Chambers and Tomlinson, 2009]) that individual ChIP studies tend to report DNA-binding of the corresponding TF to large numbers of genes. It has been proposed (in [Li et al., 2008]) that some of this DNA-binding may be due to stochastic chemical 'noise' (protein happening to be attached to DNA by chance, as opposed to being drawn to attach through electrostatic forces) or to a relatively high degree of non-specific, non-functional binding. Even with scepticism regarding these proposed features of the measured binding, it is clear that for many TFs, binding to DNA proximal to a predicted target gene does not necessarily have a direct impact on the expression of that gene. To illustrate this, correlations of

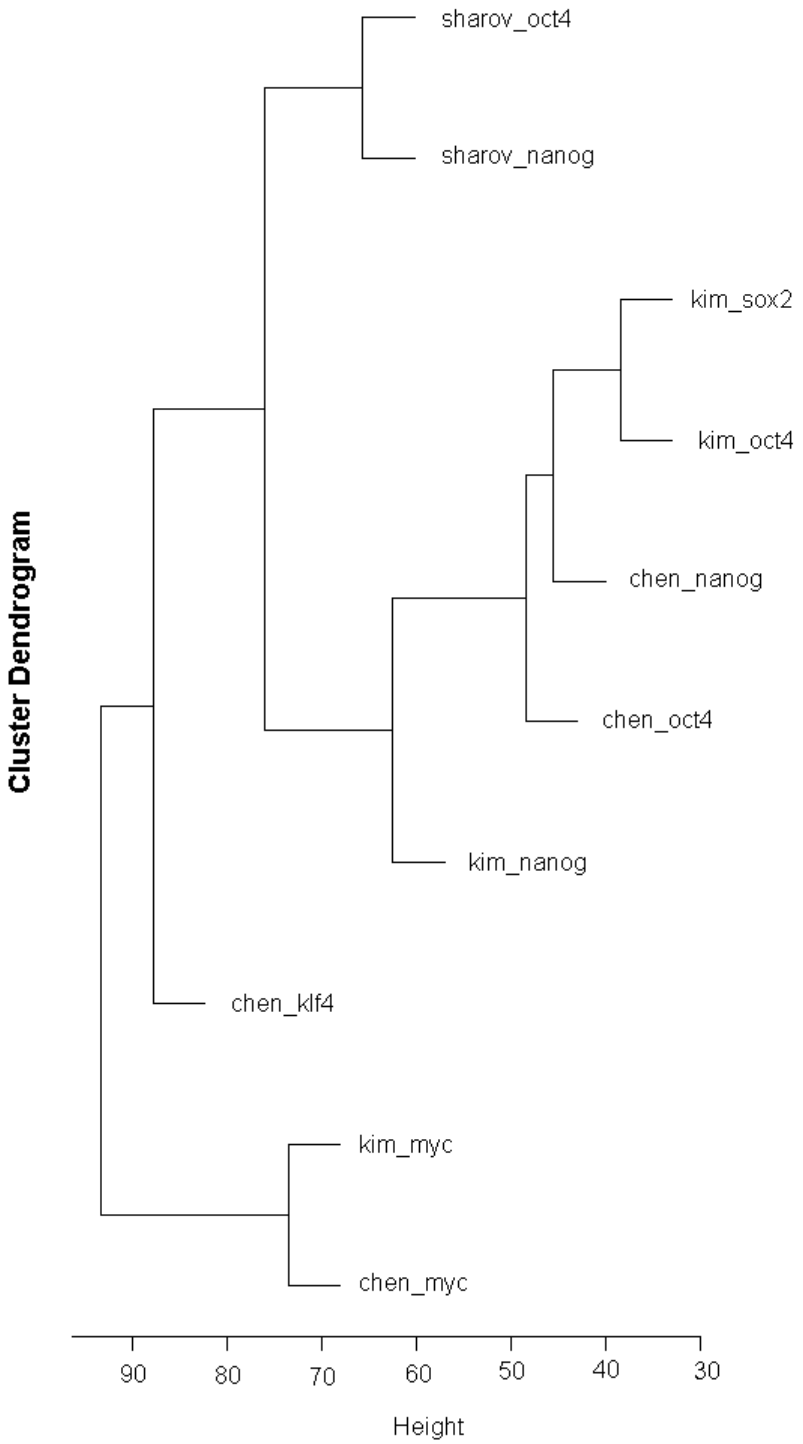


Figure 5.15: Hierarchical cluster tree showing similarities between results from different ChIP studies

expression profiles of predicted binding-targets (from ChIP studies) to their respective TFs were calculated across the large collection of gene expression data described in Section (3.5). Histograms of these correlations across the full set of targets for each of a number of TFs are shown in Fig. 5.16.

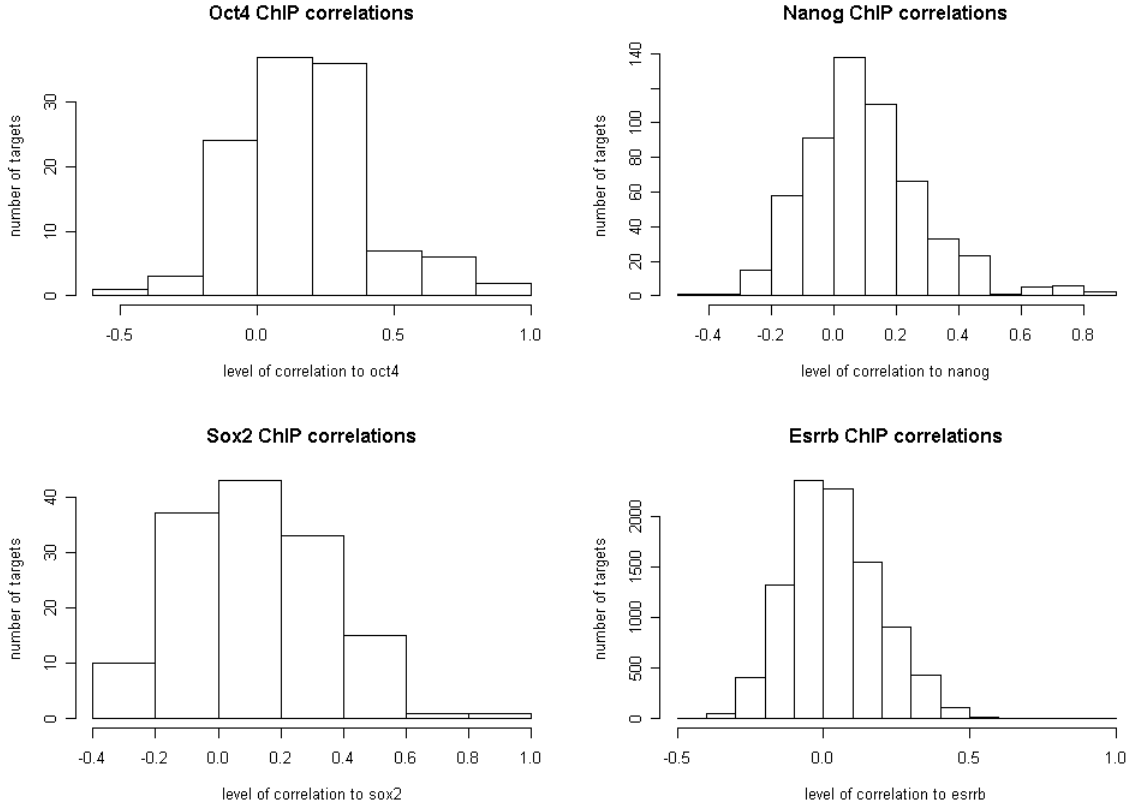


Figure 5.16: Distribution of Pearson correlation coefficients between expression of predicted binding targets and their respective TFs across large collection of gene expression data.

If a gene's expression were to be directly and exclusively regulated by one particular TF, it would be expected that the expression profiles (across any dataset) of that target gene and its regulating TF would be perfectly correlated (or anti-correlated), indicated by a correlation coefficient of 1 (or -1). It is clear from the distributions of correlation coefficients for target genes and binding TFs shown in Fig. 5.16 that such cases are extremely rare, if they exist at all (and it should be noted that proteins of each of Oct4, Sox2 and Nanog bind to their own respective promoters and auto-regulate).

It is clear from the plots in Fig. 5.16 that the majority of 'target' genes identified as having proximal DNA bound by a given TF have expression patterns almost completely uncorrelated to those of the respective TF (as indicated by the histogram peaks around $\rho = 0$ for each TF analysed in this way). However, this does not imply that such

target genes are not transcriptionally regulated in any way by the TF in question, just that they are not directly and exclusively regulated by that TF. There may be many reasons for such an observation to arise: the target gene may be expressed in certain contexts where the TF is not expressed, and in such circumstances its expression will necessarily be regulated by other TFs; or the TF in question may be one of a number of TFs that combinatorially regulate the expression of the target in such a way that different combinations of TFs have different transcriptional consequences, resulting in observations in which the expression of the target gene is directly dependent on that of the TF in question in only the subset of samples without expression of any redundant TFs; and there may be many other explanations not mentioned here. In fact, the distribution of correlations of expression of predicted targets to their binding TFs that are (near) specifically expressed in ES cells (Oct4 and Nanog) when evaluated across ES cells only shift significantly away from the $\rho = 0$ peak, as illustrated in Fig. 5.17.

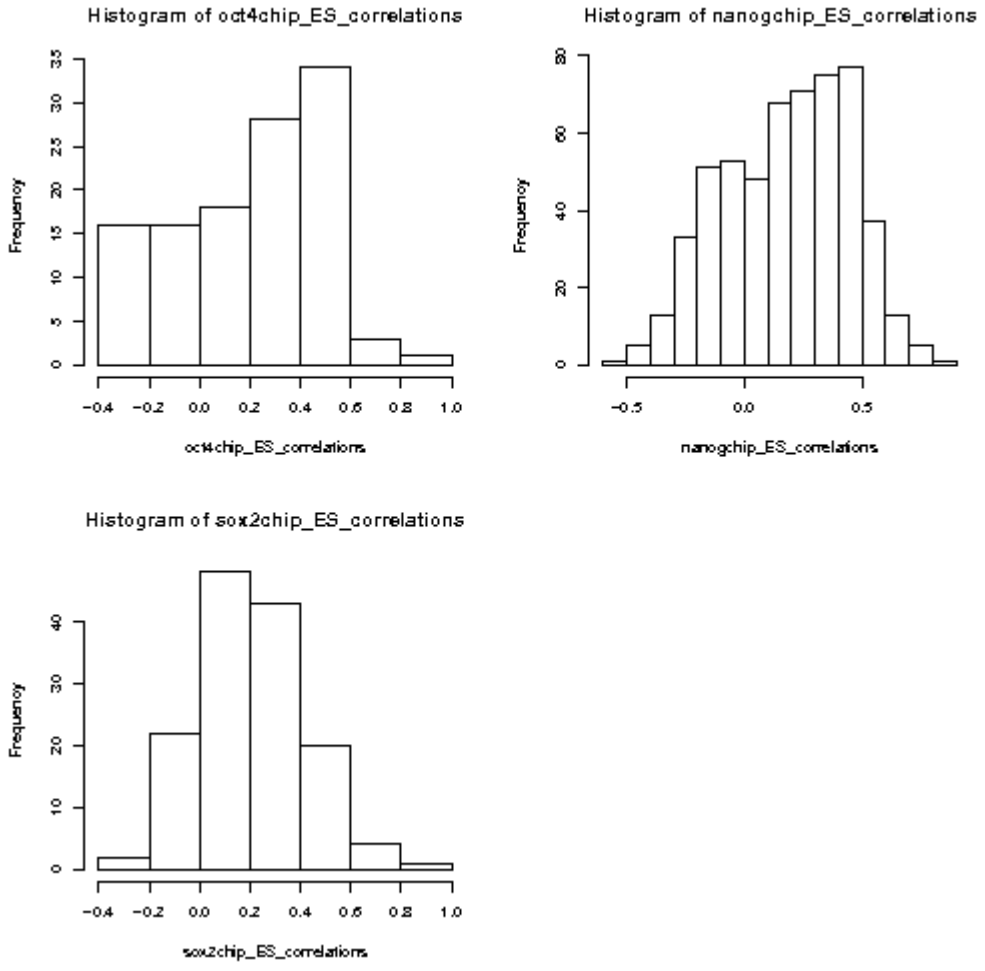


Figure 5.17: Distribution of Pearson correlation coefficients between ES cell expression levels of predicted binding targets and that of their respective TFs.

Therefore, in order to further understand the mechanisms of activity of a TF of interest, it would be useful to identify any relevant circumstances in which each target gene's expression appears to be dependent on that of the TF in question, or to identify expression patterns of targets known to be bound by a set of TFs that might be thought to interact in regulating gene expression that correspond to different combinations of expression of that set of TFs. As such circumstances may involve subsets of all the samples with gene expression data available (even if requiring expression of the TF), a meta-analysis approach based on biclustering applied to the available gene expression data would seem to be particularly appropriate for these tasks.

Comparative Enrichment of ChIP Targets

The identification of genes with a relationship in gene expression patterns to a given TF and known DNA-binding by that TF presents a real application of gene expression data meta-analysis by which approaches for such meta-analysis can be evaluated. The principle of this evaluation is that, as discussed in Section (3.6.2), if a meta-analysis of gene expression data is intended to identify transcriptional relationships involving a particular TF of interest it would be expected that a significant proportion of those genes identified with predicted transcriptional relationships on the basis of expression patterns would be bound by the TF in question, due to the fact that this will be a prerequisite for the TF actively regulating the expression of the 'related' gene.

A range of TFs for which DNA-binding data is publicly available were given as individually applied guide genes for the HBLCA approach. Ranked gene lists were obtained for each TF through integration of results from biclusters involving ES cells, as this was the biological context in which the ChIP assays were performed. Additionally, for each TF used to obtain a genelist through biclustering, a ranked list of correlated genes was obtained based on Pearson correlation coefficients for each gene's RMA-normalized gene expression values with those of the TF.

Comparative rates of recovery of the known DNA-bound targets were assessed through generating lists of the number of the top ranking genes in each list known to have proximal DNA bound by the TF in question for each increasing length of list. Plots of these increasing numbers of known DNA-binding targets identified by each method for each TF are shown in Fig. 5.18.

The comparative enrichment plots shown in Fig. 5.18 demonstrate that the genes identified by the HBLCA approach as having expression co-dependency patterns with each of a number of genes of interest seem to be generally more likely to have proximal DNA bound by the relevant gene of interest (and therefore likely to be transcriptional targets) than those identified as having similar expression patterns with each gene of interest through a large-scale correlation approach that has been shown to reveal

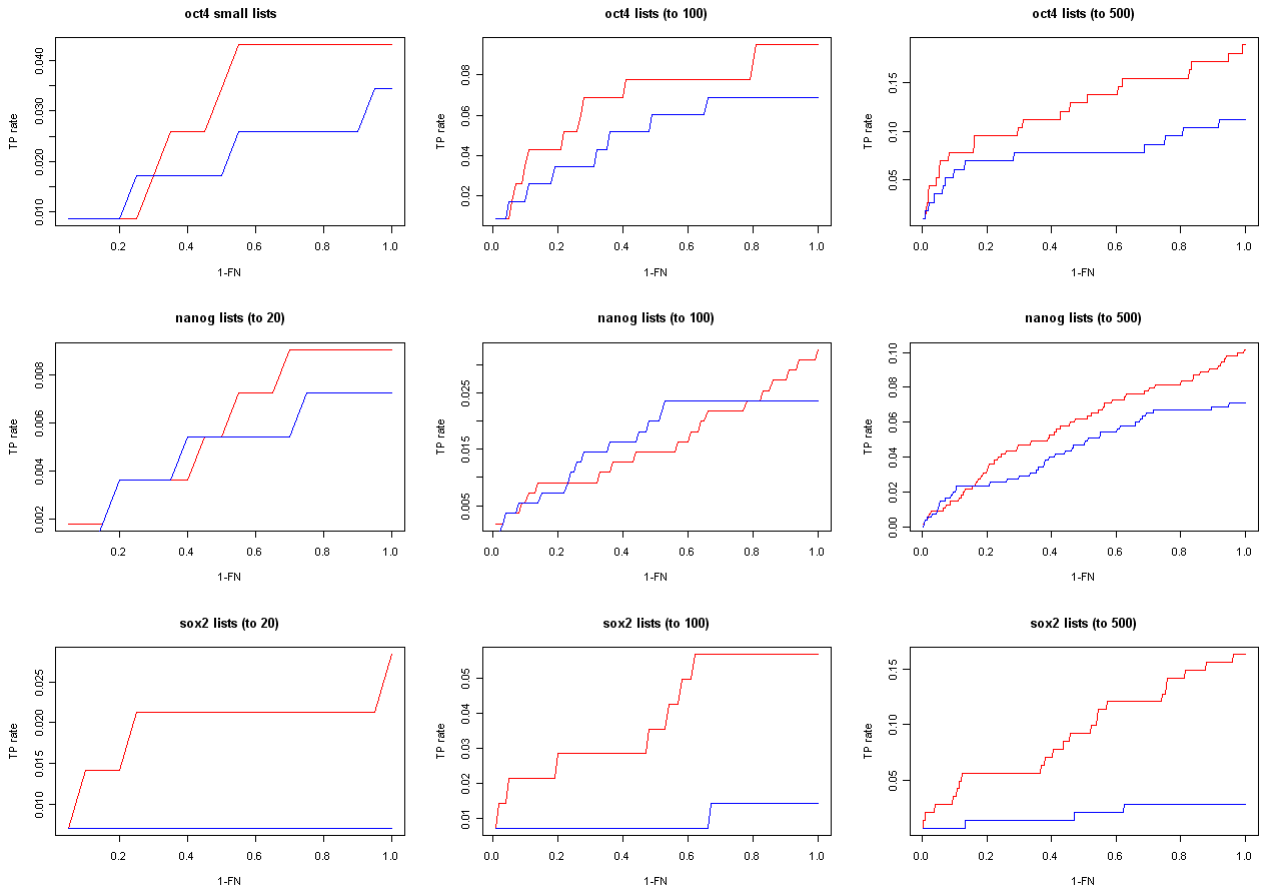


Figure 5.18: Comparative rates of discovery of DNA-binding targets identified by HBLCA (red) and correlation meta-analysis (blue). Proportion of targets present in top-ranking genelist from each method is indicated by the vertical axis, while the horizontal axis increases from left to right with increasing length of genelist obtained from the meta-analysis.

transcriptional relationships between genes (as described in the previous paragraph and demonstrated in [Day et al., 2009]).

To provide additional evidence to support the claim that the enrichment of TFs' binding targets shown in the lists of associated genes obtained through application of the HBLCA approach represents an ability to discover real transcriptional relationships between genes, enrichments for the genelists identified using the HBLCA approach were compared with those for genelists obtained through application of LIMMA [Smyth, 2004] to individual gene expression datasets involving controlled alteration of expression of the corresponding TF. Comparison datasets used for Oct4 were constructed from those relevant microarray samples from each of [Hall et al., 2009, Ivanova et al., 2006] (obtained privately and through GEO accession numbers GSE4679). Comparison datasets used for Nanog were from each of [Ivanova et al., 2006, Sharov et al., 2008] (obtained privately and through GEO accession numbers GSE4679 and GSE8617, respectively). Comparison datasets used for Sox2 were from each of [Ivanova et al., 2006, Masui et al., 2007] (obtained through GEO accession numbers GSE4679 and GSE5895, respectively). All individual datasets were normalized prior to differential expression analysis using the RMA method [Irizarry et al., 2003] as implemented in Bioconductor [Gentleman et al., 2004], and statistical tests for differential expression were performed using the `affy` and `limma` packages within Bioconductor, as described in [Smyth, 2004]. Plots showing the comparative enrichments calculated for each of these TFs are given in Fig. 5.19 (as in Fig. 5.18) for comparison of genelists from the HBLCA approach with those from an existing successful approach based on analysis of individual expression datasets. These plots show that the HBLCA approach generally identifies more genes with DNA-binding by the relevant TF through its evaluation of gene expression than standard differential expression analysis performed on data from individual, targeted transcriptional profiling experiments.

Enrichments of a TFs binding targets in the bicluster genelists suggest that real transcriptional relationships are being identified through this meta-analysis approach and as a result provide more useful validation data than many of the less transcriptionally relevant evaluations discussed in Section (2.3.6). However, they should not be taken as a direct and absolute measure of success of such an approach as it is expected both that some genes bound by a given TF would have expression profiles seemingly unrelated to that of the TF due to dependency on additional TFs (it is known that certain groups of TFs regulate the transcription of some target genes in a combinatorial manner, as discussed in [Chambers and Tomlinson, 2009]), and that some DNA-binding inferred from high-throughput ChIP data may be non-functional [Li et al., 2008]. Additionally, it is likely that some genes showing co-dependent expression patterns with the gene of interest will not be direct transcriptional targets of that gene, but may either be targets of a gene that is a direct target of the gene of interest or of a gene that regulates the

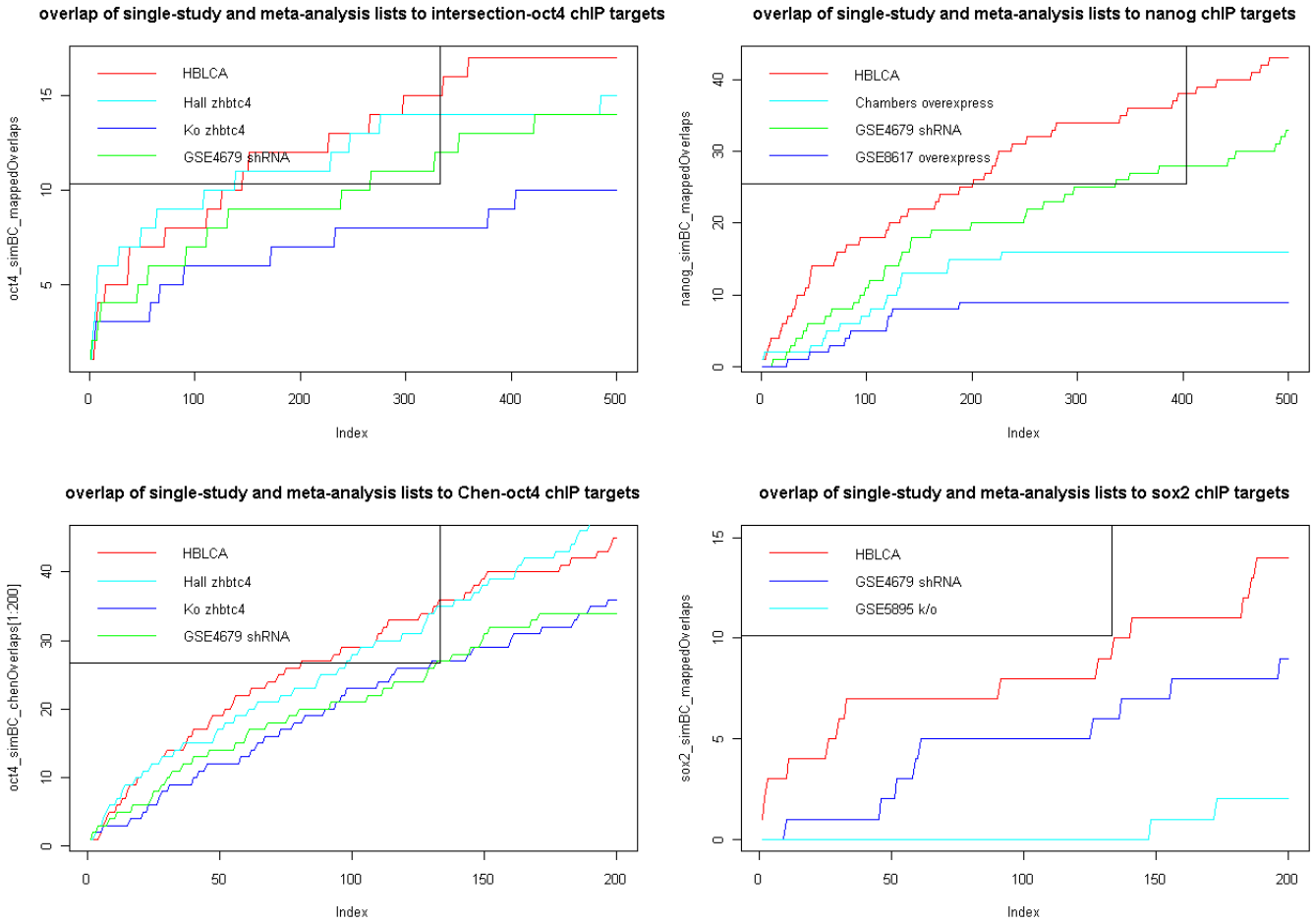


Figure 5.19: DNA-binding targets identified by HBLCA (shown in red) and standard statistical analysis of individual targeted transcriptional profiling experiments. Numbers of targets present in each ranked genelist are plotted against the length of the genelist. Genelists for individual datasets obtained through statistical assessment of differential expression via LIMMA.

expression of both the gene of interest and the gene(s) with co-dependent expression with the gene of interest. Therefore, some genes that are bound by a TF may not have related expression patterns (across a large dataset) to that TF, and some genes with clearly corresponding expression patterns (and some underlying transcriptional relationship) with the TF may not show binding by the TF. Despite this possibility of being a distorted proxy to representing a target set of known transcriptional relationships involving a TF of interest, the DNA-binding lists obtained from publicly available high throughput ChIP data do at least provide a means of testing some direct transcriptional relevance of the resulting genelists created through directed meta-analysis of gene expression data. And by such measures, the HBLCA approach presented in Section (5.1) does appear to result in lists of associated genes (for a given TF of interest) with estimated expression co-dependency scores reflecting some transcriptionally relevant patterns even more clearly than alternative approaches involving meta-analysis or controlled transcriptional profiling experiments involving genetically altered expression of the gene of interest, methods that have been shown to be effective tools in biological research.

5.2.3 Discussion

The results presented in Section (5.2) indicate that the HBLCA approach to meta-analysis successfully identifies genes with known DNA-binding or consistent gene expression patterns in relevant biological contexts, and thus likely transcriptional relationships, in its lists of genes identified as displaying gene expression co-dependency patterns involving a gene of interest in particular biological contexts. The HBLCA method achieves this at least as effectively as, and often considerably more effectively than well-established methods of generating similar predictions based on analysis of gene expression data. The levels of enrichment of respective target lists shown in this section suggest that the identification of genes with the desired expression patterns performed by the HBLCA method forms a useful tool for inferring transcriptional relationships between genes, and thus to utilise existing gene expression data for the study of biological processes.

Motivations for integrating DNA-binding data and gene expression data in order to predict transcriptional regulatory targets of particular TFs have been discussed, and a number of observations relating to the similarities between DNA-binding results obtained for the same TF in different experiments have been presented. These observations (especially those shown in Figs. 5.14 & 5.15) support the case for taking a meta-analysis approach to data from ChIP studies, in addition to taking such an approach with gene expression studies. The lack of correlation of expression across a large gene expression dataset of the majority of DNA-binding targets of a number of TFs with the respective binding TF is illustrated in Fig. 5.16, with the implication this has regarding the identification of relevant transcriptionally regulated targets of a given

TF in particular biological contexts being that an integrated meta-analysis approach incorporating meta-analysis of ChIP datasets as discussed above and a gene expression meta-analysis method based on a biclustering approach would appear to be ideally suited to utilise the available data for this target prediction task. Following on from the demonstration (in Section (5.2.2)) of the effectiveness of the HBLCA meta-analysis approach for the identification of genes with DNA-binding by a given TF on the basis of patterns observed in gene expression data, the following section presents the results of taking an integrated meta-analysis approach to identify regulatory targets of TFs particularly relevant to the transcriptional control of pluripotency.

5.3 Integration of ChIP Data with Gene Expression Meta-Analysis Results

High-throughput ChIP assay technologies have enabled the study of genome-wide DNA-binding of any given TF. The knowledge obtained through such study is potentially useful in the elucidation of mechanisms of regulation of gene expression involved in certain biological processes. However, for a number of reasons (mentioned in Sections (3.5.2) and (5.2)) DNA-binding information does not provide a complete picture explaining all aspects of the mechanisms of transcriptional control of biological processes. In order to identify predicted targets of a given TF, a number of studies reported in the literature (e.g. [Sharov et al., 2008, Loh et al., 2006, Chen et al., 2008, Kim et al., 2008, Marson et al., 2008]) have performed whole genome scale ChIP and gene expression assays and integrated the results to obtain lists of predicted target genes with both DNA-binding of the TF in question and correlated gene expression patterns with those of the TF. Following this approach, the HBLCA meta-analysis tool presented in Section (5.1.6) may be utilised in conjunction with (meta-) analysis of data from ChIP studies, resulting in an integrated meta-analysis approach for predicting transcriptional mechanisms of biological processes known to involve TFs of interest. This section presents data supporting the statement of motivation for this integrated approach and providing some insight into DNA-binding as assessed by ChIP assays, followed by the description of a number of possible methods for performing such integrated meta-analysis and, finally, results of the application of this integrated meta-analysis approach to TF-target prediction are shown.

Motivated by the observations outlined in the previous section, and by the fact that HBLCA approach presented in Section (5.1.6) has been demonstrated to be able to identify significant gene expression (co-)dependency patterns involving TFs and their DNA-binding targets (as shown in Section (5.2.2)), it was proposed that this approach be used in conjunction with meta-analysis of ChIP data to identify predicted targets of a TF: genes with both reliable DNA-binding by the TF in question and significant patterns of codependent expression with the TF across relevant subsets of samples

from a large collection of gene expression data. Examples of such integrated analysis are presented in this section, and applications of a related integrated meta-analysis approach to the identification of gene expression patterns associated with different combinations of interacting TFs are described in Section (6.3), with significant results concerning the study of transcriptional control of pluripotency.

There may be a number of ways of integrating data from multiple ChIP studies and gene expression datasets, however the simplest approach involves generating lists of high-confidence DNA-binding targets of a TF and finding which of those targets have significant expression co-dependency patterns with the TF in question, across samples representing a biological context relevant to the particular question. If there are multiple ChIP datasets available for the TF in question from relevant biological samples, again there may be a number of ways of utilising these to obtain a single list through meta-analysis, but one straightforward option is to take the intersection of the lists of targets identified by each of the individual studies available. While this will inevitably be very stringent, discounting a larger proportion of possible targets with some binding evidence as the number of available (ChIP) datasets increases, those targets that have evidence in all of the studies available ought to be high-confidence binding targets (especially as the number of datasets available increases). Given that the desired stringency and number of available studies may vary for different tasks, target lists can be obtained by ranking all those targets in any study and selecting a threshold that balances stringency and confidence against the number of targets pursued for further investigation. This approach is taken partly because the results from ChIP studies tend to be expressed as a binary output (e.g. as in [Kim et al., 2008, Marson et al., 2008]): that is, a gene is either bound by a TF in a sample or not. Especially when analysing data from multiple different technologies, this may be the best possible approach: it would be difficult to compare relative abundances of binding in samples based on essentially arbitrary measurements from different platforms.

If only a single study is available, there still may be methods to rank potential targets based on the measurements obtained. For example, a method has been developed for ChIP-seq data to rank targets according to a Transcription Factor Association Score (TFAS) [Ouyang et al., 2009]. Using a study with Oct4 DNA-binding data [Chen et al., 2008] as an example, ranking according to this TFAS results in greater correspondence to a list of 'reliable targets' obtained through the intersection of target lists from all available Oct4 ChIP studies than that expected from a random ranking of those genes identified as binding targets in the study, as demonstrated with the ROC curve plotted in Fig. 5.20.

This would suggest that the application of such approaches may prove to be effective when raw ChIP-seq data is available, especially when such raw data becomes available

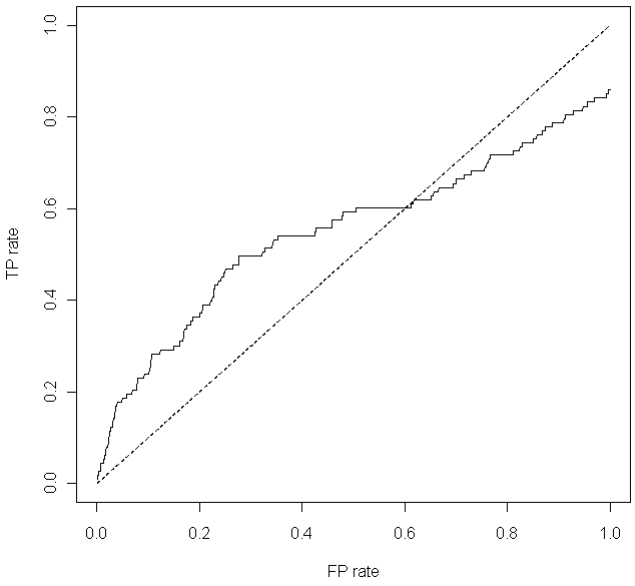


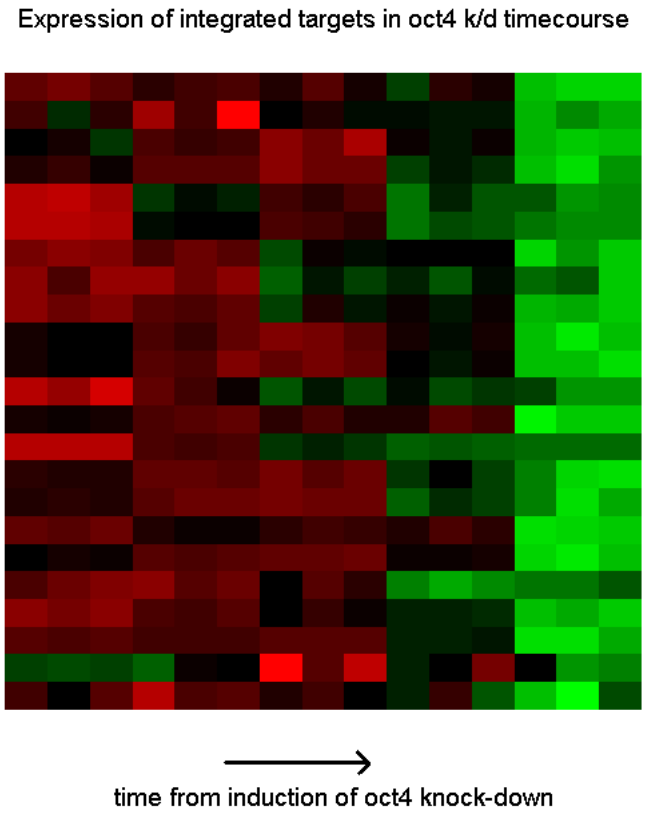
Figure 5.20: ROC for TFAS-ranked ChIP targets in a single study [Chen et al., 2008] belonging to a set of targets consistently bound in multiple studies. The proportion of consistently-bound genes present in the TFAS-ranked genelist from the [Chen et al., 2008] study is plotted against the length of genelist (corresponding to the proportion of those genes specifically identified as Oct4-bound in the [Chen et al., 2008] study that appear in the ranked genelist).

from multiple relevant studies (allowing new meta-analysis methods to be developed to take advantage of the information in the raw data). However, for the investigation of the roles of Oct4, Sox2 and Nanog in transcriptional control of pluripotency, and particularly for the prediction of regulatory targets of Oct4 presented in the following section, target lists were available from multiple studies from different platforms and so the application of a TFAS or similar scoring system to meta-analysis target lists was irrelevant.

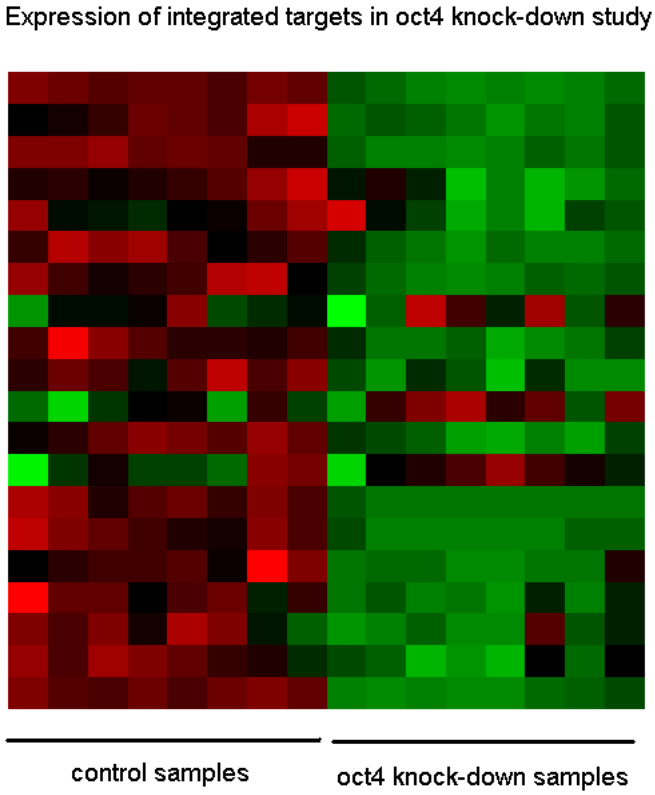
Taking the approach described above to obtain lists of Oct4 DNA-binding targets at different confidence levels (from data from multiple ChIP studies [Kim et al., 2008, Chen et al., 2008, Sharov et al., 2008, Marson et al., 2008]), the HBLCA approach was applied to identify significant gene expression co-dependency patterns between Oct4 and any of the high-confidence Oct4 DNA-binding targets. As a (fairly arbitrary) means of obtaining a list of the highest-confidence targets, genes were identified as likely Oct4 targets (in ES cells) if they were in both the set of genes with the 100 highest scoring co-dependency probabilities as evaluated by the gene expression meta-analysis algorithm using Oct4 as a guide gene and the set of genes identified as Oct4 binding targets in all the available ChIP studies. As an illustration of the results of this simple integrated meta-analysis method for target prediction, heatmaps showing the expression levels of the resulting predicted Oct4 targets in two Oct4 knock-down experiments [Hall et al., 2009, Ivanova et al., 2006] (not included in the dataset used for gene expression meta-analysis) are presented in Fig. 5.21.

These expression heatmaps show clearly that the high-confidence predicted targets obtained with the method described above, incorporating meta-analysis of multiple ChIP studies and the HBLCA method for assessing co-dependent gene expression patterns in a biological context of interest from a large collection of gene expression data, appear to be reliably transcriptionally responsive to forced changes in expression of the supposed regulating TF of interest.

To provide a further indication of the reliability of the expression co-dependency patterns identified with the HBLCA approach, lists of differentially expressed genes in a number of targeted transcriptional profiling experiments were obtained, ranked according to statistical significance of differential expression as calculated with the Empirical Bayes method from LIMMA (as described in the Section (5.2.1)). Lists of high-confidence predicted targets for Oct4 and Nanog were obtained through integrated meta-analysis as described above, although for Oct4 a larger list was obtained through selecting those genes with Oct4 DNA-binding in a particular individual study [Chen et al., 2008] featuring in the list of the 100 highest scoring genes according to the HBLCA meta-analysis approach. For one Nanog knock-down experiment [Ivanova et al., 2006] and two Oct4 knock-down experiments [Hall et al., 2009,



(a) Oct4 knock-down time-series [Hall et al., 2009]



(b) Oct4 shRNA knock-down [Ivanova et al., 2006]

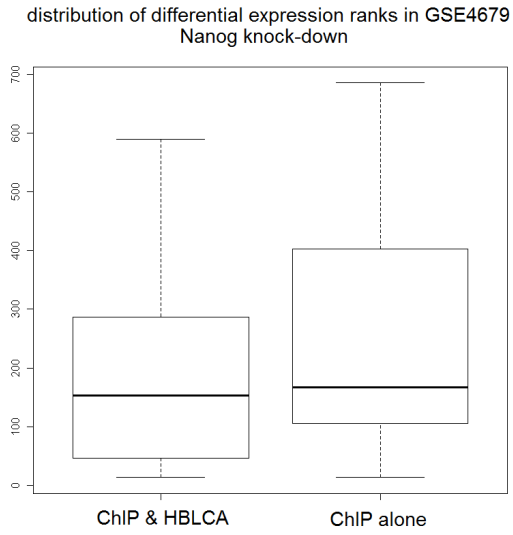
Figure 5.21: Heatmaps showing expression of predicted Oct4 targets in Oct4 knock-down experiments (each predicted target has a row of values across all replicates in the experiments, with red squares indicating relatively high expression of that gene in that sample and green samples relatively low expression)

Ivanova et al., 2006], the differential-expression significance rankings in each experiment were obtained for the relevant predicted target list and the list of genes with consistent DNA-binding across the relevant ChIP studies. Box plots are shown in Fig. 5.22 to illustrate the differences in the distribution of such rankings when the list of reliable DNA-binding targets is filtered on the basis of gene expression patterns observed in a different, large dataset (as is the case with the procedure performed here).

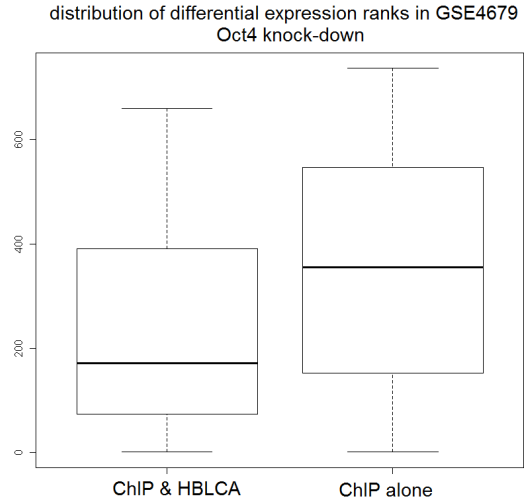
The differential expression ranking distributions shown in Fig. 5.22 show that, as expected, identification of DNA-binding targets that show patterns of expression co-dependency with the binding TF across relevant samples in a large dataset are likely to be more significantly differentially expressed upon forced alteration of the expression of that TF than those DNA-binding targets without such expression co-dependency observations.

As a further demonstration of the effectiveness of this integrated meta-analysis approach, predicted targets of another TF (Klf2) were identified through obtaining a list of genes with patterns of codependent expression with Klf2 in ES cells and selecting those genes with the strongest apparent expression co-dependency patterns that additionally showed binding by Klf4 (a related TF for which ChIP data was available). It was expected that Klf2 would share a significant proportion of targets with Klf4 due to their similarity in protein structure/function [Pearson et al., 2008], with redundancy noted in [Jiang et al., 2008]. It was noted that the level of Klf2 expression dropped significantly in two Oct4 knock-down time series microarray experiments [Hall et al., 2009, Sharov et al., 2008], therefore it was proposed that it would be interesting to see whether any of the predicted Klf2 targets also showed differential expression in these experimental datasets, especially if that differential expression corresponded to the differential expression of Klf2. Fig. 5.23 shows the expression levels of the Klf2 predicted targets (obtained using the novel integrated meta-analysis approach) throughout each time series (in terms of fold-change to the median of the control set of replicates for that experiment), with the colour of each line (gene) dependent on the final expression level of the gene. Fewer genes are shown for the plot on the right hand side, as this experiment [Sharov et al., 2008] was performed on a custom microarray platform and not all predicted targets could be mapped to appropriate probesets on this platform (conversion between the Affymetrix probeset IDs and the custom array of [Sharov et al., 2008] was performed by mapping respective identifiers to official gene symbols, as this provided greater overlap than any other available annotation).

