Integration and biological interpretation of microarray gene expression profiling data

A thesis submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

by

Suraj Menon

Department of Pathology
School of Medicine
Cardiff University

September 2009

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed:...(candidate)   Date: 11/11/09

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

Signed:...(candidate)   Date: 11/11/09

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed:...(candidate)   Date: 11/11/09

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed:...(candidate)   Date: 11/11/09

Acknowledgements

"I am your PhD supervisor. I am a resource for you to exploit."

- Prof. David Glyn Kipling (2005)

.. or something in that vein. This was one instruction(?) I did carry out heroically well, and I hope David's experience with me has not put him off saying this to his next student. I wouldn't know where to begin to thank him: he granted me the opportunity to embark on this most significant intellectual endeavour; he has taught me about the Scientific Method, critical thinking, scientific rigour, scientific caution (and paranoia), presentations, scientific writing and erm... sea squirts; he has prepared me for a career in science, and helped me take my first step into it. I have learnt much from what I affectionately (well... at least a lot of the time) refer to as David's 'flaming sword of logic'. His enthusiasm for observing and understanding life and the processes that drive it is infectious and inspirational. I would any day rate our discussions and meetings right up alongside the joys of discovery and seeing a well-planned experiment work as amongst the most enjoyable aspects of my PhD experience.

Much thanks is owed as well to my secondary supervisor, Dr. Ian Brewis. Despite his busy schedules he has always been around when I needed him, with a shoulder to lean on, and a pint to cry into. His advice and encouragement have always been very helpful, and are much appreciated.

I must of course thank the long-suffering Dr. Peter Giles, my fellow troll (obscure Dilbert reference) whom I exploited almost much as I did David. I tried my best to infuriate him with a constant barrage of questions on statistics, programming and microarray data analysis of breathtaking stupidity, and endless pleas for help ("My computer is broken."). However, he has admirably responded to these with superhuman levels of patience, stoicism and of course, helpfulness. I am also grateful to Dr Hui Sun Leong for letting me use her painstakingly compiled database of genelists.

My three-year research period was generously funded by an MRC-BP Dorothy Hodgkin Postgraduate Award, which is a life-line for international students such as myself. I am also grateful to the University of Wales for funding my writing-up period through the Lord Merthyr Scholarship, which ensured that I was not driven to the extremes of desperation as have been certain other writing-up PhD students (not-so-obscure popular culture reference).

I must also thank my parents for their unwavering love and support, and for somehow (rather unscientifically) believing that I was capable of achieving much more than I ever thought I could.

Finally, and most of all, I would like thank my wife: for her love, care and patience which helped me through the most difficult times; for leaving behind everything and everyone she has ever known to come and be with me; for giving me all the reason I will ever need to better myself and be someone.

Summary

Many different strategies have been developed for the analysis of microarray data and these have a significant influence on the level and quality of knowledge that may be achieved from a microarray-based experiment. Two such strategies are explored in this thesis.

Part A of this thesis describes explorations of a resource-efficient strategy that could allow for large-scale integration of microarray data in an unsupervised fashion. For this purpose, comparisons were carried out between a series of genelists manually extracted from the literature, representing a disparate set of microarray experiments. Initial results were highly unexpected, and are likely to have been caused by violations of the assumptions of the hypergeometric test used for assessing comparisons. Statistical modelling was found to successfully simulate these results; however the estimated net effect of these violations was found to be considerable. These findings strongly caution against the comparison of microarray experiments using their genelists.

Part B then describes the development of Gene Set Discovery (GSD), a novel methodology to perform threshold-free gene set analysis of microarray datasets without requiring sample class information. This was achieved by deriving a novel metric that allows for the selection of those gene sets that exhibit significant discrimination between samples. GSD was implemented on four microarray datasets and the results were found to be biologically plausible and/or in agreement with prior analyses of these datasets. These findings suggested that GSD could be a potentially useful tool for biological theme discovery in microarray datasets, particularly in studies of cancer where sample classification is problematic. Also described is a related methodology for extraction of informative genes from within selected gene sets, and a scheme for visualization of results in an integrated format.

Abbreviations

ALL – Acute Lymphoblastic Leukaemia

AML – Acute Myeloid Leukaemia

ANOVA – Analysis of Variance

DEG(s) – Differentially Expressed Gene(s)

EGID – Entrez Gene Identifier

EST – Expressed Sequence Tag

FDR – False Discovery Rate

GEO – Gene Expression Omnibus

GNF – Genomics institute of the Novartis research Foundation

GO – Gene Ontology

GOBP – Gene Ontology Biological Process

GSA – Gene Set Analysis

GSD – Gene Set Discovery

GSEA – Gene Set Enrichment Analysis

GSECA – Gene Set Enrichment Coherence Analysis

HUGO – HUman Genome Organization

IQR – Inter Quartile Range

KEGG – Kyoto Encyclopaedia of Genes and Genomes

MAS5 – Microarray Suite 5.0

M-GDM – Mean of the Gene Distance Matrix

M-SDM – Mean of the Sample Distance Matrix

SD-GDM – Standard Deviation of the Gene Distance Matrix

SD-GDV – Standard Deviation of the Gene Distance Vector

SD-SDM – Standard Deviation of the Sample Distance Matrix

ORA – Over-Representation Analysis

PAM – Point Accepted Mutation

RMA – Robust Multichip Average

WD/ MYX/ PLEO/ DD – Well Differentiated/ MYXoid/ PLEOmorphic/ De-Differentiated (Liposarcoma tumour subtypes)

Table of Contents

Chapter 1: Introduction

The work described in this thesis primarily concerns the implementation and development of methods with which to analyze and interpret data from microarray-based experiments. This chapter introduces some of the basic underlying concepts of microarray technology. It also aims to provide a broad overview of the analytical workflows, technical methodologies and strategies used for microarray data analysis. Because this is a constantly developing sphere of research with a varied range of available options, this chapter places emphasis on describing those concepts, methods and strategies that are relevant to work described in this thesis.

Section 1.1 examines the importance of high-throughput gene expression profiling, and the evolution of microarrays for this purpose. Section 1.2 then introduces Affymetrix GeneChip technology and the generation of gene expression data. Section 1.3 describes several different strategies for microarray experimental design and analysis, the choice of which to use depends on the aims of an experiment. Section 1.4 describes the underlying concepts of methodologies that have been popularly used to aid the biological interpretation of microarray data. In particular, this section describes methods used to link microarray data to prior biological knowledge. Section 1.5 examines the concepts and strategies used for integration of data from different microarray-based experiments. Finally Section 1.6 describes the scope and structure of work described in this thesis, as well as guidance regarding the terminology used.

1.1 Gene expression profiling using microarrays

1.1.1 The paradigm shift in molecular biology

A grand aim of molecular biology studies has been to elucidate how information coded in the genome is used for the development and maintenance of a functional living organism. Prior to the development of some of the technologies described below, most research was carried out by studying the functions of one gene at a time. However, the biochemical processes of life involve complex networks and interactions between genes and gene products, and the scope for such a 'reductionist' approach to capture these complexities is limited (Vukmirovic and Tilghman 2000).

Over the past 10-15 years, there have been several technological advancements that have allowed for molecular biology studies to be carried out using a 'holistic' approach. One of the first such developments was that of high-throughput whole-genome sequencing technologies, which has led to the sequencing of complete genomes of hundreds of organisms, including humans (Lander et al. 2001; Venter et al. 2001). While genome sequences (along with the information derived from sequence analyses) can be thought of as 'gene catalogues' representing lists of all the components of a functional genome, other high-throughput 'post-genome' technologies have been developed that allow global studies of the interactions and relationships between these components. One of the most widely used of these technologies is DNA microarrays (Pease et al. 1994; Schena et al. 1995).

Microarrays allow for global gene-expression profiling, by monitoring the levels of mRNA expressed by thousands of genes simultaneously. The mRNA complement of a cell (i.e. its transcriptome) is a major determinant of phenotype and function. Unlike the genome, it is highly dynamic, changing rapidly and dramatically both during normal cellular events (such as cell division), or in response to external stimuli (such as treatment with a drug) (Lockhart and Winzeler 2000). The rationale for the use of microarrays is that observation of the levels of gene expression (i.e. mRNA abundance), and the conditions for expression, can provide clues about the functions of genes. Patterns of expression shared by many genes can inform about broader biochemical themes and processes (such as pathways and regulatory mechanisms), as well as interactions between genes and gene products. Simultaneous observation of large numbers of genes (such as all genes of an organism) allows for identification of potentially all genes relevant to particular experimental conditions.

Microarrays have been used for a wide variety of applications, such as biomarker identification, pharmacogenomics, toxicogenomics, disease class discovery, etc. While microarrays have been developed for other purposes, such a detection of mutations ('genotyping arrays'), this thesis only concerns use of microarrays for gene expression profiling.

1.1.2 The development of microarray technology

The key principle underlying microarray technology is that complementary nucleic acid (DNA or RNA) strands hybridize to each other. This principle has formed the basis for several established molecular biology techniques, such as Southern and Northern blotting. In Southern blotting, short nucleic acid sequences are radio-labelled and used as 'probes' to hybridize to DNA sequences that have been separated on the basis of size by gel electrophoresis. The occurrence of binding is then visualized using radiation-sensitive photographic film. In Northern blotting, the probes are hybridized to mRNA instead. In both cases, the intensity of the radio-labelled probe on the film is then used as a semi-quantitative measure of the amount of DNA/RNA present, as compared to a known standard.

The use of arrays for gene expression profiling has developed from the idea of a mass parallel version of these blotting techniques (Lander 1999), with a key distinction being the immobilization of probes to a solid substrate. The first such arrays ('macroarrays') involved spotting of cDNA libraries (usually of unknown sequence) as probes onto porous nylon membranes, onto which radio-labelled mRNA was hybridized. The microarrays that are used today have evolved from these, and involve great improvements in terms of experimental efficiency and information content. Several factors have aided the development of microarrays, such as the use of non-porous solid substrate (glass slides), the use of fluorescence for detection (as opposed to radio-labelling) and the development of technologies for synthesis or deposition of probes on substrates at very high densities (Lockhart and Winzeler 2000).

1.1.3 Overview of microarray technologies

There are several different techniques that can be used to create microarrays, which in turn require different experimental workflows and data analysis pipelines. However, all these aspects share some fundamental principles. Firstly, in all cases, the nucleic acid sequences representing the probes are bound to solid surfaces (usually glass slides) at known positions, using a variety of techniques. Next, expressed mRNA is extracted from an experimental sample and converted into cDNA by reverse transcription. This is then labelled using fluorescent dyes, eluted onto the arrays and allowed to hybridize. Hybridization is detected by fluorescence following laser excitation, and the intensity of the fluorescence is used to compute an estimate of expression levels. Details of two of the most popularly used microarray technologies are described below.

1.1.3.1 cDNA ('spotted') microarrays

cDNA microarrays (Brown and Botstein 1999) are created by robotic spotting of entire cDNA/EST sequences onto glass slides at precise pre-defined points, to be used as probes. Normally, these are used to assess differential expression between two samples: cDNA reverse-transcribed from mRNA extracted from one sample is labelled with a green fluorescent dye (Cy3), and that from the other with a red fluorescent dye (Cy5).

The labelled cDNA is mixed and allowed to co-hybridize on the slide. The slide is then scanned using two different wavelengths of a laser to obtain the intensities for each dye used (two-channel detection).

1.1.3.2 High-density oligonucleotide microarrays

Oligonucleotide arrays (Lipshutz et al. 1999) are created using photolithographic techniques that allow for extremely high feature density with complete control of sequences used as probes. Typically, a set of probes comprising of unique sequences are used to represent a single gene or expressed sequence tag (EST). cDNA reverse-transcribed from the mRNA from each sample is labelled and hybridized onto separate arrays, each of which is laser-scanned separately (single channel detection).

1.1.3.3 Choice of microarrays

The choice of which microarray technology to use is decided by the needs of the researcher, and both technologies described above have advantages and disadvantages relative to each other. cDNA microarrays are typically designed and produced by the researchers themselves, and this system allows a great deal of flexibility with regards to array design and features. This also does not require prior knowledge of the sequence of the probes, which is useful for experiments on organisms for which availability of sequence data is limited. Oligonucleotide microarrays are usually designed and produced by commercial manufacturers, and require sequence information for probes. However, this removes the resource-intensive and potentially error-prone requirement for researchers to maintain and use cDNA libraries for probes.