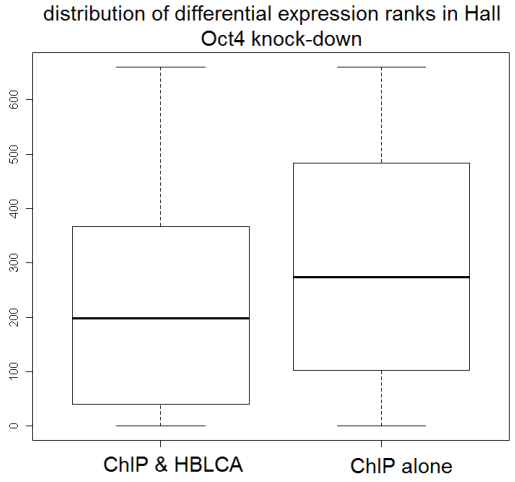
The expression profile plots shown in Fig. 5.23 indicate that the integrated meta-analysis approach may be able to identify transcriptional targets of a given TF that respond even to relatively minor fluctuations in the expression level of the TF. In addition, the examples of differential expression in distinct datasets shown with Klf2 targets



(a) Nanog shRNA study

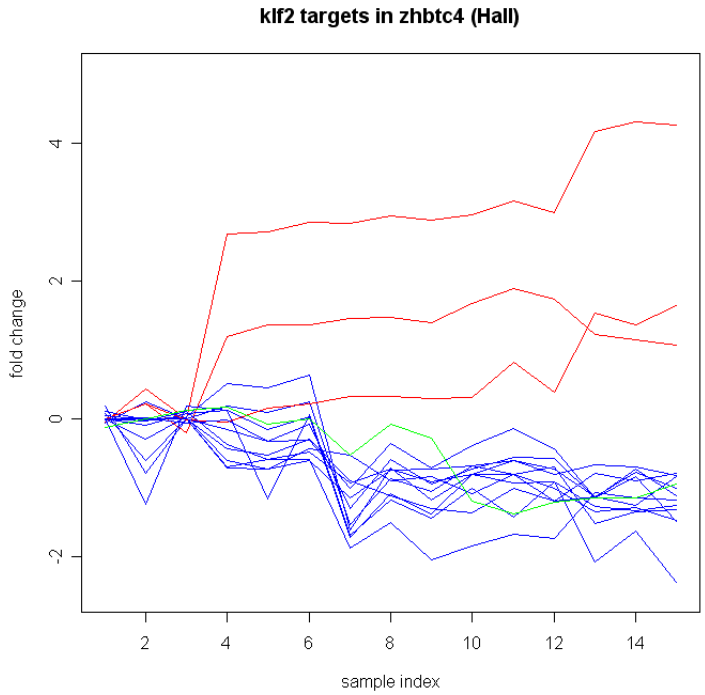


(b) Oct4 shRNA study

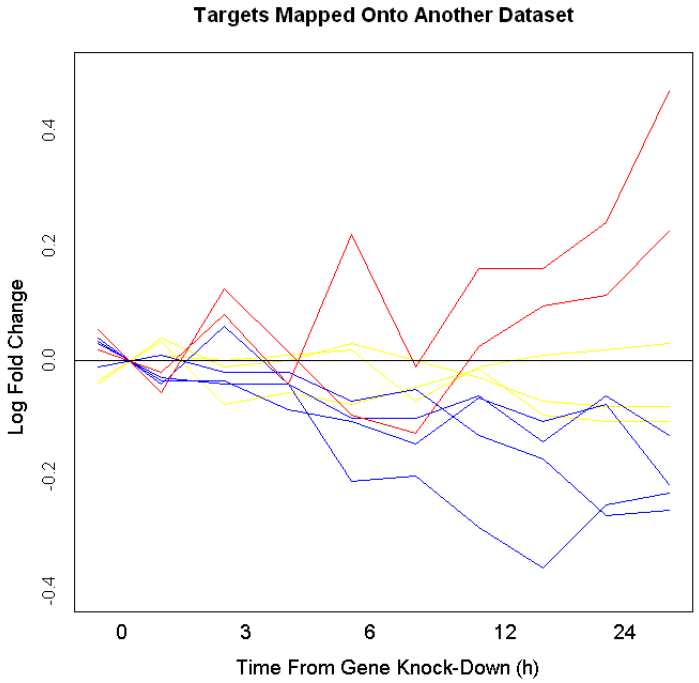


(c) Oct4 tet-inducible knock-down study

Figure 5.22: Distributions of differential expression rankings of DNA-binding targets with and without filtering according to gene expression co-dependency patterns as identified using the HBLCA approach presented in Section (5.1.4). Differentially-expressed genelists were calculated for 'held-out' datasets using LIMMA. Rank distributions show that genes with TF binding and TF expression co-dependency discovered by HBLCA, shown in the left-hand boxplot in each panel, are generally more significantly differentially expressed in the held out TF knock-down datasets than an equivalent randomly-selected set of genes with binding by the TF. Median profile across each dataset shows almost no change in expression across the respective dataset.



(a) Oct4 knock-down time-series [Hall et al., 2009]



(b) Oct4 knock-down time-series [Sharov et al., 2008]

Figure 5.23: Profile plots showing fold-change of predicted Klf2 targets in Oct4 knock-down experiments

predicted using DNA-binding data for Klf4 indicate that this novel target prediction method using integrated meta-analysis may be robust to inaccuracies in the underlying data (as represented by using DNA-binding data from a different TF in the same family). The demonstrations provided above are intended to illustrate the potential for the integrated meta-analysis approach described in this section to be applied to prediction of relevant transcriptional targets of TFs of interest in particular biological contexts so that the roles of such genes in biological processes of interest may be investigated experimentally, and that this target prediction may be performed utilising data already existent in the public domain.

5.3.1 Discussion

An integrated meta-analysis approach to the identification of a TF's regulatory targets relevant to a given biological context is described in Section (5.3.3), along with the results of its application to target prediction tasks involving Oct4 and Klf2. This novel approach has been demonstrated (in Fig. 5.22) to improve the likely significance of differential expression of predicted targets in response to a change in expression of the TF expected to regulate their expression, compared to the lists obtained by using either data source (DNA-binding or gene expression) in isolation. In addition, examples of successful application of this integrated meta-analysis approach to transcriptional target prediction are illustrated with Figs. 5.21 & 5.23, showing how high-confidence targets predicted through the integrated meta-analysis approach display differential expression in response to differential expression of the relevant TF expected to regulate their expression, even when the differential expression of the TF is not especially pronounced (as in the examples shown in Fig. 5.23).

The integrated meta-analysis approach proposed in this section has been demonstrated to be an effective means of utilising the HBLCA approach presented in Section (5.1.6) in conjunction with data from DNA-binding studies in order to predict transcriptional regulatory relationships involving TFs of interest. Such a method may have many potential applications in biological research, where prediction of transcriptional regulation of one set of genes by another may guide experimental investigation of the transcriptional control of a given biological process in such a way that expenditure of time and resources may be reduced without compromising the impact of the results of the research.

Additionally, the investigation of DNA-binding data from multiple ChIP studies presents further opportunities for the HBLCA approach to be applied to identify potentially interesting relationships between expression levels of certain TFs or combinations of TFs and the consequences in terms of expression level of genes with proximal DNA bound by these TFs. In the case of study of the transcriptional control of pluripotency, this would be especially useful as a relatively large number of TFs have been

implied to have roles in the transcriptional control of pluripotency but the precise roles of each and therefore the consequences of expression of different combinations of these TFs is largely unknown [Chambers and Tomlinson, 2009]. Sections (6.2-6.4) present such investigations of gene expression patterns involving DNA-binding targets of TFs of interest, using the HBLCA approach to meta-analysis of gene expression data.

5.4 Chapter Summary

Building on the work presented in the previous two chapters, a novel biclustering-based meta-analysis approach, HBLCA, was developed and is described in Section (5.1). This approach to meta-analysis of gene expression data focuses on the identification of expression co-dependency patterns within consistent biological contexts (as defined by the global transcriptional profile of a sample), made possible by the introduction of a probabilistic framework for estimating the significance of observed expression variations of any genes between any two sets of samples in a dataset. This represents a probabilistic framework for the concept of localised gene expression co-dependency that was introduced in Section (3.6.3). By implementing such a localised gene expression co-dependency analysis approach, the HBLCA method enables the prediction of transcriptional relationships of particular biological relevance involving genes of interest, as demonstrated in Sections (5.2 & 5.3). Additionally, this approach enables investigation of the dependence on biological contexts of observed transcriptional relationships between particular genes and of the transcriptional mechanisms by which TFs of interest may influence a characteristic phenotype, yielding insight into transcriptional relationships that is not provided by any existing gene expression data analysis tools, which is to be demonstrated in the following chapter. This novel analysis tool exists as a set of R programs that incorporate the dataset compilation and novel gene expression-state modelling transformation tools described in Sections (3.3 & 4.1.4) respectively. A bicluster visualization tool has been developed to assist with interpretation and confirmation of predicted relationships discovered by this novel analysis approach, and can be used in conjunction with the analysis approach as an additional analysis tool for investigating transcriptional relationships involving genes of interest, as demonstrated in Section (5.1.7).

It was shown through analysis of data from a number of genome-wide ChIP studies that there may be considerable between-study variation in terms of conclusions relating to target genes bound by individual TFs, and that the majority of even those genes reliably identified as targets of a given TF across multiple studies do not show clear correlation of expression with the TF in question across a large collection of gene expression data. By integrating analysis of genome-wide DNA-binding data with the output of the HBLCA tool, it was demonstrated that high-confidence regulatory targets of a TF of interest could be identified that display relevant expression patterns in 'held-

out' datasets coinciding with a change in expression level of the binding TF.

After the development of a set of gene expression analysis approaches described in this chapter, and demonstration of the success of these approaches in identifying relevant transcriptional relationships from large collections of gene expression data, the following chapter presents a number of investigations of the transcriptional control of pluripotency carried out using the analysis approaches introduced in this chapter. The results of these investigations provide further evidence for the value of the contribution to the repertoire of tools available to a biologist represented by the novel analysis approaches described in this chapter.

Chapter 6

Transcriptomic Analysis Of Pluripotency

In this final chapter describing work performed in the course of the research project, a number of examples are given to show successful application of the HBLCA approach to open problems in biological research regarding investigation of the transcriptional control of pluripotency. The results presented constitute findings that would not have been obtainable without the development and application of the analysis approaches presented in the previous chapter (for which development was, in turn, dependent on the work presented in Chapters 3 & 4).

6.1 Functional Decomposition of a List of Genes Differentially Expressed Upon Pou5f1 Knockdown

As mentioned in Section (2.1.4), Oct4 (Pou5f1) is a transcription factor that is essential for establishment and maintainance of the pluripotent state that is characteristic of mouse ES cells [Nichols et al., 1998, Niwa et al., 2000]. As part of a study performed to investigate the mechanisms involved in Oct4's transcriptional control of pluripotency, motivated partly by the observation that the majority of well-characterised Oct4 targets appear not to be essential for derivation or maintainance of ES cells, a transcriptional profiling experiment was performed by [Hall et al., 2009] using Affymetrix MOE430v2 microarrays to measure genome-wide expression levels in replicate samples obtained from Zhbtc4.1 ES cells [Niwa et al., 2000] following treatment with doxycycline (inducing rapid downregulation of Oct4 mRNA and protein in these cells) and from untreated samples. As described in [Hall et al., 2009], for Zhbtc4.1 ES cell cultures grown for 0hrs, 5hrs, 10hrs and 30hrs with doxycycline treatment, three RNA samples were extracted using the QIAGEN RNeasy Mini Kit then labelled and hybridised to separate microarrays according to the $5\mu g$ standard Genechip protocol. Following array scanning, normalized log-intensity gene expression measurements were obtained through application of RMA [Irizarry et al., 2003]. Relative expression levels as measured by the arrays were confirmed through qRT-PCR for a number of genes, as shown in [Hall et al., 2009]. This gene expression dataset provides a resource for obtaining valuable insight into the transcriptional mechanisms of control of the pluripotent state, particularly in terms of the role of Oct4.

Standard statistical analysis approaches can be used to identify genes with significant expression variation between one time point and another (or the untreated control). In order to identify genes with an immediate transcriptional response to a drop in the level of Oct4 expression, a list was obtained of the genes with the greatest measured change in expression level between the samples profiled 10hrs following dox treatment and those profiled without dox treatment (i.e. $0hr$ time point). This list of genes is included in Table 6.1. As a means of exploring a possible functional interpretation of this list of targets and identifying shared involvement of any of these targets in particular biological processes, Gene Ontology (GO) analysis was performed by using the

DAVID [Huang et al., 2009b] functional annotation tool to test statistical enrichment of *Biological Process* ontology terms within the genelist. Such GO-based functional enrichment analysis is a widely used technique for investigating functional implications of a given genelist [Huang et al., 2009a], however, such analysis is dependent on the availability of accurate and up-to-date annotations. Due to structure of the GO annotations, there is no context-specific annotation. That is to say, there is no distinction between those processes in which a given gene may be involved in only one particular biological context and those processes in which that gene may be universally involved. Additionally, these analyses test for statistical enrichment of annotation terms within a whole list of genes, which may in fact involve a heterogenous composition of smaller groups of genes associated with distinct but co-occurring biological processes. Possibly arising as a consequence of these features of standard GO enrichment analysis, there were no biological process annotation terms enriched below a FDR threshold (after Benjamini-Hochberg correction [Benjamini and Hochberg, 1995]) of 0.1 in the list of Oct4 early targets given in Table 6.1.

It was proposed that the HBLCA tool presented in Section (5.1.6) could be used to provide some means of counteracting the complicating features of functional enrichment analysis described above. HBLCA could be used to identify lists of genes with apparent expression co-dependency with each of the predicted targets in the original genelist, across relevant biological contexts (in this case, ES cells and some very similar non-ES samples). These bicluster-associated genelists for each of the targets in the original genelist could be used to provide more context-specific functional annotation through enrichment analysis of each of the associated genelists, which would additionally be more robust to inaccuracies of annotation due to the increased number of annotations from which relevant functional association is determined. Additionally, the associated genelists could be used as a means of separating the possibly heterogenous input genelist into groups of genes with similar expression patterns within the relevant biological context(s), in a similar manner to the clustering of individual bicluster genelists involved in the bicluster-evidence integration procedure described in Section (5.1.5). This section describes in further detail the way in which these steps may be performed to counteract the complicating features of functional enrichment analysis of a genelist, along with the results of application of the procedure to analysis of the Oct4 early-target genelist mentioned above. Results from similar application of the procedure to associated genelists identified using a related correlation-based method are presented for comparison, with the advantages of the biclustering-based meta-analysis approach for gene association highlighted.

6.1.1 Discovering Structure Within a Genelist Through Biclustering

If a genelist is insufficiently annotated using standard functional enrichment analysis, as was the case with the list of early Oct4 targets given in Table 6.1, lists of bicluster-

ProbeSet	Common Name	Description
1436799_at	D230005D02Rik	RIKEN cDNA D230005D02 gene
1415983_at	Lcp1	lymphocyte cytosolic protein 1
1429338_a_at	Nol9	nucleolar protein 9
1432004_a_at	Dnm2	dynamin 2
1420948_s_at	4833408C14Rik	RIKEN cDNA 4833408C14 gene
1458716_at	Dusp27	dual specificity phosphatase 27 (putative)
1421313_s_at	Cttn	cortactin
1452811_at	Atic	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
1430162_at	3830417A13Rik	RIKEN cDNA 3830417A13 gene
1420901_a_at	Hk1	hexokinase 1
1452148_at	Lrpap1	low density lipoprotein receptor-related protein associated protein 1
1422210_at	Foxd3	forkhead box D3
1438133_a_at	Cyr61	cysteine rich protein 61
1416516_at	Fscn1	fascin homolog 1, actin bundling protein (Strongylocentrotus) purpuratus)
1449231_at	Zfp296	zinc finger protein 296
1423952_a_at	Krt2-7	keratin complex 2, basic, gene 7
1427770_a_at	Slc2a3	solute carrier family 2 (facilitated glucose transporter), member 3
1427385_s_at	Actn1	actinin, alpha 1
1417945_at	Pou5f1	POU domain, class 5, transcription factor 1
1421924_at	Slc2a3	solute carrier family 2 (facilitated glucose transporter), member 3
1420998_at	Etv5	ets variant gene 5
1416515_at	Fscn1	fascin homolog 1, actin bundling protein (Strongylocentrotus) purpuratus)
1417760_at	Nr0b1	nuclear receptor subfamily 0, group B, member 1
1421749_at	—	—
1434357_a_at	Kpnb1	karyopherin (importin) beta 1
1450929_at	Zfp57	zinc finger protein 57
1426538_a_at	Trp53	transformation related protein 53
1448152_at	Igf2	insulin-like growth factor 2
1448169_at	Krt1-18	keratin complex 1, acidic, gene 18

Table 6.1: Oct4 early-response genes

associated genes for each target in the original list can be obtained by performing HBLCA meta-analysis as described in Section (5.1.6) using each target in turn as the guide gene for the analysis. In the case of the list of early Oct4 targets, this analysis was performed using the large collection of microarray data presented in Section (3.3), across only those sample groups (pre-calculated as in Section (5.1.4)) in the dataset involving samples with high levels of expression of Oct4 or Nanog (to represent ES cell samples or similar), as this was the biological context of interest in the investigation. From each such meta-analysis with a predicted target as a guide gene, a list of target-associated genes was selected as the highest-scoring 100 genes in the 'globally' integrated gene set (those from Bayesian integration performed across the full set of biclusters) given that all biclusters identified by the algorithm were deemed to be relevant. It was observed that for some of the predicted targets no significant biclusters were identified, indicating that there was not sufficient data in the dataset to distinguish a drop in the expression level of that gene in ES cells or similar samples from a general loss of the ES cell transcriptional profile. For those genes in the original target list that did give rise to biclusters and resulting associated genelists, these genelists could be used for further investigation of the original list of targets.

Functional Analysis of Bicluster Genelists

To provide an alternative to standard functional enrichment analysis of the input genelist (using statistical testing as performed by DAVID [Huang et al., 2009b]) that provides a more context-specific functional annotation and is more robust to erroneous (or missing) annotations in the GO database, the associated genelists calculated as described in the previous paragraph can be tested using standard functional enrichment analysis to identify any GO terms significantly enriched in the list of associated genes (based on gene expression patterns observed across relevant biological contexts). This was performed for the bicluster-associated genelists for the Oct4 early-response target list, resulting significant annotation associations for each gene are shown in Table 6.3.

Using those GO biological process terms associated to each of the genes in the original target genelist, statistical enrichment p-values were calculated from the hypergeometric distribution (with the number of probesets annotated with each term obtained from Affymetrix annotation data in *NetAffx* [Affymetrix,]). The results of enrichment analysis of bicluster-associated GO terms (as opposed to the GO-annotated biological process terms for each gene, which yielded no significant enrichments) are shown in Table 6.4. As such p-values were not obtained for all GO terms, multiple testing correction through the FDR family of methods (e.g. Benjamini-Hochberg [Benjamini and Hochberg, 1995]) was not possible, but multiple testing correction may be desired as the number of categories that each gene *may* have been annotated with is large (approximately $10,000$). As it is therefore unclear whether multiple testing correction is strictly necessary here (and has already been taken into account in the individual

Table 6.2:

ProbesetID	GeneSymbol	GO Category	Enrichment FDR
1416157_at	Vcl	RNA localization	2.20E-004
1416515_at	Fscn1	RNA localization	5.90E-003
1416516_at	Fscn1	RNA localization	6.00E-003
1417752_at	Coro1c	RNA processing	1.10E-002
1417945_at	Pou5f1	stem cell differentiation	3.40E-006
1418078_at	Psme3	chromatin organization	8.70E-003
1420647_a_at	Krt8	Calcium-independent cell-cell adhesion	6.00E-002
1421313_s_at	Cttn	cell cycle	3.80E-003
1421811_at	Thbs1	cell adhesion	8.00E-007
1423691_x_at	Krt8	regulation of cell proliferation	1.30E-004
1427385_s_at	Actn1	cell cycle	3.60E-002
1427408_a_at	Thrap3	M phase of mitotic cell cycle	2.20E-005
1427550_at	Peg10	cell cycle	1.90E-002
1427739_a_at	Trp53	stem cell differentiation	2.80E-005
1427770_a_at	Slc2a3	M phase of mitotic cell cycle	2.30E-004
1429338_a_at	Nol9	RNA processing	1.10E-004
1429802_at	Hsd17b14	embryonic morphogenesis	7.80E-002
1432004_a_at	Dnm2	M phase	5.10E-002
1434357_a_at	Kpnb1	nuclear transport	3.70E-005
1435989_x_at	Krt8	Calcium-independent cell-cell adhesion	6.60E-002
1448169_at	Krt18	negative regulation of cell proliferation	5.20E-002
1449578_at	Supt16h	RNA localization	6.30E-002
1449898_at	Sept1	blood vessel development	3.70E-003
1450576_a_at	Sf3a2	chromatin organization	1.40E-002
1450929_at	Zfp57	chromosome organization	8.10E-002
1451782_a_at	Slc29a1	response to DNA damage stimulus	5.10E-003
1451927_a_at	Mapk14	RNA processing	5.10E-003
1452811_at	Atic	RNA processing	3.30E-002
		M phase	6.20E-002

Table 6.3: Bicluster-associated GO annotations for Oct4 early-response targets, obtained for each target by analysis in DAVID of top ranking 100 genes most associated with the corresponding probeset according to HBLCA output.

enrichment calculations for each bicluster-associated genelist), and it was only possible to perform the highly conservative Holm method of p-value adjustment [Holm, 1979] (which may well be overly conservative [Hochberg and Benjamini, 1990]), both adjusted and un-adjusted p-values from the hypergeometric distribution enrichment calculations are given in Table 6.4 for each bicluster-associated GO term statistically enriched in the list given in Table 6.3.

It is clear from the enriched category list given in Table 6.4 that by using the HBLCA meta-analysis tool to obtain lists of genes with relevant expression pattern associations for each of the genes in an input genelist and performing functional enrichment analysis on each of these lists of associated genes to annotate the original input list, certain biological processes were identified as having significantly enriched association to the input gene list where no such enrichments were identified using a standard functional enrichment analysis technique. This application of HBLCA has resulted in significant functional signatures being identified in a target list that would not have been discovered using existing functional analysis techniques. This analysis indicates that transcriptional regulation of RNA localization, cellular response to stress and the cell cycle (M-phase), chromatin organisation and stem cell differentiation occurs when the level of Oct4 mRNA in ES cells drops. Oct4-based transcriptional regulation of cell adhesion, cytoskeletal organization and proliferation is also suggested.

Clustering List of Differentially Expressed Genes

The functional enrichment analysis approach described above resulted in significant biological process signatures being identified in a genelist where standard methods failed, due to the ability to associate genes to processes they appear to be involved in even if those annotations are not available. However, this approach does not address the other significant issue regarding standard functional enrichment analysis techniques: a genelist may well be functionally heterogeneous and as a result, biological process terms associated to a significant subset of the genes in a genelist may not be evaluated as significant due to the size of the remaining genelist not in that subset. If there is structure within the genelist involving multiple components comprising genes with similar functional associations that are linked to expression pattern similarities of those genes, it may be possible to use the bicluster-associated genelists to identify such structure and evaluate biological process signatures individually for each subset of the original list. In addition to the possible benefits in terms of the identification of functional signatures in heterogeneous genelists, it may be useful to identify structure based on expression patterns of genes in a potentially heterogeneous genelist for other purposes (such as identifying potential co-regulators for targets of a given TF, as demonstrated in Section (6.2)).

category	count	overall	probes in GO term	P-value	adjusted p
cell adhesion	3	35	1189	6.40E-002	1
chromatin organization	2	35	17	7.90E-005	0.59
regulation of cell proliferation	3	35	983	4.00E-002	1
actin cytoskeleton organization	2	35	248	1.60E-002	1
cell cycle	7	35	1456	1.10E-004	0.82
RNA localization	4	35	4	3.00E-013	2.20E-009
cellular response to stress	3	35	5	4.30E-009	3.20E-005
M phase	6	35	12	1.30E-016	9.70E-013
response to DNA damage stimulus	2	35	558	7.00E-002	1
stem cell differentiation	2	35	19	9.90E-005	0.74

Table 6.4: Enriched bicluster-associated GO annotations in Oct4 early-response target list

In a manner similar to that used in the identification of structure within a possibly heterogenous set of biclusters described in Section (5.1.5), a gene-association matrix can be constructed from the integrated biclustering results obtained through application of the HBLCA tool to an appropriate dataset, with each gene in the input list being used in turn as a guide gene. Such a gene-association matrix can be constructed with each column containing the co-dependency scores for all genes, for each gene in the original target list. Using PCA as described in Section (5.1.5), the dimensionality of this matrix with a large number of rows can be reduced to a more manageable number of principal components. To illustrate the structure that may be apparent in a list when taking this approach, the values of the first two principal components of the gene-association matrix were calculated for each gene in the Oct4 early-response target list, with the gene-association matrix created as described above from the output of the HBLCA tool. Fig. 6.1 shows each gene in the Oct4 early-response target list plotted with these values for each of the first two principal components used as x and y co-ordinates.

Following dimensionality reduction to the minimum number of principal components that explain at least a set proportion of the variation across the matrix, k-means clustering can be used in conjunction with the gap statistic (using `kmeansGap` from the R package `SLmisc`) to identify optimal clustering of the input genelist based on the gene expression meta-analysis results observed. When this process was performed on the Oct4 early-response target list, 3 significant clusters of genes were identified in the original input list. One immediately interesting observation was that even though the clustering was performed on the basis of gene expression data from the large meta-analysis compendium, each of the clusters appears to represent distinct components of expression pattern within the dataset used for the original differential expression analysis to identify the early-response targets. Fig. 6.2 shows the expression profiles across this dataset (the dataset from [Hall et al., 2009]) both in terms of the raw expression values and the GESTr-transformed expression state confidence values. It would appear that the first cluster (in green) represents a component whose expression levels are restored at the final time point, the second cluster (in blue) involves genes with expression levels that appear to stabilise after the original drop at the 10hr time point, and the third cluster (in red) seems to involve genes whose expression level continues to drop between the final two (15hr and 30hr) time points.

While this identification of structure within the Oct4 early-response target list is interesting in itself, it would also be useful to be able to associate functional characteristics (as well as the expression characteristics observed in Fig. 6.2) with each of the clusters within the input genelist.

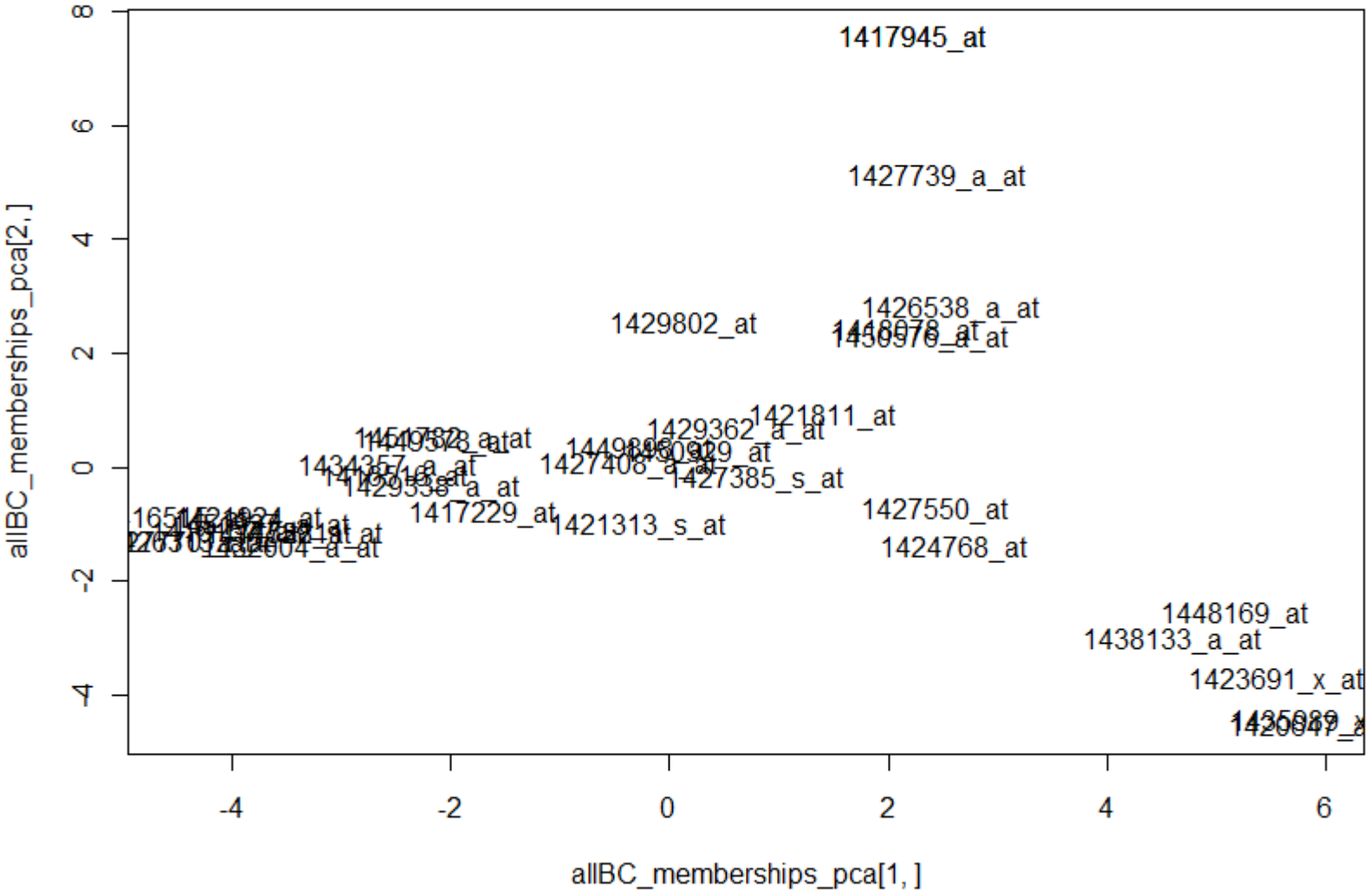
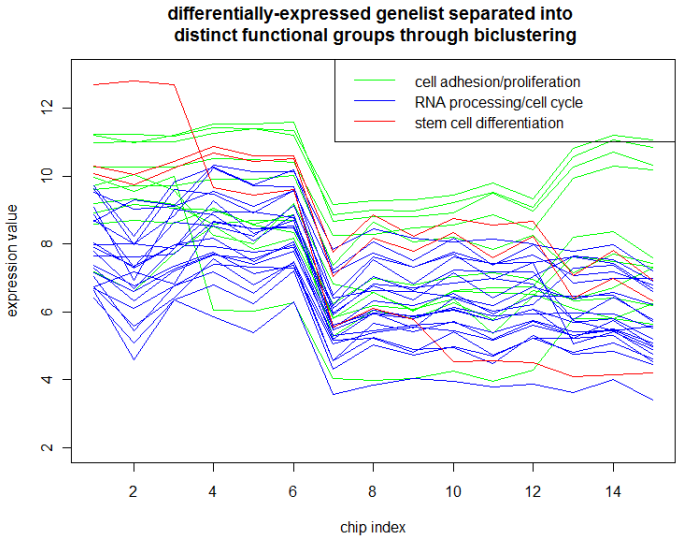
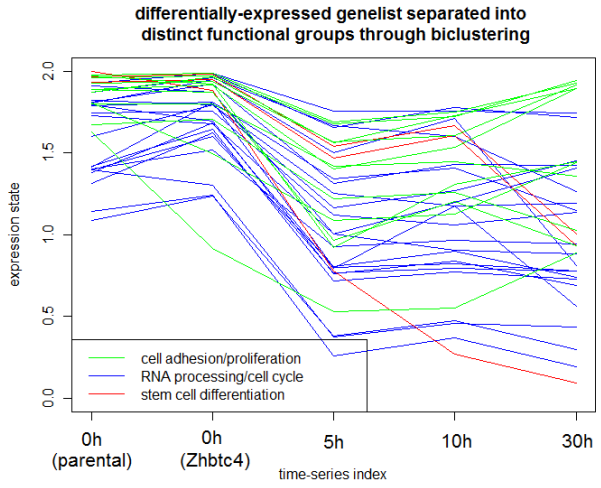


Figure 6.1: Principal component representation of Oct4 early-response target list based on meta-analysis gene expression association patterns. Affymetrix probeset IDs with most significant Oct4-induced response at 10hrs are plotted in horizontal and vertical axes according to representation of that probeset's gene associations (obtained with HBLCA) in the first two principal components of the matrix comprised of the gene associations for all Oct4-induced response genes.



(a) Raw expression values for early-response targets in Oct4 knock-down time series



(b) GESTr-transformed values for early-response targets in Oct4 knock-down time series

Figure 6.2: Profile plots showing expression patterns in [Hall et al., 2009] Oct4 knock-down time series dataset for early-response target clusters. Chip index in plot (a) indicates which of the 15 samples from the experiment the measurements correspond to: chips 1-3 represent the parental cell line (no Oct4 knock-down), chips 4-6 represent Zhbtc4 cells without dox treatment, chips 7-9 represent Zhbtc4 cells after 5h dox treatment, chips 10-12 represent Zhbtc4 cells after 10h dox treatment, and chips 13-15 represent Zhbtc4 cells after 30h dox treatment.

Functional Enrichment Analysis of Decomposed Genelist Subsets

By re-calculating the enrichment of bicluster-associated GO Biological Process terms in the Oct4 early-response target list for each cluster individually, it was apparent that genes in the target list that seemed to have similar gene expression associations (as determined by meta-analysis) were more likely to have similar functional associations (according to functional enrichment analysis of the set of bicluster-associated genes for each predicted target). The first cluster seemed to be associated with cell adhesion, chromatin organisation and regulation of proliferation; the second cluster seemed to be associated with RNA localization, cellular response to stress and M phase (cell cycle); the third cluster seemed to be associated with stem cell differentiation and gastrulation. The bicluster-associated functional enrichments calculated individually for each cluster (as opposed to the list as a whole) is given in Table 6.5.

The results given in Table 6.5 show that this approach has not only increased the significance of the biological process associations to the Oct4 early-response target genelist, it has also provided a means of associating different functions to different components within the original target list. Taken as a whole, this novel approach to functional enrichment analysis based on utilising the HBLCA approach presented in Section (5.1.6) provides an improved method for identifying biological processes associated with a genelist, and further offers a means of identifying structure within a genelist comprising heterogeneous functional components and identifying biological processes (or similar annotation terms) significantly associated with any of these components.

6.1.2 Comparison with Correlation-Based Approach

It has been demonstrated that the novel functional enrichment analysis approach presented in Section (6.1.1) has the ability to provide a number of interesting results where standard functional enrichment analysis approaches fail to offer any insight in to the biological significance of a genelist, and as such this represents a useful application of HBLCA to a real biological research problem (the investigation of functional components involved in the early transcriptional response to a drop in Oct4 expression level). However, it was proposed that the success of the novel functional enrichment analysis approach may not depend exclusively on the existence of the HBLCA tool, and that it may still be successful using a simpler correlation-based meta-analysis approach. In order for a correlation-based meta-analysis approach to be appropriate for the novel functional enrichment analysis approach, pairwise correlations between genes would have to be evaluated across only those samples in the large gene expression dataset that reflect the relevant biological context. In the case of the analysis presented here, the relevant samples involved those samples involved in the ES cell and 'near-ES' sample-groups used for the biclustering-based meta-analysis above. Having identified a relevant gene expression 'context,' associated genelists were obtained for each of the

category	group number	in group	group size	probes in category	P-value	adjusted p
cell adhesion	1	3	11	1189	2.60E-003	1
chromatin organization	1	2	11	17	7.30E-006	6.10E-002
regulation of cell proliferation	1	2	11	983	2.30E-002	1
actin cytoskeleton organization	1	2	11	248	1.60E-003	1
RNA localization	2	4	21	4	3.50E-014	2.60E-010
cellular response to stress	2	2	21	5	2.10E-006	1.60E-002
M phase	2	6	21	12	4.30E-018	3.20E-014
response to DNA damage stimulus	2	2	21	558	2.70E-002	1
cell cycle	2	4	21	1456	4.20E-003	1
stem cell differentiation	3	2	3	19	5.00E-007	3.70E-003

Table 6.5: Enriched bicluster-associated GO annotations in similar subsets of Oct4 early-response target list

members of the Oct4 early-response genelist analysed above by taking the 100 genes with the highest Pearson correlation coefficients with the input gene across the relevant samples. Unlike the biclustering-based meta-analysis approach, this correlation-based meta-analysis identifies associated genes for all inputs regardless of whether there is sufficient contrast across the provided gene expression context to distinguish the desired relationships from any other expression patterns. The DAVID functional association tool was used to identify GO Biological Process terms significantly enriched in each correlation-associated genelist. These significant correlation-associated GO terms for each of the genes in the Oct4 early-response list are given in Table 6.6.

It was noted that, although some of the input genes were associated with a number of biological processes, a markedly lower proportion of the correlation-associated genelists resulted in any significant annotations than with the bicluster-associated genelists. This is partly due to the implicit prediction by the HBLCA approach that requisite data to be able to identify the desired gene expression pattern associations is not available. The contrast between these methods is demonstrated by the proportions of 'associated genelists' obtained through each meta-analysis method that resulted in significant functional annotation, as shown in Fig. 6.3.

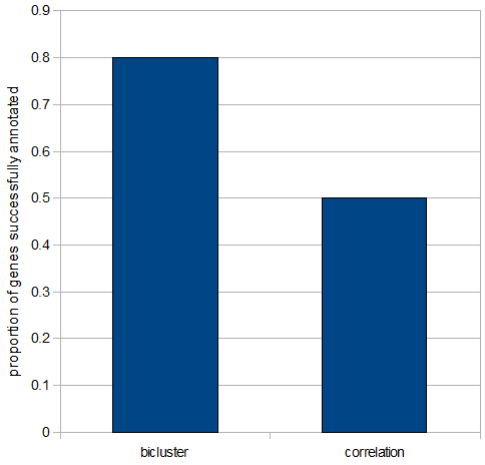


Figure 6.3: Proportion of successful functional annotation through gene expression meta-analysis associations

Having noted this less desirable property of using calculation of Pearson correlation coefficients as a meta-analysis approach in place of the HBLCA tool as incorporated into the novel functional enrichment analysis technique described in Section (6.1.1), obtaining functional associations with the list of Oct4 early-response targets was still successful, albeit to a lesser degree than demonstrated for the co-dependency analysis approach. A list of the significantly enriched biological process terms identified through statistical enrichment analysis of correlation-associated GO terms is given in Table 6.7.

Probeset	Genesymbol	GO category	FDR
1415812_at	Gsn	wound healing	1.90E-002
1416157_at	Vcl	protein localization	2.60E-003
1416515_at	Fscn1	protein localization	7.70E-003
		cell cycle	1.30E-002
1417945_at	Pou5f1	stem cell development	5.00E-004
		embryonic pattern specification	1.90E-002
		embryonic morphogenesis	6.70E-002
		gastrulation	7.30E-002
1418078_at	Psme3	RNA splicing	1.60E-002
1419018_at	Rhox6	placenta development	8.30E-003
1421313_s_at	Cttn	protein localization	1.70E-005
		cell division	6.20E-002
1421811_at	Thbs1	cell adhesion	5.20E-005
		gland morphogenesis	2.00E-003
		blood vessel development	2.30E-002
		tube development	2.50E-002
		negative regulation of cell proliferation	6.40E-002
1421813_a_at	Psap	protein localization	1.50E-003
1421924_at	Slc2a3	protein localization	1.70E-002
		M phase	4.20E-002
		cell division	4.60E-002
1422450_at	Ctnnd1	actin cytoskeleton organization	9.10E-002
		regulation of phosphoinositide 3-kinase activity	3.10E-002
1425329_a_at	Cyb5r3	protein localization	6.10E-003
1425711_a_at	Akt1	lamellipodium assembly	7.10E-002
1426313_at	Bre	protein localization	1.40E-002
1426538_a_at	Trp53	response to DNA damage stimulus	6.40E-002
		cell cycle	6.50E-002
		protein localization	7.20E-002
1427550_at	Peg10	placenta development	1.60E-007
1430162_at	3830417A13Rik	placenta development	1.10E-006
1432004_a_at	Dnm2	protein localization	4.00E-004
1434357_a_at	Kpnb1	protein localization	6.20E-003
1438133_a_at	Cyr61	cell adhesion	7.40E-004
1449578_at	Supt16h	M phase	4.00E-002
		RNA splicing	4.90E-002
1450929_at	Zfp57	negative regulation of cell proliferation	8.10E-003
		lymphocyte differentiation	2.00E-002
		positive regulation of apoptosis	6.10E-002
		embryonic morphogenesis	6.30E-002
1451897_a_at	Nbr1	protein localization	2.60E-002
		cell cycle	3.80E-002
		M phase	4.90E-002
1451927_a_at	Mapk14	protein localization	1.40E-003
		M phase	5.00E-002
		cell division	6.20E-002
1452811_at	Atic	M phase	7.00E-002
1456080_a_at	Serinc3	protein localization	1.10E-003
1459897_a_at	Sbsn	placenta development	1.00E-002

Table 6.6: Correlation-associated GO annotations for Oct4 early-response targets, obtained for each target by analysis in DAVID of top ranking 100 genes most correlated with the probeset in question across the subset of the reference gene expression dataset corresponding to ES cells and the most-similar non-ES cell samples (with similarity defined by the Euclidean distance across all probesets represented in the dataset).

category	count	overall	probes in GO term	P-value	adjusted p
protein localization	13	53	292	1.80E-017	1.30E-013
M phase	5	53	12	1.50E-012	1.10E-008
embryonic morphogenesis	2	53	31	6.20E-004	1
placenta development	4	53	105	7.40E-006	6.20E-002
cell division	4	53	676	8.20E-003	1
negative regulation of cell proliferation	2	53	387	7.60E-002	1

Table 6.7: Correlation-associated GO annotations enriched in Oct4 early-response target list

It is interesting to note that each gene expression meta-analysis approach results in different biological processes being associated with this list of Oct4 early-response targets, indicating that these different methods may identify different signatures within the input data. The same gene-association based clustering approach as described above was also applied to the correlation-associated genelists, resulting in 4 different clusters being identified within the original genelist. The results of functional enrichment analysis of each of these individual clusters in terms of the correlation-associated GO terms shown above is given in Table 6.8.

Again, this functional enrichment analysis approach involving a correlation-based meta-analysis approach to identify GO terms significantly enriched in associated genelists for each of the members of the input Oct4 early-response target list identifies some functional components not identified when using HBLCA as described in Section (6.1.1). However, when the expression patterns in the original time series dataset are observed for each of the clusters of the input list identified on the basis of correlation-associations, it would appear that there is less of a clear expression pattern representing each cluster than was the case for the clusters identified using gene association lists provided by HBLCA. Expression plots (equivalent to those shown in Fig. 6.2) are given in Fig. 6.4 for each of the 4 correlation-association based clusters.