1.2 Affymetrix GeneChip technology

Affymetrix is one of the largest commercial manufacturers of high-density oligonucleotide microarrays. All data presented in this thesis are derived from experiments carried out on this platform.

1.2.1 Experimental workflow and data generation

1.2.1.1 Array design

The Affymetrix GeneChip microarrays consist of oligonucleotide probes (25 nucleotides long) synthesised by a photolithographic process. Each gene is represented by one or more probe-sets, each comprising of 11-20 perfect-match (PM) oligonucleotide probes and 11-20 corresponding mismatch (MM) probes. The PM probes have sequences that are complementary to sequence fragments of a particular gene. The MM probes are identical to the PM probes except with a single base substitution at position 13 (out of 25). Figure 1.1 displays the Affymetrix probe-set design strategy for eukaryotic organisms. Affymetrix claims that MM probes allow for quantification of (and subsequent control for) background noise and cross-hybridization by transcripts from different genes. The sequences used to design the probes are derived from several public sequence databases, such as UniGene, RefSeq, GenBank and dbEST (Affymetrix 2001).

1.2.1.2 Sample processing

The typical experimental workflow for samples from eukaryotic organisms is described as follows: first mRNA is isolated from cells (which may be from a tissue sample or a cell line), and then reverse-transcribed into double-stranded cDNA.

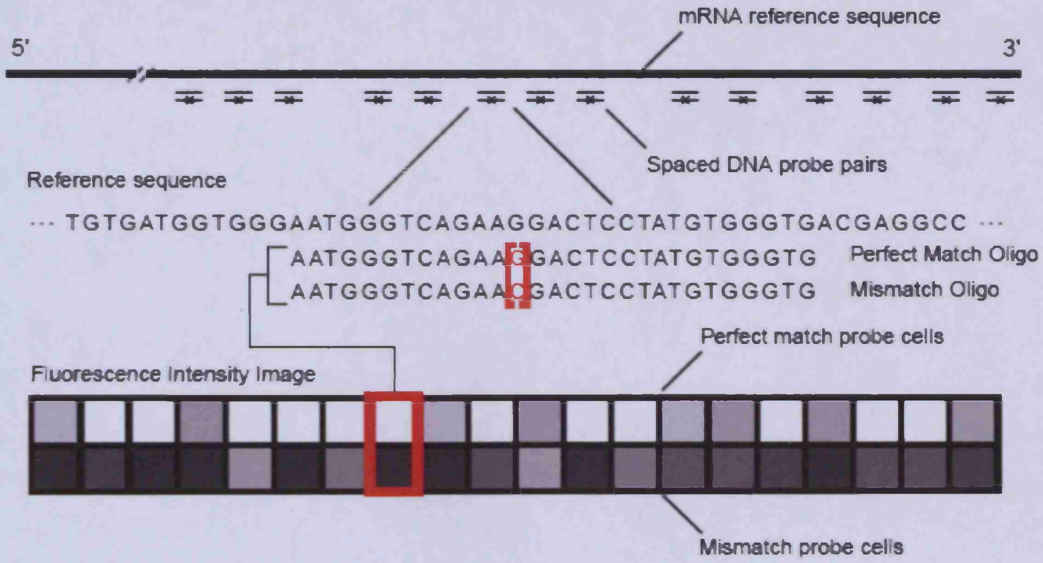

mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1.1 Affymetrix probe-set design strategy. This figure is reproduced from Figure 2b in Lipshutz et al (1999)

The next stage involves amplification of this cDNA into biotin-labelled cRNA which is then fragmented. This cRNA is then eluted over an array to allow for hybridization to occur over an extended period of time (16 hours) at optimal hybridization temperatures. It is assumed that the amounts of cRNA that hybridize to their respective probes is proportional to their relative levels within the original sample. Following this, unhybridized material is washed away and a fluorochrome (streptavidin-phycoerythrin) is added to bind to the biotin-label on the hybridized cRNA. The array is then placed in a scanner where a laser is applied to excite the fluorochrome. An image of the array is stored as a *DAT* file, recording the intensity of fluorescence for each probe in many pixels. Using software included with the scanner, a single intensity value is calculated for each probe using all pixel intensities for that probe. These probe intensity values are stored in a *CEL* file.

1.2.2 Data pre-processing

The probe intensity values contained in *CEL* files represent the 'raw data' from microarray-based analyses. However, a series of data manipulation and statistical modelling steps are usually carried out to make this data comparable within and across arrays. Such pre-processing is carried out to produce biologically meaningful data that can then be used for expression analysis. Many pre-processing methods are now available; the Affycomp (Cope et al. 2004) initiative to benchmark these methods has been used to assess nearly 90 such methods (as of July 2009). However it is still unclear as to which is the method is the 'best' (Allison et al. 2006).

Most of these methodologies have in common a three-stage approach (Bolstad et al. 2005; Gentleman and Huber 2008): one stage involves 'background correction' to control for any non-specific signal (as may be caused by cross-hybridization of non-target transcripts with similar sequences), and to identify a detection threshold. Carrying out of this stage helps make the data across an array comparable, and increases array sensitivity.

The second stage is the process of 'between-array normalization' which is performed to minimize undesirable technical variability between data across the arrays, as may be caused by differences in handling, labelling, hybridization and scanning of different arrays. This stage is necessary to make the data comparable across chips, and to ensure much of the variability between arrays is due to biological reasons (which are of interest to a researcher).

The final stage is 'summarization'; this is particularly relevant to data from Affymetrix arrays because a probe-set representing a single gene transcript on an array comprises of 11-20 different probes. This process then involves combining the multiple probe intensity values for each probe-set to produce a single gene expression value for that probe-set.

Two of most widely used pre-processing methodologies are described below.

1.2.2.1 Microarray Suite 5.0 (MAS5)

MAS5 (Affymetrix 2002) is the software developed by Affymetrix for pre-processing of microarray data, and utilizes probe intensity data for both PM and MM probes. Background correction is carried out by using the lowest 2% probe intensities for various regions of the array to calculate background values for those regions. Probe intensities are then adjusted using a weighted average of each of the background values. Between-array normalization is carried out using a scaling technique: a baseline array is selected and all other arrays are scaled to have the same mean intensity as this array. Summarization is carried out by calculating a 'Signal' value representing the expression level for each probe-set, using intensity values for all PM probes and adjusting these for intensity values of all MM probes. MAS5 also provides for each probe-set, a 'Detection Call' to indicate if the transcript represented by a probe-set is 'Present', 'Marginal' or 'Absent'.

1.2.2.2 Robust Multichip Average (RMA)

The RMA algorithm, which was developed by independent researchers (Irizarry et al. 2003), utilizes intensity values for only the PM probes. Background correction is carried out by modelling probe intensity values as the sum of a Gaussian noise component and an exponential signal component. Between-array normalization is performed by using quantile-quantile normalization to impose the same empirical distribution of intensities to each array. Summarization is based on a multi-array model using the 'median polish' algorithm to robustly estimate central tendency.

1.3 Microarray data analysis strategies and workflows

Pre-processing of Affymetrix microarray data is usually followed by one or more steps of data transformation, such as log transformation, and gene-wise mean/median centring. This yields a 'gene expression matrix' which is the starting point for all subsequent data analysis. It comprises of a matrix where the rows represent genes (probesets) and the columns represent experimental conditions (samples). The cells are filled with numbers representing the expression level of a gene within a sample.

The following sections describe categories of statistical methodologies and data mining techniques that are commonly used to analyze microarray data. These can be divided into two broad categories on the basis of whether or not they utilize information regarding the samples (Allison et al. 2006; Butte 2002; Causton et al. 2003; Dudiot and Fridyland 2003; Tarca et al. 2006). The choice of the data analysis workflow depends on the nature and design of the experiment, as well as the information desired by the researcher.

1.3.1 Supervised Analysis – Class comparison and prediction

Methodologies used to analyze microarray data can be considered to be 'supervised' if they require knowledge regarding the classes of samples. These classes usually represent two or more different experimental conditions. Sample class information may be known *a priori* during the experimental design phase (for example, normal versus diseased samples, cells treated with a chemical versus untreated cells, or different times points of a developmental process) or may be derived through unsupervised class discovery studies (see Section 1.3.1.2).

1.3.1.1 Class comparison: assessing differential gene expression

The primary objective of class comparison studies is to assess whether the expression profiles of samples representing two or more classes are different, and to identify which genes exhibit differential expression levels across these sample classes. Typically, these involve performing a statistical test to assess the significance of differences in gene expression levels across sample classes for each gene separately. A wide range of tests have been used for this purpose (Jeffery et al. 2006; Pan 2002), such as variants of the Student's t-test for two-group analyses and variants of the analysis of variance (ANOVA) test for when more than two groups are being analyzed. Genes can then be ranked on the basis of some metric derived from these tests and then selected using a pre-defined cut-off value to indicate significance (typically, a p-value of <0.05).

An issue that arises because of the performance of separate statistical tests on each of several thousand genes is that of the potentially large number of 'false positives' that would be expected. For example, when selecting genes exhibiting p-values of less than 0.05 after carrying out a statistical test for differential expression on 20,000 genes, it is expected to select as many as 1,000 genes simply by chance alone (and not for any biological reasons). For this reason, it is considered necessary to carry out some form of 'multiple testing correction', such as the Benjamini-Hochberg FDR method (Benjamini and Hochberg 1995) to increase p-values in proportion to the number of tests being performed (i.e. the number of genes being tested). Because the stringency of these tests increases with the number of genes analyzed, it has also become standard practice to carry out 'non-specific filtering' of genes to reduce this 'gene universe' size (Huber et al. 2008; Scholtens and Heydebreck 2005). These include removing genes that fail to exceed threshold levels of expression (such as those used by the MAS5 algorithm to assign Detection Calls of 'Present'), or variability (because genes exhibiting stable expression levels across sample classes are unlikely to be of interest).

1.3.1.2 Class prediction: developing sample classifiers

The objective of class prediction (also known as 'classification') techniques is to build a set of genes (a 'classifier') using samples with known classes that can classify other samples for which class information may not be available. Given a dataset with known sample classes, class prediction is typically carried out by first using a subset of this data (the 'training set') to derive a classifier. The accuracy of class prediction achieved by the classifier is then tested on another subset of the data ('validation'). Having assessed the quality of the predictions, the classifier can then be used on new datasets. Several different supervised machine learning algorithms have been used for this purpose, such as support vector machines, neural networks and decision trees (Allison et al. 2006; Butte 2002; Causton et al. 2003; Quackenbush 2001).

1.3.2 Unsupervised analysis – Class discovery

Unsupervised analysis of microarray data requires no prior knowledge of sample classes. The objective of such analyses is to 'discover' classes of genes and samples within a microarray dataset by identifying groups of genes and samples on the basis of their expression profiles. For this purpose, methods are used to find any underlying structure within the data with respect to shared patterns of gene expression.

Class discovery in microarray data was first described by Golub et al (Golub et al. 1999), who achieved automated separation of acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL), as well as differentiation between B-cell ALL and T-cell ALL, without requiring prior knowledge of these cancer classes. Indeed, class discovery has found much utility in studies of cancers, where morphological and histological methods may not provide adequate discrimination between tumour sub-types. For example, Alizadeh et al (Alizadeh et al. 2000) used hierarchical clustering to identify two previously unknown sub-groups of diffuse large B-cell lymphoma (DLBCL) samples. These new sub-groups corresponded to highly significant differences in patient survival rates and prognoses, even though the samples were not morphologically distinct. There are many other such examples in the published literature (Bittner et al. 2000; De Cecco et al. 2004; Ivshina et al. 2006; Perou et al. 2000).