As a quantitative assessment of a similar feature, the within-group correlations across the original Oct4 knock-down dataset were calculated for each of the clusters of Oct4 early-response targets as separated through HBLCA-based associations or through correlation-based meta-analysis associations. Even though there were fewer clusters in the input list identified through the HBLCA-based associations than through the correlation-based associations, which would result in an expectation of lower within-cluster correlations for the HBLCA-based association clusters, the within-cluster correlations suggest that the clusters obtained through HBLCA-based associations mapped better back to the input experiment than those obtained through correlation-based associations (see Fig. 6.5).

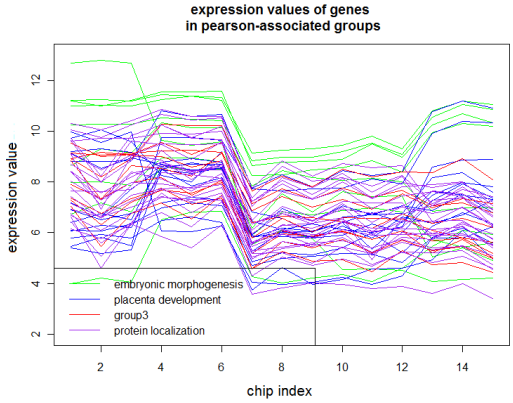
The results presented here suggest that a simple correlation-based meta-analysis approach may also be appropriate for the functional enrichment analysis approach proposed in this section, although it appears to be less successful for this purpose (in the case of the example analysis performed on a list of predicted Oct4 early-response targets) than the HBLCA approach.

6.1.3 Discussion

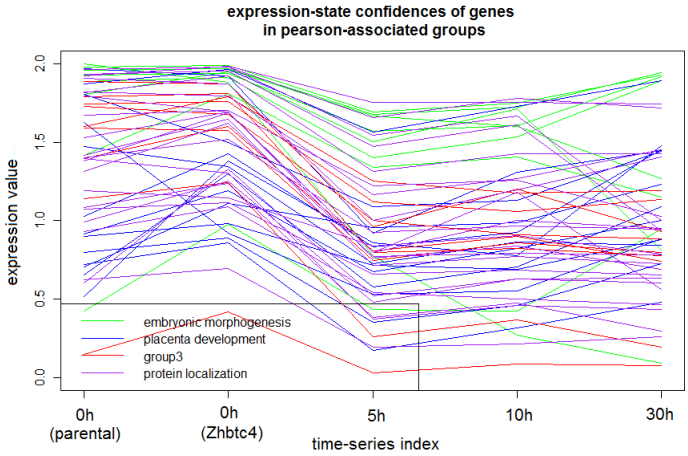
A novel approach to functional enrichment analysis has been proposed to avoid some of the potentially confounding features of standard functional enrichment analysis

category	group number	in group	group size	probes in category	P-value	adjusted p
embryonic morphogenesis	1	2	9	31	1.60E-005	0.12
placenta development	2	4	13	105	2.00E-008	1.50E-004
M phase	3	2	9	12	2.30E-006	1.70E-002
protein localization	4	12	22	292	2.60E-021	1.90E-017
cell cycle	4	3	22	1456	3.00E-002	1
M phase	4	3	22	12	2.20E-008	1.60E-004
cell division	4	3	22	676	4.20E-003	1

Table 6.8: Correlation-associated GO annotations enriched in different subsets of the Oct4 early-response targets



(a) Raw expression values for early-response targets in Oct4 knock-down time series



(b) GESTr-transformed values for early-response targets in Oct4 knock-down time series

Figure 6.4: Profile plots showing expression patterns in [Hall et al., 2009] Oct4 knock-down time series dataset for early-response target clusters. Chip index in plot (a) indicates which of the 15 samples from the experiment the measurements correspond to: chips 1-3 represent the parental cell line (no Oct4 knock-down), chips 4-6 represent Zhbtc4 cells without dox treatment, chips 7-9 represent Zhbtc4 cells after 5h dox treatment, chips 10-12 represent Zhbtc4 cells after 10h dox treatment, and chips 13-15 represent Zhbtc4 cells after 30h dox treatment.

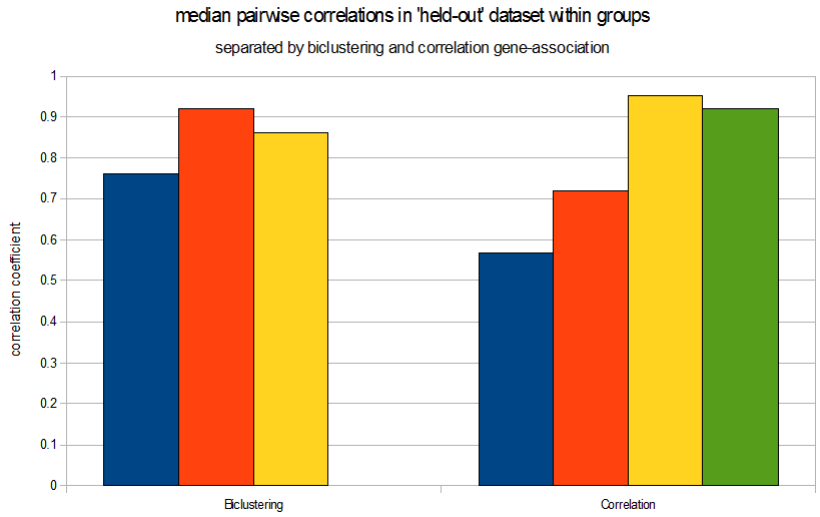


Figure 6.5: Within-cluster correlations across input dataset (from [Hall et al., 2009]) for clusters identified within a list of Oct4 early-response targets by clustering of associated genelists calculated through the HBLCA approach (left) and through a Pearson correlation-based meta-analysis approach (right). Despite there being more clusters resulting from the correlation meta-analysis approach, the genes within each cluster are not better-correlated than the genes within corresponding clusters from the HBLCA derived genelist. This indicates that the clusters from the HBLCA approach better reflext related components of the Oct4-induced transcriptional network in pluripotency.

techniques. This functional enrichment analysis approach uses gene expression meta-analysis to find associated genesets for each member of an input genelist, identifying clusters within the input genelist based on similarities between the meta-analysis gene-associations, and calculating statistical enrichments across each cluster of the input genelist in terms of functional annotation terms significantly enriched in the associated genesets for each of the members of that cluster. This analysis approach has been demonstrated to identify biological processes associated with a list of Oct4 early-response genes when standard functional enrichment analyses failed. This approach appears to be particularly successful when using the HBLCA approach presented in Section (5.1.6), identifying clusters of genes with related expression patterns in an Oct4 knock-down dataset (from which the original target list was obtained) in addition to having distinct functional association signatures. Therefore, as well as presenting the development of an improved approach to function enrichment analysis (which constitutes a significant result in its own right), this work demonstrates a successful application of HBLCA in which existing meta-analysis approaches (predominantly based on correlation and/or global expression analysis) would either be unsuitable or unable to identify the same significant biological signatures. In addition, the analysis of Oct4 early-response targets presented above has identified groups of biological processes that may be dependent in mouse ES cells on the transcriptional activity of Oct4.

6.2 Identification and Explanation of Structure Within a List of Oct4 DNA-Binding Targets

It has been observed that the majority of putative target genes with DNA-binding evidence from a given transcription factor (TF) are not directly and exclusively regulated by the TF in question, as discussed in Section (5.3.2) and reported in [Li et al., 2008]. This may be expected due at least in part to co-operative or redundant regulation of the expression of a target gene by multiple TFs, in addition to any epigenetic regulatory mechanisms that may affect the ability of the TF in question to activate transcription of the target.

Given the known significance of Oct4 in establishing and maintaining the pluripotent state [Niwa et al., 2000, Takahashi and Yamanaka, 2006, Chambers and Tomlinson, 2009], but the general lack of knowledge regarding mechanisms of regulation of genes involved in the pluripotent state, an investigation of genes bound by Oct4 was proposed in order to predict possible co-regulating partners for different subsets of Oct4's DNA-binding targets, and in such a way provide further insight into the Oct4-centred transcriptional regulatory network responsible for controlling pluripotency. This section presents the results of adopting a similar approach to that taken for functional analysis described in Section (6.1.1), with the HBLCA algorithm presented in Section (5.1.6) being used to discover groupings within a list of Oct4 DNA-binding targets on the basis of gene expression co-dependency associations.

6.2.1 Discovering Structure Within Target List Through Biclustering

Following application of the HBLCA algorithm to obtain lists of bicluster-associated genes with expression co-dependency patterns observed for each of the 'input' list of Oct4 DNA-binding targets, a gene association matrix was created similar to that described in Section (6.1.1) (although in this case a rank-based score was used). PCA was applied to the matrix of gene-association scores for the set of Oct4 consensus chIP targets in order to reduce dimensionality to a suitable input matrix for robust k-means clustering. Following PCA dimensionality reduction, the resulting gene association component score matrix was clustered using k-means clustering with the Gap Statistic to determine the optimum number of clusters. It should be noted that not all the input Oct4 DNA-binding targets were clustered in this way as a significant proportion (72%) failed to result in biclusters passing *ad hoc* filtering criteria based on visual inspection of any co-dependency patterns identified by the algorithm. This failure to result in appropriate biclusters for some of the input genes was due to a lack of suitable samples in the expression dataset that were suitably similar to ES cells but showed significant contrasts of expression of those 'failed' input genes. That some of the Oct4 DNA-binding targets were not included further in this analysis due to lack of appropriate biclusters

in no way affects the validity or interpretability of results of the subsequent analysis for those input targets that did result in identification of appropriate expression biclusters.

An initial observation of the results from this analysis of the list of consistently bound Oct4 regulatory targets was that there appear to be groups of targets that each share distinct patterns of gene expression co-dependency associations across ES cells. A graphical representation of the clustering in terms of the most significant 2 principal components of gene-association scores is given in Fig. 6.6.

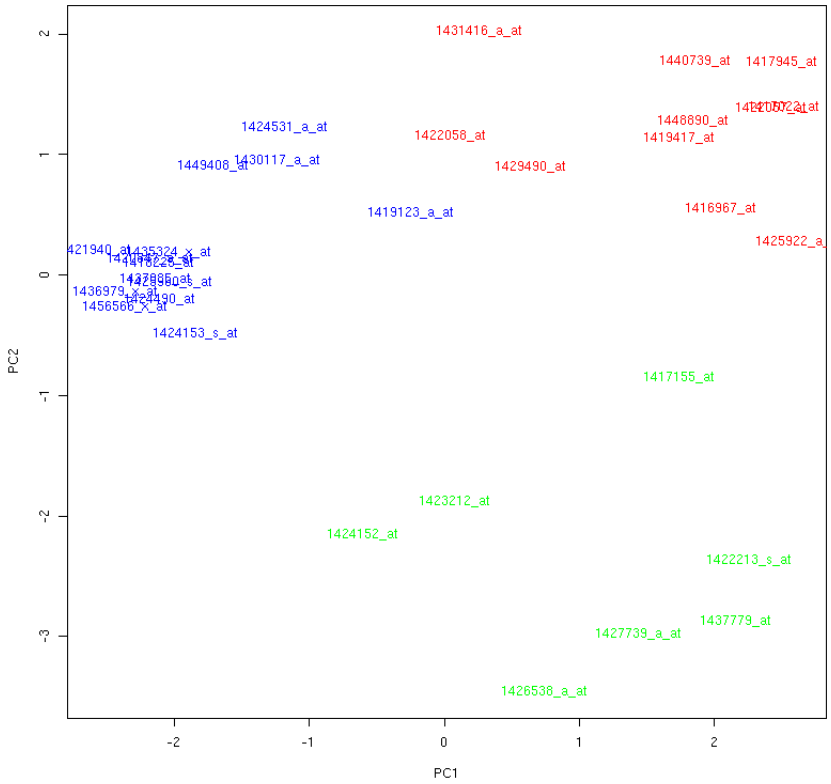


Figure 6.6: Representation of Oct4 chIP targets with satisfactory biclusters in terms of 2 principal components, coloured by group classification on the basis of k-means clustering (with gap statistic) performed on plotted principal component representation only

The clustering demonstrated in Fig. 6.6 seems to indicate that there exist clear groups of targets, each comprising genes with apparently similar expression association patterns in the reference compendium. The colours of the probeset IDs representing each Oct4 target show which of the clusters (determined by k-means clustering with gap statistic) from the PCA representation of the gene association matrix that target belongs to. A list of the genes and the corresponding grouping is provided in Table 6.9.

Probeset_ID	Gene	Group
1422057_at	Nodal	1
1422058_at	Nodal	1
1425922_a_at	Mycn	1
1431416_a_at	Jam2	1
1417945_at	Pou5f1	1
1429490_at	Rif1	1
1416967_at	Sox2	1
1440739_at	Vegfc	1
1419417_at	Vegfc	1
1448890_at	Klf2	1
1417022_at	Slc7a3	1
1418225_at	Orc2l	2
1449408_at	Jam2	2
1435324_x_at	Hmgb1	2
1436979_x_at	Rbm14	2
1456566_x_at	Rbm14	2
1430117_a_at	Zfp64	2
1424153_s_at	Sall4	2
1425960_s_at	Pax6	2
1419123_a_at	Pdgfc	2
1424531_a_at	Tcea3	2
1437085_at	D630039A03Rik	2
1424490_at	Zfp428	2
1420847_a_at	Fgfr2	2
1421940_at	Stag1	2
1427739_a_at	Trp53	3
1426538_a_at	Trp53	3
1417155_at	Mycn	3
1437779_at	Foxh1	3
1422213_s_at	Foxh1	3
1424152_at	Sall4	3
1423212_at	Phc1	3

Table 6.9: List of MOE430v2 probeset IDs corresponding to those genes with Oct4 DNA-binding evidence in all of 5 individual genome-wide chIP studies [Kim et al., 2008, Chen et al., 2008, Marson et al., 2008, Sharov et al., 2008, Loh et al., 2006] and sufficient variation in expression across ES cells for identification of co-dependently expressed genes using the HBLCA approach.

6.2.2 Association of Target Subsets With Different TFs

Although the identification of subsets within the set of Oct4 DNA-binding targets is an interesting observation in itself, and suggests that the methods employed may combine to form a useful tool for the investigation of complex transcriptional regulatory mechanisms, it would be especially interesting if some biological function or transcriptional event could be associated with each of the clearly-defined subsets in order to offer some explanation of *why* the observed structure appears. To progress towards this goal, an approach was developed to identify the individual genes providing greatest predictive power in classifying the Oct4 targets into the subsets identified as reported above. This approach involved the use of linear discriminant analysis (LDA) to identify the principal components of the gene association matrix that best separate the identified clusters, followed by analysis of the weightings of the discriminating principal components to find the genes with greatest contribution and finally evaluation of the discriminating power of each of those predicted discriminatory genes. A full description of this approach follows, using the Oct4 target list investigation as an example.

Linear Discriminant Analysis

LDA is similar to PCA, with the main difference between the methods being that in PCA the subspaces identified as principal components are those explaining the greatest variation observed between all the data objects, whereas in LDA the subspaces identified as linear discriminants are those explaining the greatest variation between specified subsets of the data objects. Given that it was a PCA representation of the gene association matrix that had been used to perform the clustering to identify classes within the set of all Oct4 targets, LDA could be performed on this PCA representation of the matrix in order to identify which of the principal components best separate the specified clusters (as this may differ from the ordering of the principal components as provided by PCA). LDA was performed using the `lda` function from the R package 'MASS' and the contribution of each principal component to discrimination between the input classes was evaluated through inspection of the scaling weights in the first linear discriminants (provided as output of the `lda` function).

This LDA of the PCA representation of the gene association matrix for Oct4 targets revealed that principal components ranked 24 and 22 in PCA showed greater discrimination between the classes than principal components 2 and 3, although the first principal component provided most discriminatory power (corresponding to observations from Fig. 6.6). Visual inspection of representations of the Oct4 targets in each combination of any 2 of the top-ranking principal components from PCA and LDA of the gene association score matrices revealed that only representations involving the first principal component resulted in separation of the input probesets into the distinct cluster groups calculated through k-means clustering using each of 2, 3 and 4 clusters

(data not shown). Therefore, if the listed Oct4 target genes are to be considered as belonging to one of only 2 classes, it is most likely that any genes discriminating between these classes will have a high weighting (either positive or negative) in the first principal component.

Principal Component and Class-Discriminating Genes

Having identified principal components of the Oct4 targets' gene association matrix that discriminate between subsets of the targets sharing similar gene association patterns, in order to obtain any biologically meaningful predictions regarding explanatory factors for the identified target subsets the discriminatory association patterns must be expressed in terms of (sets of) individual genes. If a particular principal component discriminates effectively between each subset of interest then genes with a highly significant contribution to that principal component (indicated by a high weighting, positive or negative, of that gene in the principal component) will be more likely to discriminate (on the basis of association scores) between the subsets than other genes. Such genes with the highest contributions to the best discriminating principal components can therefore be used as a starting point for the prediction of gene association classification rules on the basis of the underlying gene association scores.

As a further step towards identification of genes effectively discriminating between target subsets on the basis of expression association patterns, the gene association scores from the targets to each of a panel of potential discriminatory genes were obtained and averaged over each subset. When these average subset-association scores for each class are used as bases for a coordinate system, likely discriminatory genes will appear significantly off-diagonal. With a panel of 16 potential discriminatory genes selected on the basis of contribution to discriminatory principal components for the three subsets of Oct4 targets shown in Fig. 6.6, each of the panel genes were plotted using average association scores to the three subsets as coordinates, with the corresponding 3D scatterplot shown in Fig. 6.7.

While such a plot as that shown in Fig. 6.7 provides guidance into which genes may be effective discriminants between subsets of a gene list (in this case the list of Oct4 targets), for any biological conclusions to be made regarding co-association of a particular subset to any particular gene, that gene must be clearly associated to at least some of the genes in one subset of the gene list and not to any of the members of the input gene list belonging to one of the other subsets. If such discriminatory genes can be identified, a final analysis step is required in which any subsets with explanatory genes are pruned by removing those members of the subset without a clear association to the explanatory gene(s). While possibly somewhat protracted, the series of procedures described in this section provide a means of associating unknown genes with subsets of an input gene list, such that if an association to identified 'discriminatory genes'

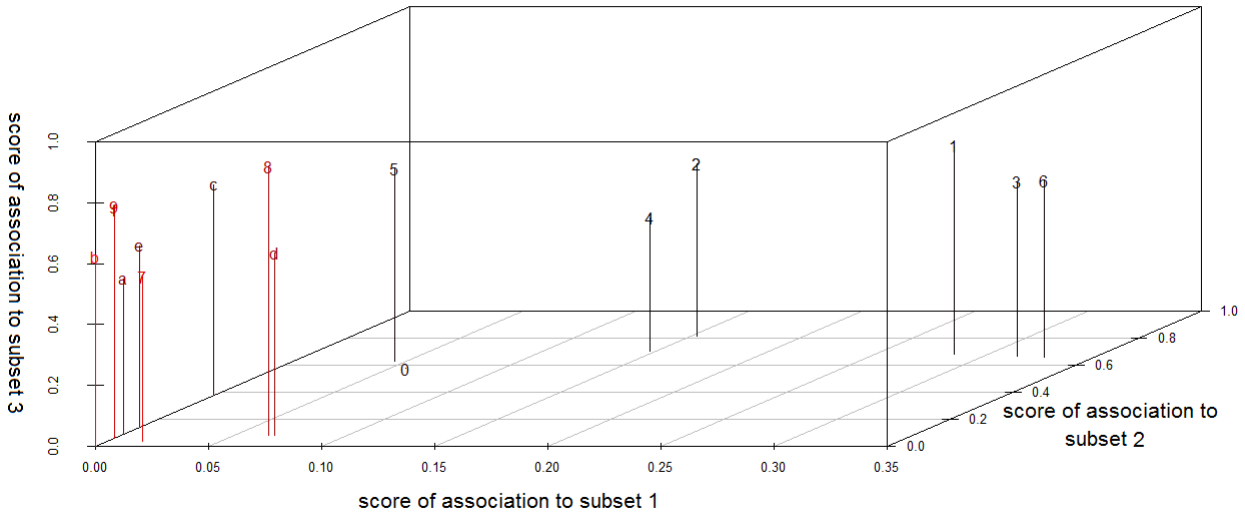


Figure 6.7: Panel of 'potential discriminant' genes (labelled with a hexidecimal index) plotted in terms of average association to the targets from each of 3 subsets of Oct4 target

is observed for any gene in the input list then that gene can be assumed to share similarities in expression pattern with the other genes in the subset 'marked' by that discriminatory gene. Furthermore, when a strong association to a single discriminatory gene is observed for each of a number of genes in a subset of the input list, sharing similar gene association patterns and therefore similar gene expression patterns, it is proposed that such a discriminatory gene will be likely to be of functional significance to that subset and to provide insight into either the consequences or the causes of that subset's separation in expression pattern from the rest of the genes in the input list.

Oct4 Target Subsets With Explanatory Genes

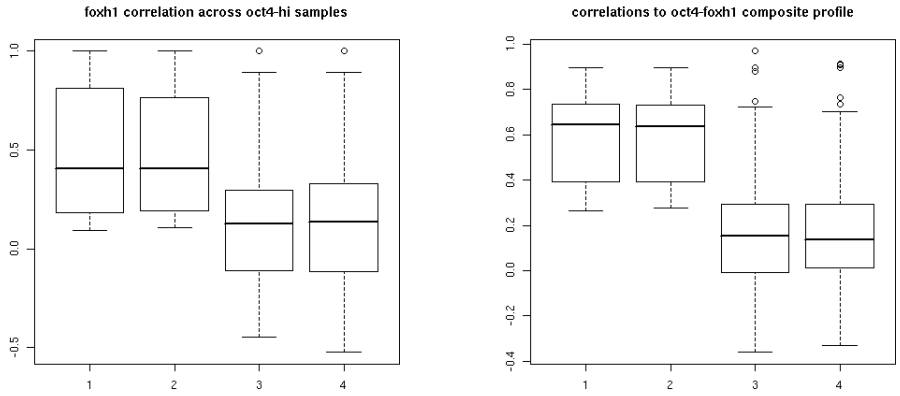
Following the analysis procedure described above, two subsets of Oct4 targets were identified, each showing clear similarity of gene association patterns (and therefore implicitly sharing similar expression patterns), for which individual genes served as clear discriminants in that strong associations were observed to all those targets within a particular subset and none of the other targets. As the genes in the first subset belong exclusively to the group of probesets plotted in blue in Fig. 6.6, and the genes in the first subset belong exclusively to the group of probesets plotted in green in Fig. 6.6 it can be seen that each of these subsets share *overall* similarity of gene association patterns and not just a shared association to the discriminatory gene. The subsets are given, along with their association scores to the discriminatory genes, in Table 6.10. Interestingly, the best discriminatory genes for the two subsets are both Oct4 targets and thus in the subsets themselves. However, an additional gene that was not on the original target list (dnmt3b) was found to show particularly strong associations to each of the members of the first subset and not to other Oct4 targets.

To demonstrate that the discriminatory genes identified through the above procedure correspond to real expression patterns of the targets, correlations of expression profile to each of the discriminatory genes were evaluated across ES and near-ES samples in a large gene expression dataset (described in Section (3.3)) for the respective subsets of Oct4 targets in comparison to all other targets. As a further observation, correlations were evaluated across the whole dataset for each of the target subsets compared with the remaining Oct4 targets to a composite expression profile based on expression of both Oct4 and the respective discriminatory gene. Plots for target subset 1 showing correlation to Foxh1 are presented in Fig. 6.8, and plots for target subset 2 showing correlation to Fgf4 are presented in Fig. 6.9.

The plots in Fig. 6.8 indicate that the genes in subset 1 are generally better correlated with Foxh1 across Oct4-expressing samples than the other Oct4 targets, and show significantly better correlation across the entire reference dataset to a composite profile based on required expression of Oct4 & Foxh1. Similar effects are shown in Fig. 6.9 for the expression association of Oct4 targets in subset 2 with Fgf4 (and Oct4 & Fgf4).

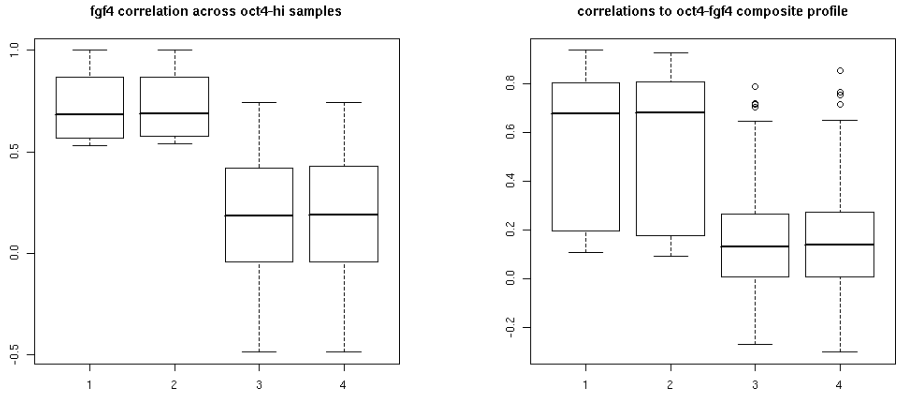
Probeset_ID	Gene	Group	Foxh1(I)	Foxh1(II)	Dnmt3b	Fgf4	Pou5f1
1437779_at	Foxh1	1	1	0.82	0.99	0	0.75
1422213_s_at	Foxh1	1	0.99	1	0.97	0	0.49
1426538_a_at	Trp53	1	0.82	0.79	0.89	0	0
1427739_a_at	Trp53	1	0.86	0	0.91	0	0.8
1424152_at	Sall4	1	0.78	0.98	0.91	0	0.58
1423212_at	Phc1	1	0.8	0.78	0.64	0	0.51
1422057_at	Nodal	2	0.16	0	0	0.94	0.88
1422058_at	Nodal	2	0.49	0	0	0.97	0.66
1440739_at	Vegfc	2	0	0	0	0.65	0.9
1419417_at	Vegfc	2	0	0	0	0.55	0.83
1448890_at	Klf2	2	0	0	0	0.81	0.92
1417022_at	Slc7a3	2	0	0	0	0.93	0.94

Table 6.10: Groups of Oct4 DNA-binding targets with associated dependency scores for a panel of 'explanatory TFs.'



(a) Foxh1 (b) Foxh1 & Oct4

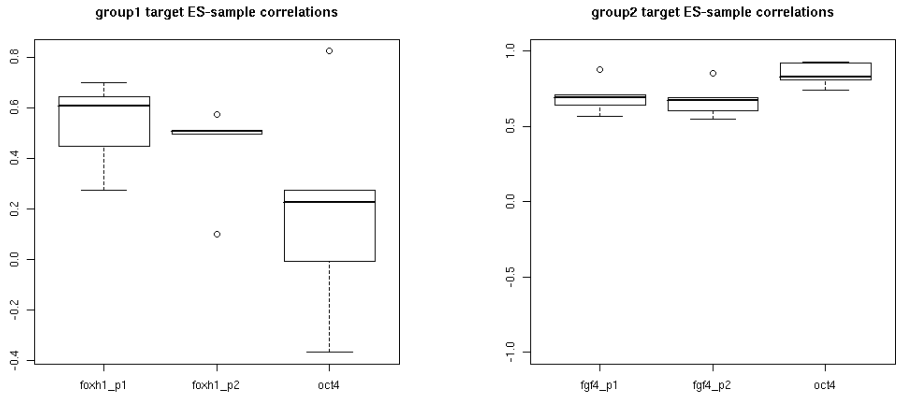
Figure 6.8: Illustration of expression correlation of Oct4 target group 1 to Foxh1. Panel (a) shows the Pearson correlation coefficients across Oct4-expression samples for the two Foxh1 probesets to each of the probesets listed in Table 6.10 as belonging to Group 1 (boxes 1 & 2) and to all of the other Oct4 targets (boxes 3 & 4). Panel (b) shows equivalent Pearson correlation coefficients for a composite profile scoring simultaneously high expression of both Oct4 and Foxh1 (on the GESTr scale), using each of the Foxh1 probesets in turn. As in panel (a), boxes (1 & 2) show Pearson correlation coefficients to Oct4 targets belonging to group 1 (listed in Table 6.10) and boxes (3 & 4) show the correlation coefficients to all other Oct4 targets from the list given in Table 6.9.



(a) Fgf4 (b) Fgf4 & Oct4

Figure 6.9: Illustration of expression correlation of Oct4 target group 2 to Fgf4. Panel (a) shows the Pearson correlation coefficients across Oct4-expression samples for the two Fgf4 probesets to each of the probesets listed in Table 6.10 as belonging to Group 2 (boxes 1 & 2) and to all of the other Oct4 targets (boxes 3 & 4). Panel (b) shows equivalent Pearson correlation coefficients for a composite profile scoring simultaneously high expression of both Oct4 and Fgf4 (on the GESTr scale), using each of the Fgf4 probesets in turn. As in panel (a), boxes (1 & 2) show Pearson correlation coefficients to Oct4 targets belonging to group 2 (listed in Table 6.10) and boxes (3 & 4) show the correlation coefficients to all other Oct4 targets from the list given in Table 6.9.

In order to confirm these associations, a comparison of correlations was performed using the respective discriminatory genes and Oct4 (the binding TF) for the targets in subset2 across a curated group of samples representing ES cells, but including Oct4 samples with knock-down. Plots are given in Fig. 6.10 to show the distributions of the correlations to each 'explanatory gene' for the probesets in the relevant subset, excluding the explanatory gene probesets as these might skew the distribution plots. As the target subset 2 contains Oct4, this presumably represents a set of targets whose expression in ES cells is predominantly regulated by Oct4. This is reflected in the correlation comparison plots in Fig. 6.10 (b), which show a roughly equal correlation across ES cells (including those with Oct4 knock-down) to Fgf4 and Oct4 for the probesets in subset 2. However, Fig 6.10 (a) shows that the probesets in subset 2 correlate to Foxh1 across ES cells, even with Oct4 knock-down, whereas only one of the subset 2 probesets (representing Phc2) correlates significantly to Oct4 across the same samples. This might imply that although Oct4 targets, this subset (comprising Foxh1, Trp53, Dnmt3b and Sall4) are regulated by some other transcription factor (possibly Foxh1 itself) that acts in conjunction with (or redundantly to) Oct4.



(a) Distribution of correlation scores for 'subset 1' probesets (excluding Foxh1) to each of two Foxh1 probesets and Oct4, evaluated across ES cells but including Oct4 knock-down samples

(b) Distribution of correlation scores for 'subset 2' probesets (excluding Fgf4) to each of two Fgf4 probesets and Oct4, evaluated across ES cells but including Oct4 knock-down samples

Figure 6.10: Comparison of expression correlation across ES cell samples for Oct4 target subset 1 to Foxh1 and Oct4 and for Oct4 target subset 2 to Fgf4 and Oct4

6.2.3 Discussion

A method has been developed for the identification of structures of gene expression association groupings within a list of genes, utilising the HBLCA approach described in Section (5.1). This method has been described through demonstration of its application to the investigation of a list of Oct4 DNA-binding targets, for which subsets of genes with different expression profiles were identified. One of these subsets involves a group

of genes with expression level in ES cells seemingly exclusively dependent on that of Oct4, and the other subset appears to be well characterised by an association with Foxh1 expression in ES cells. These results demonstrate a way in which novel tools developed through the course of this work can be utilised to gain insight into the mechanisms of transcriptional regulatory activity of a TF of interest. Furthermore, this analysis has revealed a potentially interesting observation regarding the transcriptional regulation of Oct4 DNA-binding targets in ES cells, highlighting targets with clear Oct4 expression dependency and targets with expression level seemingly dependent on Foxh1 rather than Oct4, which may suggest a role for Foxh1 as a co-regulator for a set of Oct4 targets in ES cells. Further investigation of these proposed relationships could lead to insight into the mechanisms of transcriptional regulation of maintenance and/or acquisition of pluripotency.

6.3 Investigation of Combinatorial Activity of Key Pluripotency TFs

It is widely agreed that Oct4, Sox2 and Nanog are fundamental regulators of the pluripotent state [Chambers and Tomlinson, 2009, Loh et al., 2006, Ivanova et al., 2006], but the precise mechanisms by which they control maintenance or acquisition of pluripotency are unknown [Chambers and Tomlinson, 2009]. Through genome-wide DNA-binding studies for each of these TFs, potential targets have been identified [Chen et al., 2008, Kim et al., 2008, Marson et al., 2008, Sharov et al., 2008, Loh et al., 2006]. An interesting observation has been made that the sets of DNA-binding targets for each of these TFs overlap considerably, as illustrated in Fig. 6.11 (and discussed in [Chambers and Tomlinson, 2009, Chen et al., 2008]). For example, approximately only one-third of the most consistently Oct4-bound genes across multiple chIP studies are not also reliably bound by either Sox2 or Nanog.

However, as reported in Section (5.3.2), the majority of these identified DNA-binding targets do not show clear expression patterns in relation to the binding TF, even within a restricted biological context of only ES cells. Given that the 3 TFs in question are known to bind common subsets of targets (including each other), it would be especially interesting not just to identify targets with binding evidence and expression co-dependency evidence for each TF (as described in Section (5.3.3)), but to utilise the HBLCA approach in concert with meta-analysis of data from multiple genome-wide DNA-binding studies to identify target genes that appear to be regulated by different combinations of Oct4, Sox2 and Nanog in order to obtain a deeper understanding of the mechanisms involved in transcriptional regulation of pluripotency.

While the HBLCA meta-analysis tool provides a unique opportunity to identify genes with biologically significant expression co-dependency patterns with unique combina-

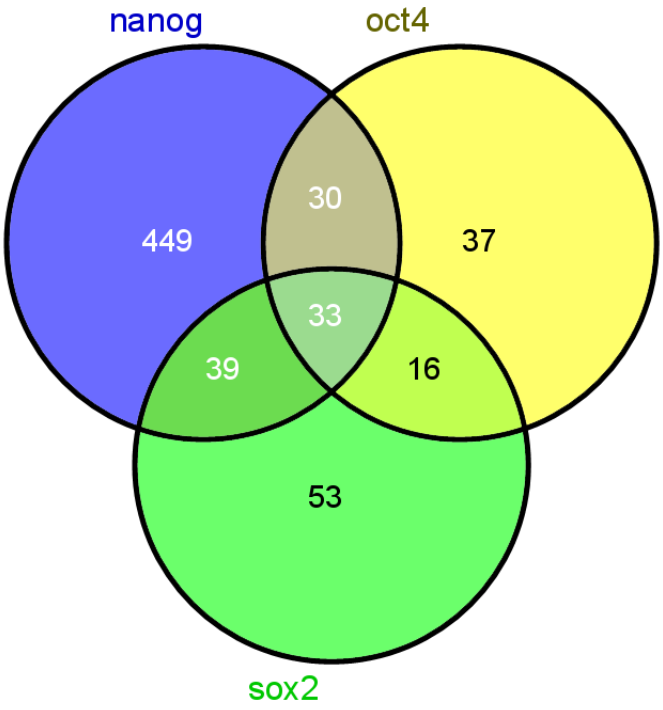


Figure 6.11: Overlap between lists of genes found to be consistently bound by each of Oct4, Sox2 and Nanog through meta-analysis of chIP experiments.

tions of Oct4, Sox2 and Nanog, in order for this to be possible it requires data with variation of expression of each combination of the controlling TFs within the biological context of interest (pluripotent cells and their immediate derivatives). The dataset used in the previous analyses and described in Section (3.3) contained insufficient data to investigate any combinations of regulation through Oct4, Sox2 and Nanog aside from Oct4 alone and all 3 TFs together. Therefore, additional data was required from samples that are near to ES cells (in terms of global transcriptional profile) but with variation of additional combinations of the 3TFs. Data was gathered from a number of studies for which the data was made available subsequent to the compilation of the dataset described in Section (3.3) and from one unpublished study (I. Chambers, personal communication). The meta-analysis dataset was augmented with these samples through the preprocessing and linear interpolation procedures described in Section (4.3). An illustration of the capacity for combinatorial expression co-dependency analysis of Oct4, Sox2 and Nanog afforded by this augmented dataset is given in Fig. 6.12 with ES cell samples and their transgenic derivatives plotted in three dimensions of Oct4, Sox2 and Nanog expression levels. It can be observed in Fig. 6.12 that there exist points in the four 'distant' corners of the scatterplot, representing each combination of high and low expression of each of Oct4 and Nanog, whilst maintaining high

Sox2 expression. This would suggest that the dataset offers potential for dissecting gene expression co-dependencies with different combinations of Oct4 and Nanog in ES cells. However, there appears to be only limited potential for discovering expression patterns involving contrasts in Sox2 expression level independent of Nanog and Oct4.

Gene expression analysis was performed using HBLCA, using each combination of the 3 TFs as guide genes, with the other TFs specified as 'contrast genes' in order to evaluate co-dependency to the unique combination (as described in Section (5.1.3)).

6.3.1 Identification of Genes with Expression Patterns Associated to Independent Combinations of Pou5f1, Sox2 and Nanog

Application of the HBLCA approach to this augmented dataset with different combinations of Oct4, Sox2 and Nanog specified as 'guide' genes and 'contrast' genes resulted in lists of genes ranked according to consistency of co-dependency of expression to each of the specified combinations. As suggested by Fig. 6.12 the dataset analysed did not include extensive collections of samples with a sufficient degree of overall similarity to ES cell samples yet covering the full range of combinations of expression levels of each of these three key pluripotency TFs. As a consequence, it would be expected that if the meta-analysis were run with low enough stringency so as to utilise contrasts involving each of these combinations, there would be some overlap between the genelists from analysis with each combination. Lists of top-scoring probesets from integration of co-dependency patterns across sets of biclusters filtered by visual inspection as described in the previous section are provided in Tables 6.11, 6.12 & 6.13, with overlaps between the lists summarised in Fig. 6.13. An illustration of the effectiveness of this approach is given in Fig. 6.14, which shows plots (generated by the visualisation tool described in Section (5.1.7)) for the biclusters identified by the meta-analysis algorithm as displaying contrasts in Nanog expression, contrasting the genes identified through HBLCA (panel on top) with those most correlated with Nanog across the whole gene expression dataset (panel on bottom). Of particular note is the fact that the genes identified by the HBLCA approach follow the general trend of Nanog in the expression levels across 'contrast' samples that are largely similar to the normal ES cell samples of the biclusters far better than the genes obtained through correlation analysis.

The overlaps in lists of genes showing expression co-dependency to each of Oct4, Nanog and Sox2 are further illustrated in Fig. 6.15, which plots the top-ranking genes from each list in three dimensions representing rank-based score in the lists of Oct4 co-dependent genes, Sox2 co-dependent genes and Nanog co-dependent genes (x,y and z axes, respectively). From this plot it is evident that the majority of the top-ranking genes in the list for any one of the TFs are also present in the lists for the other TFs, and typically at a fairly high rank. Those genes that lie near the vertices of the cube formed by the axes of Oct4, Sox2 and Nanog co-dependency may be especially interesting, as

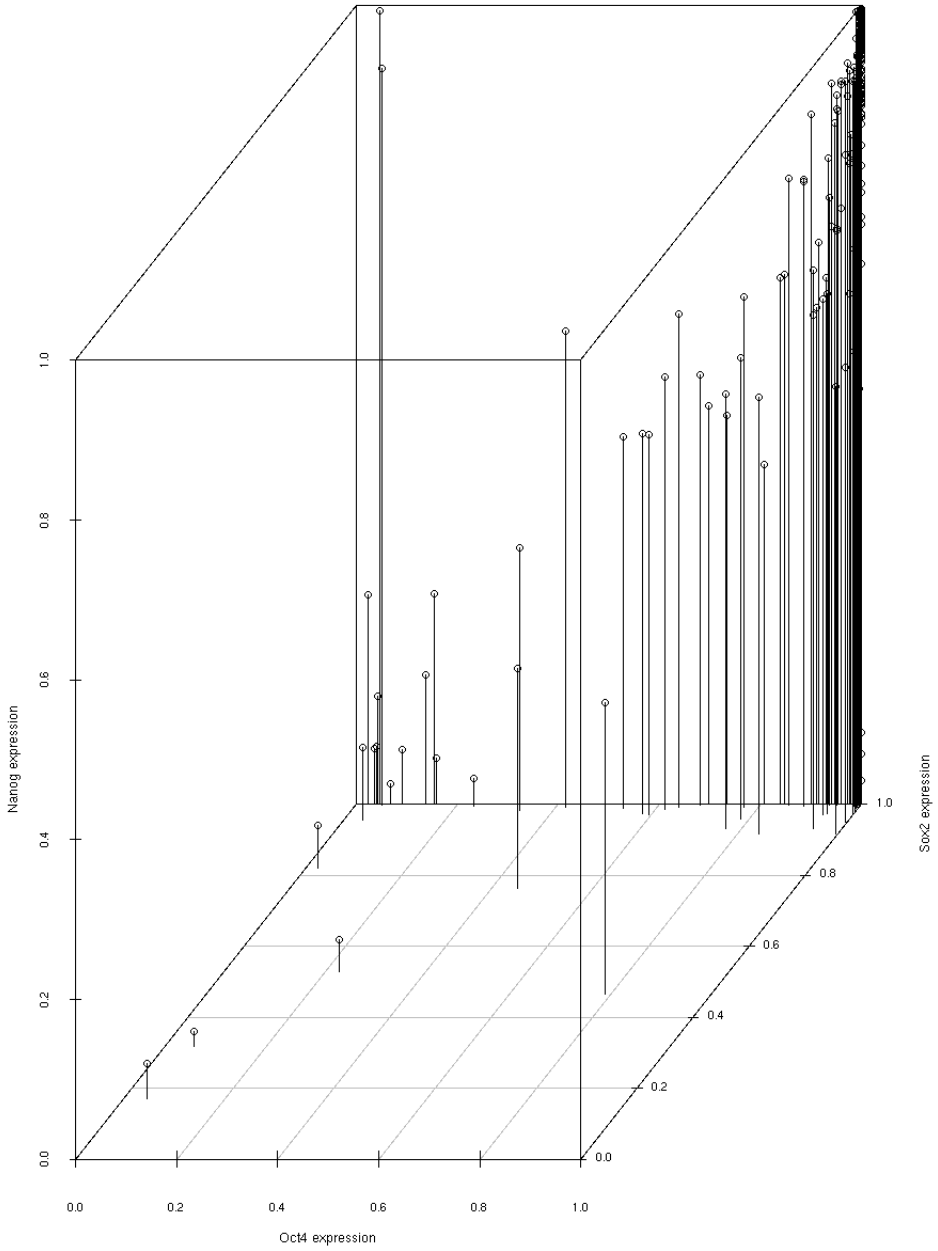


Figure 6.12: 3D representation in terms of Oct4, Sox2 and Nanog expression levels (x,y and z axes respectively) of ES cell samples and their derivatives in the augmented dataset described in Section (6.3).

Probe Set ID	Gene Symbol	Codependency Probability	Nanog Binding Site
1429388_at	Nanog	1	Kim,Marson,Sharov
1420086_x_at	Fgf4	2.64E-005	Kim,Marson,Sharov
1417760_at	Nr0b1	1.67E-005	Chen,Kim,Marson,Sharov
1449064_at	Tdh	1.09E-005	Chen,Kim,Marson,Sharov
1416043_at	Nasp	4.44E-006	
1432207_a_at	Toe1	4.36E-006	
1451158_at	Trip12	3.56E-006	
1431430_s_at	Trim59	2.32E-006	
1426653_at	Mcm3	2.12E-006	Chen,Kim
1418027_at	Exo1	2.07E-006	
1431252_a_at	Zfp655	1.63E-006	
1418470_at	Yes1	1.46E-006	
1423430_at	Mybbp1a	1.24E-006	
1416915_at	Msh6	9.40E-007	Chen,Kim,Marson,Sharov
1453604_a_at	Hbs1l	8.31E-007	Chen,Marson
1434279_at	NA	7.21E-007	
1436020_at	Zfp828	7.18E-007	
1435379_at	Urb2	7.09E-007	Sharov
1426810_at	Kdm3a	4.04E-007	Chen,Marson,Sharov
1448777_at	Mcm2	1.94E-007	

Table 6.11: Genes displaying greatest and most consistent co-dependency of expression with Nanog, independent of Sox2 and Oct4 fluctuations. Binding evidence column gives studies reporting binding of Nanog to target: Chen, Kim, Marson and Sharov indicate [Chen et al., 2008], [Kim et al., 2008], [Marson et al., 2008] and [Sharov et al., 2008], respectively.

Probe Set ID	Gene Symbol	Codependency Probability	Sox2 Binding Site
1416967_at	Sox2	1	Chen,Kim,Liu,Marson
1421749_at	Lin28	1E-004	
1424008_a_at	Rbpms2	2.82E-006	Marson
1417945_at	Pou5f1	8.53E-007	Chen,Kim,Liu,Marson
1440085_at	Eda2r	4.27E-007	
1419706_a_at	Akap12	3.88E-007	Marson
1428142_at	Etv5	2.89E-007	Kim,Marson
1425416_s_at	Psrc1	2.38E-007	
1434280_at	NA	1.91E-007	
1425042_s_at	Pelp1	1.36E-007	Marson
1438237_at	Rex2	1.30E-007	
1424801_at	Enah	8.93E-008	
1454904_at	Mtm1	7.13E-008	
1420731_a_at	Csrp2	7.04E-008	
1425565_at	Rest	5.07E-008	Chen,Kim,Liu,Marson
1449064_at	Tdh	4.64E-008	Chen,Marson
1418362_at	Zfp42	4.45E-008	Marson
1423925_at	Dhx16	3.50E-008	
1426645_at	Hsp90aa1	1.31E-008	Chen,Marson
1448692_at	Ubqln4	4.01E-009	

Table 6.12: Genes displaying greatest and most consistent co-dependency of expression with Sox2, independent of Nanog and Oct4 fluctuations. Binding evidence column gives studies reporting binding of Sox2 to target: Chen, Kim, Liu and Marson indicate [Chen et al., 2008], [Kim et al., 2008], [Liu et al., 2008] and [Marson et al., 2008], respectively.

Probe Set ID	Gene Symbol	Codependency Probability	Oct4 Binding Site
1417945_at	Pou5f1	1	Chen,Kim,Liu,Marson,Sharov
1420086_x_at	Fgf4	3.72E-010	Chen,Kim,Marson,Sharov
1449064_at	Tdh	2.17E-010	Chen,Kim,Marson,Sharov
1417760_at	Nr0b1	4.29E-011	Liu,Marson,Sharov
1419418_a_at	Morc1	5.11E-013	Chen,Marson,Sharov
1449288_at	Gdf3	3.50E-013	Chen,Liu,Marson,Sharov
1420085_at	Fgf4	6.72E-015	Chen,Marson,Sharov
1422697_s_at	Jarid2	3.92E-017	
1430139_at	Hells	3.58E-017	
1423424_at	Zic3	1.90E-017	Liu,Marson,Sharov
1448562_at	Upp1	2.11E-019	Chen,Marson
1421151_a_at	Epha2	1.96E-019	Chen,Marson
1427953_at	Fanci	2.79E-020	Liu,Marson
1424008_a_at	Rbpms2	1.16E-020	Marson,Sharov
1438237_at	Rex2	9.02E-021	
1456515_s_at	Tcfl5	8.80E-021	
1422058_at	Nodal	6.97E-021	Chen,Kim,Liu,Marson,Sharov
1427238_at	Fbxo15	4.99E-021	
1439065_x_at	Gm13152	3.04E-021	
1419234_at	Helb	1.86E-021	Chen,Marson

Table 6.13: Genes displaying greatest and most consistent co-dependency of expression with Oct4, independent of Nanog and Sox2 fluctuations. Binding evidence column gives studies reporting binding of Oct4 to target: Chen, Kim, Liu, Marson and Sharov indicate [Chen et al., 2008], [Kim et al., 2008], [Liu et al., 2008], [Marson et al., 2008] and [Sharov et al., 2008], respectively.

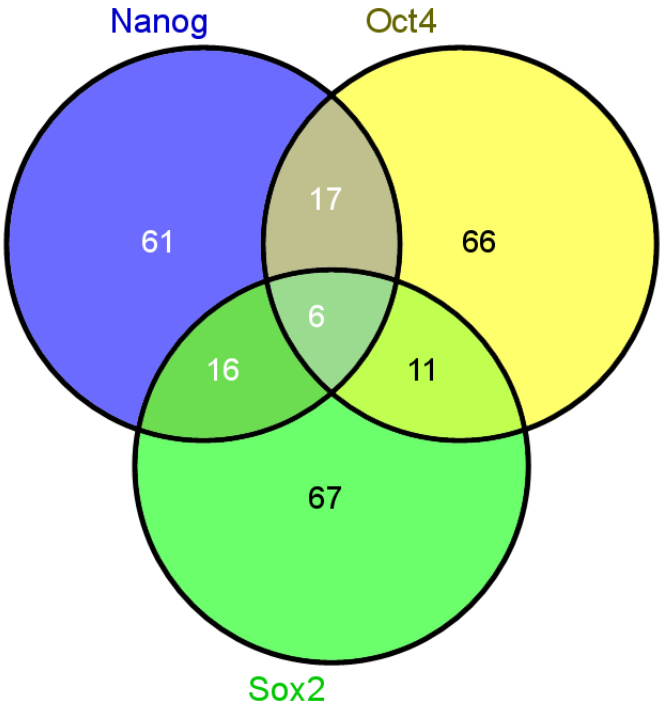
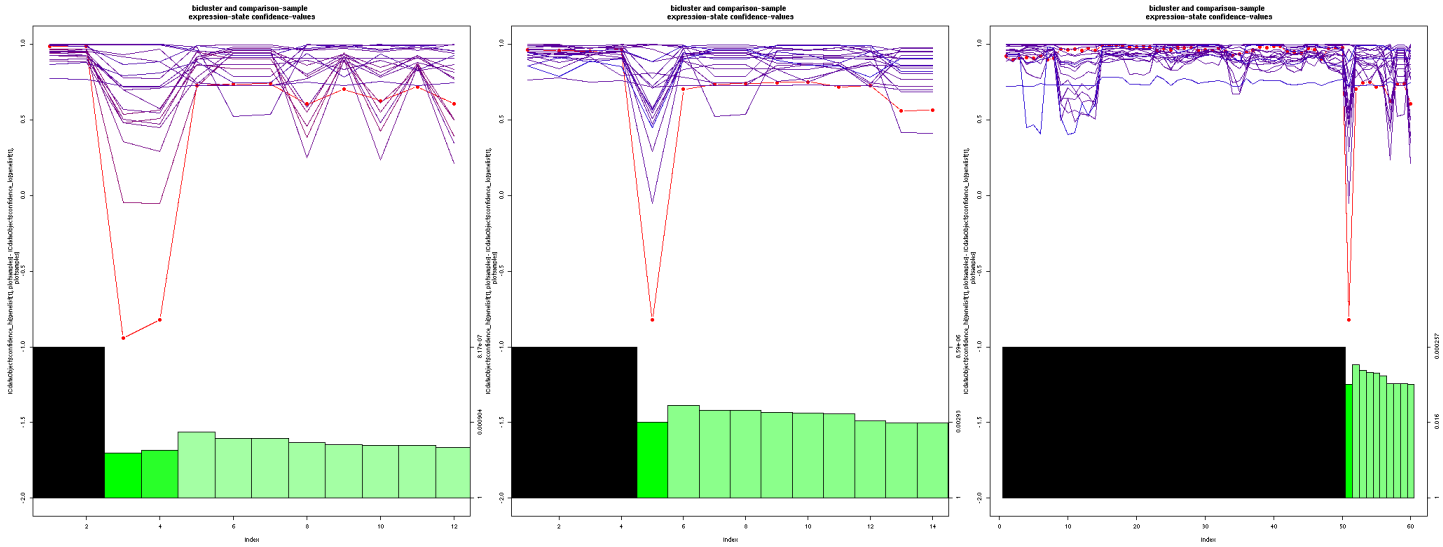
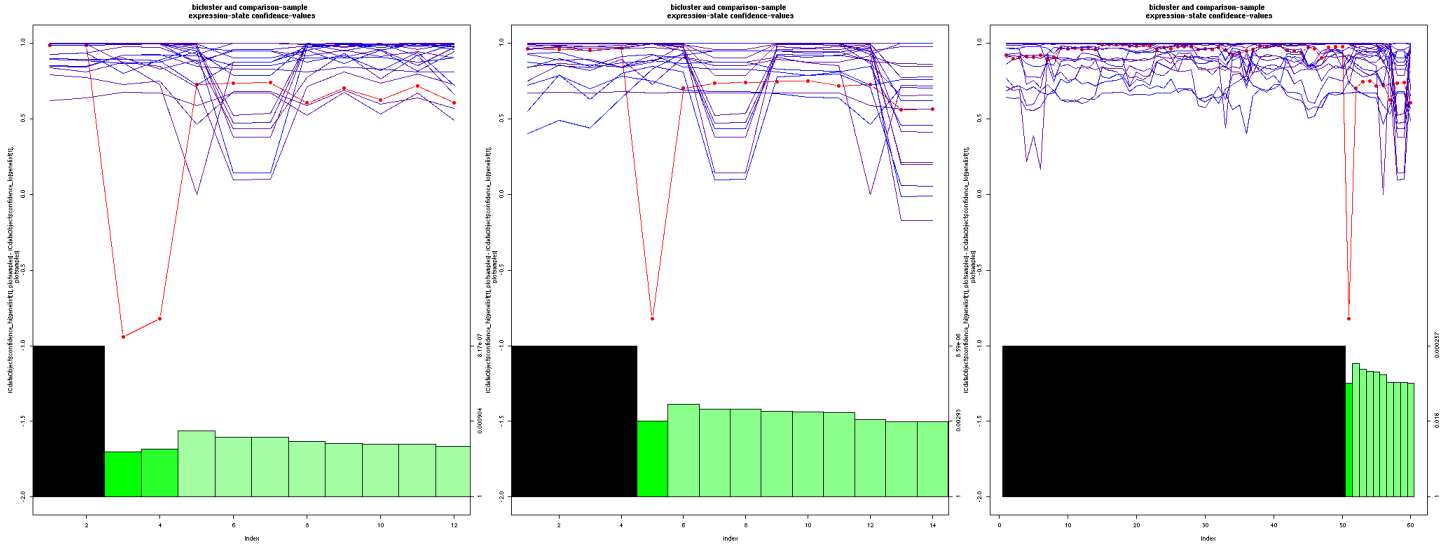


Figure 6.13: Overlaps between lists of genes displaying co-dependency on each of Oct4, Nanog and Sox2 independent of the other two TFs.



(a) Nanog-dependent expression in ES cells for genes associated with Nanog through HBLCA.



(b) Nanog-dependent expression in ES cells for genes associated with Nanog through correlation analysis

Figure 6.14: Comparison of Nanog-dependent expression in ES cells for genes identified by HBLCA and by global correlation analysis.

these would be hypothesised to be associated with expression of a particular combination of the TFs. A summary of these genes is presented in Table 6.14, also including genes arising uniquely from application of the HBLCA approach using a combination of multiple 'guide' genes.

Probe Set ID	Gene Symbol	TF Dependency
1449266_at	Mecp2	Oct4
1456515_s_at	Tcfl5	Oct4
1426645_at	Hsp90aa1	Sox2
1416967_at	Sox2	Sox2
1421749_at	Lin28	Sox2
1431252_a_at	Zfp655	Nanog
1422697_s_at	Jarid2	Oct4 (Nanog)
1438237_at	Rex2	Oct4, Sox2
1439065_x_at	Gm13152	Oct4, Sox2
1419706_a_at	Akap12	Oct4, Sox2
1451158_at	Trip12	Sox2, Nanog
1418027_at	Exo1	Sox2, Nanog
1454904_at	Mtm1	Sox2, Nanog

Table 6.14: Genes displaying co-dependency to unique combinations of Oct4, Sox2 and Nanog.

Owing to the overlap seen in the lists of genes with co-dependency observed by HBLCA to each of Oct4, Sox2 and Nanog independent of the other two TFs, an association score was produced for each TF for each potential target gene appearing in any of the target lists. Using the resulting TF-target association scores, lists of TF-specific targets were produced to identify genes that are likely to be regulated uniquely by one of Oct4, Sox2 or Nanog. Lists of genes most uniquely associated to particular combinations are provided in Tables 6.15, 6.16 & 6.17. The unique co-dependency of the genes in each of these lists is illustrated in Fig. 6.16, in which the genes are plotted in three co-dependency-score dimensions as Fig. 6.15.

It would be expected that the TFs would regulate the expression of their targets primarily through DNA-binding dependent mechanisms, and therefore the most interesting predictions of expression dependency (and thus critical transcriptional regulatory relationships) would be those involving particular combinations of Oct4, Sox2 and Nanog with expression association to genes which are bound by (at least, but not necessarily exclusively) those TFs displaying the expression co-dependency with the potential target. It may be possible that the mechanism of regulation might involve inhibition of binding of another TF to the target gene's promoter or enhancer regions, but as there is presently no data available to support such a hypothesis, genes with these associations are left out of the final target list. This final list of targets implicated

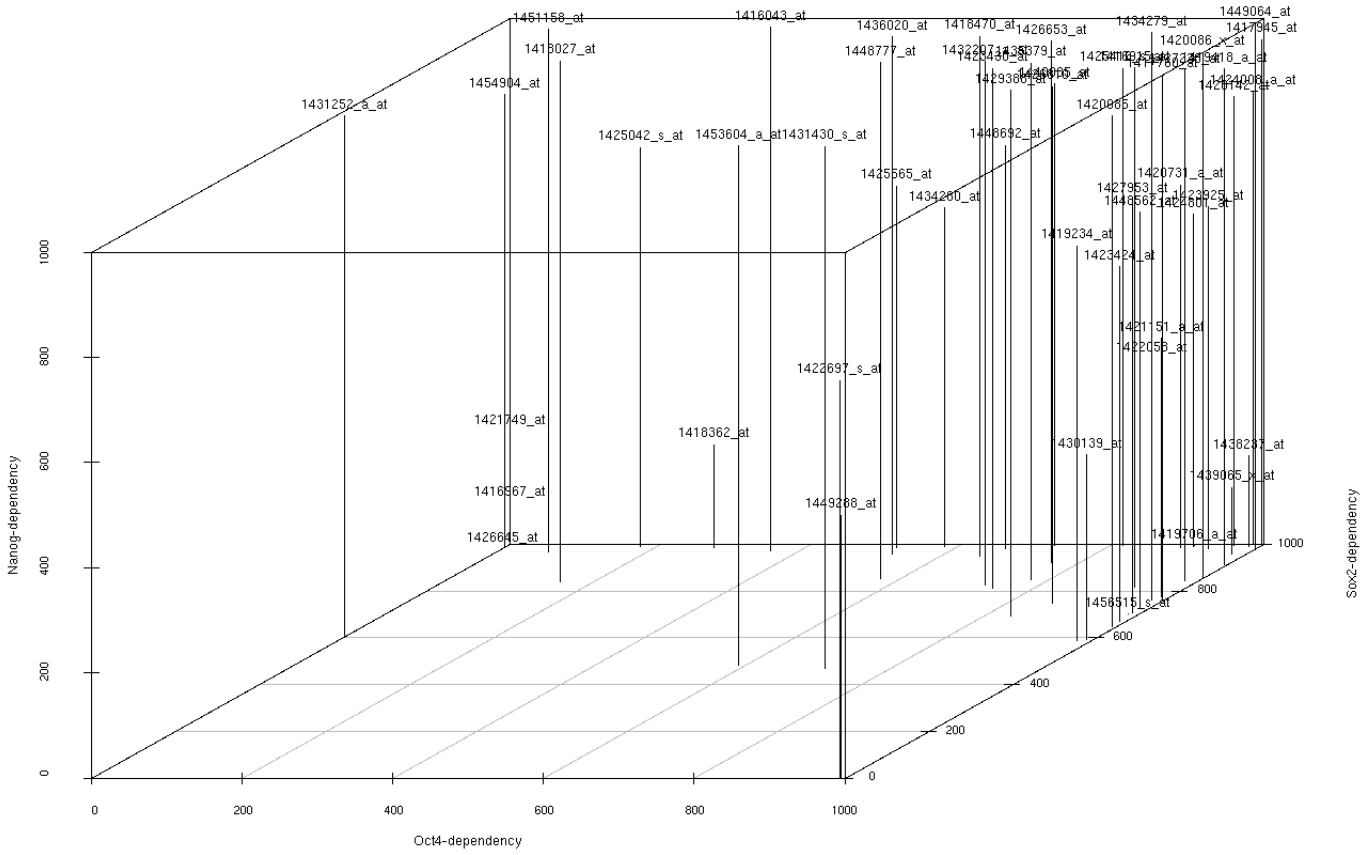


Figure 6.15: Top-ranking members of each list of genes with expression co-dependency evident for Oct4, Nanog and Sox2, plotted in three dimensions according to rank-based score for the genes in each of the three lists.

Probe Set ID	Gene Symbol	Oct4 Binding Site
1436837_at	Mael	
1429154_at	Slc35f2	Chen,Marson
1457314_at	L1td1	Kim,Marson
1430134_a_at	Yars2	Chen,Marson
1424784_at	Gm13139	Kim,Marson
1423465_at	Frrs1	
1452098_at	Chtf18	
1423289_a_at	1810029B16Rik	
1433478_at	Parl	
1416492_at	Ccne1	
1434239_at	Rrp12	
1418435_at	Mkrn1	Chen,Marson,Sharov
1436728_s_at	Rtel1	Marson
1433789_at	Snhg3	
1416687_at	Plod2	
1420113_s_at	2410022L05Rik	Kim
1428315_at	Ebna1bp2	
1422922_at	Recql4	
1423064_at	Dnmt3a	
1433692_at	Nat10	Chen,Marson,Sharov

Table 6.15: Genes displaying unique co-dependency to Oct4. Binding evidence column gives studies reporting binding of Oct4 to target: Chen, Kim, Liu, Marson and Sharov indicate [Chen et al., 2008], [Kim et al., 2008], [Liu et al., 2008], [Marson et al., 2008] and [Sharov et al., 2008], respectively.

Probe Set ID	Gene Symbol	Sox2 Binding Site
1426645_at	Hsp90aa1	Chen,Marson
1417845_at	Cldn6	Kim
1433651_at	Wtip	
1451320_at	Arhgap8	Kim,Marson
1422418_s_at	Supt4h1	
1417656_at	Mybl2	Chen,Kim,Marson
1418488_s_at	Ripk4	
1456615_a_at	Bptf	
1416364_at	Hsp90ab1	Kim,Marson
1438957_x_at	Cds2	
1418761_at	Igf2bp1	
1454159_a_at	Igfbp2	Kim,Marson
1429291_at	Psmd1	
1434328_at	Rpl15	
1435448_at	Bcl2l11	Marson
1416967_at	Sox2	Chen,Kim,Liu,Marson
1432393_a_at	Thg1l	
1426953_at	Hmgxb4	
1418351_a_at	Dnmt3b	Marson
1460325_at	Pum1	

Table 6.16: Genes displaying unique co-dependency to Sox2. Binding evidence column gives studies reporting binding of Sox2 to target: Chen, Kim, Liu and Marson indicate [Chen et al., 2008], [Kim et al., 2008], [Liu et al., 2008] and [Marson et al., 2008], respectively.

Probe Set ID	Gene Symbol	Nanog Binding Site
1423325_at	Pnn	
1422967_a_at	Tfrc	
1422994_at	Pikfyve	Sharov
1422135_at	Zfp146	Kim,Sharov
1423234_at	Psmd5	
1421462_a_at	Lepre1	
1426370_at	Far1	
1433897_at	AI597468	
1452220_at	Dock1	Kim,Marson
1420475_at	Mtpn	
1431096_at	Ints8	
1448399_at	Tax1bp1	Sharov
1453949_s_at	Lypla1	Marson,Sharov
1416423_x_at	Ssb	
1424569_at	Ddx46	
1449504_at	Kpna1	
1430575_a_at	Tpp2	
1440894_at	Tmtc3	Sharov
1420251_at	NA	
1421940_at	Stag1	Kim,Marson,Sharov

Table 6.17: Genes displaying unique co-dependency to Nanog. Binding evidence column gives studies reporting binding of Nanog to target: Chen, Kim, Marson and Sharov indicate [Chen et al., 2008], [Kim et al., 2008], [Marson et al., 2008] and [Sharov et al., 2008], respectively.

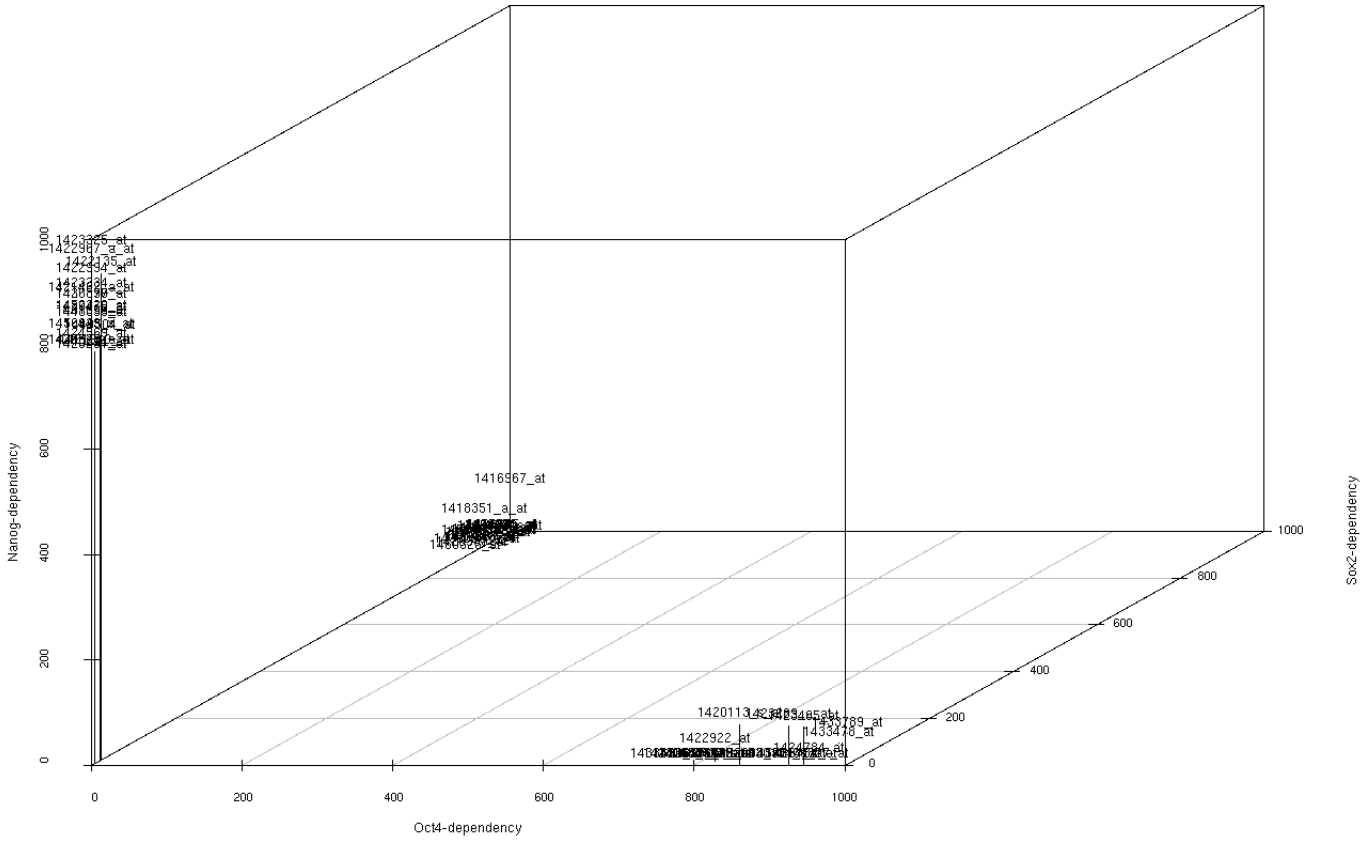


Figure 6.16: Genes with expression co-dependency evident to only one of Oct4, Nanog or Sox2, plotted in three dimensions according to rank-based score for the genes in each of the three lists.

in the pluripotency network through association with particular combinations of Oct4, Sox2 and Nanog is given in Table 6.18.

Probe Set ID	Gene Symbol	TF Dependency
1429154_at	Slc35f2	Oct4
1457314_at	L1td1	Oct4
1430134_a_at	Yars2	Oct4
1424784_at	Gm13139	Oct4
1418435_at	Mkrn1	Oct4
1436728_s_at	Rtel1	Oct4
1420113_s_at	2410022L05Rik	Oct4
1433692_at	Nat10	Oct4
1449266_at	Mecp2	Oct4
1419706_a_at	Akap12	Oct4&Sox2
1426645_at	Hsp90aa1	Sox2
1417845_at	Cldn6	Sox2
1451320_at	Arhgap8	Sox2
1417656_at	Mybl2	Sox2
1416364_at	Hsp90ab1	Sox2
1454159_a_at	Igfbp2	Sox2
1435448_at	Bcl2l11	Sox2
1416967_at	Sox2	Sox2
1418351_a_at	Dnmt3b	Sox2
1422994_at	Pikfyve	Nanog
1422135_at	Zfp146	Nanog
1452220_at	Dock1	Nanog
1448399_at	Tax1bp1	Nanog
1453949_s_at	Lypla1	Nanog
1440894_at	Tmtc3	Nanog
1421940_at	Stag1	Nanog

Table 6.18: Genes with TF binding and unique expression co-dependency to combinations of Oct4, Sox2 and Nanog.

As a final interesting observation, it was noted that in the cluster analysis of Oct4 biclusters, one bicluster had a significantly different targets to the others. This bicluster involved only iPS cell samples, as opposed to ES cells. The genes that appear to be Oct4 dependent in iPS cells but not ES cells might be specifically involved in the reprogramming process and yet not directly involved in the maintenance of pluripotency. A list of such genes is provided in Table 6.19. It is predicted that those genes identified in this list that additionally have Oct4 binding sites may have an Oct4-dependent role in the acquisition of pluripotency arising from the reprogramming process.

Probe Set ID	Gene Symbol	Oct4 Binding Site
1426936_at	Gm6958	
1429932_at	4930566F21Rik	
1430368_s_at	1700019D03Rik	Chen,Kim,Marson,Sharov
1431865_a_at	Zfp819	Kim,Liu,Marson,Sharov
1434917_at	Cobl	Chen,Kim,Marson,Sharov
1436419_a_at	1700097N02Rik	
1436799_at	Enox1	
1438861_at	Bnc2	
1444051_at	1700019D03Rik	
1452004_at	Calca	Marson,Sharov
1452063_at	Zbtb8a	Marson,Sharov

Table 6.19: Genes displaying expression co-dependency with Oct4 uniquely to iPS samples. Binding evidence column gives studies reporting binding of Oct4 to target: Chen, Kim, Liu, Marson and Sharov indicate [Chen et al., 2008], [Kim et al., 2008], [Liu et al., 2008], [Marson et al., 2008] and [Sharov et al., 2008], respectively.

6.3.2 Discussion

Through application of meta-analysis approaches developed through the course of this work, using large collections of gene expression data together with collections of genome binding data, it has been possible to identify genes that appear to be targets of Oct4, Sox2 and Nanog with expression dependency on different combinations of these key regulators of pluripotency. Such genes are likely to be involved in the acquisition of phenotypes (which may involve a loss of pluripotency) associated with loss of expression of each of these key TFs that are typically expressed at high levels in all pluripotent cells. This analysis also reveals targets whose expression in cells that have gained pluripotency through reprogramming would be expected to be critically dependent on expression of each of Oct4 and Sox2. Thus, this analysis has provided insight into possible transcriptional mechanisms through which pluripotency may be maintained, and provides a basis for guiding the design of functional studies to investigate the roles of these key TFs and their predicted targets in the acquisition, maintenance and loss of pluripotency, which may have implications regarding the utility of pluripotent cells as tools underpinning the study of developmental biology and disease, drug screening or for regenerative medical therapies.

6.4 Investigation of Myc-Activated Gene Expression Program

cMyc is an important transcription factor involved in the development of a number of cancers [Chang et al., 2008, Lutz et al., 2002]. It was one of the 4 factors shown to induce pluripotency in mouse fibroblasts [Takahashi and Yamanaka, 2006] and subsequent studies have demonstrated its relevance to the transcriptional control of pluripotency through evidence suggesting it greatly enhanced rates of reprogramming of somatic cells to an induced-pluripotent state [Nakagawa et al., 2008]. Despite being a transcription factor of considerable interest, much of the mechanisms responsible for these important biological roles of cMyc are as yet unknown. For example, it has been proposed that the role of cMyc in reprogramming involves facilitating binding of the other reprogramming factors (Oct4, Sox2 and Klf4) through modification of the epigenome of the somatic origin cells prior to induction of pluripotency [Sridharan et al., 2009], consistent with observation that Myc proteins (and cMyc in particular) have global effects increasing the accessibility of chromatin [Knoepfler et al., 2006]. Under such assumptions of a mechanism that has the capacity to reverse existing epigenetic modifications, it would be expected that at least some of the regulatory targets of such a TF would show similar expression dependencies on the TF regardless of biological context.

With a large compendium of gene expression data, the novel meta-analysis approaches developed through this work offer the opportunity to investigate the likely

validity of such claims regarding the mechanisms of cMyc transcriptional regulation, in terms of differences in gene expression co-dependencies of cMyc's DNA-binding targets as observed across different biological contexts. An investigation of gene expression dependencies of the DNA-binding targets of cMyc using the HBLCA meta-analysis approach is described in the following section, with expression observations shown to provide some insight into the mechanisms of the transcriptional regulatory activity of cMyc.

Recently, considerable interest has been shown in the association of gene expression signatures between ES cells and cancers, and the influence of cMyc on these signatures. Examples of this work are provided in [Ben-Porath et al., 2008] and [Wong et al., 2008], which both use the 'module map' approach based on gene set expression analysis described in [Segal et al., 2004]. A very recently published study investigates this relationship between cMyc and gene expression in ES cells and cancer in further detail [Kim et al., 2010a]. As the relationship between cMyc and pluripotency is of particular interest, an investigation into the expression patterns of predicted cMyc targets and ES cell expressed genes in cMyc-expressing and ES cell samples was carried out using the HBLCA technique.

6.4.1 Integrated Analysis to Identify Myc Targets

Using an implementation of HBLCA with cMyc as a guide gene, a set of biclusters involving similar samples with expression contrasts of cMyc were obtained. This set of biclusters was filtered on the basis of visual inspection of the discovered expression patterns displayed by the bicluster visualisation tool described in Section (5.1.7). The resulting set of reliable biclusters was grouped according to similarity of resulting gene-associations of each bicluster, with the gap statistic being used to identify the optimal number of clusters of biclusters, in a similar manner to the clustering performed on the list of Oct4 differential expression targets described in Section (6.1.1). A bayesian integration approach was used to identify genes with consistent expression co-dependency patterns observed across multiple biclusters from the whole set of all reliable biclusters, in addition to genes with such patterns consistent across each subset of the biclusters identified through the similarity-based clustering. Each of these processes was performed as described in Section (5.1).

A set of genes with proximal cMyc-binding was obtained from publicly available high-throughput chIP datsets (mapping of bound sequences to genes was performed by S. Morfopoulou). This set of putative binding targets of cMyc was used to obtain a set of high-confidence cMyc targets through integration with the results of gene expression meta-analysis, as described in Section (5.3.3).

Distribution of cMyc Targets' Expression

As a preliminary step towards analysis of cMyc target expression patterns, samples in the large collection of gene expression data described in Section (3.3) were assigned sample type classifications on the basis of thorough investigation of the records associated with the individual datasets from which the raw data had been obtained. Using these sample type classifications as the basis for a colour-coded annotation bar, heatmaps could be produced to show the expression level of any gene represented on the dataset's microarray platform across all the samples in the dataset, with quick inference of the biological context represented by any particular measurement. Just such a heatmap, showing the expression level of the bound cMyc targets with most highly correlated expression profiles to cMyc, is presented in Fig. 6.17.

The expression profiles shown in Fig. 6.17 show that some of the genes bound by cMyc are also well correlated with its expression across this large dataset. While correlation across large datasets has been shown to be an effective means of identifying genes with related biological roles [Day et al., 2009], such correlations may arise from genes being expressed in the same biological contexts, analogous to the sample-specific expression effects discussed in Section (3.6.3). As this class of expression relationship does not provide any direct information regarding expected dependency of gene expression (and thus transcriptional relationships) with the gene of interest, the investigation of cMyc target distribution presented here focuses on targets with expression co-dependencies in different biological contexts, even if the overall correlation of these targets' profiles to that of cMyc is not particularly pronounced. This effect is demonstrated in Fig. 6.18, showing expression profiles for cMyc binding targets with clear cMyc expression co-dependency in at least some biological contexts, that are relatively poorly correlated with cMyc in terms of the global expression profile across the dataset.

While the expression heatmaps shown in Figs. 6.17 & 6.18 illustrate effectively the biological contexts in which particular cMyc binding targets are expressed, it is difficult to explore expression dependencies in such a plot, owing to the fact that that would require detailed examination of the expression trends within subsets of the dataset that may appear too small for effective scrutiny. However, this information is the basis of the bicluster evaluation employed in HBLCA. Therefore, to investigate the distribution of cMyc target dependencies on cMyc, and therefore likely transcriptional relationships, across different biological contexts, more detailed analysis of the output from the biclustering meta-analysis was performed.

Distribution of Targets' cMyc Co-Dependency Across Biological Contexts

As a means of providing some insight into the specificity of high-confidence observations of expression co-dependency with cMyc, across a range of biological contexts, a

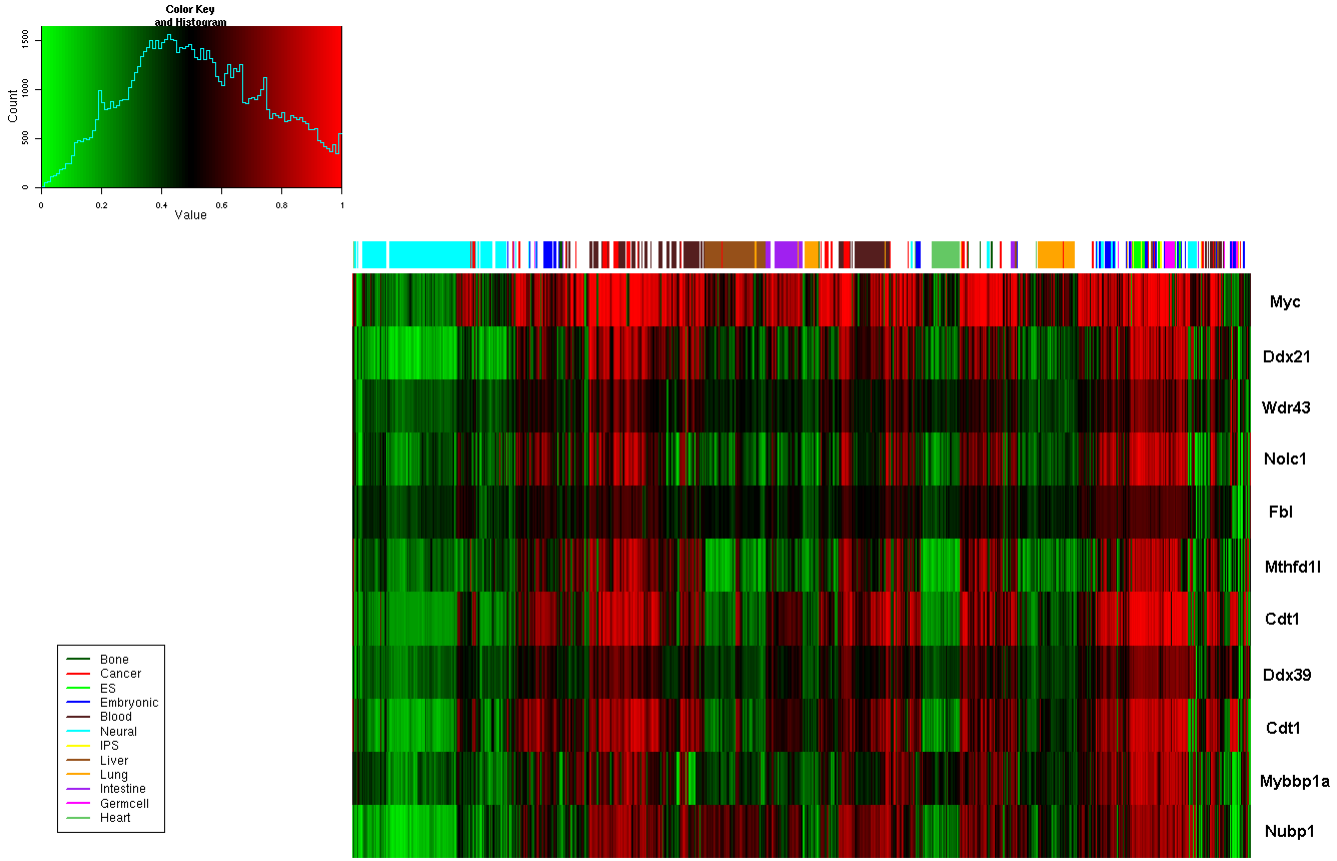


Figure 6.17: Heatmap showing GESTr-transformed gene expression state values for cMyc correlated targets across all samples in a large collection of microarray data. cMyc measurement is included in the top row of the heatmap, and the colourbar above the heatmap indicates the broad biological context classification of each sample (with key in the bottom left of the figure).

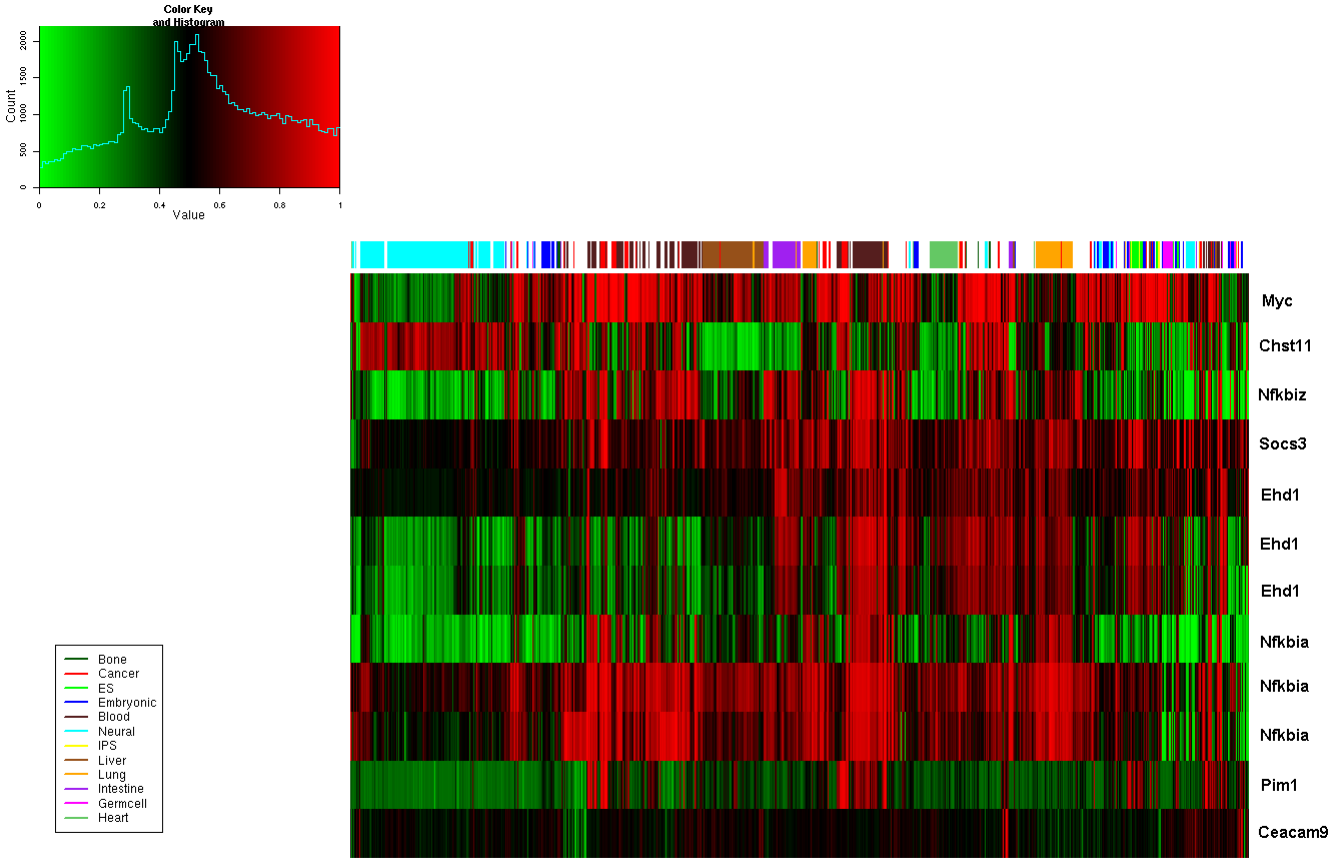
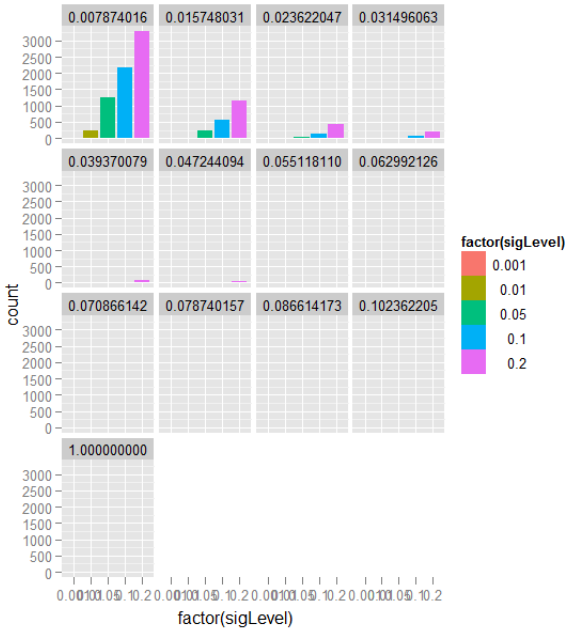


Figure 6.18: Heatmap showing GESTr-transformed gene expression state values for high confidence cMyc targets (from integrated meta-analysis) across all samples in a large collection of microarray data. cMyc measurement is included in the top row of the heatmap, and the colourbar above the heatmap indicates the broad biological context classification of each sample (with key in the bottom left of the figure).