Several different techniques have been used for class discovery in microarray data, such as clustering and self-organizing maps. Described below are two of the most popularly used clustering methodologies:

1.3.2.1 Hierarchical clustering

A key concept used by hierarchical clustering techniques is that of 'distances'. These represent quantifications of the dissimilarity between any pair of genes or samples in a microarray dataset, based on their expression profiles. For this purpose, genes can be considered to be points in M-dimensional space, for an experiment with M number of samples (Kuruvilla et al. 2002). Similarly, samples can be considered to be points in N-dimensional space, where N represents the number of genes being considered. A number of distance metrics can be derived from this model, using, for example, the Euclidean distance between any two points, or the vector angle (as the cosine distance). One such popularly used distance measure is the Pearson's correlation distance, which is equivalent to using the vector angle for mean-normalized data (Eisen et al. 1998).

Hierarchical clustering algorithms then use these distances to build a tree ('dendrogram') to represent the hierarchical structure of the data. The 'nodes' represent genes or samples and the 'branch' lengths are based on the pre-calculated distances. 'Divisive' hierarchical clustering methods start off by considering all objects (genes or samples) to be part of a single cluster and divide this into further sub-clusters. This is iterated by considering each sub-cluster separately till all objects are separated from each other. 'Agglomerative' hierarchical clustering methods start by considering each gene/sample to be separate clusters. The most similar objects are then considered to be a single cluster and the distances between this cluster and all other objects are re-calculated. The object closest to the new cluster is added to this cluster; this is iterated till all objects are grouped in a single cluster. Using a pre-defined distance, or one decided by the user having inspected of the dendrogram, the tree can be 'cut' at certain points to define the final clusters.

One technique that has widely been used for cluster analysis of microarray data (and in microarray data visualization schemes such as 'heatmaps') is agglomerative hierarchical clustering utilizing the Unweighted Pair Group Method with Arithmetic mean (UPGMA) method. This uses the average distance between every point in one cluster to every point in another cluster as the measure of cluster distance ('average linkage').

1.3.2.2 K-means clustering

K-means is a 'partitioning' clustering technique that differs from hierarchical clustering in that it does not produce a hierarchical structure of objects, does not require pre-calculation of all pair-wise distances between objects, but does require a user-defined number of clusters (K). This is carried out by random (or heuristic) assignment of all objects to K number of clusters. The distances between each object and cluster center ('centroid') are calculated, and each object is re-assigned to the cluster with the nearest centroid. This is iterated by recalculation of centroids (of the newly formed clusters) until the centroids stabilize or a pre-defined number of iterations is achieved.

1.4 Biological interpretation of microarray data

The results derived from the strategies and data analysis pipelines described in Section 1.3 may be sufficient for certain microarray-based experiments. For example, having derived a list of genes that exhibit significantly differential expression levels across disease and non-disease samples, these genes may then be used as candidate biomarkers for the disease. However, in most cases, a natural progression would be to attempt to interpret these results in terms of their underlying biology. This could provide further insight into the biological mechanisms that are relevant to an experiment.

One of the most widely used strategies to achieve this is to investigate microarray data in the context of 'biological themes'. Such themes include biochemical pathways and processes, and can be represented as sets of genes known *a priori* to be relevant to any particular theme ('gene sets'). Investigation of microarray data in terms of biological themes can be termed as 'gene set analysis' (GSA), and currently there are many options for this purpose: Huang et al have identified and reviewed as many as 68 different methodologies (Huang da et al. 2009). Key to the utility of these methodologies has been the development of publicly available databases that store information regarding biological themes in an electronic format that allows for automated analyses. This section first describes these databases, and then introduces two broad categories of GSA techniques.

1.4.1 Gene annotation databases

The Gene Ontology database (Ashburner et al. 2000) comprises of annotation data for genes of a wide range of species. This was created by using a controlled vocabulary to describe and represent *a priori* biological knowledge regarding these genes and their products. The components of this vocabulary ('GO terms') each have a unique identifier, and are grouped into three different categories, representing different types of information for any particular gene and gene product. The are biological processes, e.g. "apoptosis" (GO:0006915); molecular functions, e.g. "kinase binding" (GO:0019900); and cellular components, e.g. "plasma membrane" (GO:0005886). Each of these GO terms can be thought of as biological themes; thus any theme can be represented as the set of genes that are annotated with a particular GO term.

Other sources of gene sets that have been used as biological themes by GSA methodologies include biological pathway annotation databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004) and Biocarta (BioCarta 2005).

A key difference between GO and other sources of gene sets is that GO includes a framework of relationships between terms. Typically, GO terms representing larger, more general themes are considered to be 'parents' of GO terms representing smaller, more specific themes. For example, all genes that are annotated with the GO term "neurogenesis" (GO:0048699) are also annotated with the parent of this term, which is "nervous system development" (GO:0007399), but not the other way around. Parent-child relationships between GO terms are arranged in the form of directed acyclic graphs.

1.4.2 'Threshold-based' gene set analysis

Many of the earliest methodologies to carry out GSA were Over-Representation Analysis (ORA) techniques that attempted to identify 'enrichment' of biological themes within lists of 'interesting' genes that had been derived by analysis of microarray data ('genelists'); for example, a list of genes found to exhibit significantly different levels of expression across two experimental conditions, as assessed by a t-test. A considerable number of such tools are now available, and have been reviewed by Khatri (Khatri and Draghici 2005), Huang (Huang da et al. 2009) and Rivals (Rivals et al. 2007). Typically, these methods first identify the number of genes annotated to a particular biological theme that are present in the genelist from an experiment, and then use statistical tests to assess if this observed number is significantly greater than what might be expected by chance. Statistical models commonly used by these methods include the hypergeometric distribution, Fisher's exact test, the χ-squared test and the binomial distribution (Huang da et al. 2009; Khatri and Draghici 2005; Rivals et al. 2007).

ORA-based methods for GSA have been referred to as 'threshold-based', to indicate that these methods test genelists that have been derived using some statistical threshold (for example, using a p-value cut-off of <0.05 after testing all genes for differential expression). This is the key aspect that differentiates ORA-based GSA methods from the category of GSA methods that is described below.

1.4.3 'Threshold-free' gene set analysis

This category of GSA methods does not require selection of genelists; rather many of them require as input a list of all genes considered in an experiment (i.e. all genes represented on a microarray) ranked according to their adherence to some pre-defined pattern of expression across all samples. As there is no implementation of a statistical cut-off to identify genelists for further investigation, these methods have been termed 'threshold-free'.

A particular issue regarding the use of genelists by the threshold-based methods described in the previous section is that there is bias for selection of genes that show the greatest levels of differential expression (for example, in terms of fold-change) into genelists. Such a strategy may fail to detect structure within the dataset that could be of interest to a researcher: for example, if a significant proportion of genes associated with a particular pathway show consistent changes of expression levels across experimental conditions, this pathway is likely to be of interest to a researcher. However if these changes are relatively small in magnitude, it is likely that these genes may not be selected into genelists and detection of this pathway may be missed altogether (Ben-Shaul et al. 2005; Breitling et al. 2004; Huang da et al. 2009; Nam and Kim 2008; Subramanian et al. 2005). Indeed, for small and noisy datasets, few or no genes may exceed threshold levels of significance. Another issue is the arbitrary nature of setting thresholds for selection of genelists – changing threshold values could lead to different results from ORA-based GSA methods (Pan et al. 2005). Threshold-free methods attempt to mitigate these issues by doing away with the need for selection of a genelist.

Many such methods are now available (Nam and Kim 2008), and one of the most popular of these is Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005). This methodology can be summarized as follows: all genes on an array are first ranked according to their differential expression across two sample classes using some metric. The positions of genes associated with any one gene set are then identified in this ranking. To test whether the genes for this gene set are enriched toward the top or bottom of the ranked list, an Enrichment Score (ES) is calculated, which is equivalent to a weighted Kolmogorov-Smirnov-like statistic. The significance of the ES value is estimated by comparison with a null distribution of ES values which is calculated by permuting sample class labels and recalculating ES values for that gene set for each permutation.

1.5 Integration of microarray data from different experiments

Microarrays have proven to be highly popular tools for biological research, and this is reflected in the development of publicly available databases containing raw data derived from hundreds of microarray based experiments such as the Gene Expression Omnibus (GEO) (Edgar et al. 2002) and ArrayExpress (Parkinson et al. 2007). The previous section described how further biological insight could be obtained by the integration of the results of a microarray-based experiment with other sources of information, such as biological annotation. The availability of data from hundreds of microarray-based experiments then allows for the prospect of another level of integration: that of data from different microarray experiments. Such integration could provide opportunities for improved sensitivity and validation of the results of microarray experiments, as well as deeper biological insight than may be achieved through the analysis of a single microarray dataset in isolation.

1.5.1 The need for integration of microarray data

A significant issue regarding the results of any single microarray based experiment considered in isolation is that microarrays “sacrifice specificity for scale” (Troyanskaya 2005): the expression data derived from any experiment involves that of several thousands of genes, measured over a relatively small number of samples. The sophisticated statistical methodologies used to analyse microarray data offer limited control of noise and technical variation, often at the cost of sensitivity of detection. Furthermore, validation of the results derived from any microarray based experiment is required to be carried out using experimental procedures such as RT-PCR and Northern blotting.

1.5.1.1. Validation of results of microarray based experiments

Validation of the results derived from statistical analyses of microarray data (for example, a list of differentially expressed genes) using experimental procedures is expensive and resource-intensive (Kim and Park 2004). Furthermore, such validation is typically carried out only on a non-randomly selected (and thus potentially biased) subset of these results (Moreau et al. 2003).

The integration of the results of any particular microarray based experiment with those of other biologically similar experiments could provide a resource-efficient and objective validation of these results. The rationale for such a strategy is that it could allow for control of laboratory and platform-specific effects. For example, if a group of genes is found to be differentially expressed in each of several similar experiments which have been carried out in different laboratories, with different experimental protocols and on different platforms, then this provides strong evidence of the biological relevance of these genes and of the reliability of these results (Aggarwal et al. 2006; de Magalhaes et al. 2009; Hu et al. 2005; Hwang et al. 2004; Keegan et al. 2007; Moreau et al. 2003; Schlicht et al. 2004; Warnat et al. 2005; Xu et al. 2005; Zhou and Gibson 2004).

1.5.1.2 Increasing sample size to achieve greater sensitivity

The key feature of statistical techniques that are used to identify relevant genes within a microarray dataset is to distinguish between 'true' biological variations and undesirable technical variations. However, due to the high levels of noise inherent to microarray datasets, the power to detect genes that exhibit changes in expression that are biologically relevant but low in magnitude is diminished. The statistical power of these analyses can be increased by increasing the sample size of an experiment; however, this is also limited by the costs of running more arrays as well as availability of samples. Integration of several different analogous microarray datasets could increase the power to detect differentially expressed genes by increasing the sample size for an experiment

(Choi et al. 2004; Grutzmann et al. 2005; Hamid et al. 2008; Hu et al. 2005; Kim and Park 2004; Mulligan et al. 2006; Wang et al. 2004).

1.5.2 Methodologies and experimental strategies for integration of microarray data

1.5.2.1 Meta-analysis

'Meta-analysis' refers to the use of statistical techniques to combine the results of several different experiments. The first such meta-analysis of microarray data was carried out by Rhodes et al, who reanalyzed four prostate cancer datasets to determine genes that were differentially expressed in all the datasets (Rhodes et al. 2002). Meta-analyses have subsequently been carried out successfully in many different studies, where they have been shown to provide significant improvements in terms of the reliability of results as well as sensitivity of statistical tests, as compared to analysis of single datasets in isolation. These include studies of gastric cancer (Aggarwal et al. 2006), pancreatic cancer (Grutzmann et al. 2005), breast cancer (Smith et al. 2008), lung cancer (Parmigiani et al. 2004), alcohol consumption (Mulligan et al. 2006), and *Drosophila* circadian rhythms (Keegan et al. 2007).

While meta-analytical techniques combine the results of analyses of different microarray datasets carried out separately, many studies have involved combining of the datasets themselves followed by a single analysis of the combined dataset (Borozan et al. 2008; Choi et al. 2003; Hu et al. 2005; Stevens and Doerge 2005; Wang et al. 2004; Warnat et al. 2005; Xu et al. 2005). While these studies have also described significant advantages as compared to analyses of single datasets in terms of reliable results, their particular strength is the increased power to detect relevant genes due to the considerable increases in sample size.

1.5.2.2 Cross species integration of microarray data

The development of databases such as Homologene (Wheeler et al. 2008), Resourcerer (Tsai et al. 2001) and Inparanoid (O'Brien et al. 2005), which store relationships between homologous genes across a wide range of species in an electronic format that is accessible and useable, has allowed for the performance of cross-species integration of microarray data.

The principle that core biological networks and pathways are evolutionarily conserved across species is the basis for the use of model organisms for the study of human diseases (for example, mouse models of cancer). On this same basis, cross-species integration of microarray data could be a powerful tool for validation of microarray data (for example, if similar sets of genes are found to be relevant within similar microarray experiments carried out on diverse species, it unlikely that these genes were selected due to chance or technical effects), as well as help understand evolution of these processes (Lee et al. 2005; McCarroll et al. 2004; Zhou and Gibson 2004).

Cross-species integration of microarray data have successfully been carried out in many studies, such as those of aging (de Magalhaes et al. 2009; McCarroll et al. 2004; Wennmalm et al. 2005), liver cancer (Fang et al. 2005; Lee et al. 2004), lung cancer (Sweet-Cordero et al. 2005), breast cancer (Chan et al. 2005), prostate cancer (Ellwood-Yen et al. 2003; Schlicht et al. 2004) and COPD (DeMeo et al. 2006).

1.5.2.3 Experimental integration of microarray data

Integrative analysis of microarray data has not only been used for the purposes of validation or increasing sample size: often, such data integration may be an exploratory (hypothesis-generating) or a confirmatory (hypothesis-driven) experiment in itself:

Chang et al (Chang et al. 2004) used fibroblast gene expression profiles to derive a set of genes representing wound healing and applied it to several different cancer datasets;

their subsequent findings supported their initial hypothesis of a link between wound healing and cancer.

Rhodes et al (Rhodes et al. 2004) carried out a large scale meta-analysis of ~40 microarray based studies of cancer to derive a common transcriptional profile for neoplastic transformation that was shared across a wide range of cancer types, regardless of cell of origin.

Using an approach termed 'comparative functional genomics', Lee et al integrated data from microarray based experiments of human hepatocellular carcinoma with those of several mouse models of the disease (Lee et al. 2004); using unsupervised class discovery techniques (such as those described in Section 1.3.2), they were able to identify 'best-fit' mouse models for the disease.

Sweet-Cordero et al integrated human and mouse microarray data (Sweet-Cordero et al. 2005) using GSEA and showed firstly that the KrasLA mouse model could successfully represent only human lung adenocarcinoma (as opposed to other lung cancers), and secondly that a link between KRAS2 mutations and human lung adenocarcinomas could only be established by integration with the mouse model data.

1.6 Notes for readers

1.6.1 Thesis scope and structure

The work described in this thesis has been divided into two parts: A and B. Part A comprises of Chapters 2-5, which describe research involving cross-platform and cross-species integration of microarray data using lists of differentially expressed genes. Part B comprises of Chapters 6-8, which describe the development and implementation of Gene Set Discovery (GSD), a novel methodology enabling performance of theme-based functional analysis of microarray data in an unsupervised fashion.

All research described in this thesis involves data derived from experiments carried out on the Affymetrix commercial microarray platform. However, all findings can, in principle, be applied to data derived from any other microarray platform.

1.6.2 Aims

Part A of this thesis (Chapter 2-5) describes explorations of the concept of integrating microarray datasets using solely their genelists. This prospect is particularly of interest because it is resource-efficient enough to allow large scale comparisons of many different datasets in an unsupervised fashion, which could in theory lead to the discovery of unexpected links between experiments. However, many studies have shown low levels of similarity between genelists even from very similar experiments. The main aim of Part A of this thesis is thus to explore whether, such a strategy could still be of use to researchers. This was carried out by carrying out comparisons between a large set of genelists, including cross-platform and cross-species comparisons. A secondary aim was to observe whether comparisons between genelists from species that are evolutionarily distant could yield biologically meaningful results, and thus to assess the utility of such an approach.

Part B of this thesis describes the development of the GSD, a novel methodology which could allow for GSA of microarray datasets to be carried out in an unsupervised manner. The aims of these explorations were to firstly develop an appropriate metric to quantify the information content for any set of genes using simulated datasets with known information types and levels. Secondly, this approach would require to be validated on real-world microarray datasets to assess whether the method yielded biologically meaningful and useful results. Other aims included developing a metric that could allow for extraction of informative genes from within gene sets, and a visualization scheme that could present the diverse types of information involved to the user in an intuitive integrated format.

1.6.3 Terminology

The terms 'microarray', 'array' and 'chip' have been used interchangeably throughout this thesis. The term 'GeneChip' refers specifically to Affymetrix microarray platforms.

The term 'genelist' has been used to denote a set of genes that is derived experimentally through statistical analysis of microarray data, for example after testing for differential expression of genes.

The term 'gene set' has been used to indicate a set of genes that can be derived from gene annotation databases, such as Gene Ontology (GO) (Ashburner et al. 2000), and typically represent biological themes such as pathways. These gene sets have been created using prior knowledge of the biological functions of these genes. The term 'gene set analysis (GSA)' has been used to refer to any methodology involving the study of biological themes, represented as gene sets, within microarray datasets. This included all threshold-based and threshold-free methods (see Section 1.4). The term ‘gene set enrichment analysis (GSEA)’ refers to a particular method to carry out threshold-free GSA (Subramanian et al. 2005). ‘Gene set discovery (GSD)’ is an unsupervised threshold-free GSA technique, the development and implementation of which is described in Part B of this thesis.

Part A:

Integration of microarray based experiments using lists of differentially expressed genes

Chapter 2: Strategies for large-scale integration of microarray data

2.1 Introduction

2.1.1 Unsupervised integration of microarray datasets

Section 1.5 introduced the concept of integration of microarray data, that is comparison of microarray datasets from several different experiments, which could have been carried out in different laboratories, on different microarray platforms and on different species. Many different examples were cited to illustrate the utility of such an approach, both as a tool for validation and increasing the statistical power of analyses, as well as to provide deeper biological insight that may be achieved solely by the analysis of a single experiment. However, as Finocchiaro et al note, the majority of such integrative analyses select the datasets to be integrated using a 'supervised' approach: researchers often select the datasets that they wish to compare based on prior hypotheses that there is some common underlying biology between them (Finocchiaro et al. 2005). This is certainly the case for all the examples cited in Section 1.5. Indeed, careful selection of which datasets could be integrated is considered to be a pre-requisite for such approaches (Ramasamy et al. 2008).

However, as with most supervised methods, the possibilities for novel discovery may be somewhat limited. A more efficient utilization of the large amounts of information contained within a microarray dataset might be an unbiased exploratory analysis involving the comparison of the dataset with a diverse, unselected collection of datasets created without any prior assumption as to biological links between them. For example, having carried out a microarray-based experiment, a researcher might ask the question, "Which other experiments is my experiment similar to?" Carrying out an exploratory analysis such as described above could then provide the researcher with an unbiased way of finding other datasets not only with *expected* similarity (for example another very similar experiment), but crucially, this could also lead to *unexpected* links being found. The latter possibility is of particular significance as this could potentially lead to new discoveries and biological insight.

2.1.2 Using lists of differentially expressed genes as representatives of microarray experiments

As described in Section 1.1, a microarray-based experiment typically yields data from a few to several hundred arrays, each of which may contain expression data for thousands of genes, and analysis of such 'raw' microarray datasets is resource intensive. While most integrative studies of microarray data involve large-scale re-processing of raw experimental datasets, the numbers of datasets reanalyzed have been limited, since each of the datasets is pre-selected, usually on the basis of some hypothesized biological similarity.

However, an unsupervised meta-analysis of an experiment with, for example, all datasets available on a public repository like GEO is likely to involve a very significant computational workload. For example, even if we were to concentrate on solely the 21434 Affymetrix hgu133a samples contained in GEO (as of 10th February 2009), with each array containing 22283 probesets, this would involve mining of 477,613,822 data points. The computational workload is also paralleled by a significant requirement for manual intervention during the process of microarray data analysis, which presents a particularly significant problem for scalability.

These considerations are a particularly acute problem for exploratory analyses, where it is expected that the majority of comparisons would not yield interesting data. It is then difficult to justify carrying out such a large number of complex and resource-intensive analyses where most results will not be of interest. This then raises the need for a 'quick and easy' exploratory analysis involving a first-pass filtering of potentially interesting links between experiments, with the assumption that once found, these can then be explored in greater detail by analyzing the raw datasets.

One potential solution involves data reduction and summarization: comparisons could then be carried out between summaries of datasets rather than between entire raw datasets. A popular summary of a microarray experiment that could be used for this purpose is the list of 'interesting' genes created during the analysis of microarray data, usually after tests for differential expression (See Section 1.3.1.1). Datasets containing potentially millions of gene expression values could then be reduced to a few hundred gene identifiers. This is, in fact, the workflow adopted by several groups (Cahan et al. 2005; Finocchiaro et al. 2005; Newman and Weiner 2005; Yi et al. 2007), who have created databases of experimentally-derived genelists for the purpose of comparison. The basis of this workflow is the argument that similarity between two genelists could reflect similarity between the corresponding experiments, in turn reflecting some shared biology (Rubin 2005).

However, the use of genelists to compare microarrays is a controversial prospect. Several studies have shown that even similar experiments exhibit little overlap between genelists (Cahan et al. 2007; Cheadle et al. 2007; Ein-Dor et al. 2005; Jeffery et al. 2006; Manoli et al. 2006; Tan et al. 2003). These have been attributed to various factors, such as differences in laboratories, experimental protocols, microarray platforms and data analysis strategies. Studies have shown that similarity in the results of similar microarray based experiments can be induced by standardization of experimental protocols and data analysis algorithms (Bammler et al. 2005; Irizarry et al. 2005; Larkin et al. 2005). This then creates doubts regarding the reliability of using solely genelists to compare microarray experiments; the genelists archived in the databases cited above represent a diverse set of experiments carried out in a wide range of different laboratories, on different platforms and species, and created using different statistical methodologies.

The primary aim of the work described in Part A of this thesis was thus to explore if the comparison of microarray experiments using genelists could be a feasible and reliable method to find links between disparate experiments in an unsupervised fashion, in light of the issues described above. It was intended to achieve this by carrying out comparisons between a diverse set of genelists derived from a number of different experiments (examining different biological themes) carried out in different laboratories, using different statistical methodologies. Examination of the results of these comparisons could then be carried out using both global (for example, by carrying out unsupervised hierarchical clustering [see Section 1.3.2.1] of genelists using a standardized measure of similarity, and examining the clusters for any dominant biological themes), and local (for example, focused examination of experiments found to have significantly similar genelists to detect shared underlying biology) strategies to assess whether these results were biologically meaningful and of use to researchers. Secondary aims included assessing how far across evolutionary distances could cross-species comparisons be performed while still deriving biologically meaningful results.

This chapter, in particular, details the development of strategies to carry out comparisons between lists of differentially expressed genes. Two major issues are addressed, the first being the translation of genelists across chips and species: for example, how could a list of human genes and a list of *C. elegans* genes be made comparable? The second is that of assessing the statistical significance of the overlap between any two genelists, which would enable the detection of 'real' biological similarities as opposed to overlaps that could have been caused just by chance.

2.2 Explorations and Results

2.2.1 Conversion of gene identifiers across array-types and species

The following explorations will use expression microarrays from Affymetrix to illustrate the principles and concepts, although similar issues would be faced with arrays produced by other companies. Affymetrix currently manufactures expression arrays for a wide number of species, ranging from humans and popular model animals like mice, rats and zebrafish, to plants and prokaryotes (see Table 2.1). Integration of lists of differentially expressed genes, especially across array-types and species, would require ensuring that the genelists are comparable. As explained in Section 1.2, the basic units of gene expression data within an Affymetrix array are the probesets, each of which has a unique ID label, and it is (usually) these identifiers that comprise lists of differentially expressed genes.