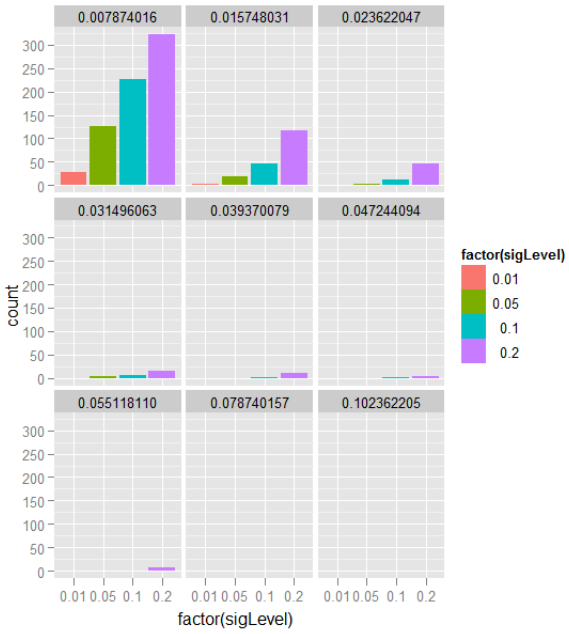
specificity measure was defined for each gene that achieves at least a specified rank in any bicluster as the proportion of the other biclusters in which that gene is again observed to achieve at least that rank. This specificity level was evaluated for a range of maximum-rank thresholds, and the distribution of the proportions of biclusters containing each gene is shown in Fig. 6.19 (first with all significant bicluster genes and then with only those significant bicluster genes for which there exists DNA-binding evidence from cMyc) for the significance thresholds corresponding to particular rank thresholds. Significance levels were evaluated from rankings by the discrete cumulative distribution of the ranks (i.e. probability that a randomly chosen gene would have better rank) and scaled to account for multiple testing by multiplying the resulting p-value by the number of lists considered minus one.

An immediate observation from Fig. 6.19 is that in general, it appears rare for genes with clear cMyc expression co-dependency in one bicluster to show a clear cMyc expression co-dependency in many other biclusters, with genes even at low ranks rarely appearing in more than 3% of the discovered biclusters. This may not be particularly surprising, on account of the fact that cMyc is expressed in a wide range of biological contexts for which the epigenetic landscapes and the sets of expressed TFs may differ considerably, and so this observation would not necessarily be inconsistent with the hypothesis that cMyc acts to increase chromatin availability, as the resulting dependently expressed genes would also depend on which TFs are expressed in those contexts. These findings suggest that there is not a significant cohort of ubiquitously cMyc-dependent genes, although it does not rule out the chance that there might exist cMyc binding targets which are transcriptionally active wherever cMyc is expressed but are not dependent on cMyc for expression. Owing to the fact that there would be no evidence to support any proposed role of cMyc in regulation of transcription of such targets, with the datasets and analysis tools currently available it would be more interesting to focus on those genes which show a clear expression co-dependency with cMyc (which can be identified through HBLCA).

In order to explore further the observed expression-association distribution patterns, a large collection of randomly sampled pairs of biclusters was created and the rank in each bicluster genelist of a randomly chosen gene present in at least one of the lists was recorded. This set of ranking-pairs was used to establish the effectiveness of a gene's rank in one bicluster being used to predict the same gene's rank in any other (randomly chosen) bicluster. Predictive models such as this were used to assess the impact on bicluster distribution of different effects such as the difference between biological contexts represented by the two randomly chosen biclusters, whether or not there is evidence for cMyc binding the gene, or the level of expression of the gene in ES cells. The first approach taken for this predictive analysis was to classify the genes' ranks in each of the two biclusters into bins, for each pair observation, and to plot the



(a) Proportion of biclusters including significant cMyc expression co-dependency of genes significant in any bicluster



(b) Proportion of biclusters including significant cMyc expression co-dependency of genes with cMyc DNA-binding and significant expression co-dependency in any bicluster

Figure 6.19: Bicluster-specificity of genes with significant cMyc expression co-dependency observed. Distribution of numbers of genes which appear significantly highly ranked (to a threshold determined by 'sigLevel') in any bicluster, separated into different 'universality' categories with the proportion of biclusters in which those genes are significantly ranked given by the number at the top of the corresponding panel.

distribution of the numbers of genes from the random sample with rank classification for the first bicluster across the different rank classifications for the second bicluster. Fig. 6.20 shows these results, which reinforce the initial observation made in the previous paragraph that genes with a clear cMyc expression co-dependency in one bicluster (or biological context) are unlikely to show a similar co-dependency in any other bicluster (or biological context).

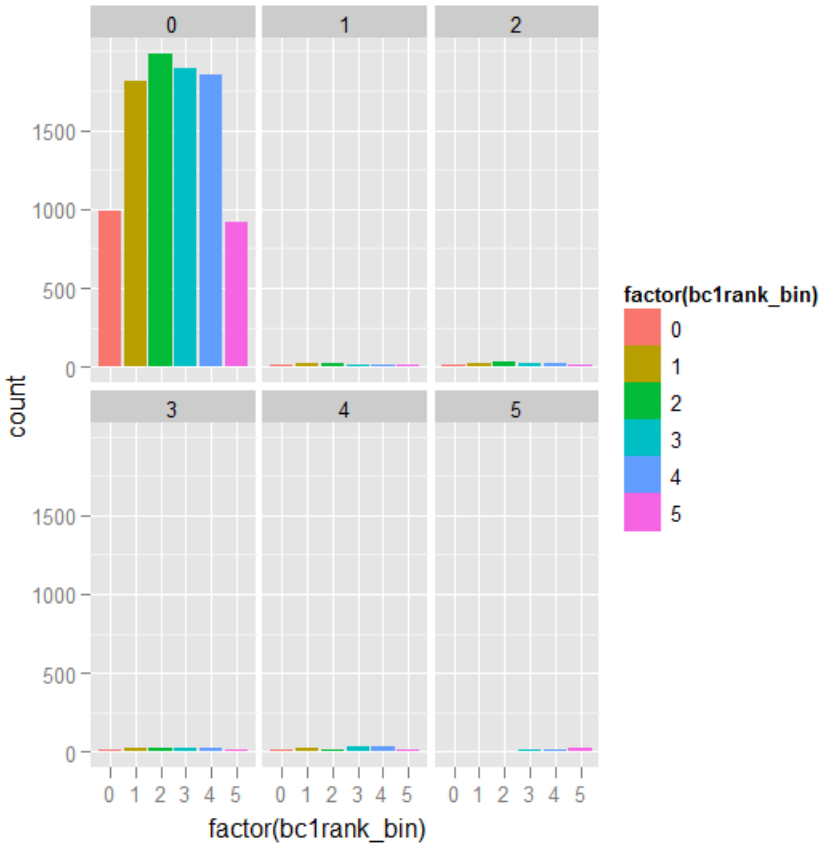


Figure 6.20: Binned distributions of cMyc-associated genes from pairs of biclusters, rank in first bicluster indicated by the colour of histogram, rank in second bicluster indicated by the panel in the plot. A lower bin number represents a better rank, with the exception of bin number '0' which indicates that the gene wasn't in the second bicluster's corresponding genelist.

Following the removal of the bicluster rank pairs for which the gene was absent from one of the biclusters, a correlation between the ranks of each bicluster did emerge. Fig. 6.21 shows the distribution of ranks in the second bicluster of a pair for bins into which a gene is classified on the basis of rank in the first bicluster of the corresponding bicluster pair. This shows a general trend for genes with a better rank in the first bicluster to have a better rank in the second bicluster. It might therefore be inferred that there is a class of targets for which a cMyc expression dependency is consistent across different biological contexts, but in order to rule out the possibility that the

observed correspondence between bicluster genelist ranks of this class of cMyc targets might solely be caused by measurements from pairs of near-identical biclusters, and to relate the observed effect to some biological meaning regarding the distribution of cMyc transcriptional regulatory activity, the bicluster-pair rank relationships would have to be stratified according to the similarity of the two biclusters in the pair. When this is performed, as shown in Fig 6.22, the trend becomes less obvious as the dissimilarity between biclusters increases, indicated in the plot by less clear diagonal progression of higher histogram bars from top-left to bottom-right of each column of panels as the column is further to the right of the figure (representing greater dissimilarity between the two biclusters of each pair). It is difficult to draw any concrete conclusions from this plot, particularly given the low numbers of genes involved that were found in the genelists for both biclusters of a randomly-sampled pair, but the output from this analysis does seem to support the earlier observation that cMyc does not appear to have a substantial cohort of target genes for which it will always activate transcription.

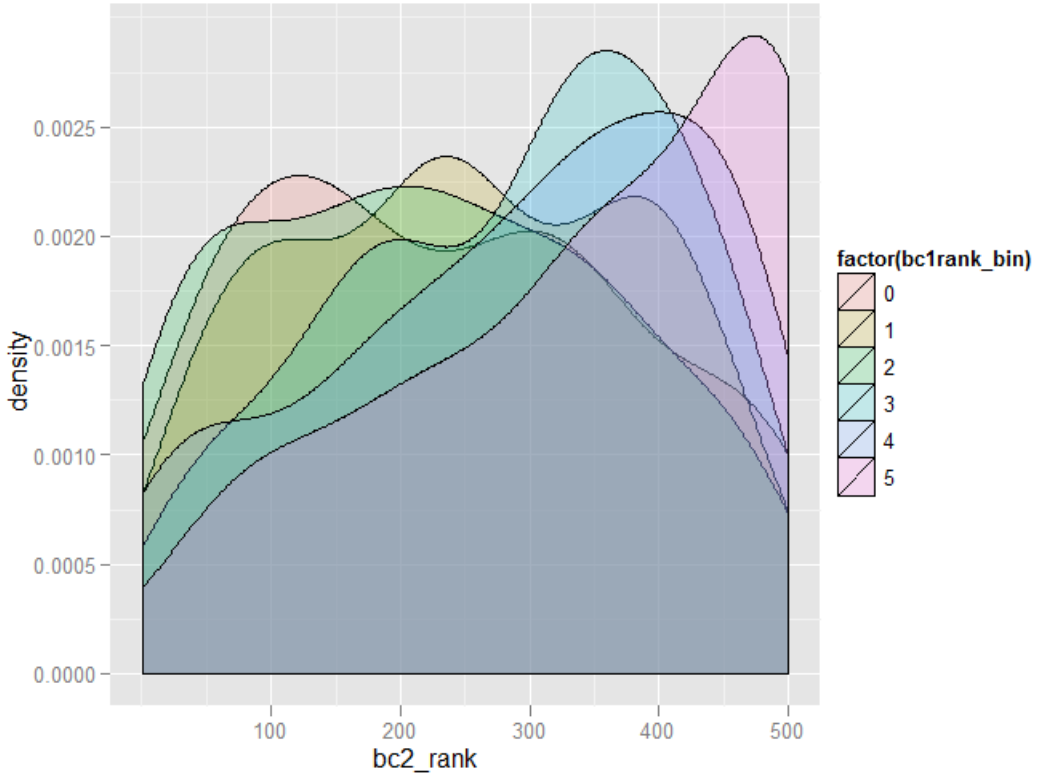


Figure 6.21: Distributions of ranks in the genelist from a randomly chosen bicluster for genes with given classes of rank in the genelist from another randomly chosen bicluster. Each coloured area under a curve represents the distribution of ranks in the secondly randomly chosen bicluster for a class of gene from the first randomly chosen bicluster, where lower class numbers represent better ranks of the gene in question as appearing in the first randomly chosen bicluster's genelist.

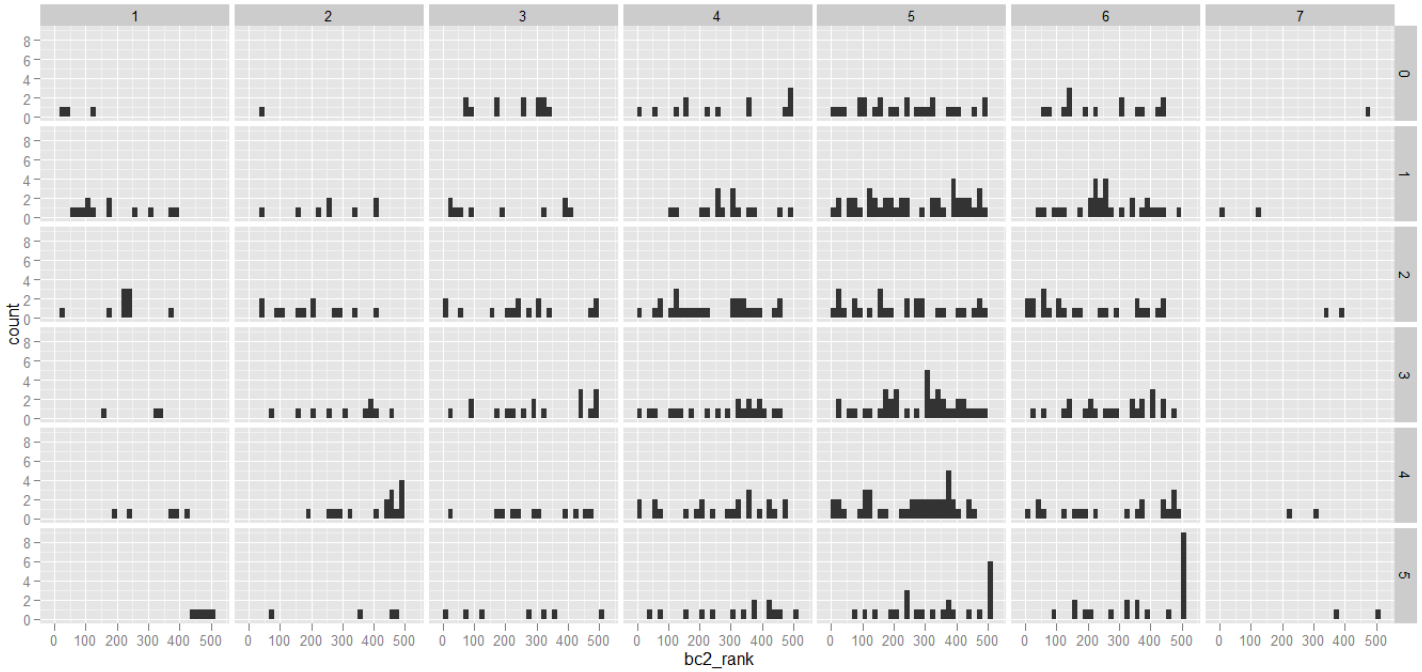


Figure 6.22: Distributions of bicluster genelist ranks for genes in randomly sampled pairs of biclusters, stratified according to dissimilarity of the samples from each bicluster in the pair. Columns of panels represent increasing bicluster sample dissimilarity from left to right, rows of panels represent worse rank in the first bicluster of the pair from top to bottom. Ranks in the second bicluster increase from left to right within each panel. Therefore, a relationship of a gene's rank in one bicluster to its rank in another bicluster is indicated by a trend in a column for the distribution density indicated by the individual bars in each panel to shift from the left-hand side in the top panel of a column to the right-hand side in the bottom panel of the column. This relationships seems to be apparent only in the leftmost two columns, which correspond to bicluster pairs involving relatively similar samples (from the perspective of global transcriptional profiles).

In an attempt to identify any cMyc target genes with evidence for cMyc expression dependency across a broad range of biological contexts and to evaluate further the extent to which such expression patterns can be found across existing data, a novel visualisation approach was adopted. Following the identification of a set of reliable cMyc biclusters (as described at the beginning of this section), this set of biclusters was used as the basis for analysis of the range and strength of cMyc co-dependency of any set of genes of interest. Visualisation of co-dependency patterns was performed by creating a heatmap of cMyc co-dependency likelihood estimates from each bicluster, for each of the genes in the genelist. Interpretation of the biological significance of the displayed patterns was enhanced through association of a biological context label with each of the biclusters and inclusion of a colourbar above the heatmap in a similar manner to that described for the expression level heatmaps shown earlier in this section. Fig. 6.23 provides an example of such a visualisation, here showing the cMyc co-dependency estimates for cMyc targets identified using HBLCA. To provide a contrast, an equivalent visualisation for genes that have the most strongly correlated expression profiles to that of cMyc is given in Fig. 6.24. These plots show that the HBLCA approach helped to identify a set of genes with cMyc expression co-dependency across a broad range of biological contexts, that would not be identified through simple correlation approaches (as demonstrated with expression profile heatmaps shown in Figs. 6.17 & 6.18). To obtain sets of high-confidence targets, the results of gene expression meta-analysis approaches were integrated with results from DNA-binding studies. Distributions of cMyc expression co-dependency scores for these high-confidence targets based on HBLCA and on correlation meta-analysis are shown in Figs. 6.25 & 6.26, respectively. Bicluster visualisations (produced by the tool described in Section (5.1.7)) are given in Fig. 6.27 for the set of relatively broadly cMyc co-dependently expressed genes identified through application of HBLCA, confirming the presence of expression patterns suggested by the co-dependency heatmaps. It may be worth noting that these broad targets do not display notable cMyc-dependent expression in a number of the biclusters, yet these biclusters do show such dependency for other genes as illustrated in Fig. 6.28. This further supports the claim that those cMyc targets with expression most dependent on expression levels of cMyc differ across biological contexts.

A further observation regarding the distribution of cMyc expression co-dependency of targets of cMyc DNA-binding was revealed through the 'subset-integration' of biclusters performed as part of the HBLCA approach. Three subsets of biclusters were identified, and the consistency of the top-ranking targets according to cMyc expression co-dependency scores is illustrated in Fig. 6.29. This further demonstrates the ability of the HBLCA approach to discover relationships in the expression levels of genes and to associate these with relevant biological contexts.

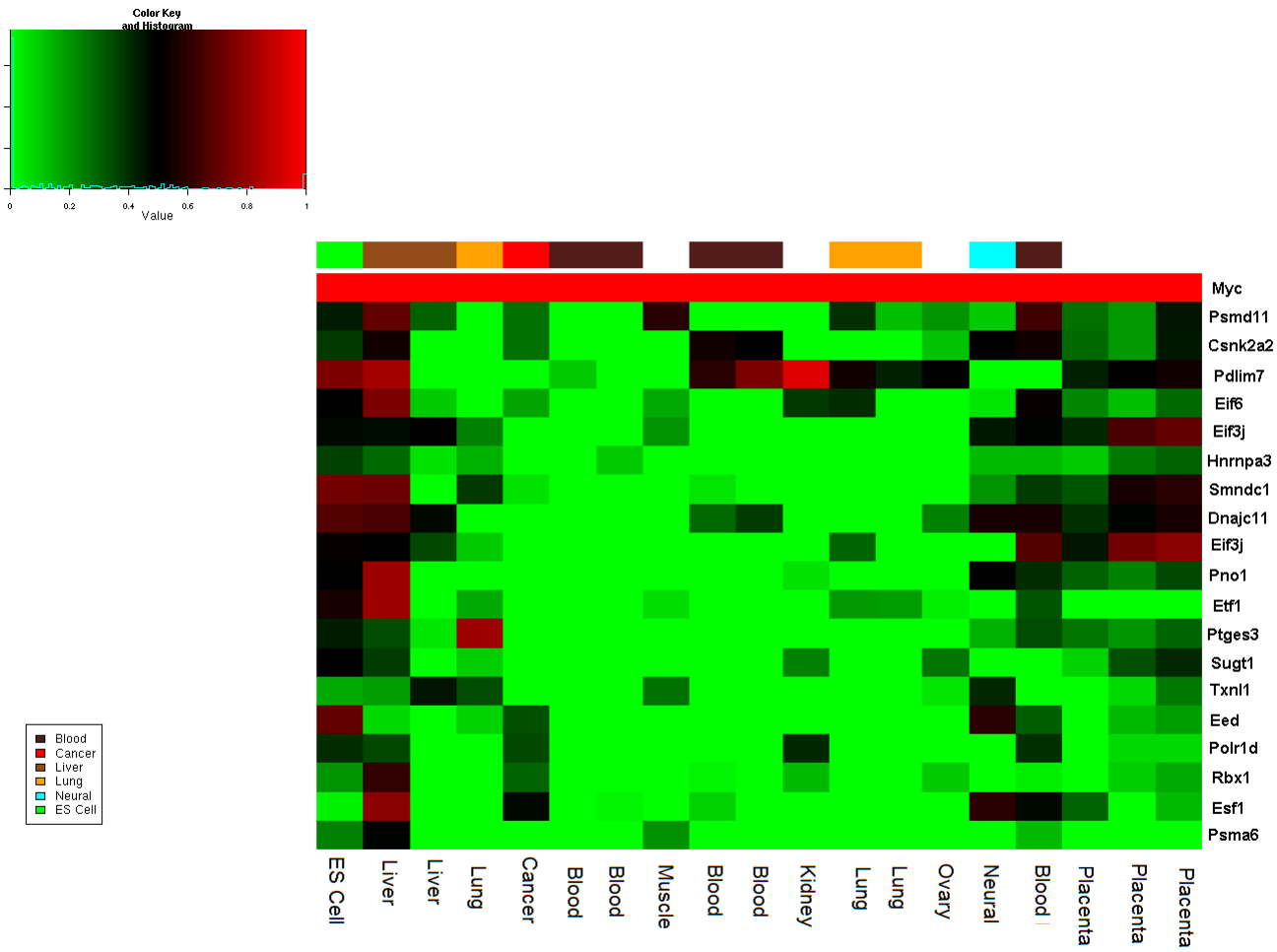


Figure 6.23: Heatmap of rank-based scores from cMyc co-dependency probability estimates for genes identified by HBLCA as displaying cMyc co-dependent expression. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 1000 or lower in the bicluster's co-dependency estimate genelist (green).

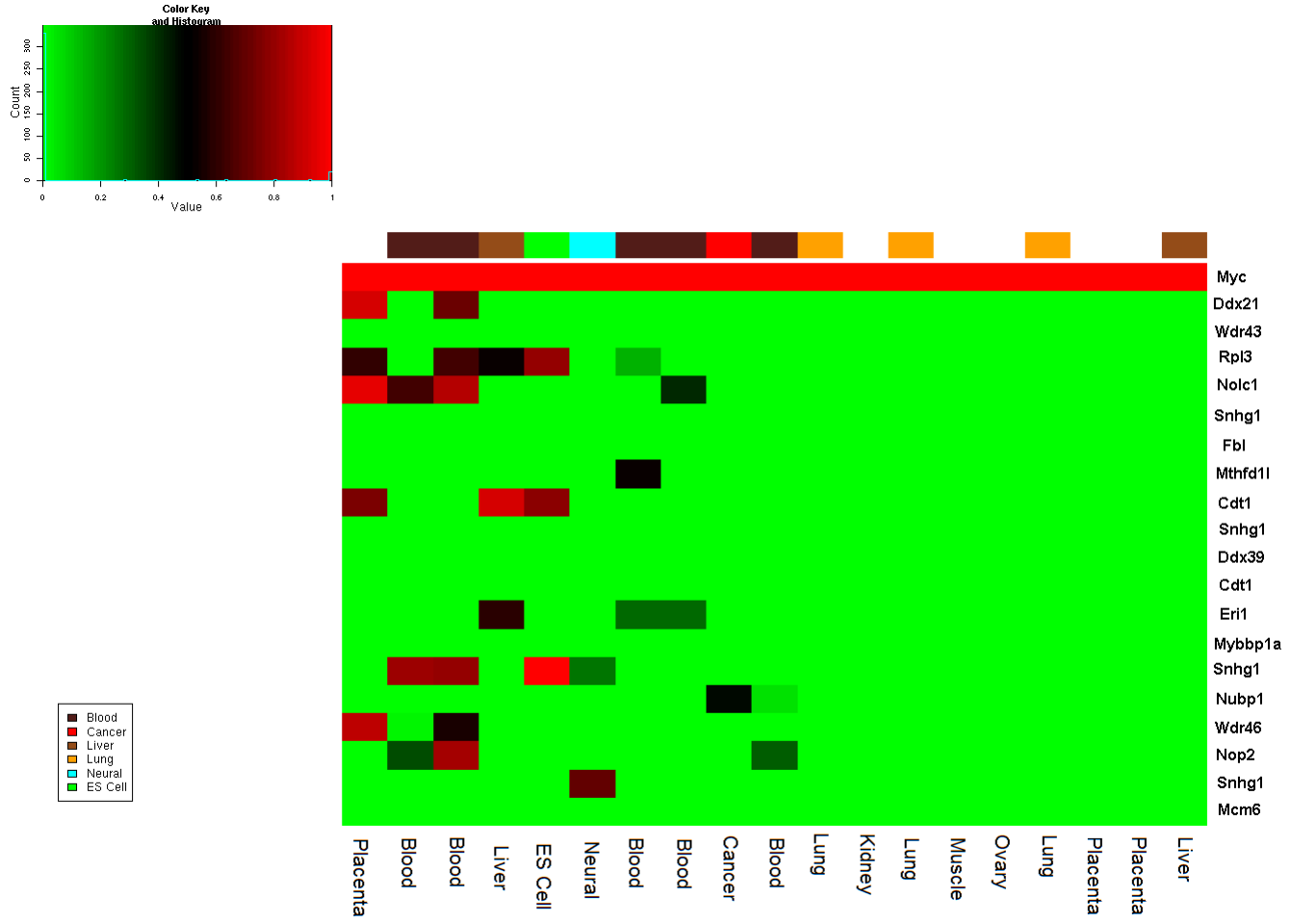


Figure 6.24: Heatmap of rank-based scores from cMyc co-dependency probability estimates for genes with expression profiles most correlated to that of cMyc. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 1000 or lower in the bicluster's co-dependency estimate genelist (green).

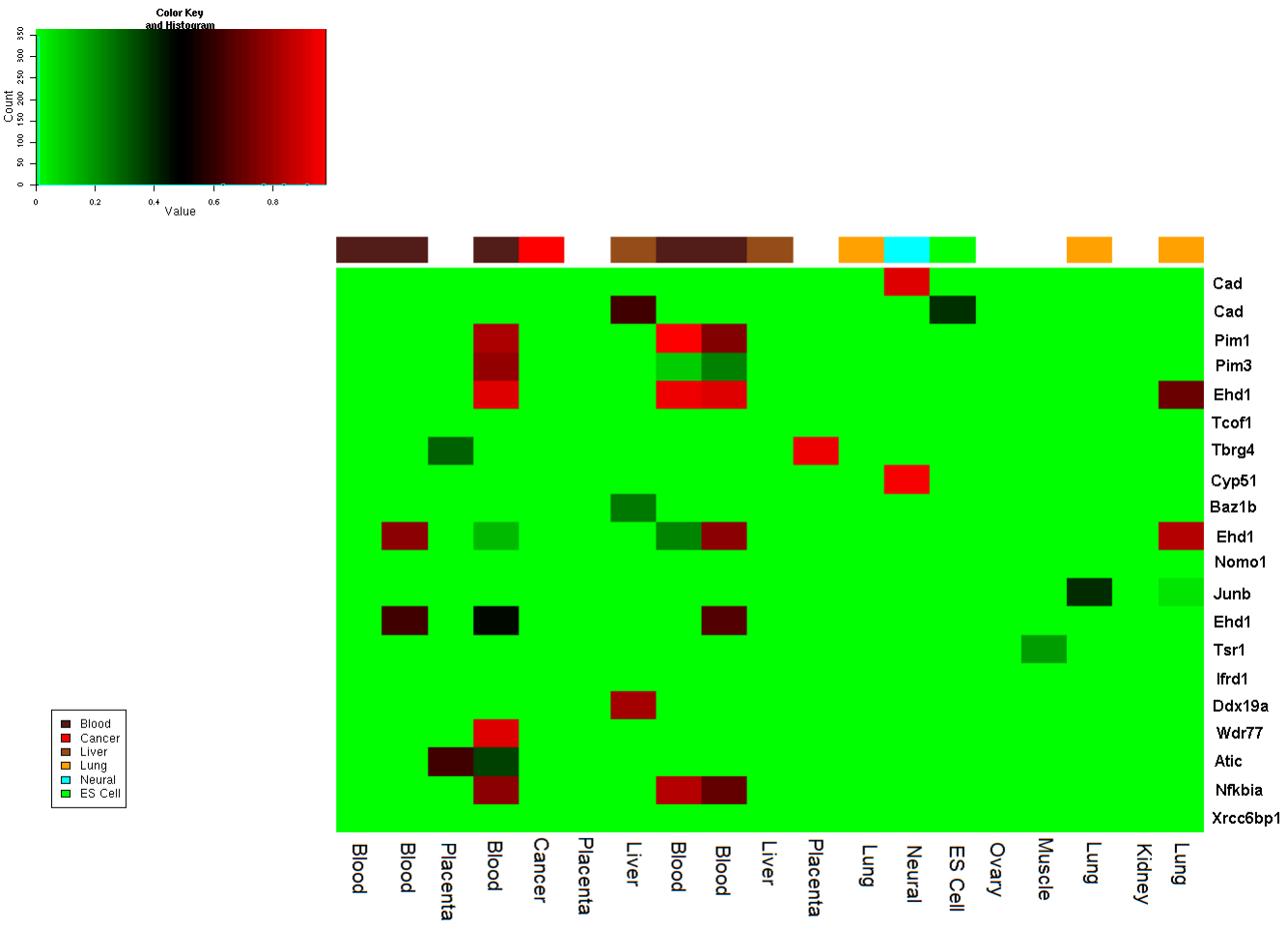


Figure 6.25: Heatmap of rank-based scores from cMyc co-dependency probability estimates for cMyc targets identified by HBLCA. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 100 or lower in the bicluster's co-dependency estimate genelist (green).

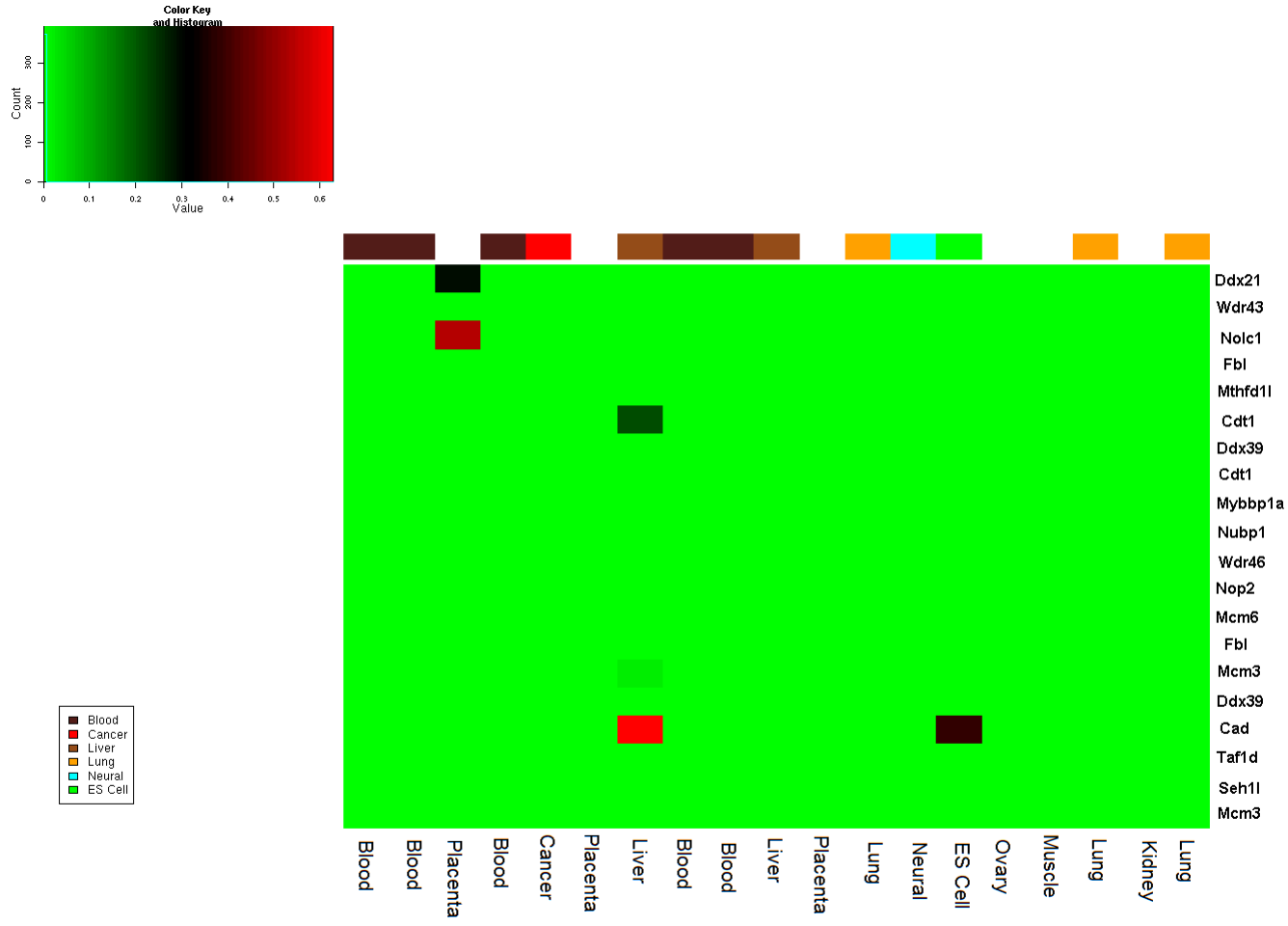


Figure 6.26: Heatmap of rank-based scores from cMyc co-dependency probability estimates for cMyc targets with expression profiles most correlated to that of cMyc. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 100 or lower in the bicluster's co-dependency estimate genelist (green).

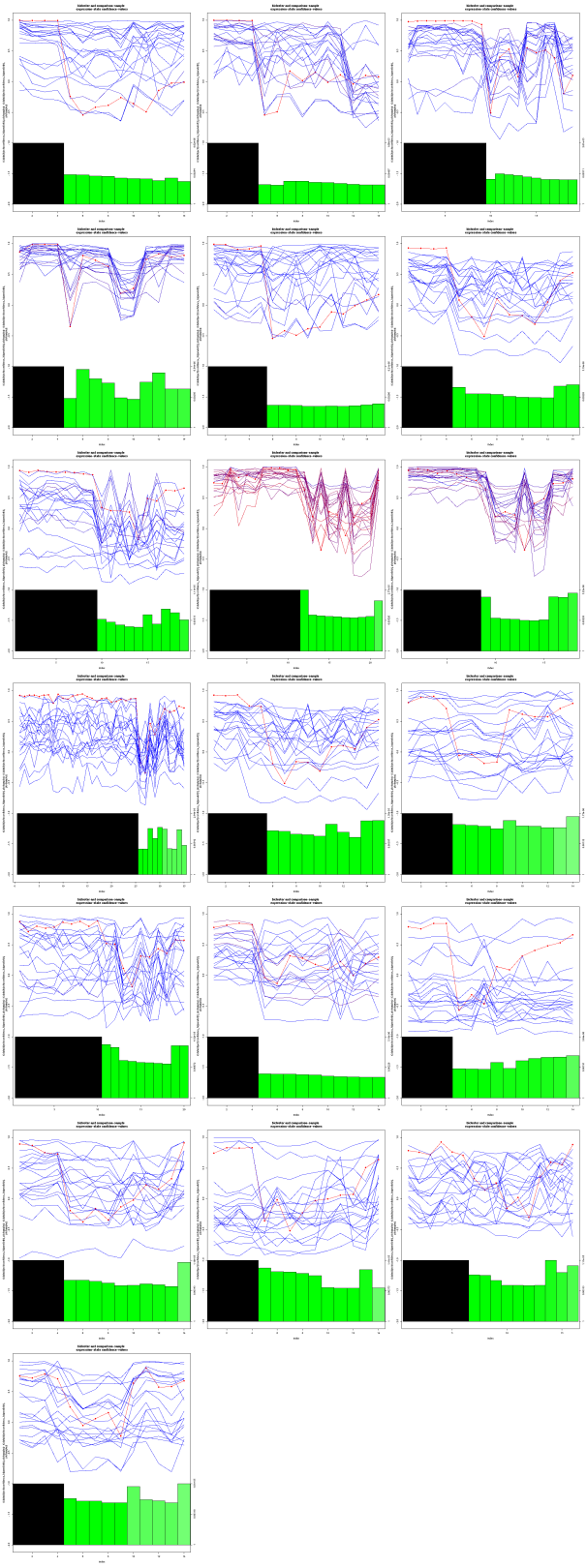


Figure 6.27: Bicluster plots of cMyc targets with relatively broad range of cMyc expression co-dependency, as shown in Fig. 6.25. cMyc expression level is shown with red dotted line, for additional information regarding these plots see Section (5.1.7).

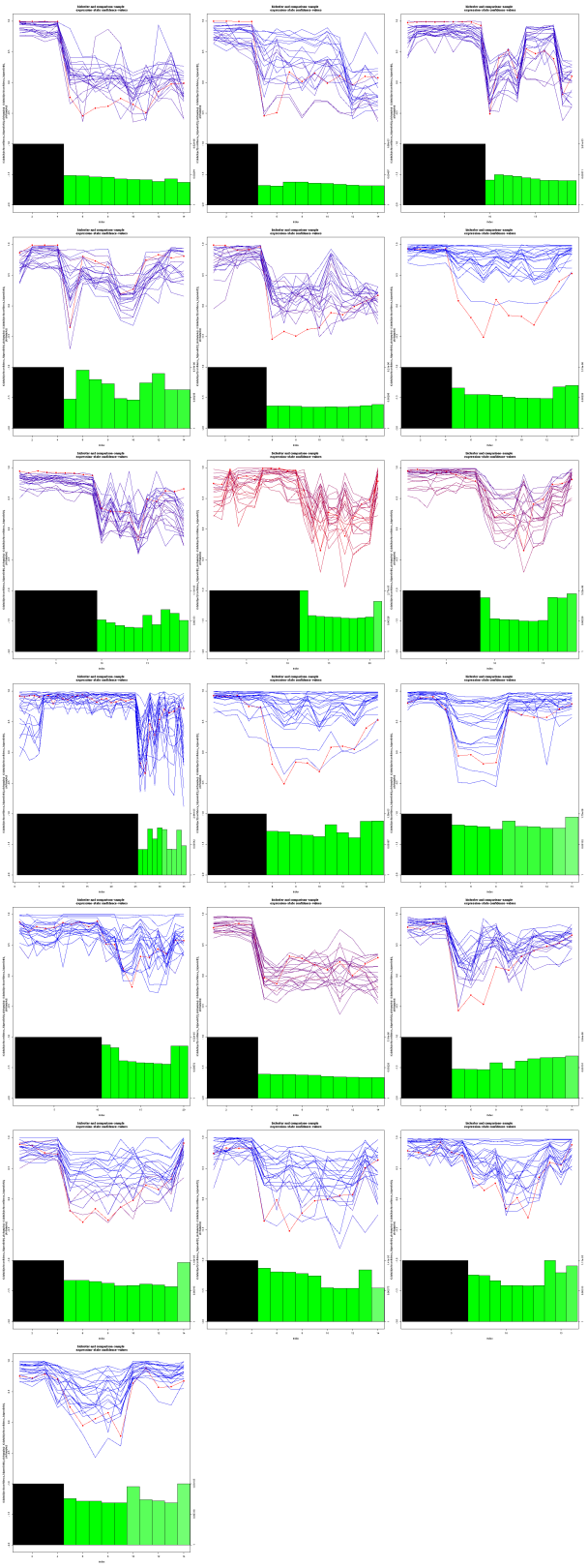
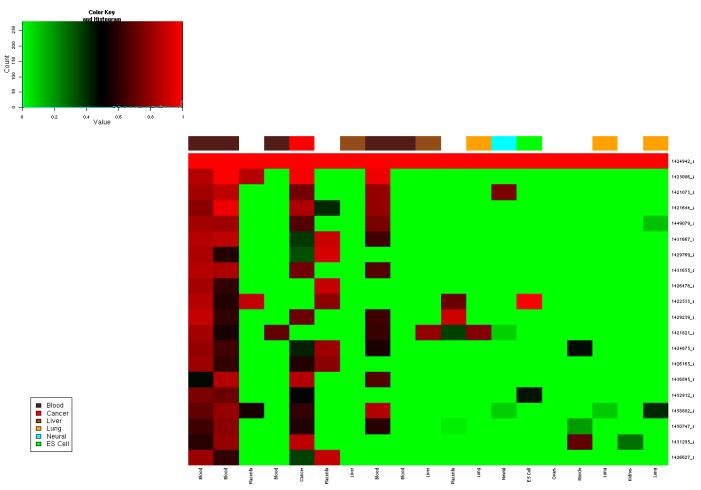
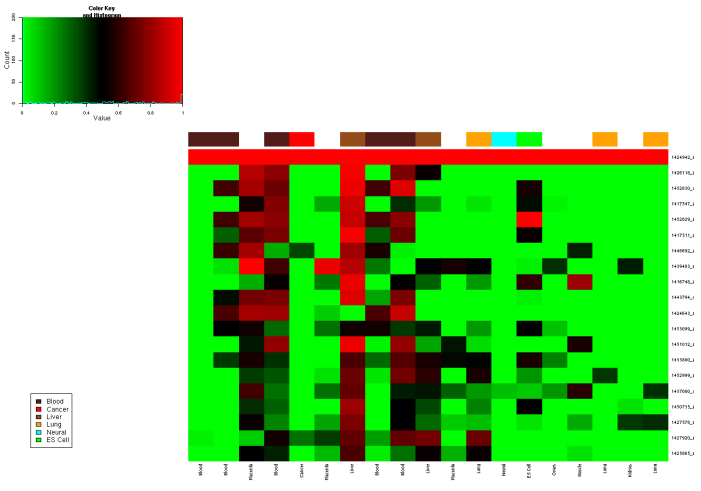


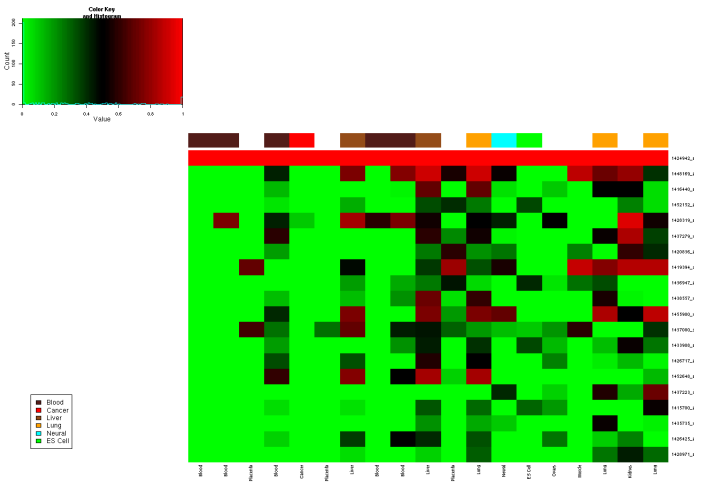
Figure 6.28: Bicluster plots of cMyc targets with most cMyc-dependent expression for each bicluster. Each panel corresponds to a bicluster, and each involves a different set of genes. cMyc expression level is shown with red dotted line, for additional information regarding these plots see Section (5.1.7).



(a) Genes consistently co-dependent across first subset of cMyc biclusters



(b) Genes consistently co-dependent across second subset of cMyc biclusters



(c) Genes consistently co-dependent across third subset of cMyc biclusters

Figure 6.29: Heatmaps showing context-dependent cMyc co-dependency of expression of cMyc targets identified as consistently co-dependent across subsets of cMyc biclusters.

6.4.2 A Myc-Dependent 'Stem Cell-Like' Expression Program

It has been proposed that cMyc promotes an 'ES cell-like transcription pattern' in [Sridharan et al., 2009], and a more concretely defined 'core ES cell program' was proposed in [Wong et al., 2008] to be induced by cMyc expression in cancer cells. It was considered pertinent to apply the novel analysis approaches applied in Section (6.4.1) to investigate the relationship between ES cell expression and cMyc co-dependency, for those genes predicted to be regulated by cMyc as a result of DNA-binding and gene expression co-dependency evidence, and for the proposed cMyc-induced ES cell associated transcriptional signature.

Firstly, investigation of the link between cMyc expression and ES cell associated expression of genes is explored using localised co-dependency analysis. A set of genes is identified as being co-dependently expressed with cMyc in a range of biological contexts, show DNA-binding by cMyc and show higher expression in ES cells than most other biological contexts. Secondly, cMyc expression co-dependency of the genes from the cMyc-induced 'core ES cell module' proposed in [Wong et al., 2008] is analysed. This proposed module is separated into apparently cMyc-responsive and cMyc-unresponsive subsets, and the ES cell associated expression of these subsets is shown.

Expression of Myc Targets in Embryonic Stem Cells

For a trivial examination of the ES cell expression patterns of cMyc targets, the distributions of both low and high expression state values were calculated across ES samples and non-ES cMyc-expressing samples, for those high confidence cMyc targets identified as described above. Fig. 6.30 shows a comparison of these distributions, from which it is clear that genes with proximal cMyc binding sites are generally expressed at a higher level in ES cells than other cMyc-expressing samples. This observed effect may be an artefact due to the fact that the cMyc binding data come from chIP experiments performed in ES cells, and so would be guaranteed to be in accessible chromatin (and thus available to be expressed) in ES cells but not necessarily any other contexts. This observation is not inconsistent with the view that cMyc might promote an ES cell-like transcriptional program, but owing to this potential bias and the incompleteness of this result in explaining ES cell related cMyc-induced expression, further investigation of this hypothesis was warranted.

Global Expression of Myc Stem Cell Targets

If cMyc activated an ES cell-like transcription pattern as suggested in [Sridharan et al., 2009], it would be expected that genes commonly associated with the ES cell (pluripotent) phenotype would have a tendency to be expressed in non-ES cell samples where cMyc was expressed. For a given set of genes that characterise the proposed ES cell-like transcription pattern, it is relatively straightforward to investigate the presence of the

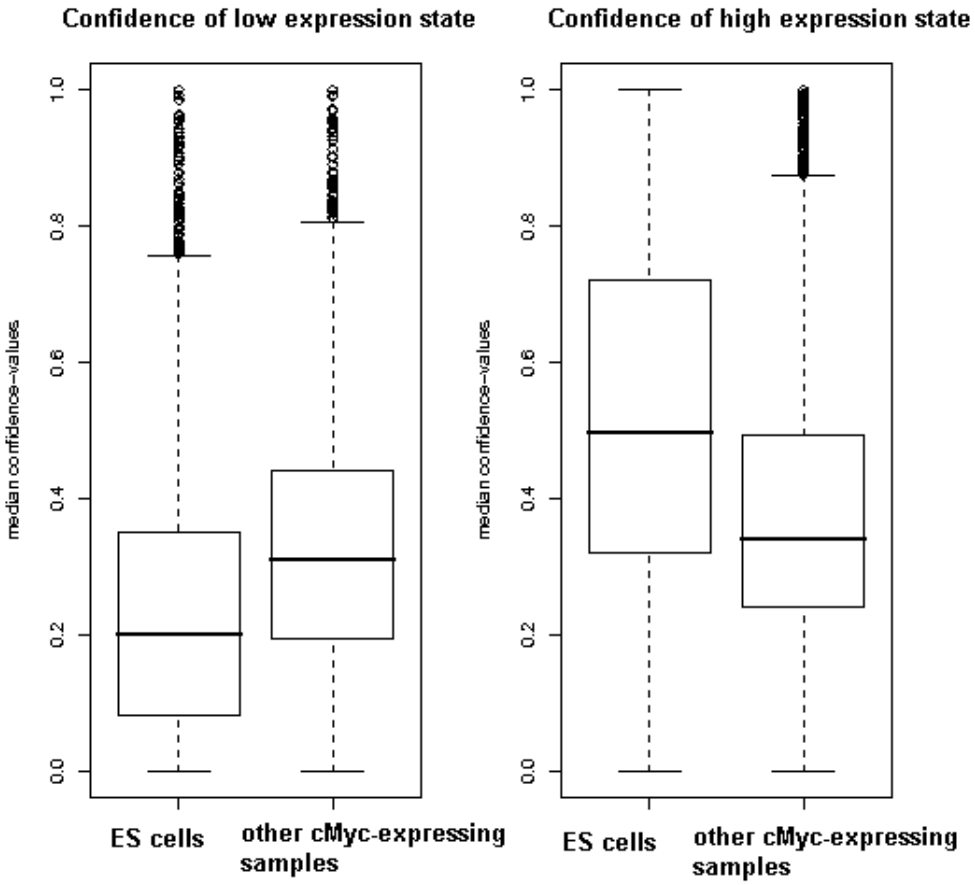


Figure 6.30: Distributions of expression-state confidences for genes with cMyc DNA-binding evidence from chIP studies, shown across ES samples only and across all other cMyc-expressing samples.

expected expression patterns by examining an annotated heatmap (with a row showing cMyc expression level) such as those shown in Figs. 6.17 & 6.18. However, defining *a priori* a list of genes that characterise an ES cell-like expression program is not obvious: it may be only a certain component of the set of genes commonly associated with pluripotency that is the target of cMyc-induced expression. Therefore, a number of different sets of genes with different ES cell and cMyc related expression characteristics are explored in this analysis.

Firstly, a simple definition of ES cell associated genes involves those which show ES cell specific expression. The theory that such genes are likely to be associated with the pluripotent phenotype has been demonstrated through characterisation of the 'ECATs' (Embryonic stem Cell Associated Transcripts) [Mitsui et al., 2003]. Thus, a panel of genes was selected through identifying probesets that show the greatest difference between the mean expression level across samples from the large gene expression dataset described in Section (3.3) annotated as ES cells and the mean expression level across all other samples in the dataset. The expression profiles for this panel of genes are shown in Fig. 6.31. The heatmap in Fig. 6.31 confirms that the genes with expression most specifically limited to ES cells, and therefore likely to be associated to the pluripotent phenotype, do not seem to show any cMyc-dependent expression outside ES cells.

For cMyc to induce the expression of a set of genes, it may be hypothesised that it would be more likely for those genes to have evidence of proximal cMyc DNA-binding. Therefore, an alternative panel of cMyc-bound ES cell associated genes was defined as those with probesets showing greatest ES cell specific expression and requiring a cMyc binding site association in the chIP list. It may be noted that this panel of genes includes the key pluripotency gene Sox2. The expression patterns of these genes are shown in Fig. 6.32, where it can again be observed that the majority of these genes do not show any cMyc-associated expression. However, a number of these genes (Dis3, Igf2bp1 & Igf2bp3) are clearly expressed outside ES cells only when cMyc is expressed. It is not obvious how these genes would be characterised in terms of the continuous spectrum representing the trade-off between ES cell specific expression and cMyc-expressing sample expression. At the opposite end of this spectrum are the genes with consistently high expression in ES cells, cMyc binding evidence and high correlation to cMyc expression. Expression profiles for a panel comprising genes best meeting these criteria is shown in Fig. 6.33.

The examples shown in Figs. 6.32 & 6.33 highlight a critical issue regarding the definition and characterisation of a cMyc-induced ES cell transcriptional program from gene expression data. As cMyc is expressed in ES cells at consistently high levels, there is nothing to distinguish genes with expression that is consistently coincident with cMyc from genes whose expression in ES cells is critical to the pluripotent phenotype

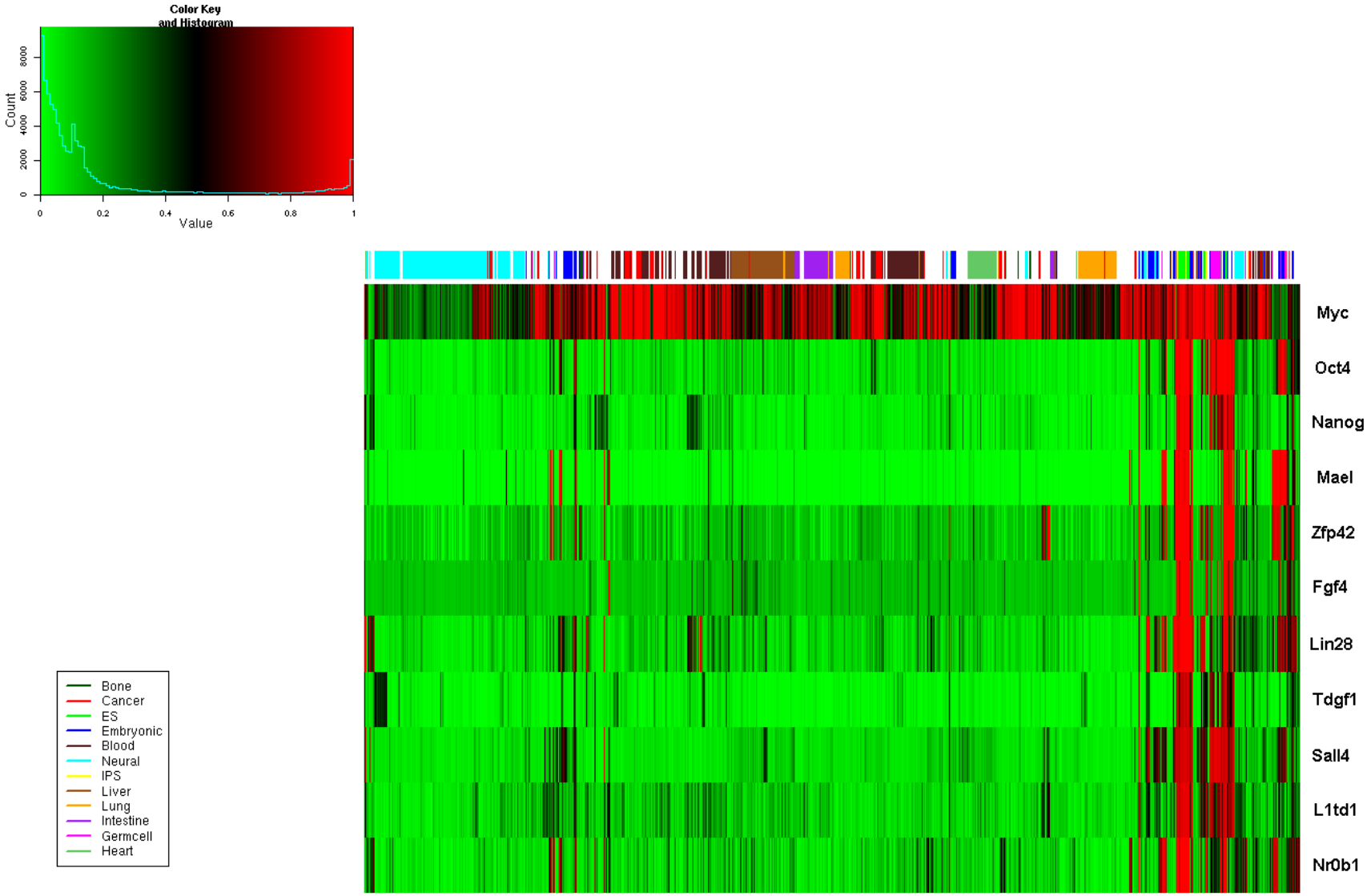


Figure 6.31: Heatmap showing expression profiles (from GESTr-transformed values) of most 'ES cell specifically expressed' genes in a large collection of processed microarray data. cMyc expression is shown in the first row of the heatmap.

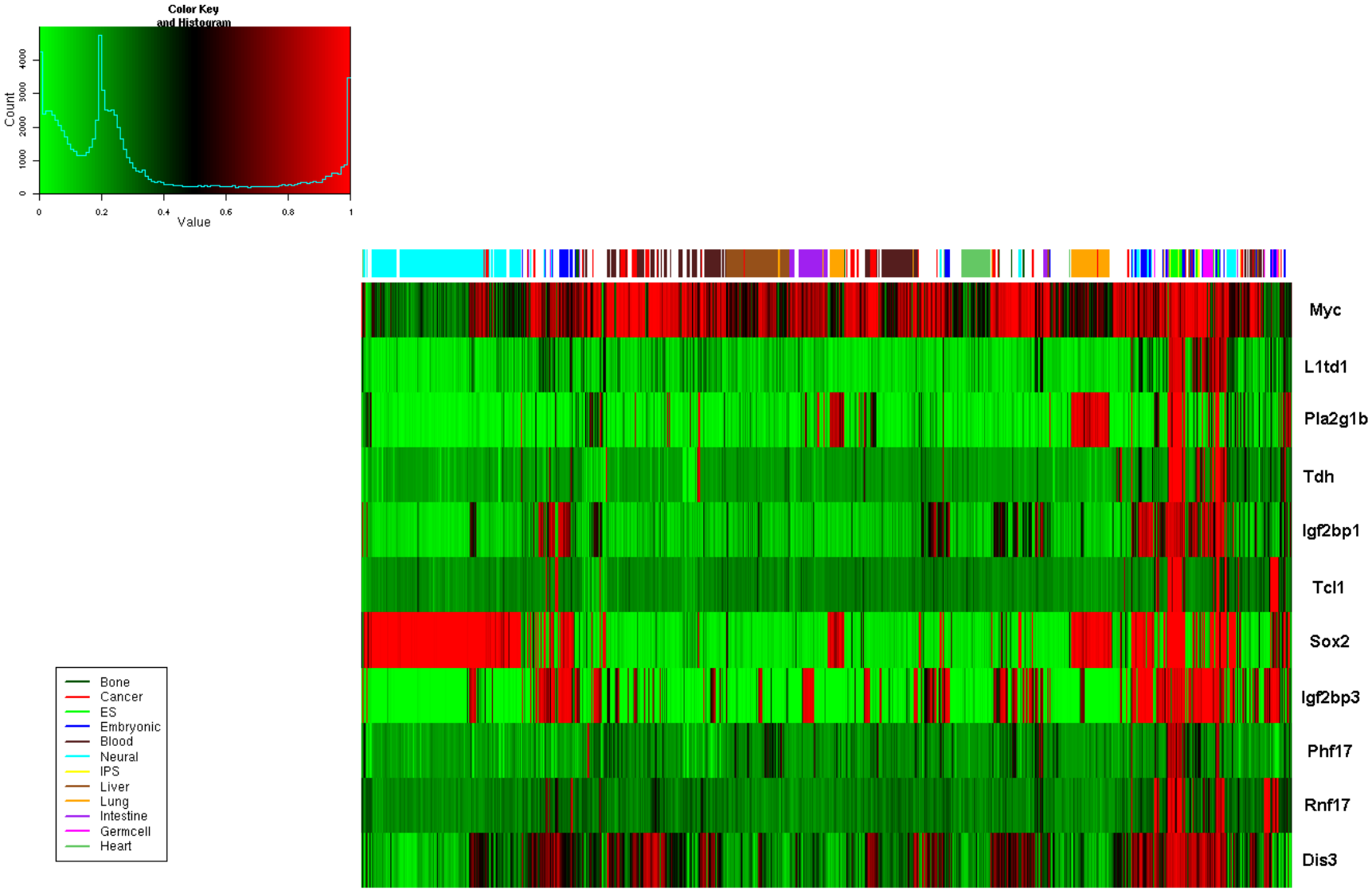


Figure 6.32: Heatmap showing expression profiles (from GESTr-transformed values) across a large collection of processed microarray data of most 'ES cell specifically expressed' genes with evidence for proximal cMyc binding sites. cMyc expression is shown in the first row of the heatmap.

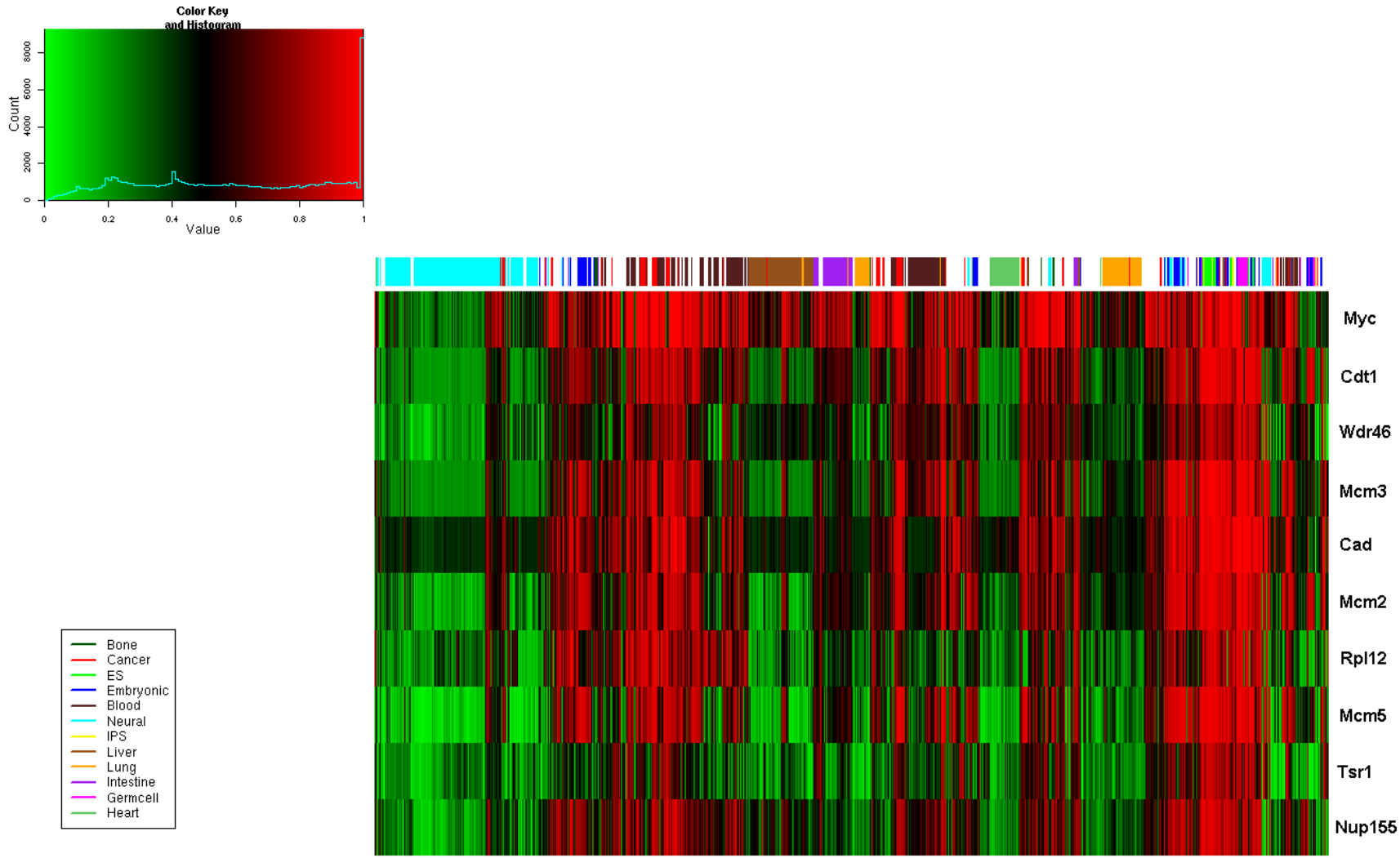


Figure 6.33: Heatmap showing expression profiles (from GESTr-transformed values) of genes with most cMyc-correlated expression across a large collection of processed microarray data and consistently high expression in ES cell samples. cMyc expression is shown in the first row of the heatmap.

of ES cells and whose expression outside ES cells may be induced by cMyc. However, the HBLCA approach provides some opportunity for investigation of cMyc-dependent expression of pluripotency-associated genes, especially when used in conjunction with the DNA-binding data.

Given that genes typically associated with pluripotency (ECATs) did not appear to show cMyc-dependent expression, even those with cMyc binding sites, an alternative set of ES cell specific cMyc target candidates was identified through simultaneous optimisation of evidence for broad cMyc expression co-dependency, consistently high expression relatively specific to ES cells, and evidence for proximal cMyc binding sites. It is hypothesised that this list would constitute cMyc regulatory targets that are associated with the pluripotent phenotype.

As utilised earlier in this section, the biclustering meta-analysis approach described in Section (5.1) provides a means for exploration of the range of biological contexts in which a set of genes display expression co-dependency with a particular gene of interest. Firstly, application of the novel visualisation technique demonstrated in Figs. 6.23 & 6.24 was used to confirm that the most ES-specifically expressed cMyc-bound genes do not display any significant cMyc expression co-dependency outside ES cells, as shown in Fig. 6.34. This approach was also used to visualise cMyc dependency of the genes in the panel of cMyc regulatory targets associated with the pluripotent phenotype (described in the previous paragraph), with the heatmap shown in Fig. 6.35. This figure illustrates that the majority of these ES cell associated cMyc target candidates do show cMyc co-dependency in a range of biological contexts, most notably in ES cells, blood (macrophages and B cells), liver and placenta. To demonstrate the ES cell association of these genes, their expression level distributions across ES cells and all other samples in the collection of gene expression data used for meta-analysis are shown in Figs. 6.36 (for GESTr transformed expression state values) & 6.37 (for RMA-normalized expression values). The expression of these candidate genes in ES cells is shown to be consistently towards the top of their range of expression levels across a large collection of data, and it is relatively rare for these genes to be expressed at such high levels outside ES cells. However, from Fig. 6.35 it can be seen that these genes show cMyc dependency in a range of biological contexts, therefore in addition to the fact that there is evidence to suggest that there exist cMyc binding sites proximal to each of these genes, these may well constitute the best available set of cMyc target genes associated with pluripotency.

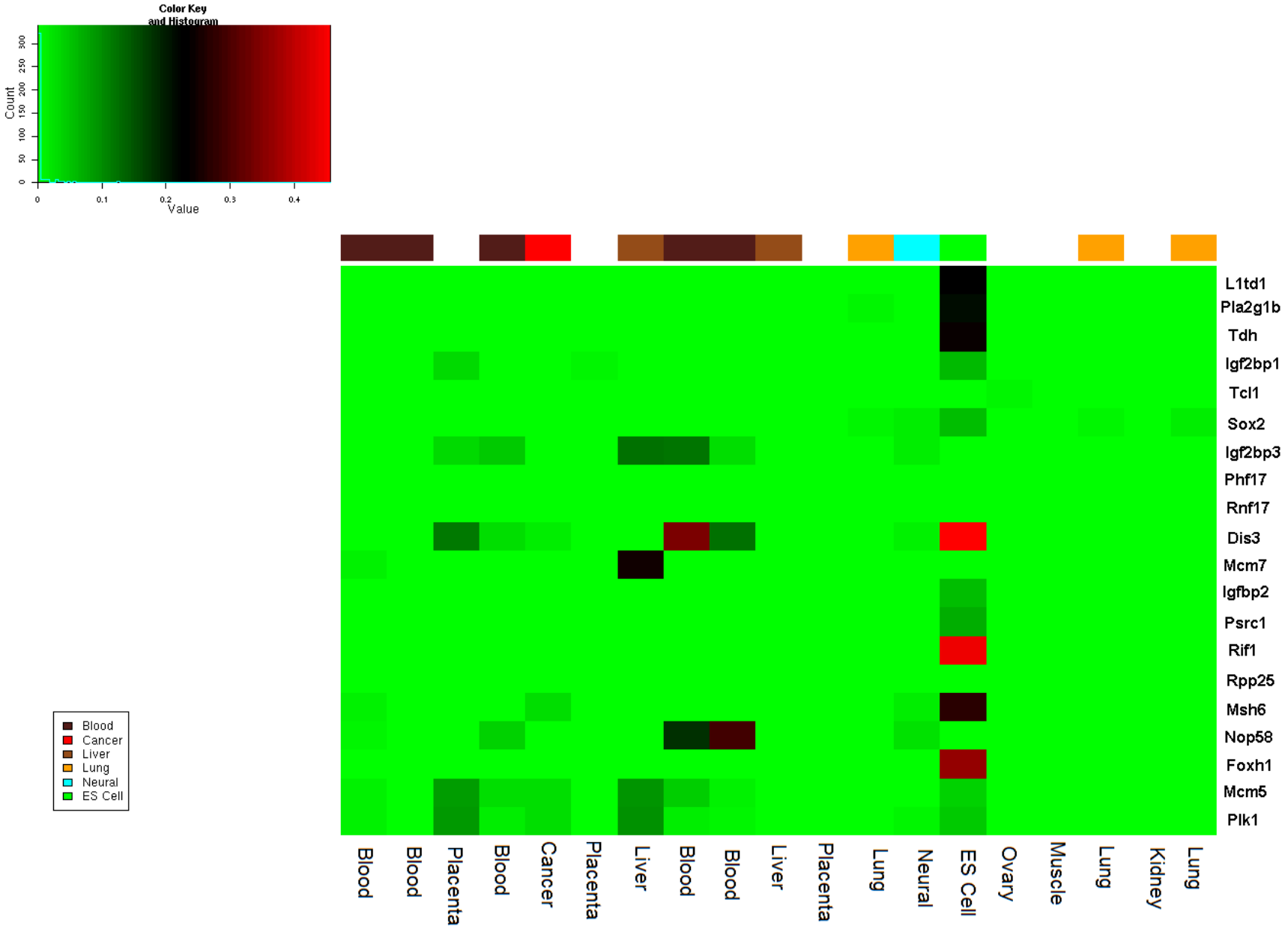


Figure 6.34: Heatmap of cMyc co-dependency probability estimates for cMyc-bound genes most specifically expressed in ES cells. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context.

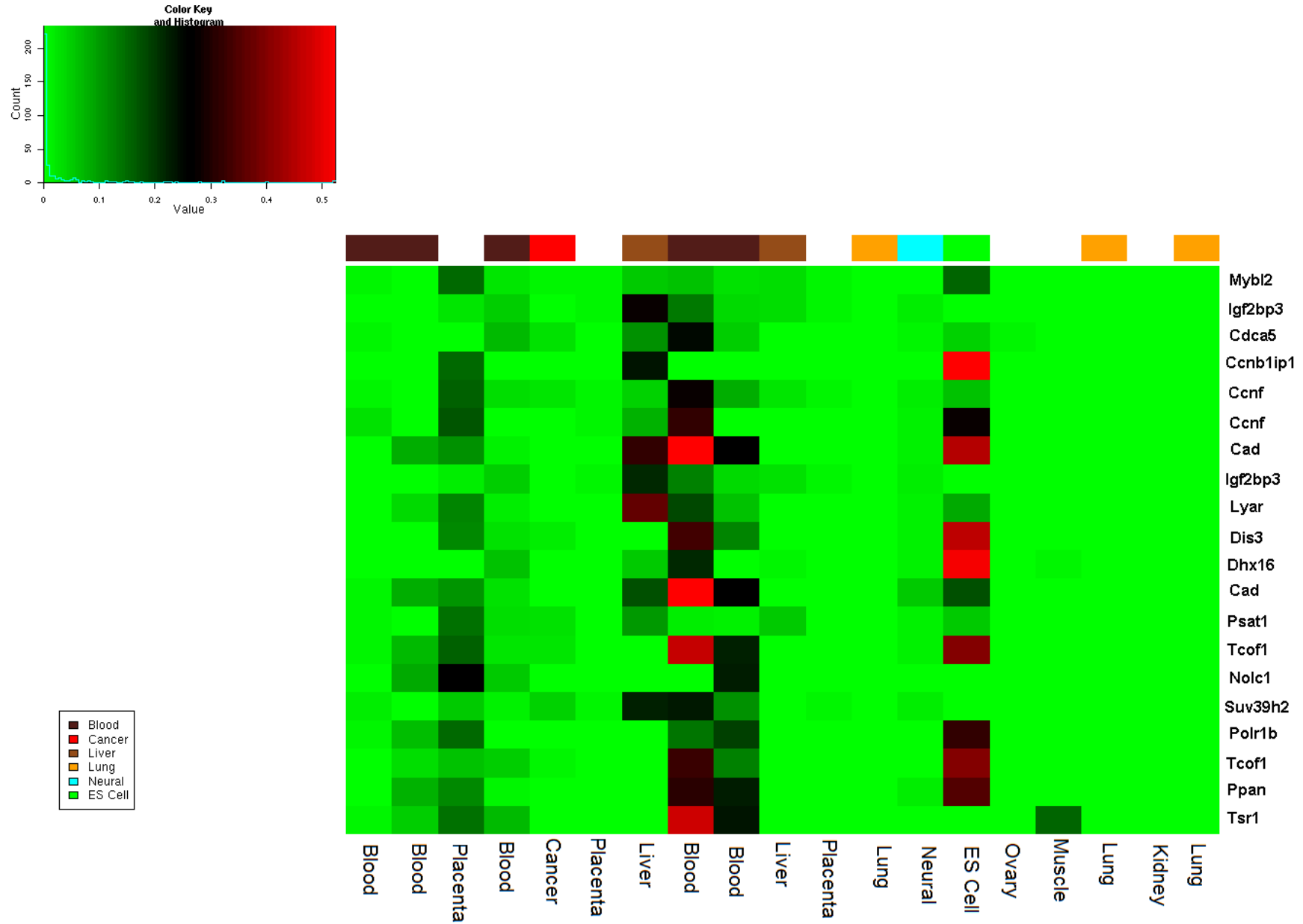


Figure 6.35: Heatmap of cMyc co-dependency probability estimates for genes with DNA-binding by cMyc, broad cMyc co-dependency and high expression relatively specific to ES cells. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context.

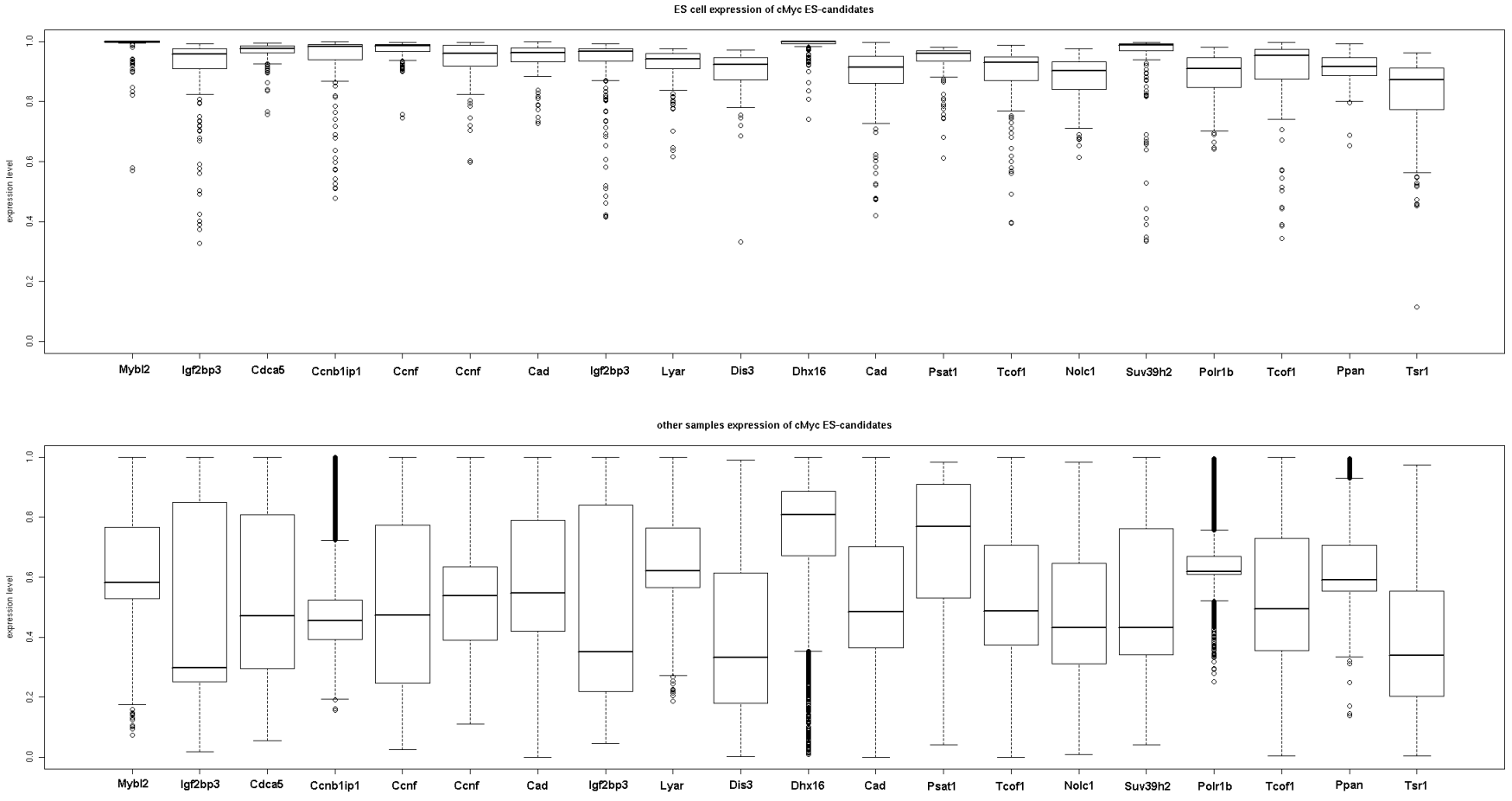


Figure 6.36: Distributions of GESTr-transformed expression state values for pluripotency-associated cMyc target candidates, shown in the top panel across ES cells only and in the bottom panel across all other samples in the dataset.

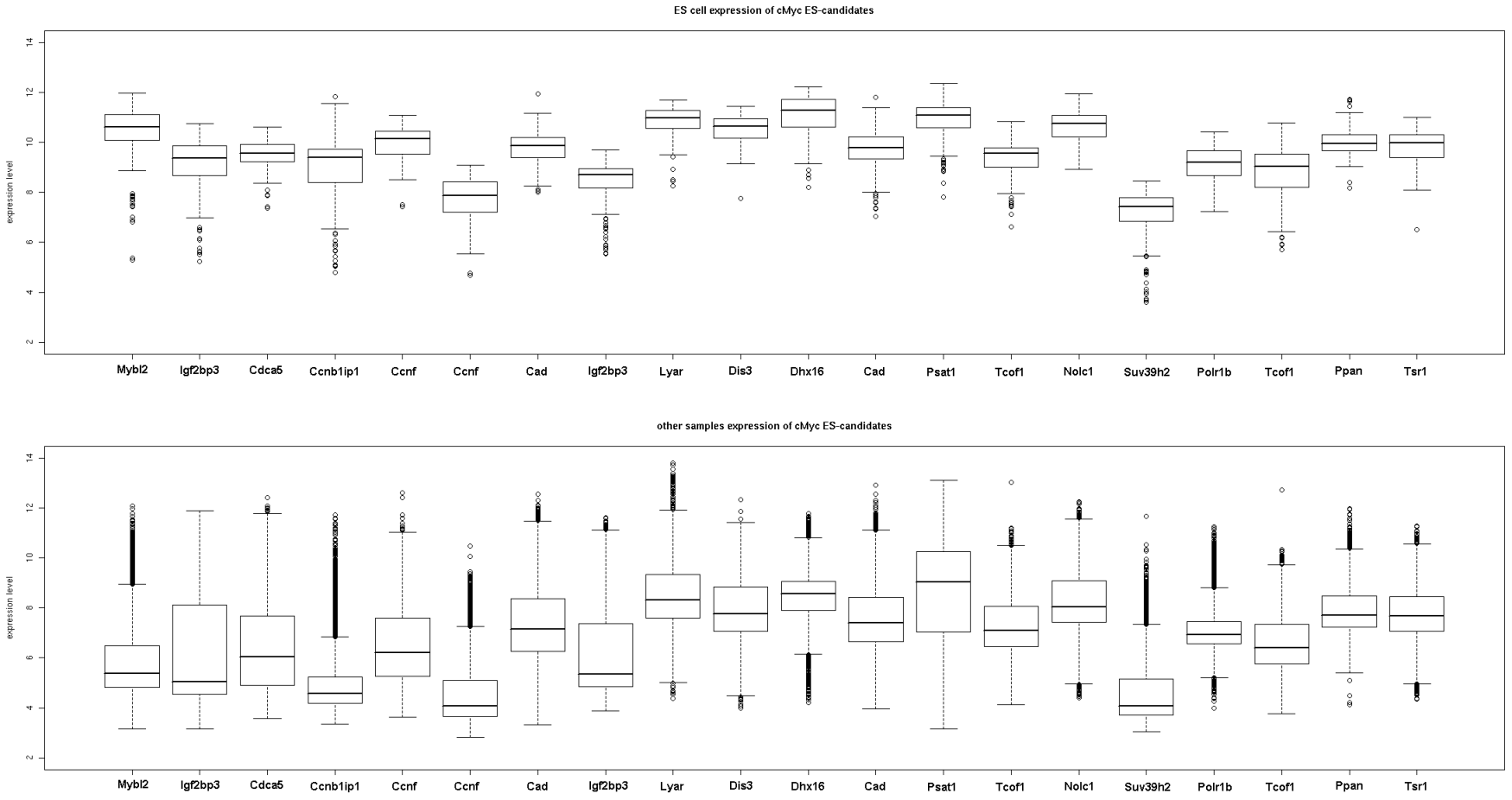


Figure 6.37: Distributions of RMA-normalized expression measurements for pluripotency-associated cMyc target candidates, shown in the top panel across ES cells only and in the bottom panel across all other samples in the dataset.

cMyc Co-dependency of 'ES Cell-Like' Transcriptional Module

It was proposed in [Wong et al., 2008] that cMyc was responsible for activating an ES cell like expression program, defined by a 'module map.' An 'ES cell module' was created in [Wong et al., 2008] based on each of a mouse and a human gene expression data compendium, with each set consisting of a set of genes belonging to gene sets consistently expressed at a higher level in ES cell studies compared to differentiated cells, as determined by gene set expression analysis described in [Segal et al., 2004]. A 'core ES cell module' was defined in [Wong et al., 2008] as the intersection of the mouse and human modules, that is the genes belonging to both lists, although it was noted that this core ES cell module did not include the critical pluripotency genes Oct4 and Nanog. The ES cell module was reported to be enriched in cancer datasets with liver, breast, prostate and gastric cancers generally showing higher expression of a number of the module's genes when compared with their respective normal counterpart tissues. However, it is apparent from the data shown in [Wong et al., 2008] that none of the proposed set of ES cell like genes appear to be consistently differentially expressed across the majority of the cancers (compared to their normal counterpart tissues), nor was the general over-expression of the module consistently observed within any one cancer type. Interestingly, no analysis was performed on the gene level across different datasets to see whether or not it was different components of the module that were responsible for the enrichment signature in different datasets. Interestingly, it was shown in [Wong et al., 2008] that expression of the majority of the ES cell like module genes increased in mouse epidermis with forced expression of cMyc and that expression of the majority of the ES cell like module genes were expressed at a higher level in explanted human keratinocytes expressing cMyc (in addition to Ras and IκB) than E2F3 or GFP. To investigate further the results of the [Wong et al., 2008] study, localised co-dependency meta-analysis was proposed to investigate the association of cMyc expression with expression of each of the genes in the ES cell like module, across a wide range of biological contexts.

Applying the analysis approach used in the previous part of Section (6.4.2), cMyc co-dependency was evaluated for each of the genes in the ES cell like module across a range of biological contexts. The co-dependency heatmap corresponding to cMyc co-dependency scores for the ES cell like module genes is provided in Fig. 6.38. It can be observed from the cMyc co-dependency estimates given in Fig. 6.38 that some of the genes in the core ES cell like module show cMyc co-dependent expression in some biological contexts, but a large proportion of the genes in the module show almost no cMyc co-dependency in any of the biological contexts analysed. It should also be noted that there seems to be relatively little consistent cMyc co-dependency observed for any genes across the range of biological contexts shown.

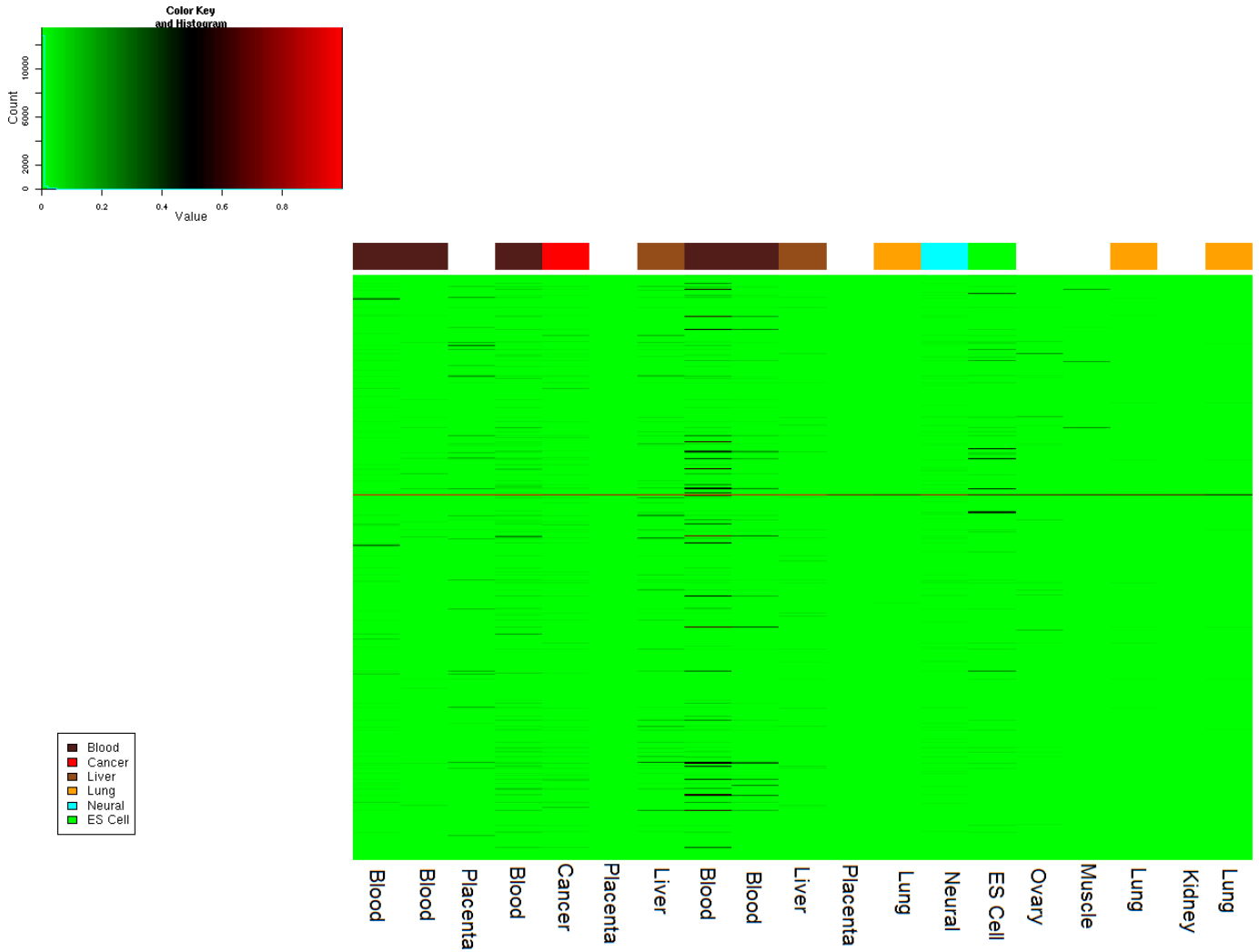


Figure 6.38: Heatmap of cMyc co-dependency probability estimates for genes in the 'core ES cell-like module.' Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context.