The first question addressed was whether these probeset IDs are shared across the different types of Affymetrix arrays. For this purpose, lists representing all probeset IDs for several chips were created and compared. Table 2.2 represents a selection of some of the most popular Affymetrix arrays for several different species and the number of probeset IDs that are shared between them. As can be seen, there is no overlap of probeset IDs across different species. Even within a species, there may be little or no overlap of probeset IDs between two different array-types, such as the human hgu133a and hgu95a arrays: these differ because they represent different generations of Affymetrix human arrays, and were created using different UniGene build versions (Affymetrix 2001). Thus, most genelists from microarray experiments carried out on different array-types (in particular, from different species) and comprising solely of probeset IDs are not directly comparable: any attempt to do so would result in no overlap between genelists. This then raises the need for some conversion of probeset IDs prior to cross-chip genelist comparisons.

(1) Species	(2) Affymetrix GeneChip Expression Analysis Arrays	(3) Bioconductor name	(4) Number of samples on GEO	(5) Number of series on GEO
Homo sapiens	Human Genome U133 Plus 2.0 Array	hgu133plus2	21434	743
	Human Genome U133 Set Array - A	hgu133a	19982	639
	Human Genome U95 Set Array -A	hgu95a/ av2	5264	275
	Human Genome U133 Set Array - B	hgu133b	4420	109
	Human Genome U133A 2.0 Array	hgu133a2	2092	109
	Human HG-Focus Target Array	hgfocus	1935	47
Mus musculus	Mouse Genome 430 2.0 Array	mouse4302	10388	798
	Murine Genome U74 Version 2 Set MG-U74A	mgu74av2	5490	435
	Mouse Expression Array 430A and Mouse Genome 430A 2.0 Array	moe430a/a2	4797	378
	Mouse Expression Array 430B	moe430b	957	77
Rattus norvegicus	Rat Genome 230 2.0 Array	rat2302	3107	168
	Rat Genome U34 Array Set RG-U34A	rgu34a	3047	142
	Rat Expression Set 230 Array RAE230A	rae230a	2184	102
	Rat Genome U34 Array Set RG-U34B	rgu34b	456	10
Drosophila melanogaster	Drosophila Genome Array	drosgenome1	1088	93
	Drosophila Genome 2.0 Array	drosophila2	736	54
Arabidopsis thaliana	Arabidopsis ATH1 Genome Array	ath1121501	4325	330
	Arabidopsis Genome Array	ag	134	19
Yeast spp (S. cerevisiae; S.pombe)	Yeast Genome S98 Array YG-S98	ygs98	1489	111
	Yeast Genome 2.0 Array	yeast2	417	25
Other eukaryotes	Soybean Genome Array	soybean	3029	17
	C.elegans Genome Array	celegans	452	25
	Zebrafish Genome Array	zebrafish	423	37
	Chicken Genome Array	chicken	338	23
	Wheat Genome Array	wheat	315	11
	Rhesus Macaque Genome Array	rhesus	300	23
	Maize Genome Array	maize	249	16
	Rice Genome Array	rice	248	21
	Porcine Genome Array	porcine	237	12
	Xenopus laevis Genome Array	xenopuslaevis	198	20
	Bovine Genome Array	bovine	197	16
	Barley Genome Array	barley1	184	13
Escherischia coli	E. coli Antisense Genome Array	ecoliasv2	672	38
	E. coli Genome 2.0 Array	ecoli2	204	26
Other prokaryotes	Pseudomonas aeruginosa Array	paeg1	287	33
	S. aureus Genome Array	saureus	153	17

Table 2.1 A selection of Affymetrix gene expression analysis arrays (in previous page). The shortened names in column 3 are those used in the Bioconductor annotation packages for the respective arrays (arrays will be referred to by these names for the rest of the thesis). Column 4 represents the number of arrays for which data is available on the Gene Expression Omnibus, while Column 5 displays the number of series of chips (where a series usually represents a set of arrays from the same experiment), as of 10th February 2009.

Translations of genelists can be made possible by utilizing the biological annotation that is associated with each probeset. Such annotation is available from sources such as the Affymetrix NetAffx annotation files (Liu et al. 2003) and the Bioconductor (Gentleman et al. 2004) annotation packages, which incorporate various annotation sources including the former (see Material and Methods). For example, while the hgu133a and hgu95a arrays contain distinct sets of probeset IDs, these can refer to the same genes. Thus, the hgu133a 222152_at and the hgu95a 37569_at probesets both represent the PDCD6 programmed cell death 6 gene. Thus genelists from experiments carried out on two different array-types (but representing the same species) can be made comparable by converting probeset IDs into species-specific gene identifiers.

While several types of annotation are available for each probeset (for example, Unigene IDs, RefSeq IDs, Entrez Gene IDs and gene symbols), Entrez Gene IDs (EGIDs) were selected as the biological annotation of choice. There are two main reasons for this. First is that EGIDs (and gene symbols) are probably the most biologically-intuitive units of annotation, being gene-centric in focus, while the others are sequence-centric. The second reason is technical: more probesets are annotated with EGIDs than any other source (see Material and Methods), and probesets that are not annotated with an EGID also have no other annotation (data not shown). As a result, conversion of probeset IDs into EGIDs results in the least loss of information.

	hgu133a	hgu133plus2	hgu95a	mouse4302	moe430a	mgu74av2	rat2302	rae230a	rgu34a	xenopuslaevis	zebrafish	drosgenome1	drosophila2	celegans	ag	ath1121501
hgu133a	100	100	1.2	0	0	0	0	0	0	0	0	0	0	0	0	0
hgu133plus2	41	100	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
hgu95a	2.1	2.1	100	0	0	0	0	0	0	0	0	0	0	0	0	0
mouse4302	0	0	0	100	50.2	0	0	0	0	0	0	0	0	0	0	0
moe430a	0	0	0	100	100	0	0	0	0	0	0	0	0	0	0	0
mgu74av2	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
rat2302	0	0	0	0	0	0	100	51.1	0	0	0	0	0	0	0	0
rae230a	0	0	0	0	0	0	100	100	0	0	0	0	0	0	0	0
rgu34a	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
xenopuslaevis	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
zebrafish	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
drosgenome1	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
drosophila2	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
celegans	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
ag	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
ath1121501	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Table 2.2 Probeset IDs that are shared across various Affymetrix array-types. Numbers represent the percentage of (non-control) probeset IDs of row-wise arrays that are also found in column-wise arrays. Grey cells denote comparisons between arrays from the same species.

However, conversion of probeset IDs to EGIDs alone would be insufficient translation for comparisons between genelists from experiments carried out on different organisms. EGIDs are species-specific and different species are each annotated with a distinct set of EGIDs. For example, the human GAPDH gene has the EGID of 2597, while the mouse homolog (Gapdh) has the EGID of 14433. There is thus a need for the conversion of the EGIDs of genes from one of the species into the EGIDs of the corresponding homologous genes of the other species. This could be carried out by using the homologous relationships between genes across different species that are recorded electronically in databases such as Homologene (Wheeler et al. 2008), Resourcerer (Tsai et al. 2001), and Inparanoid (O'Brien et al. 2005) (See Materials and Methods).

Direct comparison of lists of probeset IDs is sometimes possible, such as when comparing genelists from experiments carried out on the same array-type, or in those instances when all probesets of one chip are present in the other (for example the hgu133a and hgu133plus2 arrays, where the former is a subset of the latter). One complexity that the use of probesets introduce is that such analyses would be vulnerable to biases caused by genes which are represented by more than one probeset. For example, consider a gene that is represented in an array by 3 probesets. If this gene is differentially expressed in two experiments, we might expect all three probesets to be present in the genelists of both these experiments. Thus, in a comparison of these two genelists, the contribution of this gene to the overlap size would be three units (probesets), rather than the desired contribution of one unit (gene/EGID).

Figure 2.1 represents an exploration of all genes (represented as EGIDs) present on several arrays, and the number of probesets they are annotated to.

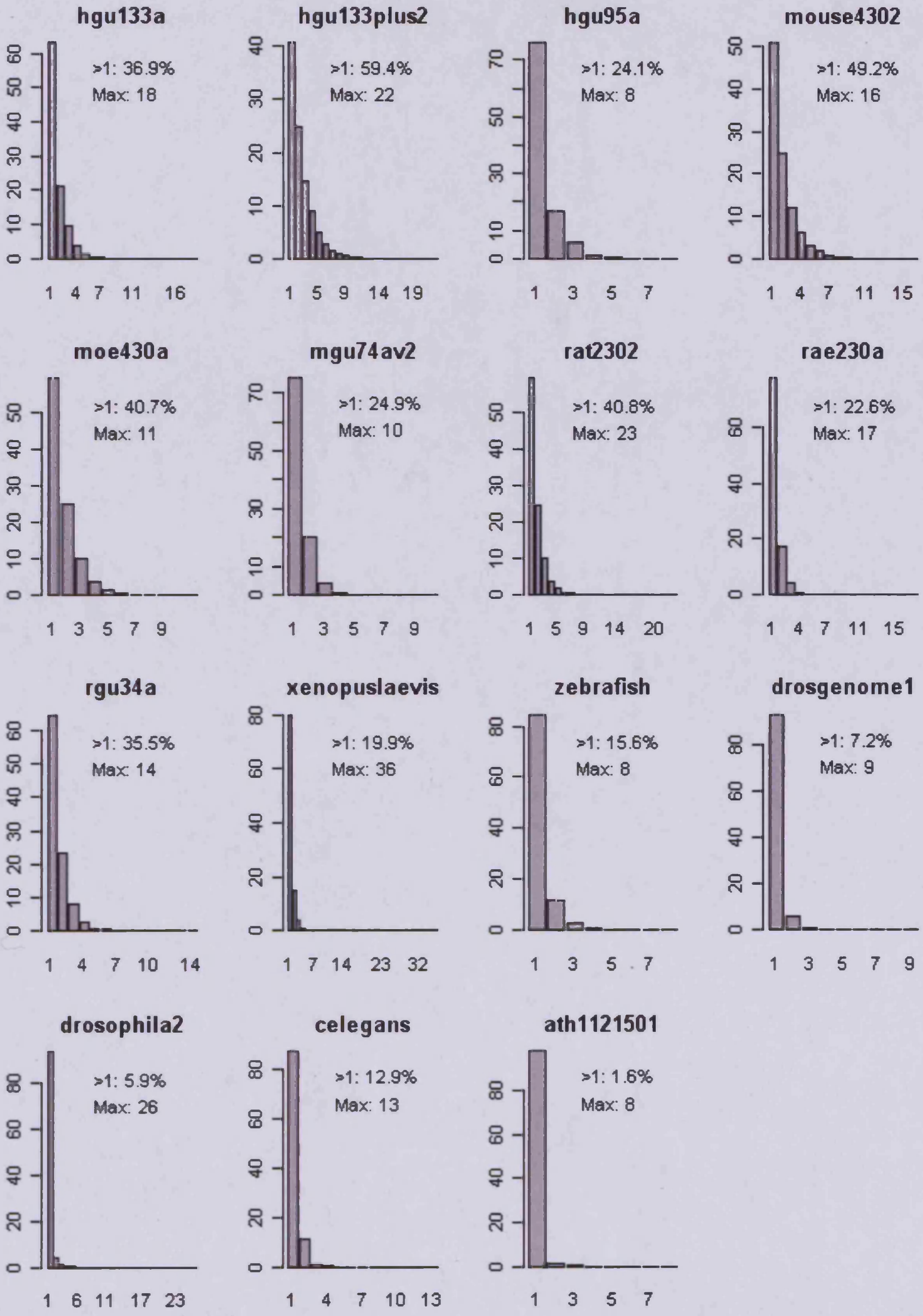

hgu133a
>1: 36.9%
Max: 18
0 10 20 30 40 50 60
1 4 7 11 16
hgu133plus2
>1: 59.4%
Max: 22
0 10 20 30 40
1 5 9 14 19
hgu95a
>1: 24.1%
Max: 8
0 10 30 50 70
1 3 5 7
mouse4302
>1: 49.2%
Max: 16
0 10 20 30 40 50
1 4 7 11 15
moe430a
>1: 40.7%
Max: 11
0 10 20 30 40 50
1 3 5 7 9
mgu74av2
>1: 24.9%
Max: 10
0 10 30 50 70
1 3 5 7 9
rat2302
>1: 40.8%
Max: 23
0 10 20 30 40 50
1 5 9 14 20
rae230a
>1: 22.6%
Max: 17
0 20 40 60
1 4 7 11 15
rgu34a
>1: 35.5%
Max: 14
0 10 20 30 40 50 60
1 4 7 10 14
xenopuslaevis
>1: 19.9%
Max: 36
0 20 40 60 80
1 7 14 23 32
zebrafish
>1: 15.6%
Max: 8
0 20 40 60 80
1 3 5 7
drosgenome1
>1: 7.2%
Max: 9
0 20 40 60 80
1 3 5 7 9
drosophila2
>1: 5.9%
Max: 26
0 20 40 60 80
1 6 11 17 23
celegans
>1: 12.9%
Max: 13
0 20 40 60 80
1 4 7 10 13
ath1121501
>1: 1.6%
Max: 8
0 20 40 60 80
1 3 5 7

Figure 2.1 Distribution of Entrez Gene IDs (EGIDs) according to the number of probesets annotated to them. The x-axis denotes the number of probesets annotated to an EGID. The y-axis denotes the number of EGIDs in each category as percentages of total EGIDs on an array. In the legend, the upper number is the percentage of EGIDs having >1 probesets annotated to them; the lower number is the maximum number of probesets annotated to one EGID in that array.

As can be observed, as many as 59.4% of all EGIDs on an array can be linked to more than one probeset (as seen in hgu133plus2), and as many as 36 probesets can be annotated with the same EGID (as seen in xenopuslaevis).

These figures suggest that biases caused by the presence of more than one probeset representing a single gene are likely to occur. For this reason, it was decided that the strategy for comparing genelists from experiments carried out on the same array-type would require an initial step of converting lists of probeset IDs into unique lists of EGIDs to which those probesets are annotated to.

2.2.2 The need for assessment of significance of overlaps between genelists

A very simple metric to assess biological similarities between genelists would be the number of shared genes. For example, if genelist A and genelist B share x number of genes, and genelist A and genelist C share y number of genes, and $x>y$, then we might infer that genelist A shows greater biological similarity with genelist B than with genelist C. However, the size of this overlap is likely to be affected not only by real biological similarities between the genelists, but also by systematic effects that depend on the lengths of the genelists being compared, and the size of the gene universe. The gene universe here refers to the set of all genes from which the genelists were selected.

For example, if the genelists being compared are derived from experiments carried out on the same array-type, the gene universe for this comparison might be all the genes present on that array. These systematic effects exerted by genelist length and gene universe size on the size of overlaps between genelists were explored as follows.

The effect of genelist size would be exerted such that, when the universe size is constant, the larger are the genelists being compared, the larger we would expect that overlap to be. Consider a comparison of genelists A and B, where genelist B consists of 10% of genes on the entire array. Then, we would expect it to contain, on average, 10% of the genes in genelist A by chance alone. If genelist B were to have 20% of genes on the array, we would expect it contain, on average, 20% of the genes in genelist A and so on.

To explore this effect, the following simulation was carried out: a set of 20,000 arbitrary and unique identifiers was created, representing all the genes present on an imaginary array. From this, a total of 1000 identifiers were selected randomly and without replacement, representing a reference genelist of 1000 genes. Similarly, a series of 'test genelists', of sizes ranging from 10 to 20,000 identifiers were also selected. Each of these test genelists was compared to the reference genelist and the size of overlap for each comparison was recorded, and these have been displayed as grey dots in Figure 2.2(a).

The positive linear relationship between the overlap size and length of test genelists is clearly observed, and the observed overlap sizes vary around the expected overlap sizes (represented by the red line). This expected overlap size is calculated by the formula $\frac{(L1 * L2)}{N}$ where $L1$ and $L2$ are the lengths of the genelists being compared while N represents the size of the gene universe. Thus, we find that in a gene universe of fixed size, the size of overlap between any two genelists is directly proportional to the lengths of the genelists.

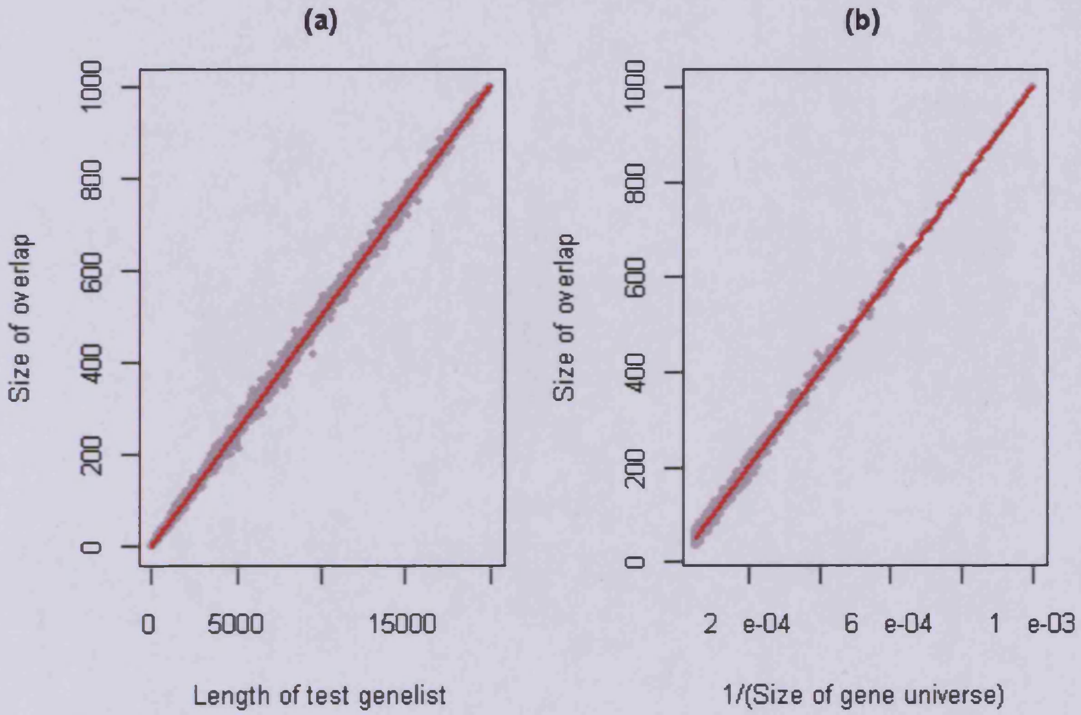


Figure 2.2 - Systematic effects of (a) genelist length and (b) gene universe size on the size of overlap between genelists. Figures are based on simulations carried out using a synthetic array represented by a set of unique and arbitrary identifiers. (a) Overlap sizes observed during comparisons of 'genelists' (created by random selection of identifiers from a synthetic array without replacement) of various sizes from a single synthetic array of constant size, plotted against the lengths of those genelists. (b) Overlap sizes observed when 'genelists' of constant size (selected randomly and without replacement) from synthetic arrays of various sizes are compared, plotted against the inverse of the gene universe size (i.e. the number of genes present in the synthetic arrays). The grey points represent the observed overlap sizes, while the red lines represent the expected overlap sizes.

The formula to calculate the expected overlap size also suggests that the size of the gene universe is inversely proportional to the size of the expected overlap, i.e. the larger the size of the universe, the smaller would be the expected overlap between any two genelists, and vice versa. To explore this, sets of unique and arbitrary identifiers, representing a series of imaginary arrays of sizes ranging from 1000 to 20,000 genes was created. From each of these arrays, two 'genelists' of 1000 genes (identifiers) each were randomly selected (without replacement) and compared. The number of genes in each overlap was recorded and is displayed as the grey dots in Figure 2.2(b).

A positive linear relationship is observed between the overlap sizes and the inverse of the gene universe sizes, and again, the observed overlap sizes vary around the expected overlap sizes. Thus, it is found that during comparisons of genelists that are of constant lengths, but are selected from gene universes of varying sizes, the overlap size is inversely proportional to the size of the gene universe.

These findings suggest that the size of overlap between genelists alone would not be a suitable indicator of the biological relatedness between any two genelists, because it would be not be possible to distinguish biological effects from the systematic effects of genelist and universe sizes.

For the purposes of genelist comparison, the systematic effects of universe size can be controlled for by keeping the gene universe size constant. There are several ways to achieve this: for genelists from the same array-type, the universe size for all comparisons could be all the genes on the chip; for comparisons between genelists from different array-types, the gene universe could be only those genes that are present on both array-types (and the genelists would be filtered to reflect this); for cross-species comparisons, the gene universe could consist only of genes for which homologues are present on both arrays (and the genelists would be filtered accordingly). However genelist sizes vary greatly and these steps would not address this.

2.2.3 Selection of metrics for the assessment of the significance of overlaps between genelists

The size of the overlap between any two genelists will in part depend on the lengths of genelists involved. This therefore raises a need for a metric to assess the significance of the observed overlap, that is a metric that indicates if an observed overlap is any greater than would be expected by chance alone (and preferably to indicate by how much). Deviations of the observed overlap from the expected overlap can then be argued to have occurred due to real biological similarities between genelists with the magnitude of the deviation giving some indication of the degree of biological similarity (or dissimilarity).

Three metrics were initially chosen as potential candidates to assess the significance of overlaps between genelists. These were fold change, the binary similarity index and the hypergeometric distribution, and were tested as follows. A set of unique and arbitrary identifiers, representing a synthetic array of 20,000 genes, was created. From this, a reference genelist of 1000 genes was created by random selection of genes without replacement. Then, a set of test genelists, varying in length from 100 to 20,000 were created, again by random selection of genes without replacement. Each test genelist was compared to the reference genelist and the overlap size was recorded; all three metrics were calculated for each comparison. Each metric was then assessed for its dependency on genelist length and the results of these analyses are presented below.

2.2.3.1 Fold change

Fold change is a simple and intuitive metric, which is calculated as the observed overlap size divided by the expected overlap size, and provides the magnitude of the observed overlap size relative to the size of the expected overlap. Thus, a fold change value of 1 would indicate that the observed and expected overlap sizes are the same, while a fold change value of 2 would indicate that the observed overlap is twice the expected size of overlap, and so on. In natural space, the scale of fold change values is asymmetrical (for example, a ten-fold increase in an observed value with respect to the expected value yields a fold change value of 10, while a ten-fold decrease would yield a value of 0.1), and for this reason, fold change values are usually converted to log space. Thus a log fold change value of zero indicates that the observed value is equal to the expected value, while positive and negative log fold change values indicate observed values that are greater and lesser than the expected values respectively.

Fold change values were obtained from the simulations described above, and these were transformed using the natural log. Figure 2.3(a) shows the histogram of these log fold change values. As the comparisons were carried out between lists of 'genes' that had been selected in a random fashion, it is expected that the majority of comparisons should show no appreciable difference between the observed and expected values. Indeed, the median of this distribution is a log fold change value of zero.

Figure 2.3(b) shows a Q-Q plot where quantiles of the distribution of log fold change values are plotted against those for a normal distribution. This was created to investigate whether the log fold change values are normally distributed. As can be observed, the distribution of log fold change values tends towards non-normality as the values become higher and lower than the median value of zero. Further light is shed in Figure 2.3(c), where the log fold change values are plotted against the lengths of the test genelists used in each comparison. It is observed that while the values vary around the median of zero, regardless of the length of the test genelists, the variability of log fold change values is greater for shorter test genelists than for longer test genelists.