Based on the cMyc co-dependency analysis, the core ES cell like module was divided into subsets of cMyc-responsive and cMyc-unresponsive genes. The subset of cMyc-responsive genes from the module were ranked according to the degree to which consistent co-dependency with cMyc was observed across a range of biological contexts. Co-dependency heatmaps are presented in Figs. 6.39 & 6.40 for the top ranking genes from the cMyc-responsive subset of the core ES cell like module and for the whole cMyc-unresponsive subset, respectively. These plots demonstrate that it was possible to extract a component of the proposed core ES cell module that appeared to be more reliably co-dependently expressed with cMyc than the module as a whole, however this analysis reveals that there were no genes from the module (other than cMyc itself) with evidence for co-dependent expression across a particularly broad range of biological contexts. It should be noted that, in a recent study, expression of the core ES cell module proposed in [Wong et al., 2008] was shown in a collection of mouse leukaemia models to be consistently related to neither the expression of ES cell associated genes nor the expression of a set of of cMyc-interacting genes [Kim et al., 2010a].

Following on from the co-dependency analysis, Pearson correlation with expression levels of cMyc across a large collection of gene expression data was calculated for each gene from the two defined subsets of the core ES cell like module. Shown in Fig. 6.41, this correlation analysis confirmed that the genes with more cMyc co-dependent expression also showed generally better-correlated expression with that of cMyc. Interestingly, the genes in the cMyc-responsive subset of the module seem in general to be more specifically expressed at a high level in ES cells than the cMyc-unresponsive subset, as shown in Fig. 6.42. As the sample groups used for this comparison have very similar average levels of cMyc expression, it is highly unlikely that the observed difference in specificities between the cMyc-responsive and cMyc-unresponsive subsets is due only to cMyc-induced expression. So, not only has the core ES cell like module proposed in [Wong et al., 2008] as being cMyc induced been shown to include only a minority of genes with evidence for cMyc co-dependent expression across a large, diverse collection of gene expression data, but it has also been shown that this minority of genes is generally more specifically expressed in ES cells than the rest of the genes in the module. Additionally, there is statistically significant (hypergeometric test $p = 1.07 * 10^{-7}$) overlap between the genes from cMyc-responsive subset of the ES cell like module and genes from the broad cMyc-induced ES cell associated candidate genes shown in Fig. 6.35, indicating some correspondence between the ES cell associated cMyc candidate analysis presented earlier in Section (6.4.2) and the narrowing down of the [Wong et al., 2008] cMyc-induced ES cell like module described here.

6.4.3 Discussion

The results presented in Section (6.4) demonstrate a number of ways in which HBLCA can be utilised to study the specificity of the expression of a TF's targets on the ex-

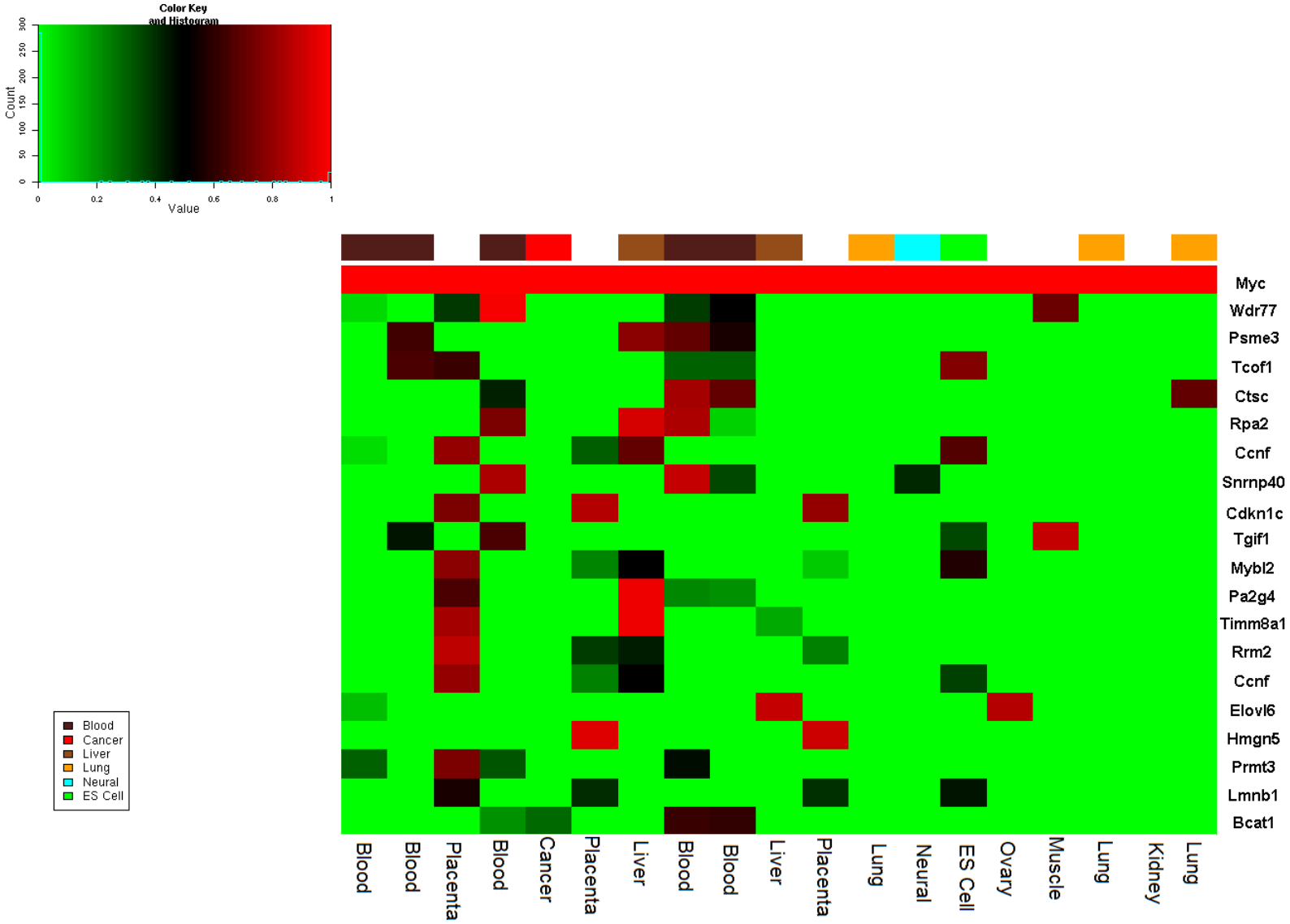


Figure 6.39: Heatmap of rank-based scores from cMyc co-dependency probability estimates for subset of core ES cell like module genes with broadest cMyc co-dependency. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 1000 or lower in the bicluster's co-dependency estimate genelist (green).

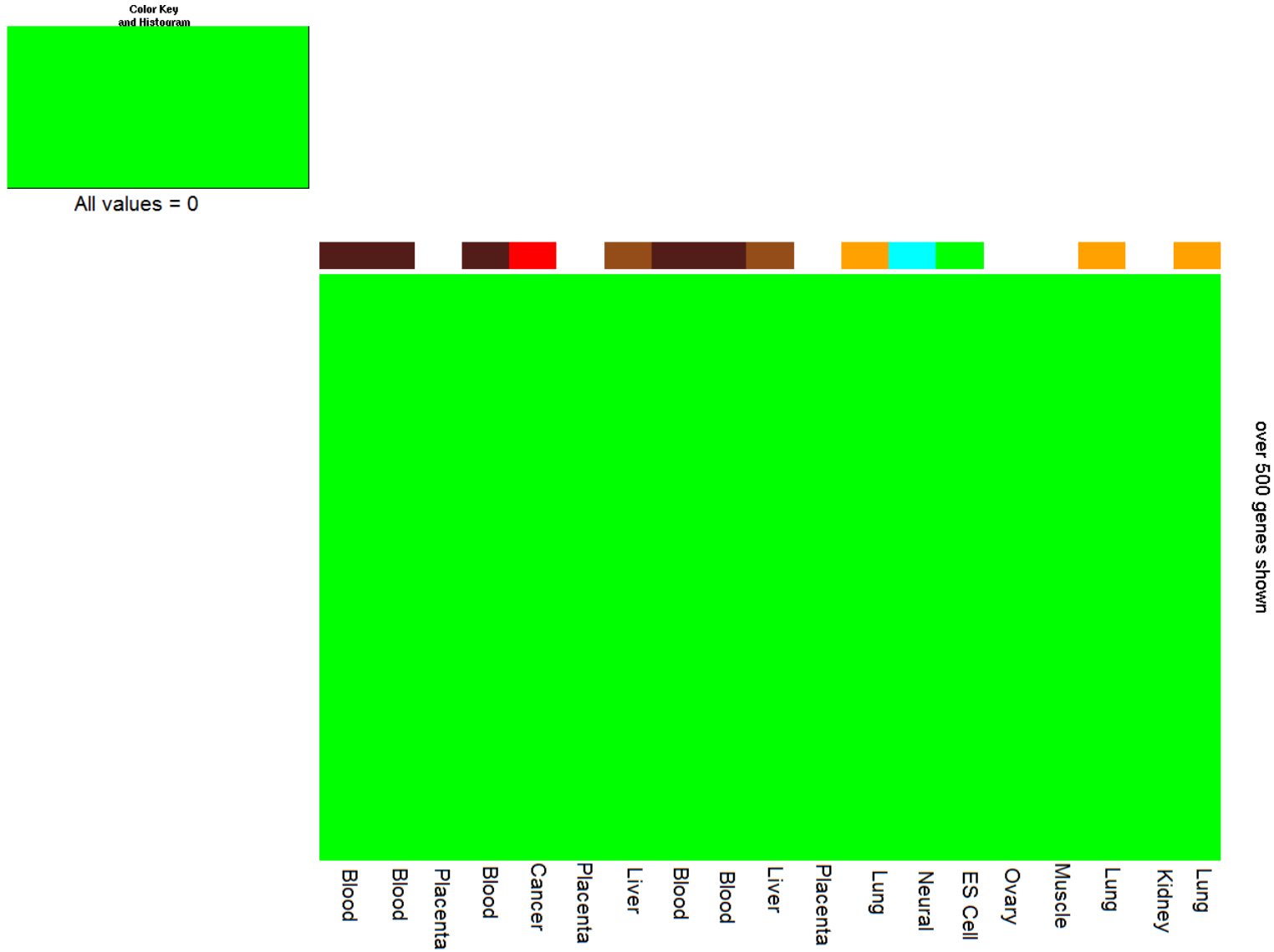


Figure 6.40: Heatmap of rank-based scores from cMyc co-dependency probability estimates for subset of core ES cell like module genes without evident cMyc co-dependency. Codependency scores are shown across a set of reliable biclusters, where each bicluster displays clear cMyc expression contrast within a given biological context. Scores range from rank 1 in a bicluster's co-dependency estimate genelist (red) to rank 1000 or lower in the bicluster's co-dependency estimate genelist (green).

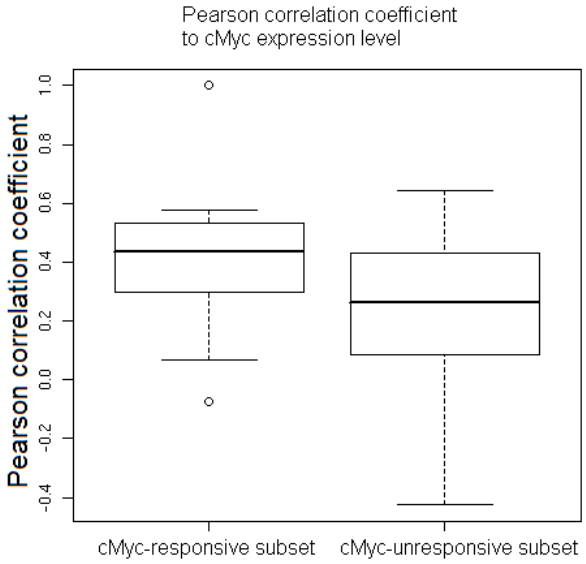


Figure 6.41: Distribution of cMyc Pearson correlation coefficients across each subset of the core ES cell like module. Correlation coefficients calculated for each gene in each subset of the module, comparing expression level of the gene to that of cMyc across all samples in the meta-analysis gene expression dataset used throughout this chapter.

pression levels of that TF, and to gain insight into possible transcriptional mechanisms through which a TF may influence observed phenotypes of interest in different biological contexts. The distribution of cMyc target dependencies was assessed across biclusters showing clear cMyc expression variation amongst globally transcriptionally similar sets of samples, ultimately suggesting that targets regulated by cMyc appear to be largely context-specific and that no global cMyc-dependent targets (that is, genes that are ubiquitously dependent on cMyc for expression) were identified. The difference between genes with globally correlated expression profiles across a large and diverse dataset and genes showing repeatedly associated expression within a restricted biological context was highlighted in Figs. 6.17 & 6.26, the latter utilising a novel visualization tool introduced for the evaluation of the distribution of co-dependencies in expression levels observed between a panel of genes and a gene (or particular combination of genes) of interest. Furthermore, it was demonstrated that the HBLCA approach was used to identify a set of cMyc targets whose expression was more likely to be dependent on cMyc within any of the examined biological contexts than those targets whose expression is most correlated with that of cMyc across a large collection of gene expression data.

In light of the proposition that cMyc may initiate an 'ES cell-like transcriptional program' [Sridharan et al., 2009], the novel meta-analysis approach was used in conjunction with other novel tools that utilise the output of this meta-analysis to investigate the relationship between expression levels in ES cells and cMyc dependent expression

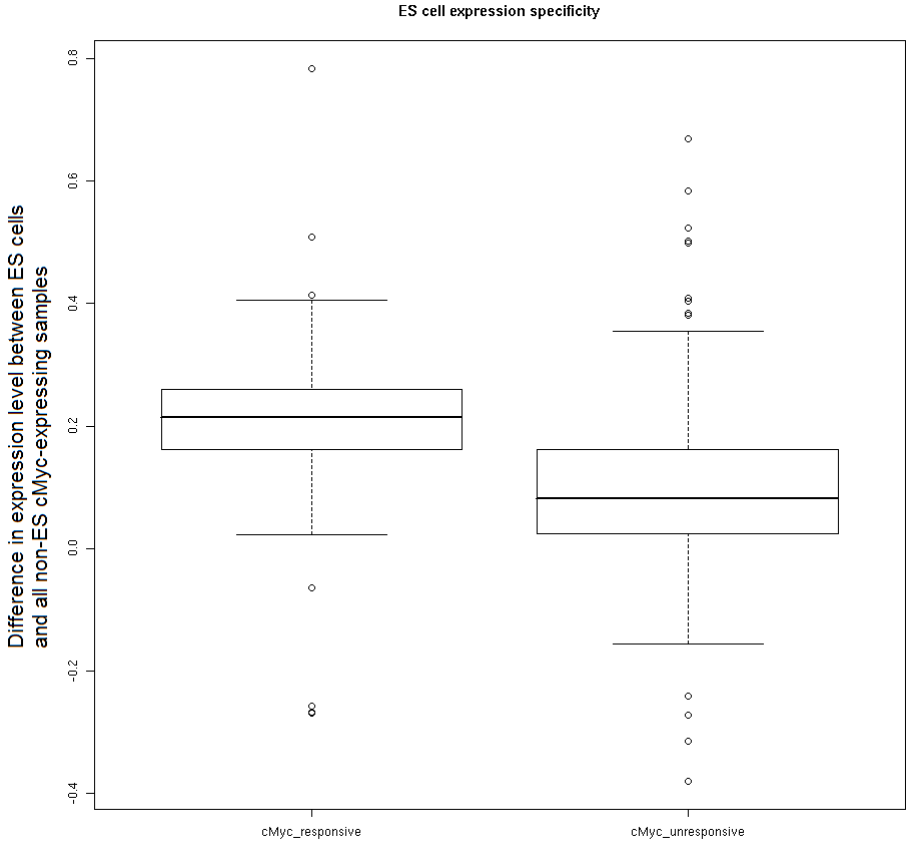


Figure 6.42: Distribution of ES cell expression specificity scores shown for each subset of the core ES cell like module. Specificity scores calculated as difference between mean GESTr-transformed expression state value of gene across ES cells and mean value across all cMyc expressing non-ES cell samples, from large collection of gene expression data.

in other cells/tissues. It was demonstrated that genes with high expression levels most specifically associated with the pluripotent state (i.e. 'ECAT' like genes[1]) did not seem to show much expression association with cMyc outside ES cells, even for those genes bound by cMyc. However, HBLCA was used to identify a set of cMyc targets with relatively broad cMyc expression dependency and high expression relatively specific to ES cells. It may be interesting to explore the functional roles of these putative pluripotency-associated cMyc targets and the phenotypic consequences of their expression outside ES cells.

Additionally, the 'core ES cell like module' presented in [Wong et al., 2008] was filtered according to localised co-dependency analysis into cMyc-responsive and cMyc-unresponsive subsets. The cMyc-responsive subset was shown to display evidence for cMyc co-dependent expression in some biological contexts, was shown to be better correlated with cMyc expression across a large and diverse collection of gene expression data. Interestingly, this cMyc-responsive subset of the [Wong et al., 2008] ES cell like module was shown to be more specifically expressed at a high level in ES cells than the remainder of the module, and to overlap significantly with the set of ES cell associated and relatively broadly cMyc co-dependent cMyc targets. While there does not appear to be a universal cMyc-induced transcriptional signature evident in the large gene expression data collection analysed, the localised co-dependency meta-analysis approach performed here, using the tools described in Chapter 5, has been shown to highlight more reliably induced components from within proposed transcriptional signatures.

[1] genes with high levels of expression relatively specific to ES cells, although these are not necessarily the ECATs proposed in [Mitsui et al., 2003]

6.5 Chapter Summary

A number of investigations into the transcriptional control of pluripotency were performed using the novel analysis tools, summarised in the following paragraphs numbered 1-4. These investigations offer insights into the transcriptional regulation of pluripotency that may be tested experimentally in the future, and further demonstrated ways in which these tools may be applied to the utilisation of publicly available data for gaining insight into transcriptional mechanisms involved in the control of biological processes of interest.

1) By applying functional enrichment analysis to the output of the HBLCA approach with a list of genes differentially expressed in ES cells upon Oct4 knock-down provided as input 'guide' genes, a number of biological processes were identified as being significantly affected on a transcriptional level by the knock-down of Oct4. Of particular significance is the fact that no such significantly affected biological processes were identified when the same functional enrichment analysis techniques were applied directly to the list of input 'guide' genes. Additionally, the output of the HBLCA tool was used to identify subsets of related genes within the input list, each of which showed a characteristic expression pattern in the original Oct4 knock-down dataset and was associated with specific biological processes. This analysis therefore implicated a number of sets of genes as each possibly representing a different Oct4-induced transcriptional component controlling particular biological processes including cell adhesion, which has not previously been demonstrated to have an Oct4-dependent role in the maintenance of pluripotency.

2) Using the output of the HBLCA tool applied with a list of reliably Oct4-bound target genes as inputs, subsets of targets with related gene expression patterns across ES cells were identified. A list of genes clearly dependent on Oct4 in ES cells was produced, suggesting that these genes may be key Oct4-dependent pluripotency-associated genes. Another list of genes was identified with expression levels in ES cells seemingly unrelated to those of Oct4, but which showed correspondance of expression level with that of Foxh1. These genes may represent a novel component of Oct4 target genes for which transcription is redundantly or dominantly co-dependent on another factor.

3) The HBLCA approach was used to investigate combinatorial transcriptional regulation in ES cells involving the key pluripotency TFs Oct4, Sox2 and Nanog. These TFs show considerable overlap of DNA-binding target genes, but it was possible to identify genes with apparent dependencies on particular combinations of these TFs. These lists may help provide a basis for the decomposition of complex transcriptional regulatory mechanisms involved in the maintenance and acquisition of pluripotency.

4) The association of levels of expression of the reprogramming factor and oncogene cMyc and its DNA-binding targets was investigated across a broad range of biological contexts using HBLCA and co-dependency heatmaps. It was demonstrated in Section (6.4.1) that the cMyc DNA-binding targets with most cMyc-dependent expression in any given biological context are unlikely to show cMyc-dependent expression in any other biological context. However, HBLCA was used to identify cMyc-bound targets with apparent cMyc-dependent expression across a broad range of biological contexts. It was demonstrated that these genes were not typically associated with pluripotency (or stem cells), in contrast to the hypothesis presented in [Wong et al., 2008] and alluded to in [Sridharan et al., 2009]. To investigate this hypothesis further, cMyc-dependency of expression of ES cell specific cMyc-bound targets (and of the genes suggested in [Wong et al., 2008] as representing a cMyc-induced 'stem cell-like' signature) was evaluated across a broad range of cMyc-expressing biological contexts, which revealed that these genes do not appear to represent a universal cMyc-induced 'stem cell-like' expression program.

The previous chapter described a novel framework developed for analysis of transcriptional regulation using large collections of existing gene expression data. This chapter demonstrates the way in which this framework provided a platform for the development of a number of tools for the investigation of transcriptional regulatory mechanisms of TFs (or biological processes) of interest. These tools have been utilised to offer insights into mechanisms of transcriptional control of pluripotency, which provide hypotheses for experimental verification and demonstrate the utility of these novel tools for biological research.

Chapter 7

Final Discussion

Following the work presented throughout this thesis a discussion is given in this chapter summarising the work that has been performed and the results obtained, both in terms of significance and relevance to the fields of computational biology, transcriptomics and stem cell research. Following a recap of the motivations for meta-analysis of gene expression data and for the adaptation of the biclustering paradigm to this task, the first section below presents conclusions of research carried out to investigate practical considerations and limitations of the application of biclustering to the study of transcriptional relationships through meta-analysis of large collections of gene expression data. The second section of this chapter provides a summary of the novel methods for transcriptional pattern mining through gene expression data meta-analysis that have been developed through the course of this work. The third section presents the findings of a number of investigations into real biological questions relevant to current state of the art theory in stem cell research that were carried out using the approaches to study transcriptional relationships that were developed through the course of this work. The results of these investigations are discussed in terms of insights provided into the transcriptional control of pluripotency, in addition to the context of providing demonstrations of the relevance of successful application of these tools for biological research. The fourth section of this chapter, preceeding an overall summary to conclude this thesis, discusses open questions that remain despite the work carried out and those that have arisen as a direct consequence of this work. Further work plans are outlined as recommendations for investigations that might effectively explore the remaining open questions and provide more detailed insight into the topics discussed in this thesis.

7.1 Study Of Application Of Biclustering To Meta-Analysis Of Gene Expression Data

At the outset, the primary goals of this research were to find ways of utilising large collections of gene expression data to gain insight into the mechanisms of transcriptional regulation of mammalian biological processes and apply effective techniques to investigate the transcriptional control of pluripotency. While simple global correlation-based methods have been shown to be effective tools for prediction of some transcriptional relationships through the analysis of large gene expression datasets [Day et al., 2009], it is expected that as the biological domains represented in expression data compendia diversify, the range of transcriptional relationships remaining possible to identify with such methods will decrease to a point where it is only possible to infer relationships between genes that are simultaneously and/or exclusively regulated by particular TFs regardless of biological context and epigenetic state reflected in that biological context (the necessary biological conditions to give rise to ubiquitously correlated expression profiles).

It has therefore been proposed by a range of source (e.g. [Cheng and Church, 2000, Tanay et al., 2002, Owen et al., 2003, Prelic et al., 2006]) that as the range of biological contexts and conditions represented in whole-genome expression datasets increases (as is expected to be the case when more and more data is accumulated) it becomes increasingly more appropriate to apply data mining techniques that attempt to identify localised patterns in gene expression datasets, in which certain patterns in expression levels of a set of genes (assumed to imply transcriptional relationships of a particular nature between those genes) are observed across only a subset of the samples in the whole dataset. This principle of localised pattern mining in datasets is the fundamental principle behind the paradigm of two-way clustering [Hartigan, 1975], first applied to analysis of gene expression datasets as 'biclustering' [Cheng and Church, 2000]. Although a large number (hundreds) of methods for identification of biclusters in gene expression datasets have been described in the literature, there have been very few such methods intended for application to large datasets due to the computational complexity of the general bicluster search problem. It is important to consider when discussing biclustering methods that there are two components to the biclustering process: the definition of a suitable bicluster (that is, the nature of the localised patterns intended to be identified with the analysis) and the search mechanism employed to find the best such biclusters in the dataset. While the search mechanism is predominantly a practical consideration for which there may be a wide range of satisfactory solutions, it is the manner in which biclusters are defined that will have a dramatic impact on the transcriptional relationships uncovered through a biclustering analysis of gene expression data (and therefore the practical utility of the results of such an analysis) and yet this seems to have received relatively little emphasis in the enormous number of published research articles concerning biclustering of gene expression data. Therefore, the study of biclustering that was carried out for this work had two principal points of focus:

1. implementation of a flexible framework to identify customarily-defined biclusters in very large collections of gene expression data
2. investigation of different bicluster definitions in the context of large-scale gene expression meta-analysis and the transcriptional relationships represented by such biclusters when discovered in real data

7.1.1 Flexible Biclustering On A Large Scale

Through a reformulation of the general biclustering problem that was possible due to assumptions applicable to the application of biclustering to discovery of transcriptional regulatory patterns in gene expression data, it was possible to reduce significantly the complexity of the biclustering problem so that applications to large scale meta-analysis might be feasible (as described in Section (3.4)). After a combinatorial biclustering approach was developed for identification of customarily defined biclusters but shown to be impractically slow for large scale application, a heuristic search approach was

implemented using a genetic algorithm, resulting in an algorithm (described in Section (3.4.4)) for identification of customarily-defined biclusters in especially large collections of gene expression data. Part of the motivation for developing such a framework for biclustering analysis was that existing biclustering approaches were not suitable for large scale application due to computational limitations. As demonstrated in Section (3.5.1), the IslandCluster approach was scalable for application to the largest collections of data available at the time (and considerably beyond this scale), unlike any existing biclustering methods in common use. This result indicated that IslandCluster could be used to provide the first real insights into the practical considerations of adopting a biclustering principle for the investigation of transcriptional relationships through meta-analysis of large collections of gene expression data.

Following the evaluation of scalability of IslandCluster, it was considered essential to evaluate the success of the algorithm in identifying the intended patterns in datasets. For this reason, artificial datasets were constructed and rates of successful recovery of known bicluster patterns implanted into data were measured. Section (3.5.2) presents the findings of this evaluation, showing that IslandCluster does indeed identify the intended bicluster patterns with a high rate of success. While it is impossible to have an analogous evaluation scenario where all suitable biclusters in real data are known, in order to test the ability to recover biclusters in real data, comparitive enrichment analysis was performed on the IslandCluster method. This enrichment analysis (presented in Section (3.5.3)) demonstrated that even with a naive model of bicluster desirability, IslandCluster identifies biclusters with functional biological significance to a similar degree to the best existing alternative biclustering methods, but has the unique advantages of scalability to very large datasets and flexible definition of the bicluster expression patterns to be identified.

After a modification of the search mechanism of IslandCluster to adopt a multi-niche crowding [Cedeno and Vemuri, 1996] approach that enables simultaneous identification of diverse biclusters, this MNC-BGA was used to evaluate the manner in which bicluster definitions are reflected in the transcriptional relationships it is possible to predict using the results of the biclustering meta-analysis of gene expression data.

7.1.2 Investigating The Impact Of Bicluster Definition On Utility Of Gene Expression Meta-Analysis Results

With a method to implement discovery of biclusters of different definitions and a naive bicluster model to provide a baseline for performance evaluations, the consequences of adopting particular definitions of bicluster model were evaluated in terms of the practical output of meta-analysis of large gene expression datasets for the study of transcriptional regulation of biological processes.

It was observed that a naive model of bicluster tends to result in discovery of biclusters comprising a very large number of genes, with considerable overlap between the genelists corresponding to different biclusters. It was therefore considered advantageous to provide a means of modelling the probability of discovering any particular bicluster by chance and to score each gene's contribution to the bicluster pattern according to its chance of being present in a randomly-chosen bicluster (thus offering a greater degree of specificity for the uncovered expression patterns). An entropy-based bicluster definition was introduced in Section (3.6.2) to address these issues, and subsequent evaluation demonstrated that incorporating this more sophisticated model of bicluster desirability offers a greater degree of specificity of discovered relationships and provides a means of ranking component genes in these relationships that results in a significant improvement in terms of the ability of the biclustering meta-analysis approach to predict functional association and TF binding relationships between genes.

A further observation made as a result of application of the entropy-based MNC BGA to the investigation of transcriptional regulation of biological processes was that while most useful questions in this research involve the identification of expression patterns involving particular genes of interest in particular biological contexts, in many cases involved in the control of pluripotency, the most significant biclusters concerning the relevant biological context involve the same groups of genes with little to distinguish between the roles of a large number of TFs. As a response to these observations, an alternative bicluster model was proposed for directed biclustering to identify genes with apparent co-dependency of expression with a gene of interest across the biological context represented in the bicluster, thereby distinguishing (as far as possible) between expression patterns relating to particular genes in a particular context, rather than only identifying compnonents of context-specific expression programs. The advantages in practical terms of adopting such a bicluster model were demonstrated through comparitive enrichment analysis for functional annotation association and for prediction of DNA-binding by a TF of interest. These results are shown in Section (3.6.3) along with illustrations that the directed guide gene-dependent bicluster models result in improvement in the rate of specifically-relevant relationships being uncovered for genes with similar context-specific expression.

7.2 Methods Developed

The development of a tool for prediction of transcriptional relationships of interest, led by progressive evaluation in terms of applicability to real problems in biological research, involved the conception and implementation of a number of component methods that could be applied together in a novel meta-analysis approach in order to achieve the desired analysis. As some of these components may have application outwith the context of meta-analysis of gene expression data to predict transcriptional relationships,

these components are summarised individually below, along with the tools developed for investigation of mechanisms of transcriptional control of biological processes.

7.2.1 Expression-State Modelling

To assist the comparability of expression measurements for different genes, a method was developed that transforms gene expression measurement values from a large reference dataset into a unified scale based on inference of biological states of expression from models of the distribution of each gene's expression values. This transformation (GESTr, described in Section (4.1.4)) provides the first reported means of considering expression levels of different genes on a unified scale representing biological states of expression that adapts to the different distribution patterns of each gene.

Through analysis of the variation within sets of multiple probesets on an Affymetrix GeneChip platform (with each set mapping to the same gene), it was demonstrated that the GESTr method successfully transforms gene expression measurements into a unified scale without introducing additional non-biological variation. These results were presented in Section (4.2.1).

The primary purpose of introducing the GESTr approach was to improve biclustering meta-analysis tools through preprocessing of the data into a form from which any single value (representing expression of a particular gene in a particular sample) can be used to make an inference regarding the biological significance of that expression level, and that such an inference could be universally applicable for a given value regardless of which gene's expression was being measured in which sample. 'Semi-artificial' datasets were created through calculated permutations of a real dataset in order to evaluate the effect of adopting different preprocessing methods on recovery of implanted 'genes of interest' using a simple bicluster model. Results of this analysis (given in Section (4.2.2)) demonstrated that the GESTr improves biclustering performance over a discretization-based preprocessing method shown (in Section 4.1.2) to be superior to those adopted by a number of successful biclustering methods, and even improves upon a related bicluster discovery (gene prioritization) approach using the raw expression values as input. Additionally, a large number of biclusters were produced from real datasets using raw data, discretized data and GESTr-transformed data as input for equivalent bicluster evaluation methods: examination of the properties of these biclusters revealed that the GESTr facilitated effective large-scale bicluster-based gene expression data mining.

It was noted that this data transformation may have applications other than the improvement of large-scale biclustering approaches, for example to provide a means of interpreting data from individual gene expression studies in the context of all profiled expression levelsof each gene in a large reference collection of data. An approach was developed (TranSAM, described in Section (4.3.1)) for discovery of genes that are

significantly differentially expressed in a novel dataset in the context of their profiled expression range across a large collection of biologically diverse samples. In an example application, this approach was shown to identify targets in an Oct4 knock-down dataset with DNA-binding by Oct4 with greater success than the widely used SAM approach to identification of differentially expressed genes, supporting the claim that the GESTr data preprocessing method may have utility in the interpretation of gene expression measurements from individual experiments and the identification of genes with *biologically* (rather than purely statistically) significant changes in expression between samples of interest. At the time of writing, a number of research projects are being carried out involving work utilising the output of GESTr processing of large gene expression datasets, further supporting the claim that this data transformation may have applications outside the realm of biclustering. However, the main goal of this work was to perform large-scale meta-analysis of gene expression data for prediction of transcriptional relationships between genes involved in the control of pluripotency, and in this context the GESTr method successfully enabled the development of biclustering-based approaches to such meta-analysis that achieved this goal.

7.2.2 Grouping of Microarray Samples to Represent Distinct Biological Contexts

For the identification of context-specific gene expression patterns in large datasets, it was considered helpful to have a pre-computed set of groups of samples, each reflecting a particular biological context. By grouping similar samples on the basis of global transcriptional profiles, the notion of a consistent biological context could be determined solely on the basis of data actually present in the dataset, the same values that will be used to generate any hypotheses on the basis of observed results of large-scale meta-analysis. Furthermore, as these groups have generally consistent annotations (as shown in Section (5.1.4)) it becomes relatively straightforward to interpret the biological contexts across which any discovered pattern can be observed.

An approach was developed to obtain such a set of groups of similar samples via estimation of a cumulative probability function for the distribution of number of samples with fully interconnected similarity from each of a large number of random samplings from the dataset. With such distributions calculated for a range of similarity thresholds, sample groups involving a sufficiently large number of similar samples to have an estimated probability of occurring by chance below a set significance threshold can be identified for the given set of similarity thresholds. The results of this process are essentially hierarchical groups of similar samples, some of which are subsets of others (as illustrated in Section (5.1.4)). This approach is adaptable to any dataset (with no prerequisites) as the probability models are derived exclusively from the data itself.

While this approach has a clear application for dramatically reducing the search space for any biclustering-based application for meta-analysis of large gene expression datasets, there may be other potential uses for a method to identify a comprehensive set of groups of similar samples from a large dataset. For example, these groups might be used in conjunction with annotations to find samples that are not only described as coming from similar biological contexts, but for which it is known that the data reflect such similarity. In such a way, misannotations can clearly be corrected and unknown relationships of samples on the basis of global transcriptional programs may be discovered. However, the real advantage of such a method is that of providing the ability to search through the groups (or any relevant subset of the groups) in order to identify some desired gene expression pattern of interest in samples known to share consistent general transcriptional programs. Examples of such applications include general marker discovery and discriminant analysis to find marker profiles distinguishing between biological contexts of interest, although many other potential uses for the output of this method may exist.

7.2.3 Localised Co-Dependency Analysis For Gene Expression Data

As the primary goal of this work was to perform large-scale meta-analysis of gene expression data for the prediction of context-specific transcriptional relationships, motivating a biclustering based approach, but no existing biclustering algorithms were appropriate for this purpose (as demonstrated in Section (3.5.1)), it is perhaps unsurprising that one of the main achievements of this work was the development of a novel gene expression meta-analysis approach specifically designed to facilitate the prediction of transcriptional relationships involving key genes (or combinations of genes) of interest across specified (or subsequently identified) biological contexts. The HBLCA approach, incorporating the probabilistic framework for localised co-dependency analysis introduced in Section (5.1) and utilising the GESTr method for preprocessing input gene expression datasets, was developed as a result of progressive evaluation and redevelopment of bicluster models in terms of the practical utility of the output of the corresponding meta-analysis approaches in helping to provide a greater understanding of the transcriptional regulation of biological processes, particularly those relating to the control of pluripotency.

The HBLCA approach was evaluated in terms of comparitive success in example meta-analysis tasks against existing successful methods for prediction of transcriptional relationships through meta-analysis of gene expression data and against standard differential expression analysis methods applied to individual datasets. The results of these evaluations, given in Section (5.2), demonstrate that the HBLCA approach generally predicts transcriptional relationships between genes with known DNA-binding and/or consistent patterns of differential gene expression displayed across 'held-out' datasets profiling samples of the relevant biological contexts better than established methods

for generation of similar predictions from gene expression data. Furthermore, HBLCA can identify transcriptional relationships occurring across defined subsets of a large gene expression dataset (providing sufficient data is present to infer such relationships) reflecting biological contexts of interest, with sufficient success to be a useful tool for the utilisation of existing gene expression data for further study of the transcriptional control of biological processes. This claim is further supported by the investigations of transcriptional control of pluripotency presented in Chapter 6 and the analyses of DNA-binding and regulation of expression involved in the work presented in Sections (5.3 & 6.4), in which an implementation of HBLCA was used to gain insight into transcriptional mechanisms involved in the regulation of biological processes.

As a general tool for investigating mechanisms of transcriptional control of biological processes using existing gene expression data, especially when used in conjunction with alternative techniques for investigating the biological processes in question, this approach has the potential for widespread application in successfully advancing biological research.

7.2.4 Integrative Analysis

With increasing availability of genome-scale DNA binding datasets in the public domain comes the opportunity to combine analysis of such data with analysis of relevant gene expression data in an attempt to identify genes that are likely transcriptional regulatory targets of TFs of interest. As illustrated in Section (5.2.2), the DNA binding data alone do not explain all gene expression patterns involving the bound targets and respective binding TFs (even when considering only samples of the same context as those in which the DNA-binding studies were performed), as would be expected due to combinatorial transcriptional regulatory mechanisms involving multiple TFs and the effects of context-specific epigenetic modifications. Additionally, it was shown in Section (5.2.2) that there can be considerable variation in the results of different studies measuring genome-scale DNA binding of a TF, even when the data comes from similar experimental technologies. Therefore, an integrated meta-analysis approach involving meta-analysis of multiple DNA binding datasets for a given TF of interest and directed meta-analysis of gene expression data is recommended for the identification of genes that are both reliably bound by the TF of interest and show significant and consistent expression co-dependency across the biological context in question, as such genes are likely to be the relevant targets of the given TF in the particular biological context of interest.

A simple integrated analysis approach was presented in Section (5.3), along with the results indicating that this approach can reliably identify (with a high rate of success) genes that are consistently differentially expressed in 'held-out' knock-down studies involving variation of the expression of the TF in question, and are thus likely to be direct

transcriptional targets of the TF of interest. In the investigations relating to transcriptional control of pluripotency presented in Sections (6.3 & 6.4) this integrated analysis approach (incorporating the HBLCA approach) was used to make novel observations regarding possible mechanisms involved in the control of pluripotency, supporting the claim that such an integrated analysis approach as this may also have widespread application in biological research.

7.3 Study Of The Transcriptional Control Of Biological Processes

The methods summarised in the previous section were developed in order to provide means of gaining insight into transcriptional regulatory mechanisms through interrogation of large collections of gene expression data. Application of an ensemble of these methods can results in the identification of gene expression co-dependency patterns relevant to a particular gene (or set of genes) of interest across a particular (set of) biological context(s). The identification of such transcriptional relationships relevant to a particular biological investigation may have a wide range of potential uses to offer insight into different aspects of mechanisms of transcriptional control of biological processes of interest. The following section provides an overview of a number of investigations into different aspects of the control of pluripotency that were described in the previous chapter. These investigations illustrate a range of different ways in which the novel methods presented in this thesis can be applied to provide insight into different aspects of transcriptional regulatory mechanisms, with these applications representing research methods in themselves, which would not have been possible without the HBLCA meta-analysis approach. The results of these investigations clearly demonstrate that the methods developed through the course of this work have useful applications to biological research, through the analysis of large collections of data.

7.3.1 Results Relating To Pluripotency

Functional Analysis of Oct4-Dependently Expressed Genes

A set of genes whose expression level decreases significantly upon Oct4 knock-down in a mouse ES cell microarray experiment [Hall et al., 2009] was obtained. These genes are likely to be directly influenced by the expression of Oct4 and thus involved in the maintenance of the ES cell pluripotent state. Through application of the HBLCA approach using each of these genes as a 'guide' it was possible to associate biological processes with these genes through functional enrichment analysis of the resulting genelists. The input genes were separated into three subsets, with each subset sharing similar output from the HBLCA tool. Interestingly, each of these subsets was characterised by a unique expression profile in the Oct4 knock-down microarray dataset and was statistically enriched for different sets of biological processes. Particularly noteworth is

the fact that application of standard functional enrichment analysis techniques to the original input genelist did not reveal any biological processes significantly affected on a transcriptional level by the knock-down of Oct4.

The analysis performed as described in Section (6.1) implicated a number of sets of genes that respond to Oct4 knock-down in ES cells and may represent distinct functional components of the Oct4-mediated pluripotent state of ES cells. One such subset of genes appears to characterise a role of cell adhesion in this phenotype, which has not previously been demonstrated to be involved in pluripotency.

Oct4 Target Genes Identified with Different Expression Dependencies

Lists of genes with expression patterns in ES cells associated to individual Oct4 binding targets were obtained by applying the HBLCA tool with a set of targets reliably bound by Oct4 as input 'guide' genes. Analysis of these associated genelists enabled the identification of subsets within this list of Oct4 binding targets that shared expression dependency patterns. One such subset involved genes with clear Oct4-dependent expression in ES cells, which may represent key pluripotency genes directly regulated by Oct4. Another subset showed expression in ES cells seemingly unrelated to the level of Oct4, but did show a correspondance of expression level to that of Foxh1. This may represent a novel component of Oct4 target genes expressed in ES cells, for which transcription may be redundantly or dominantly regulated by another factor, which may be Foxh1.

Combinatorial Analysis of Regulation by Oct4, Sox2 and Nanog

Using HBLCA it was possible to identify genes with apparent expression dependencies on particular combinations of Oct4, Sox2 and Nanog: TFs with considerably overlapping sets of DNA-binding targets. The binding targets identified with apparently specific dependency of expression to a particular combination of expression of Oct4, Sox2 and Nanog may serve as a means to decompose the complex and highly interconnected transcriptional regulatory network controlling pluripotency.

Transcriptional Regulatory Activity of cMyc

HBLCA was used to investigate the cMyc-dependency of expression of cMyc DNA-binding targets across a broad range of biological contexts. It was demonstrated that the majority of cMyc-bound genes that show cMyc-dependent expression in any biological context show cMyc-dependent expression specifically in that context. Thus, cMyc does not appear to induce expression of a consistent set of targets across the full range of biological contexts in which it is expressed.