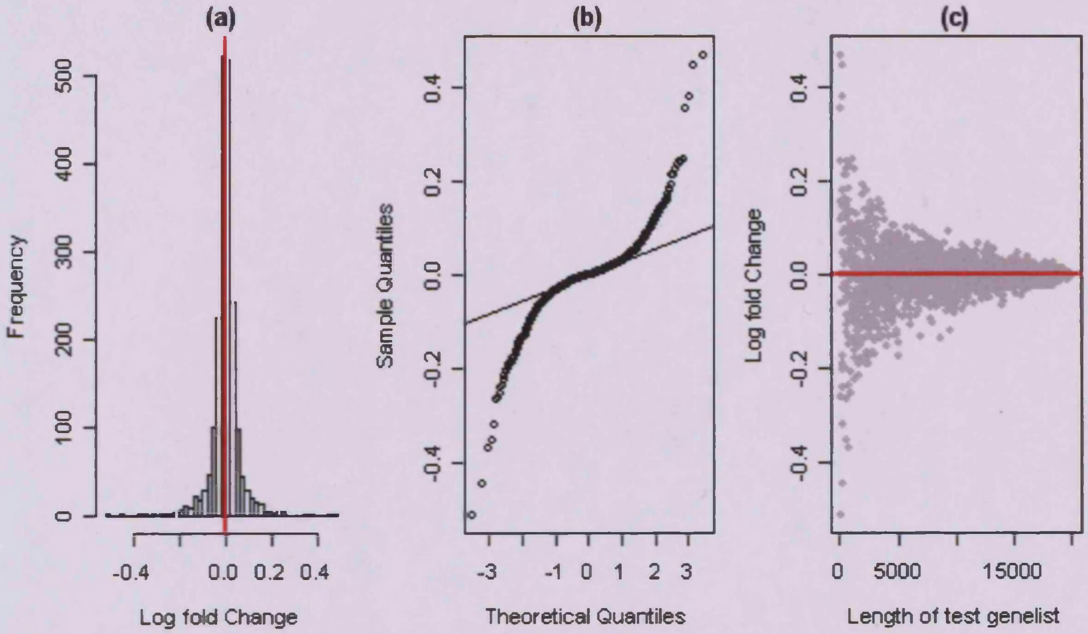


Figure 2.3 Log Fold Change values from simulation experiments. (a) Histogram of log fold change values obtained from comparisons carried out between a reference 'genelist' and a set of test 'genelists' of varying lengths from a synthetic array represented as a set of unique and arbitrary identifiers. Reference and test genelists were selected randomly and without replacement. (b) Q-Q plot for the log fold change values (Y-axis) versus the normal distribution (X-axis). (c) Log fold change values plotted against the length of test genelists. Red lines in all plots indicate the median of the distribution (i.e. zero).

This observed relationship between the variability of log fold change values and the length of test genelists is due to instability at low genelist lengths because of data granularity. For example, consider two short genelists that have an expected overlap value of 5. An observed overlap value of one gene greater than that expected value (i.e. 6) would then yield a log fold change value of 0.18. Now consider two long genelists that are expected to have an overlap size of 500. In this case, an observed overlap size of one gene greater than expected (i.e. 501) would yield a log fold change value of 0.002.

In effect, this property is undesirable as this makes the setting a cut-off log fold change value (for example, 2) to indicate significance problematic: such cut-offs may be too liberal for comparisons between short genelists and too conservative for comparisons between longer genelists. For this reason, log fold change values were not explored further.

2.2.3.2 Binary similarity index

The next metric tested was the binary similarity index, which is calculated as the size of the overlap divided by the total number of unique genes present in at least one of the two genelists. In set theory terms would translate as the size of intersect divided by the size of the union of the two genelists (see Materials and Methods).

Figure 2.4(a) shows the histogram of binary similarity values obtained from the simulations described above, while Figure 2.4(b) shows the Q-Q normality plot for the same and these indicate a skewed and non-normal distribution of values. More importantly, as can be observed in Figure 2.4(c), there is a strong, positive, non-linear relationship between the binary similarity values and the size of the genelists.

Due to this undesirable property, binary similarity indices were also not explored any further. It is noted that Cahan et al (Cahan et al. 2005) use this metric in the LOLA database not as the sole indicator of significance, but as a measure of “concordance”. Values for variance and p-values are provided to interpret the concordance values.

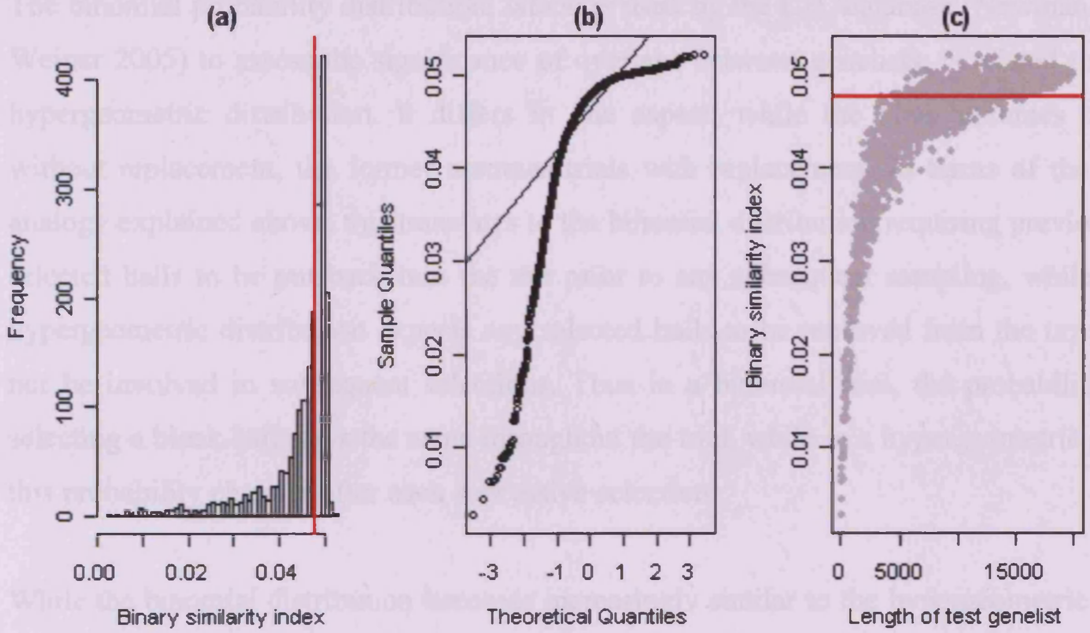


Figure 2.4 Binary similarity values from simulation experiments. (a) Histogram of binary similarity values obtained from comparisons carried out between a reference 'genelist' and a set of test 'genelists' of varying lengths from a synthetic array represented as a set of unique and arbitrary identifiers. Reference and test genelists were selected randomly and without replacement. (b) Q-Q plot for the binary similarity values (Y-axis) versus the normal distribution (X-axis). (c) Binary similarity values plotted against the length of test genelists. Red lines in all plots indicate the median of the distribution.

2.2.3.3 Hypergeometric and Binomial distributions

The final metrics tested were derived from the hypergeometric probability distribution, which can be described as follows: consider an urn (which in our case represents a microarray chip, i.e. the gene universe) filled with balls (genes), some of which are coloured black (genelist A), while the rest are coloured white. From this urn a certain number of balls (genelist B) are selected. The hypergeometric distribution then predicts the probabilities of the number of black balls among those selected (i.e. the overlap between genelists A and B).

The binomial probability distribution, which is used by the L2L database (Newman and Weiner 2005) to assess the significance of overlaps between genelists, is related to the hypergeometric distribution. It differs in one aspect; while the latter assumes trials without replacement, the former assumes trials with replacement. In terms of the urn analogy explained above, this translates to the binomial distribution requiring previously selected balls to be put back into the urn prior to any subsequent sampling, while the hypergeometric distribution expects any selected balls to be removed from the urn and not be involved in subsequent selections. Thus in a binomial trial, the probability of selecting a black ball stays the same throughout the trial, while in a hypergeometric one, this probability changes after each successive selection.

While the binomial distribution becomes increasingly similar to the hypergeometric one as the size of the gene universe increases in relation to the size of the genelists being compared, the hypergeometric distribution is a theoretically better choice for these purposes. This is because sampling with replacement, as in binomial trials, would imply occasions where a gene is present in a genelist more than once; this does not happen, because each gene is required to be unique within a genelist. Thus, the binomial distribution was excluded from further analysis on theoretical grounds. Newman et al, regarding the L2L database, concede that for the purposes of comparing genelists, the hypergeometric distribution is “more accurate”, but did not select this for use in their database as it is “more difficult to calculate” (Newman and Weiner 2005).

Two metrics can be derived from the hypergeometric probability distribution. The first of these is a Z-score, which is the observed overlap that has been standardized, taking into account the lengths of the genelists being compared and the gene universe size. As an ‘effect size’, it is conceptually similar to fold change, but is calculated differently (see Materials and Methods).

An overlap that yields a Z-score of zero indicates an overlap of expected size, while positive and negative Z-scores indicate overlap sizes that are greater and lesser than the expected overlap sizes respectively. The magnitude of Z-score values represents the extent of deviation of the observed overlap sizes from the expected overlap sizes. Hypergeometric Z-scores were calculated for the simulations described above.

Figure 2.5(a) shows the histogram of Z-scores derived from the simulation studies. This distribution is centered on a median Z-score of zero, indicating that on average, the overlap sizes are no different from what is expected by chance alone.

The Q-Q plot in Figure 2.5(b) indicates the Z-score distribution is very similar in shape to the normal Gaussian probability distribution. Also, like a standardized normal distribution, it is centered on a median of zero and has a standard deviation of ~1. One of the properties of a normal distribution is that ~95% of values can be found within the range of the median ± 2 times the standard deviation of the distribution. Indeed, it observed that 1899 of the 2000 calculated Z-scores (i.e. 94.95%) fall within this range.

This is a useful property, and can be exploited to set cut-off values for significance. For example, a cutoff value of the median + 2 times the standard deviation of the distribution would imply selection of Z-scores representing overlap sizes that have only a 2.5% probability of having occurred by chance alone. Also, as can be seen in Figure 2.5(c), there is no observable relationship between Z-scores and the lengths of the test genelists.

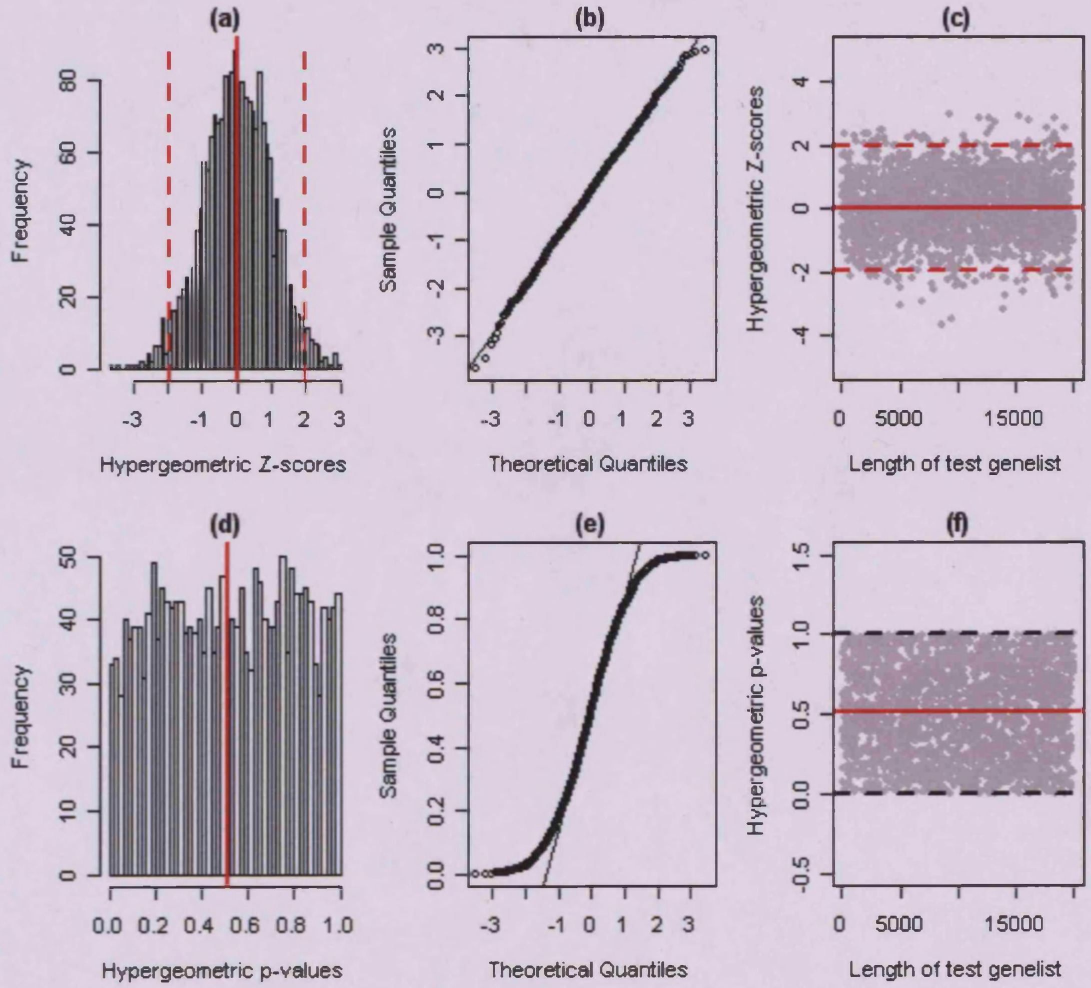


Figure 2.5 Hypergeometric Z-scores and p-values from simulation experiments. (a) and (d) show, respectively, the hypergeometric Z-scores and p-values obtained from comparisons carried out between a reference 'genelist' and a set of test 'genelists' of varying lengths from a synthetic array represented as a set of unique and arbitrary identifiers. Reference and test genelists were selected randomly and without replacement. (b) and (e) represent Q-Q plots for the Z-score and p-value distributions respectively (Y-axes) versus the normal distribution (X-axes). (c) and (f) show the Z-score and p-values respectively, plotted against the lengths of test genelists. Red lines in all plots indicate the medians of the respective distributions. Broken red lines in (a) and (c) represent median ± 2*sd of the Z-score distribution. Broken black lines in (f) represent the theoretical minimum and maximum values of the p-value distribution.