The hypothesis proposed in [Wong et al., 2008, Sridharan et al., 2009] that cMyc induces a 'stem cell like' transcriptional program was investigated using the HBLCA tool. It was demonstrated that the cMyc-bound targets with most ES cell specific expression do not appear to show cMyc-dependent expression outside ES cells. These would be the genes most obviously associated with a cMyc-induced stem cell like transcriptional signature. A set of genes was identified using HBLCA that did show relatively ES cell specific expression and cMyc-dependent expression in a range of non-ES biological contexts (including blood, liver and placenta). The genes proposed in [Wong et al., 2008] to represent the cMyc-induced stem cell like signature were separated into apparently cMyc-responsive and cMyc-unresponsive subsets. Interestingly, the cMyc-responsive subset were generally expressed at a higher level in ES cells than in other cMyc-expressing samples in a large dataset. Such genes with relatively specific ES cell expression and relatively broad cMyc co-dependent expression may represent a cMyc-induced stem cell like transcriptional signature, but they do not seem to be universally cMyc-dependent and comprise at most only a small fraction ($< 10\%$) of the supposed cMyc-induced, ES cell like module described in [Wong et al., 2008].

7.3.2 Generalisability Of Approach

While the results obtained through the investigations described in Chapter 6 are significant in offering insight into the transcriptional mechanisms involved in controlling pluripotency, they are also significant in terms of illustrating a means of utilising the novel meta-analysis tools for investigations with application to a potentially wide range of areas in biological research.

- The results of Section (6.1) demonstrate that the meta-analysis tools presented in this thesis may provide a means for alternative, context-specific functional enrichment analysis that offers advantages over existing techniques. These existing techniques have had wide-ranging application in biological research, implying that the approach described in Section (5.4) may have equally wide-ranging application.

- The analyses described in Sections (6.1 & 6.3) illustrate ways in which the output of the HBLCA tool can be used to decompose an input genelist into subsets with unique sets of shared expression co-dependencies. These subsets may share similar function, as with the subset genelists identified in Section (6.1). These approaches may be used to identify components within any complex expression pattern related to a gene or biological process of interest.

- The application of HBLCA to the identification of target genes with observed expression dependency on specific combinations of Oct4, Sox2 and Nanog was described in Section (6.3). A similar approach could be taken to investigate any

complex transcriptional regulatory network in which a number of key regulators may share significantly overlapping sets of targets, provided relevant DNA-binding data and gene expression data is available.

- The techniques used in the investigation of cMyc transcriptional regulatory activity may be applied to analyse the biological context specificity of observed or predicted transcriptional relationships. This can be achieved through visualising expression co-dependency significances evaluated with the HBLCA approach.

The development of novel meta-analysis tools presented in this thesis has afforded the ability to identify genes with expression co-dependency patterns relating to a gene or combination of genes of interest, across a broad range of individual biological contexts. The HBLCA approach to gene expression analysis has potentially far-reaching applications in biological research, both directly and in conjunction with further analysis approaches such as those utilised in the investigations presented in Chapter 6.

7.4 Open Questions & Further Work

This thesis describes an investigation of meta-analysis of gene expression data, primarily relating to the application of biclustering to this task. Theory developed in this work provides a foundation for the effective adaptation of the general biclustering paradigm to gene expression meta-analysis. This work also introduces a new concept in the analysis and utilisation of collections of gene expression data, notably that of context-specific co-dependency of expression.

The results of evaluation of different approaches proposed in this thesis have illustrated some of the potential pitfalls and considerations required specifically when considering meta-analysis of large, diverse collections of gene expression data. As the work presented in this thesis represents the first such investigation reported, the theoretical foundation provided will enable further investigation of the application of biclustering approaches to gene expression meta-analysis, and of both theoretical and practical aspects of gene expression co-dependency analysis.

7.4.1 Extending Scope of Analysis Approaches and Observations

Due to the emphasis on development of theory in this work that was required to enable successful practical application of the intended meta-analysis techniques, limitations in the scope of the applications considered in this work provide a number of obvious avenues for future work:

- **Other Organisms:** all analyses in this work concern gene expression in the mouse. It is assumed that the issues concerning gene expression meta-analysis that have been characterised and overcome in this work apply similarly to the use

of large, diverse collections of gene expression data from other organisms to study their transcriptional regulatory mechanisms. This has not been demonstrated in this thesis, and so it may prove worthwhile to utilise the novel approaches for applications to investigate transcriptional regulatory mechanisms in other organisms. In fact, such application may reveal further utilities of the developed analysis approaches. Examples might include the study of conservation or evolution of regulatory mechanisms, or highlighting species-specific and species-independent transcriptional relationships between genes of interest. However, there is a very wide range of potential applications of the developed approaches within mouse molecular biology such that this work would still have considerable impact on future biological research, even in the unlikely event that utility of the novel analysis approaches turned out to be mouse-specific.

- **Other Measurement Technologies:** Due to the abundance of data from Affymetrix microarray platforms, particularly in contrast to any other technologies, the meta-analysis approaches presented in this thesis were devloped and studies in the context of Affymetrix microarray data only. It might therefore afford further utility from these approaches and increase the scope of their possible application if they were to be generalised or adapted to data from alternative technologies. The relative utility of such application may be minimal while the body of available gene expression data is so dominated by Affymetrix microarray platforms, but this might change as next-generation sequencing technologies are increasingly applied to measuring gene expression. It should be noted that the models of expression data developed through this work and utilised in the HBLCA approach are based on arbitrary gene expression values and adapt to the observed distributions, so any required modifications should be limited to the incorporation of alternative measurement error models into the GESTr process and possibly also in the modelling of gene-variation significance estimator parameters that is required for co-dependency analysis. These modifications reflect relatively small components of the overall analysis approaches used in this work, which suggests that the majority of the conclusions presented in this thesis would not be platform-dependent.

- **Multi-Platform Analysis:** for ease of comparability of data from different samples and different experiments, all the gene expression meta-analysis presented in this thesis was performed on data from the same microarray platform. While publicly available data from this single platform constitutes a very large, diverse collection of comparable gene expression measurements (from $O(10,000)$ samples), greater utility might be afforded by the HBLCA approach if it could be adapted to apply to collections of gene expression data from multiple platforms. It is anticipated that this might constitute a significant but potentially rewarding piece of work, involving the evaluation and development of modifications to the

GESTr modelling approaches that may be required for cross-platform analysis.

7.4.2 Optimisation of Novel Analysis Approaches

When each analysis method described in this thesis was applied to a real task for biological research, complications were discovered that led to further approaches increasingly optimised towards the task at hand. It is therefore highly unlikely that the most sophisticated methods for meta-analysis described in this thesis are truly optimal for their intended task, especially considering the relatively broad scope of the intended tasks. It may well be that when applied to particular investigations it is discovered that certain parameter settings (such as those governing stringency of contrasts considered significant, consistently high expression considered necessary for bicluster inclusion, or distance allowed between bicluster and comparison samples) may give particularly good or poor results. It may also be that certain subtleties of particular applications of gene expression co-dependency analysis may be discovered during the course of subsequent investigations utilising the novel approaches presented in this thesis. It may either be pertinent to utilise such observations to improve the existing models employed in these analysis techniques, or to develop derivative methods for specific applications. However, it should also be noted that the methods used for the biological research applications presented in this thesis have been shown to be effective enough to have considerable utility in a range of applications.

7.4.3 Usability: Developing Interfaces

The work presented in this thesis focused on the development of meta-analysis approaches that could be used to study practicalities of pattern mining in large gene expression datasets, and ultimately to facilitate development of effective tools for studying transcriptional regulation through meta-analysis of gene expression data. As such, it was sufficient for any novel methods to be implemented in tools that could be used by the developer alone. It was therefore unnecessary to develop any simple, user-friendly interfaces for these tools for the scope of this work. However, now that analysis approaches involving the GESTr and HBLCA have been demonstrated to have application to investigations in biological research, it would greatly facilitate the application of these tools if simple, easy to use interfaces were vailable. R software packages have been compiled and incorporated into bash shell scripts for application of the GESTr method to new data collections and to datasets when a transformed compendium exists (e.g. for analysis with the TranSAM method). A suite of `R` functions was developed to apply the HBLCA approach to appropriately preprocessed data from within the `R` environment. However, to assist the use of these methods by biological investigators without familiarity to the R programming environment, a graphical interface would be most appropriate. It is proposed that such a graphical user interface for the biclustering meta-analysis tools, provided as a web service running on top of compiled

and preprocessed data collections, would enable widespread adoption of these methods and facilitate the incorporation of co-dependency analysis into many areas of biological research.

7.4.4 Further Application of Novel Analysis Approaches

A number of investigations were performed in this work to explore transcriptional regulatory mechanisms controlling pluripotency. The purpose of these investigations was twofold: firstly to generate and test hypotheses regarding mechanisms of activity of key TFs involved in pluripotency, and secondly to demonstrate that the gene expression analysis tools developed through the course of this work can be used to gain insight into biological processes through utilisation of existing data. While outside the scope of this work, experimental validation of the observations made in Chapter 6 make for a set of obvious follow-up tasks to the work presented in this thesis:

- **Section (6.1)** - investigation of the functional roles in ES cells of members of the subsets of genes differentially expressed upon Oct4 knock-down would consolidate the predictions made in Section (6.1). If confirmed these observations could lead to greater understanding of the complex regulatory mechanisms in balance in ES cells, and could possibly suggest new means of manipulating pluripotent cells.

- **Section (6.2)** - experimental validation of the lack of response to Oct4 differential expression in ES cells of the subset of Oct4 targets with 'Foxh1-like' expression patterns would need to be performed to confirm this prediction. It might also prove rewarding to investigate the consequences of altering the expression level of Foxh1 or the associated Oct4 targets listed in Section (5.6) in ES cells. This may provide insight into non-Oct4 dependent processes involved in the maintenance or loss of pluripotency.

- **Section (6.3)** - analysis described in Section (6.3) resulted in prediction of genes likely to be regulated in ES cells by specific combinations of the TFs Oct4, Sox2 and Nanog. High-confidence targets were identified with consistent evidence of DNA-binding by the relevant TFs, and so further confirmation of a predicted unique dependency on a particular combination of TFs may require potentially complicated combinatorial knock-down experiments. In the absence of such experimental data, further investigation of any available information regarding the identified target genes may add further weight to these predictions and provide some context regarding the possible significance of the different TF-target relationships predicted.

- **Section (6.4)** - confirm predicted lack of cMyc-response of highlighted members of the supposed cMyc-induced 'stem cell like' transcriptional signature. This could be achieved through forced overexpression of cMyc in a range of tissues and

quantifying mRNA levels of the target genes proposed in [Wong et al., 2008] that are predicted as a result of the analysis described in Section (5.5) to be unresponsive to cMyc expression in more than a restricted set of biological contexts.

As well as further investigation and validation of the observations presented in Chapter 6, as mentioned earlier the analysis tools described in this thesis may find use in many areas of biological research. Any additional successful applications of these tools to research problems will help to build a body of evidence to support the assertion that the theoretical foundations for these tools developed through the course of this work are sound, and that the implementations provided are effective.

7.5 Conclusion

The work described in this thesis covers the study, development and application of pattern mining approaches for assessing the evidence of transcriptional relationships, in the form of co-dependency of gene expression, in large collections of gene expression data. As part of this work, but of free-standing interest, approaches for modelling states of gene expression represented by large collections of potentially arbitrary measurements were developed and evaluated. Results presented in this thesis constitute the first analysis of applying biclustering methods to meta-analysis of diverse collections of gene expression data, highlighting issues concerning the dominance of bicluster expression patterns by widely-expressed genes and cell-type specific gene expression programs. An approach to modelling biological states of gene expression on a universal scale, known as the *GESTr*, is described in this thesis. The GESTr represents the first such universal model of gene expression state, and the utility of the approach is demonstrated in the facilitation of effective meta-analysis of collections of gene expression data and the identification of genes with biologically significant differential expression in a single microarray dataset. The concept of localised gene expression co-dependency analysis was introduced in this thesis, providing the opportunity to assess variations in the expression level of any pair of genes across sets of samples in the context of the similarity (or dissimilarity) of those samples. A probabilistic framework for this gene expression co-dependency was built, and is described in this thesis. This work enabled the development of a set of tools for the utilisation of large collections of diverse gene expression data to analyse evidence for transcriptional relationships involving particular genes or biological contexts of interest. A range of techniques were used to apply the HBLCA tool to study the transcriptional control of pluripotency, providing insight into cMyc-induced transcriptional signatures, identifying regulatory targets of different combinations of the TFs Oct4, Sox2 and Nanog, and identifying functional signatures associated with the differentiation of ES cells upon loss of Oct4 expression. In addition to suggesting avenues for further experimental investigation of the transcriptional control of pluripotency, the results of applying these analysis tools

to open questions in biological research suggest that the concepts introduced in this thesis and the methods developed to use these concepts for extracting relevant information from large collections of gene expression data provide effective means for the investigation of mechanisms of transcriptional regulation of biological processes. It is hoped that the novel concepts and analysis tools described in this thesis, constituting advances both in the theory of biological data analysis and in the set of bioinformatics tools available to researchers, might be further refined and adopted by the research community to afford valuable insight into the transcriptional regulatory mechanisms involved in a wide range of biological processes.

Bibliography

[Affymetrix,] Affymetrix. https://www.affymetrix.com/analysis/netaffx/index.affx.

[Ajiferuke et al., 2006] Ajiferuke, I., Wolfram, D., and Famoye, F. (2006). Sample size and informetric model goodness-of-fit outcomes: a search engine log case study. *Journal of Information Science*, 32(3):212–222.

[Allison et al., 2006a] Allison, D., Cui, X., Page, G., and Sabripour, M. (2006a). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65.

[Allison et al., 2006b] Allison, D., Cui, X., Page, G., and Sabripour, M. (2006b). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65.

[Alter et al., 2000] Alter, O., Brown, P., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97(18):10101–10106.

[Ambrosetti et al., 1997] Ambrosetti, D., Basilico, C., and Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by sox2 and oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Molecular and Cellular Biology*, 17:6321–6329.

[Arnold and Robertson, 2009] Arnold, S. and Robertson, E. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nature Reviews Molecular Cell Biology*, 10:91–103.

[Ashburner et al., 2000] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

[Assou et al., 2007] Assou, S., Carrour, T. L., Tondeur, S., Ström, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Reme, T., Hugnot, J.-P., Gasca, S., Hovatta, O., Hamamah, S., Klein, B., and Vos, J. D. (2007). A meta-analysis of human embryonic

stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells*, 25(4):961–973.

[Avilion et al., 2003] Avilion, A., Nicolis, S., Pevny, L., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on sox2 function. *Genes & Development*, 17:126–140.

[Ball et al., 2003] Ball, C., Sherlock, G., and Brazma, A. (2003). Funding high-throughput data sharing. *Nature Biotechnology*, 12:1179–1183.

[Bar-Joseph et al., 2003] Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N., Yoo, J., Robert, F., Gordon, D., Fraenkel, E., Jaakkola, T., Young, R., and Gifford, D. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21:1337–1342.

[Barkow et al., 2006] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. (2006). Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283.

[Barlow and Sherman, 1972] Barlow, P. and Sherman, M. (1972). The biochemistry of differentiation of mouse trophoblast: studies on polyploidy. *Journal of Embryology and Experimental Morphology*, 27:447–465.

[Baughman et al., 2009] Baughman, J., Nilsson, R., Gohil, V., Arlow, D., Gauhar, Z., and Mootha, V. (2009). A computational screen for regulators of oxidative phosphorylation implicates slirp in mitochondrial rna homeostasis. *PLoS Genetics*, 5(8):e1000590.

[Ben-Dor et al., 2004] Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2004). Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of Computational Biology*, 10(3-4):373–384.

[Ben-Porath et al., 2008] Ben-Porath, I., Thomson, M., Carey, V., Ge, R., Bell, G., Regev, A., and Weinberg, R. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumours. *Nature Genetics*, 40(5):499–507.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300.

[Bergmann et al., 2003] Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67:031902.

[Berriz et al., 2003] Berriz, G., King, O., Bryant, B., Sander, C., and Roth, F. (2003). Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504.

[Bhutani et al., 2010] Bhutani, N., Brady, J., Darmian, M., Sacco, A., Corbel, S., and Blau, H. (2010). Reprogramming towards pluripotency requires aid-dependent dna demethylation. *Nature*, 463:1042–1047.

[Brambrink et al., 2008] Brambrink, T., Foreman, R., Welstead, G., Lengner, C., Wernig, M., Suh, H., and Jaenisch, R. (2008). Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell*, 2(2):151–159.

[Brazma et al., 2003] Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.-A. (2003). Arrayexpress - a public repository for microarray gene expression data at the ebi. *Nucleic Acids Research*, 31(1):68–71.

[Bryan et al., 2005] Bryan, K., Cunningham, P., and Bolshakova, N. (2005). Biclustering of expression data using simulated annealing. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*.

[Buehr et al., 2008] Buehr, M., Meek, S., Blair, K., Yang, J., Ure, J., Silva, J., McLay, R., Hall, J., Ying, Q.-L., and Smith, A. (2008). Capture of authentic embryonic stem cells from rat blastocysts. *Cell*, 135(7):1287–1298.

[Cahan et al., 2005] Cahan, P., Ahmada, A., Burkea, H., Fua, S., Laib, Y., Floreac, L., Dharkera, N., Kobrinskia, T., Kalea, P., and McCaffrey, T. (2005). List of lists-annotated (lola): a database for annotation and comparison of published microarray gene lists. *Gene*, 360:78–82.

[Caldas et al., 2009] Caldas, J., Gehlenborg, N., Faisal, A., Brazma, A., and Kaski, S. (2009). Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153.

[Campbell et al., 2007] Campbell, P., , Perez-Iratxeta, C., Andrade-Navarro, M., and Rudnicki, M. (2007). Oct4 targets regulatory nodes to modulate stem cell function. *PLoS ONE*, 6:e553.

[Cedeno, 1995] Cedeno, W. (1995). *The multi-niche crowding genetic algorithm: analysis and applications*. PhD thesis, Universty of California Davis.

[Cedeno and Vemuri, 1996] Cedeno, W. and Vemuri, V. (1996). Genetic algorithms in aquifer management. *Journal of Network and Computer Applications*, 19(2):171–187.

[Cedeno et al., 1994] Cedeno, W., Vemuri, V., and Slezak, T. (1994). Multi-niche crowding in genetic algorithms and its application to the assembly of dna restriction-fragments. *Evolutionary Computation*, 2(4):321–345.

[Cedeno et al., 1995] Cedeno, W., Vemuri, V., and Slezak, T. (1995). Multiniche crowding in genetic algorithms and its application to the assembly of dna restriction-fragments. *Evolutionary Computation*, 2:321–345.

[Chakraborty and Maka, 2005] Chakraborty, A. and Maka, H. (2005). Biclustering of gene expression data using genetic algorithm. *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8.

[Chambers et al., 2003] Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, 113(5):643–655.

[Chambers et al., 2007] Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. *Nature*, 450:1230–1234.

[Chambers and Smith, 2004] Chambers, I. and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*, 23:7150–7160.

[Chambers and Tomlinson, 2009] Chambers, I. and Tomlinson, S. (2009). The transcriptional foundation of pluripotency. *Development*, 136:2311–2322.

[Chang et al., 2008] Chang, T.-C., Yu, D., Lee, Y.-S., Wentzel, E., Arking, D., West, K., Dang, C., Thomas-Tikhonenko, A., and Mendell, J. (2008). Widespread microrna repression by myc contributes to tumorigenesis. *Nature Genetics*, 40:43–50.

[Chen and Stoeckert, 2007] Chen, G. and Stoeckert, S. J. C. (2007). Clustering of genes into regulons using integrated modeling-cogrim. *Genome Biology*, 8:R4.

[Chen et al., 2008] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V., Wong, E., Orlov, Y., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H., Yeo, Z., Narang, V., Govindarajan, K., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117.

[Cheng et al., 2007] Cheng, K., Law, N., Siu, W., and Lau, T. (2007). Bivisu: software tool for bicluster detection and visualization. *Bioinformatics*, 23(17):2342–2344.

[Cheng and Church, 2000] Cheng, Y. and Church, G. (2000). Biclustering of expression data. *Proceedings of the Conference on Intelligent Systems For Molecular Biology*, pages 93–103.

[Chia and Karuturi, 2010] Chia, B. and Karuturi, R. (2010). Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms for Molecular Biology*, 5(23).

[Choe et al., 2005] Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005). Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16.

[Choi et al., 2003] Choi, J., Yu, U., Kim, S., and Yoo, O. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(S1):i84–i90.

[Cloonan et al., 2008] Cloonan, N., Forrest, A., Kolle, G., Gardiner, B., Faulkner, G., Brown, M., Taylor, D., Steptoe, A., Wani, S., Bethel, G., Robertson, A., Perkins, A., Bruce, S., Lee, C., Ranade, S., Peckham, H., Manning, J., McKernan, K., and Grimmond, S. (2008). Stem cell transcriptome profiling via massive-scale mrna sequencing. *Nature Methods*, 75(7):613–619.

[Conlon et al., 2006] Conlon, E., Song, J., and Liu, J. (2006). Bayesian models for pooling microarray studes with multiple sources of replications. *BMC Bioinformatics*, 7:247.

[Day et al., 2009] Day, A., Dong, J., Funari, V., Harry, B., Strom, S., Cohn, D., and Nelson, S. (2009). Disease gene characterization through large-scale co-expression analysis. *PLoS ONE*, 4(12):e8491.

[De Jong, 1975] De Jong, K. (1975). *An analysis of the behaviour of a class of genetic adaptive systems*. PhD thesis, University of Michigan.

[Ding et al., 2009] Ding, L., Paszkowski-Rogacz, M., Nitzsche, A., Slabicki, M., Heninger, A.-K., de Vries, I., Kittler, R., Junqueira, M., Shevchenko, A., Schulz, H., Hubner, N., Doss, M., Sachinidis, A., Hescheler, J., Iacone, R., Anastassiadis, K., Stewart, A., Pisabarro, M., Caldarelli, A., Poser, I., Theis, M., and Buchholz, F. (2009). A genome-scale rnai screen for oct4 modulators defines a role of the paf1 complex for embryonic stem cell identity. *Cell Stem Cell*, 4:403–415.

[Draghici et al., 2006] Draghici, S., Khatri, P., Eklund, A., and Szallasi, Z. (2006). Reliability and reproducibility issues in dna microarray measurements. *Trends in Genetics*, 22(2):101–109.

[Dudoit et al., 2002] Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12:111–139.

[Edgar et al., 2002] Edgar, R., Domrachev, M., and Lash, A. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

[Eisen et al., 1998] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868.

[Engelmann et al., 2008] Engelmann, J., Schwarz, R., Blenk, S., Friedrich, T., Seibel, P., Dandekar, T., and Müller, T. (2008). Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation. *Bioinformatics and Biology Insights*, 2:265–280.

[Evans and Kaufman, 1981] Evans, M. and Kaufman, M. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292:154–156.

[Falcon and Gentleman, 2007] Falcon, S. and Gentleman, R. (2007). Using gostats to test gene lists for go term association. *Bioinformatics*, 23(2):257–258.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874.

[Ferri et al., 2004] Ferri, A., Cavallaro, M., Braida, D., Cristofano, A. D., Canta, A., Vezzani, A., Ottolenghi, S., Pandolfi, P., Sala, M., DeBiasi, S., and Nicolis, S. (2004). Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development*, 131:3805–3819.

[Fleige et al., 2007] Fleige, A., Alberti, S., Gröbe, L., Frischmann, U., Geffers, R., Müller, W., Nordheim, A., and Schippers, A. (2007). Serum response factor contributes selectively to lymphocyte development. *Journal of Biological Chemistry*, 282(33):24320–24328.

[Fraley and Raftery, 1999] Fraley, C. and Raftery, A. (1999). Mclust: software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306.

[Fraley and Raftery, 2002] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

[Gasch et al., 2000] Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., and Brown, P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257.

[Gentleman et al., 2004] Gentleman, R., Carey, V., Bolstad, D. B. B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: open software development of computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

[Gilbert, 2006] Gilbert, S. (2006). *Developmental Biology*. Sinauer Associates, Inc.

[Goldberg, 1989] Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Reading: Addison-Wesley.

[Goldstein, 2006] Goldstein, D. (2006). Partition resampling and extrapolation averaging: approximation methods for quantifying gene expression in large numbers of short oligonucleotide arrays. *Bioinformatics*, 22(19):2364–2372.

[Grothaus et al., 2006] Grothaus, G., Mufti, A., and Murali, T. (2006). Automatic layout and visualization of biclusters. *Algorithms for Molecular Biology*, 1(15).

[Gyorffy et al., 2009] Gyorffy, B., Molnar, B., Lage, H., Szallasi, Z., and Eklund, A. (2009). Evaluation of microarray preprocessing algorithms based on concordance with rt-pcr in clinical samples. *PLoS ONE*, 4(5):e5645.

[Hajkova et al., 2002] Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maari, O., Reik, W., Walter, J., and Surani, M. (2002). Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of Development*, 117(1-2):15–23.

[Hall et al., 2009] Hall, J., Guo, G., Wray, J., Eyres, I., Nichols, J., Grotewold, L., Morfopoulou, S., Humphreys, P., Mansfield, W., Walker, R., Tomlinson, S., and Smith, A. (2009). Oct4 and lif/stat3 additively induce kruppel factors to sustain embryonic stem cell self-renewal. *Cell Stem Cell*, 5:597–609.

[Harbron et al., 2007] Harbron, C., Chang, K., and South, M. (2007). Refplus: an r package extending the rma algorithm. *Bioinformatics*, 23(18):2493–2494.

[Hartemink et al., 2001] Hartemink, A., Gifford, D., Jaakola, T., and Young, R. (2001). Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Proceedings of SPIE*, 4266(132).

[Hartigan, 1972] Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129.

[Hartigan, 1975] Hartigan, J. (1975). *Clustering algorithms*. John Wiley & Sons, New York, NY, USA.

[Hastie et al., 2000] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., and Brown, P. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):0003.

[Hibbs et al., 2007] Hibbs, M., Hess, D., Myers, C., Huttenhower, C., Li, K., and Troyanskaya, O. (2007). Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, 23(20):2692–2699.

[Hibbs et al., 2009] Hibbs, M., Myers, C., Huttenhower, C., Hess, D., Li, K., Caudy, A., and Troyanskaya, O. (2009). Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Computational Biology*, 5(3):e1000322.

[Hochbaum, 1996] Hochbaum, D. (1996). *Approximate algorithms for NP-hard problems*. PWS Publishing Company, Boston, MA.

[Hochberg and Benjamini, 1990] Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818.

[Hochedlinger and Plath, 2009] Hochedlinger, K. and Plath, K. (2009). Epigenetic reprogramming and induced pluripotency. *Development*, 136:509–523.

[Hochreiter et al., 2010] Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Sanden, S. V., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H., Shkedy, Z., and Clevert, D.-A. (2010). Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527.

[Holland, 1975] Holland, J. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

[Hong and Breitling, 2008] Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382.

[Hong et al., 2006] Hong, F., Breitling, R., McEntee, C., Wittner, B., Nemhauser, J., and Cory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

[Huang et al., 2009a] Huang, D., Sherman, B., and Lempicki, R. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

[Huang et al., 2009b] Huang, D., Sherman, B., and Lempicki, R. (2009b). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4:44–57.

[Husmeier, 2003] Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17):2271–2282.

[Huttenhower et al., 2009] Huttenhower, C., Mutungu, K., Indik, N., Yang, W., Schroeder, M., Forman, J., Troyanskaya, O., and Coller, H. (2009). Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274.

[Ihmels et al., 2004] Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcriptional modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003.

[Irizarry et al., 2006] Irizarry, R., Cope, L., and Wu, Z. (2006). Feature-level exploration of a published affymetrix genechip control dataset. *Genome Biology*, 7(8):404.

[Irizarry et al., 2003] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.

[Ivanova et al., 2006] Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. (2006). Dissecting self-renewal in stem cells with rna interference. *Nature*, 442:533–538.

[Jiang et al., 2008] Jiang, J., Chan, Y.-S., Loh, Y.-H., Cai, J., Tong, G.-Q., Lim, C.-A., Robson, P., Zhong, S., and Ng, H.-H. (2008). A core klf circuitry regulates self-renewal of embryonic stem cells. *Nature Cell Biology*, 20:353–360.

[Jupiter and VanBuren, 2008] Jupiter, D. and VanBuren, V. (2008). A visual data mining tool that facilitates reconstruction of transcription regulatory networks. *PLoS ONE*, 3(3):e1717.

[Katz et al., 2006] Katz, S., Irizzary, R., Lin, X., Tripputi, M., and Porter, M. (2006). A summarization approach for affymetrix genechip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics*, 7:464.

[Kim et al., 2008] Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061.

[Kim et al., 2010a] Kim, J., Woo, A., Chu, J., Snow, J., Fujiwara, Y., Kim, C., Cantor, A., and Orkin, S. (2010a). A myc nework accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell*, 143:313–324.

[Kim et al., 2010b] Kim, M., Cho, S., and Kim, J. (2010b). Mixture-model based estimation of gene expression variance from public database improves identification of differentially expressed genes in small sized microarray data. *Bioinformatics*, 26(4):486–492.

[Kleinsmith and Pierce, 1964] Kleinsmith, L. and Pierce, G. (1964). Multipotentiality of single embryonal carcinoma cells. *Cancer Research*, 24:1544–1551.

[Klemm and Pabo, 1996] Klemm, J. and Pabo, C. (1996). Oct-1 pou domain-dna interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes & Development*, 10:27–36.

[Knoepfler et al., 2006] Knoepfler, P., Zhang, X.-Y., Cheng, P., Gafken, P., McMahon, S., and Eisenman, R. (2006). Myc influences global chromatin structure. *The EMBO Journal*, 25:2723–2734.

[Kunath et al., 2007] Kunath, T., Saba-El-Leil, M., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). Fgf stimulation of the erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development*, 134:2895–2902.

[Lazzeroni and Owen, 2002] Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, 12:61–86.

[Lee and Batzoglou, 2003] Lee, S. and Batzoglou, S. (2003). Application of independent component analysis to microarrays. *Genome Biology*, 4(11):R76.

[Lemmens et al., 2006] Lemmens, K., Dhollander, T., Bie, T. D., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., Moor, B. D., and Marchal, K. (2006). Inferring transcriptional modules from chip-chip, motif and microarray data. *Genome Biology*, 7:R37.

[Levy and Hill, 2005] Levy, L. and Hill, C. (2005). Smad4 dependency defines two classes of transforming growth factor β target genes and distinguishes tgf-β-induced epithelial-mesenchymal transition from its antiproliferative and migratory responses. *Molecular and Cellular Biology*, 25(18):8108–8125.

[Lewandowski, 2001] Lewandowski, M. (2001). Conditional control of gene expression in the mouse. *Nature Reviews Genetics*, 2(10):743–755.

[Li et al., 2009] Li, G., Ma, Q., Tang, H., Paterson, A., and Xu, Y. (2009). Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research*, 37(15):e101.

[Li, 2002] Li, K.-C. (2002). Genome-wide coexpression dynamics: theory and application. *PNAS*, 99(26):16875–16880.

[Li et al., 2004] Li, K.-C., Liu, C.-T., Sun, W., Yuan, S., and Yu, T. (2004). A system for enhancing genome-wide coexpression dynamics study. *PNAS*, 101(44):15561–15566.

[Li et al., 2008] Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D., Iyer, V., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C., Chu, H., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S., Knowles, D., Gingeras, T., Speed, T., and Eisen, M. (2008). Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biology*, 6(2):365–387.

[Liebermeister, 2002] Liebermeister, W. (2002). Linear models of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60.

[Liu et al., 2008] Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G., and Kang, J. (2008). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Research*, 18:1177–1189.

[Lockhart et al., 1996] Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittman, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

[Loh et al., 2006] Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.-Y., Sung, K., Lee, C., Zhao, X.-D., Chiu, K.-P., Lipovich, L., Kuznetsov, V., Robson, P., Stanton, L., Wei, C.-L., Ruan, Y., Lim, B., and Ng, H.-H. (2006). The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, 38:431–440.

[Lutz et al., 2002] Lutz, W., Leon, J., and Eilers, M. (2002). Contributions of myc to tumorigenesis. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1602(1):61–71.

[Madeira and Oliveira, 2004] Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

[Marson et al., 2008] Marson, A., Levine, S., Cole, M., Frampton, G., Brambrink, T., Johnstone, S., Guenther, M., Johnston, W., Wernig, M., Newman, J., Calabrese, J., Dennis, L., Volkert, T., Gupta, S., Love, J., Hannett, N., Sharp, P., Bartel, D., Jaenisch, R., and Young, R. (2008). Connecting microrna genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533.

[Martin, 1975] Martin, G. (1975). Teratocarcinomas as a model system for the study of embryogenesis and neoplasia. *Cell*, 5:229–243.

[Martin, 1981] Martin, G. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *PNAS*, 78(12):7634–7638.

[Massy, 1965] Massy, W. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256.

[Masui et al., 2007] Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A., Ko, M., and Niwa, H. (2007). Pluripotency governed by sox2 via regulation of oct3/4 expression in mouse embryonic stem cells. *Nature Cell Biology*, 9(6):623–635.

[McNeish, 2004] McNeish, J. (2004). Embryonic stem cells in drug discovery. *Nature Reviews Drug Discovery*, 3:70–80.

[Mikkelsen et al., 2008] Mikkelsen, T., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B., Jaenisch, R., Lander, E., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature*, 454:49–55.

[Mirkin, 1989] Mirkin, B. (1989). *Mathematical classification and clustering*. Kluwer Academic Publishers.

[Mitchell et al., 1992] Mitchell, M., Forrest, S., and Holland, J. (1992). The royal road for genetic algorithms: fitness landscapes and ga performance. *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*.

[Mitra and Banka, 2006] Mitra, S. and Banka, H. (2006). Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39(12):2464–2477.

[Mitsui et al., 2003] Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein nanog is required for maintainance of pluripotency in mouse epiblast and es cells. *Cell*, 113(5):631–642.

[Miura et al., 2004] Miura, T., Brandenberger, R., Mejido, J., Luo, Y., Yang, A., Joshi, B., Ginis, I., Thies, R., Amit, M., Lyons, I., Condie, B., Itskovitz-Eldor, J., Rao, M., and Puri, R. (2004). Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood*, 103(8):2956–2964.

[Moreau et al., 2003] Moreau, Y., Aerts, S., Moor, B. D., Strooper, B. D., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, 19(10):570–577.

[Murali and Kasif, 2003] Murali, T. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*, pages 77–88.

[Murray and Keller, 2008] Murray, C. and Keller, G. (2008). Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell*, 132:661–680.

[Myers et al., 2006] Myers, C., Barrett, D., Hibbs, M., Huttenhower, C., and Troyanskaya, O. (2006). Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187.

[Nakagawa et al., 2008] Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without myc from mouse and human fibroblasts. *Nature Biotechnology*, 26:101–106.

[Ng et al., 2003] Ng, S.-K., Tan, S.-H., and Sundararajan, V. (2003). On combining multiple microarray studies for improved functional classification by whole-dataset feature selection. *Genome Informatics*, 14:44–53.

[Nguyen and Rocke, 2002] Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

[Nichols et al., 1998] Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klebe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the pou transcription factor oct4. *Cell*, 95(3):379–391.

[Niwa et al., 2000] Niwa, H., Miyazaki, J., and Smith, A. (2000). Quantitative expression of oct-3/4 defines differentiation, dedifferentiation or self-renewal of es cells. *Nature Genetics*, 24:372–376.

[Okita et al., 2007] Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, 448:313–317.

[Ouyang et al., 2009] Ouyang, Z., Zhou, Q., and Wong, W. (2009). Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *PNAS*, 106(51):21521–21526.

[Owen et al., 2003] Owen, A., Stuart, J., Mach, K., Villeneuve, A., and Kim, S. (2003). A gene recommender algorithm to identify coexpressed genes in c. elegans. *Genome Research*, 13:1828–1837.

[Oyanagi et al., 2001] Oyanagi, S., Kubota, K., and Nakase, A. (2001). Application of matrix clustering to web log analysis and access prediction. In *WEBKDD 2001Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, pages 13–21.

[Pearson, 2008] Pearson, R. (2008). A comprehensive re-analysis of the golden spike data: towards a benchmark for differential expression methods. *BMC Bioinformatics*, 9:164.

[Pearson et al., 2008] Pearson, R., Fleetwood, J., Eaton, S., Crossley, M., and Bao, S. (2008). Kruppel-like transcription factors: a functional family. *The International Journal of Biochemistry & Cell Biology*, 40(10):1996–2001.

[Perez-Iratxeta et al., 2005] Perez-Iratxeta, C., Palidwor, G., Porter, C., Sanche, N., Huska, M., Suomela, B., Muro, E., Krzyzanowski, P., Hughes, E., Campbell, P., Rudnicki, M., and Andrade, M. (2005). Study of stem cell function using microarray experiments. *FEBS Letters*, 579:1795–1801.

[Prelic et al., 2006] Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

[Quackenbush, 2001] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427.

[Ramasamy et al., 2008] Ramasamy, A., Mondry, A., Holmes, C., and Altman, D. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine*, 5(9):1320–1332.

[Remenyi et al., 2003] Remenyi, A., Lins, K., Nisson, L., Reinbold, R., Schöler, H., and Wilmanns, M. (2003). Crystal structure of a pou/hmg/dna ternary complex suggests differential assembly of oct4 and sox2 on two enhancers. *Genes & Development*, 17:2048–2059.

[Rocke and Durbin, 2001] Rocke, D. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6):557–569.

[Rossant, 2008] Rossant, J. (2008). Stem cells and early lineage development. *Cell*, 132:527–531.

[Santamaria et al., 2008] Santamaria, R., Theron, R., and Quintales, L. (2008). Bicoverlapper: a tool for bicluster visualization. *Bioinformatics*, 24(9):1212–1213.

[Sato et al., 2003] Sato, N., Sanjuan, I., Heke, M., Uchida, M., Naef, F., and Brivanlou, A. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. *Developmental Biology*, 260(2):404–413.

[Schena et al., 1995] Schena, M., Shalon, D., David, R., and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Segal et al., 2004] Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 36(10):1090–1098.

[Sharov et al., 2008] Sharov, A., Masui, S., Sharova, L., Piao, Y., Aiba, K., Matoba, Y., Xin, L., Niwa, H., and Ko, M. (2008). Identification of pou5f1, sox2, and nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*, 9:269.

[Shendure and Hanlee, 2008] Shendure, J. and Hanlee, J. (2008). Next-generation dna sequencing. *Nature Biotechnology*, 26(10):1135–1145.

[Simmen et al., 2007] Simmen, F., Xiao, R., Velarde, M., Nicholson, R., Bowman, M., Fujii-Kuriyama, Y., Oh, S., and Simmen, R. (2007). Dysregulation of intestinal crypt cell proliferation and villus cell migration in mice lacking kruppel-like factor 9. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, 292(6):G1757–69.

[Slonim and Yanai, 2009] Slonim, D. and Yanai, I. (2009). Getting started in gene expression microarray analysis. *PLoS Computational Biology*, 5(10):e1000543.

[Smith et al., 1988] Smith, A., Heath, J., Donaldson, D., Wong, G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature*, 336:688–690.

[Smyth, 2004] Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):article3.

[Sohal et al., 2008] Sohal, D., Yeatts, A., Ye, K., Pellagatti, A., Zhou, L., Pahanish, P., Mo, Y., Bhagat, T., Mariadason, J., Boultwood, J., Melnick, A., Greally, J., and Verma, A. (2008). Meta-analysis of microarray studies reveals a novel hematopoietic progenitor cell signature and demonstrates feasibility of inter-platform data integration. *PLoS ONE*, 3(8):e2965.

[Sperger et al., 2003] Sperger, J., Chen, X., Draper, J., Antosiewicz, J., Chon, C., Jones, S., Brooks, J., Andrews, P., Brown, P., and Thomson, J. (2003). Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. *PNAS*, 100(23):13350–13355.

[Sridharan et al., 2009] Sridharan, R., Tchieu, J., Mason, M., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. *Cell*, 136:364–377.

[Stalteri and Harrison, 2007] Stalteri, M. and Harrison, A. (2007). Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC Bioinformatics*, 8(13).

[Su et al., 2002] Su, A., Cooke, M., Ching, K., Hakak, Y., Walker, J., Wiltshire, T., Orth, A., Vega, R., Sapinoso, L., Moqrich, A., Patapoutian, A., Hampton, G.,

Schultz, P., and Hogenesch, J. (2002). Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99(7):4465–4470.

[Suarez-Farinas et al., 2005] Suarez-Farinas, M., Noggle, S., Heke, M., Hemmati-Brivanlou, A., and Magnasco, M. (2005). Comparing independent microarray studies: the case of human embryonic stem cells. *BMC Genomics*, 6:99.

[Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–676.

[Tamayo et al., 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation. *PNAS*, 96(6):2907–2912.

[Tanay et al., 2002] Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(S1):S136–S144.

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285.

[Thanos and Maniatis, 1995] Thanos, D. and Maniatis, T. (1995). Virus induction of human ifnb gene expression requires the assembly of an enhanceosome. *Cell*, 83:1091–1100.

[Thomson, 1998] Thomson, J. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*, 282:1145–1147.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, 63(2):411–423.

[Tijms, 2004] Tijms, H. (2004). *Understanding probability: chance rules in everyday life*. Cambridge University Press, Cambridge UK.

[Toern and Zilinskas, 1989] Toern, A. and Zilinskas, A. (1989). *Global optimization*. Lecture Notes in Computer Science, Berlin: Springer.

[Tusher et al., 2001] Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121.

[van den Bosch et al., 2007] van den Bosch, H., Bünger, M., de Groot, P., van der Meijde, J., Hooiveld, G., and Müller, M. (2007). Gene expression of transporters

and phase i/ii metabolic enzymes in murine small intestine during fasting. *BMC Genomics*, 8:267.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63.

[Warnat et al., 2005] Warnat, P., Eils, R., and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265.

[Wennmalm et al., 2005] Wennmalm, K., Wahlestedt, C., and Larsson, O. (2005). The expression signature of *in vitro* senescence resembles mouse but not human aging. *Genome Biology*, 6:R109.

[Whitley, 1994] Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85.

[Williams et al., 2003] Williams, D., Cai, M., and Clore, G. (2003). Molecular basis for synergistic transcriptional activation by oct1 and sox2 revealed from the solution structure of the 42-kda oct1-sox2-hoxb1-dna ternary transcription factor complex. *Journal of Biological Chemistry*, 279:1449–1457.

[Wong et al., 2008] Wong, D., Liu, H., Ridky, T., Cassarino, D., Segal, E., and Chang, H. (2008). Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*, 2:333–344.

[Wu et al., 2007] Wu, W.-S., Li, W., and Chen, B. (2007). Identifying regulatory targets of cell cycle transcription factors using gene expression and chip-chip data. *BMC Bioinformatics*, 8:188.

[Ying et al., 2003] Ying, Q.-L., Nichols, J., Chambers, I., and Smith, A. (2003). Bmp induction of id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with stat3. *Cell*, 115(3):281–292.

[Ying et al., 2008] Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. *Nature*, 453:519–524.

[Yul et al., 2007] Yul, J., Vodyanik, M., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J., Tian, S., Nie, J., Jonsdottir, G., Ruotti, V., Stewart, R., Slukvin, I., and Thomson, J. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.

[Zhang et al., 2007] Zhang, J., Ji, Y., and Zhang, L. (2007). Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23(21):2903–2909.

[Zhang et al., 2008] Zhang, X., Zhang, J., Wang, T., Esteban, M., and Pei, D. (2008). Esrrb activates oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *Journal of Biological Chemistry*, 283:35825–35833.

[Zilliox and Irizarry, 2007] Zilliox, M. and Irizarry, R. (2007). A gene expression bar code for microarray data. *Nature Methods*, 4(11):911–913.