The second metric that can be calculated using the hypergeometric distribution is a p-value, which is a measure of the probability that an observed overlap size could have occurred by chance alone. These were calculated for the above simulations assuming a one-sided test i.e. testing only for how much greater an observed overlap is than the expected size, and not the other way around (see Materials and Methods). The resultant p-value distribution is centred on a median value of ~0.5, and ranges from the minimum and maximum possible p-values of 0 and 1 respectively (Figure 2.5(d)). As expected, setting a p-value cut-off of 0.05 (i.e. the theoretical value which no more than 5% of comparisons should yield lower p-values for, if the assumptions of the distribution are not violated), results in the filtering of 102 of the total 2000 comparisons (i.e. 5.1%). Also, no relationship is observable between the p-values and the lengths of the test genelists (Figure 2.5(f)).

Figure 2.6 illustrates the relationship shared between hypergeometric Z-scores and p-values. Thus comparisons that yield high Z-scores yield low p-values. Theoretically, setting a Z-score cut-off value of 2 (from a Z-score distribution with a median of zero and standard deviation of 1) is equivalent in effect to setting a p-value cut-off of 0.025 (i.e.2.5%), as these parameters would only be exceeded by 2.5% of comparisons between randomly created lists of genes, and the empirically derived distributions appear to be in agreement with this (broken red lines in Figure 2.6 represent these cut-off values).

Thus, because hypergeometric Z-scores and p-values appear not to be influenced by the lengths of genelists being compared, when the universe size is constant, both metrics are candidates for the purpose of comparison of genelists and were taken forward into the explorations described in subsequent chapters.

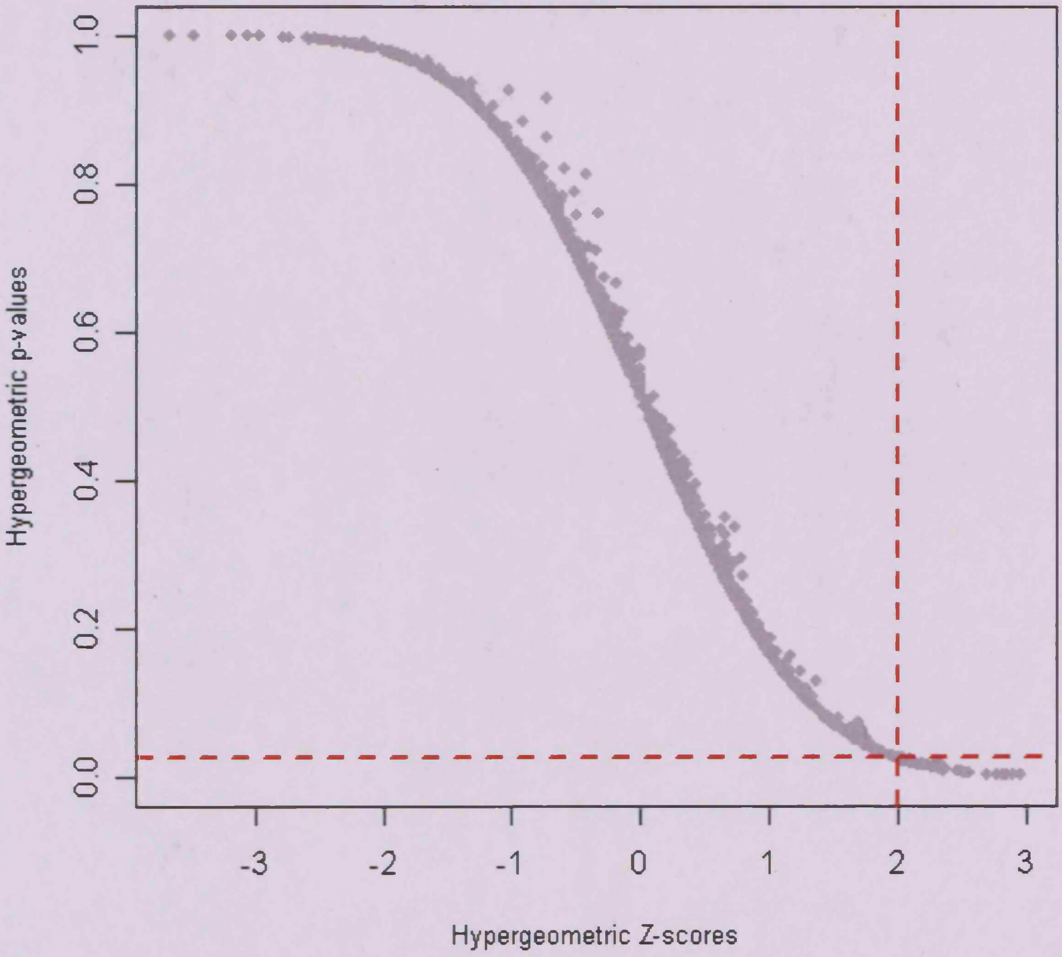


Figure 2.6 Relationship between hypergeometric Z-scores and p-values. Values represent the same distributions illustrated in Figure 2.5. Broken red lines indicate the theoretically equivalent Z-score (vertical) and p-value (horizontal) values of 2 and 0.025 respectively.

2.3 Discussion

The increasing use of microarray technology has led to massive amounts of experimental data being deposited in public databases like NCBI's GEO (Edgar et al. 2002). Access to this experimental data has allowed for integrative analyses of microarray experiments which involve the comparison of several datasets at the same time. This approach allows opportunities for knowledge validation and discovery, and several examples have been highlighted (see Section 1.5).

However, in such analyses, datasets to be analyzed are selected in a supervised fashion, i.e. the selection involves prior hypotheses and knowledge regarding shared underlying biology between the experiments being compared. This approach, while valid, is restricted in scope and might be more optimally utilized by using unsupervised approaches to data integration, for example, comparison of one experiment with, instead of a small pre-selected set of experiments, a large and non-selected set of experiments, such as found in a public database. Such an unsupervised approach has the potential to find unexpected links that could provide new knowledge and insight regarding the experiments in question (Finocchiaro et al. 2005).

However, such an approach would be highly resource-intensive. For this reason, there has been an interest in the comparison of summaries of an experiment, rather than comparisons between entire experimental datasets. Several groups have advocated the use of lists of differentially expressed genes as suitable summaries for this purpose, using the logic that similarity between genelists might imply similarity between the originating experiments (Cahan et al. 2005; Finocchiaro et al. 2005; Newman and Weiner 2005; Rubin 2005; Yi et al. 2007). Thus ranking genelists by order of similarity to a test genelist could provide researchers with prioritization of experiments with which to carry out more rigorous integrative analyses.

However, the reliability of such an approach is questionable, given the frequent observation of low levels of similarity between genelists derived from even very similar experiments (Cahan et al. 2007; Cheadle et al. 2007; Ein-Dor et al. 2005; Jeffery et al. 2006; Manoli et al. 2006; Shen et al. 2008; Tan et al. 2003). For this reason, one of the aims of this project is to assess whether this approach could be useful for researcher by carrying out all comparisons between genelists derived from a diverse array of experiments, carried out in a range of different laboratories and on different species, and examining the results. In this chapter, the strategies for carrying out genelist integration were studied and developed.

To minimize the potential confounding effects that could arise from differences between microarray platforms (for example, due to differences in experimental protocols, probe sequences, etc.), research focussed on integration of only those genelists that were derived from experiments carried out using Affymetrix microarrays. Affymetrix is one of the largest commercial microarray manufacturers and considerable numbers of experiments have been carried out on this platform. Furthermore, they represent standardized experimental protocols and technical aspects (such as probe sequences). However, the general principles derived should be broadly applicable to genelists derived from experiments carried out on other platforms.

Affymetrix currently provides expression arrays for a wide range of species, and the probeset identifiers differ across arrays for each species. Identifiers for a particular gene may also differ within a species; particularly between arrays from different generations (see Section 2.2.1). Comparison of genelists from different arrays within a species can be facilitated by converting probeset identifiers into Entrez Gene IDs for that species. This translation was also included in the strategy for comparison of genelists from the same array-type, because of the potential bias caused by groups of 2 or more probesets that represent the same gene. Cross-species genelist comparisons can in principle be carried out by converting one genelist into a set of homologous genes from the organism on which the other experiment was carried out on.

It was shown that simple overlap size between genelists is not a suitable metric with which to measure links between experiments, as it is sensitive to the systematic effects of genelist length and universe size (see Section 2.2.2). While the latter does not theoretically pose problems with regards to genelist comparison because the universe size remains constant, the variability of genelist length is a more difficult issue.

Three metrics of similarity between genelists were tested for sensitivity to genelist length (see Section 2.2.3). Both fold change and binary similarity coefficients were found to be sensitive. However, the hypergeometric distribution yielded two metrics: p-values and Z-scores, which were found to have no dependency on genelist length. Thus, these metrics were selected for further explorations in subsequent chapters. The hypergeometric distribution has been popularly used in Over-Representation Analysis of genelists (Huang da et al. 2009; Khatri and Draghici 2005), which detects the enrichment of one set of genes (such as a pathway or GO term) within another (the experimentally-derived genelist). Finocchiaro et al also used this for assessing comparisons between experimentally derived genelists (Finocchiaro et al. 2005).

Chapter 3: Comparison of lists of differentially expressed genes using the hypergeometric test

3.1 Introduction

The previous chapter introduced the concept of comparing lists of differentially expressed genes from microarray experiments, as opposed to resource-intensive comparative analyses based on the integration of primary experimental datasets. The studies described revealed general principles regarding the annotation levels at which to perform comparisons of genelists, together with information about the behaviour of key potential metrics to assess similarity between genelists. Whereas in Chapter 2 explorations were based on lists of randomly selected genes (represented as unique and arbitrary identifiers), in this chapter genelist comparison strategies are applied to a set of real experimentally-derived lists of differentially expressed genes.

One might expect that genelists would be most comparable if they were created using similar normalization techniques and statistical tests, with standardized parameters and cut-offs (Bammler et al. 2005; Irizarry et al. 2005; Larkin et al. 2005). Carrying out manual re-analysis of data from potentially hundred of experiments to derive genelists is unfeasible, and the current data storage paradigms of public repositories (in particular, with respect to experimental design) prohibit the automation of such a procedure. Genelists were thus extracted manually from published scientific literature. Initially the database consisted of genelists derived from experiments carried out on the Affymetrix hgu133a array, which was at that time the most popular commercial microarray. More genelists derived from experiments carried out on a range of array-types and species were then provided by Miss Hui Sun Leong (Department of Pathology, Cardiff University) in our laboratory. Table 3.1 provides an overview of the final database.

Array-types	**Species**	**Number of genelists**
hgu133a	*Homo sapiens*	38
hgu133plus2	*Homo sapiens*	19
mouse4302	*Mus musculus*	20
rat2302	*Rattus norvegicus*	11
drosgenome1	*Drosophila melanogaster*	20
celegans	*Caenorhabditis elegans*	11
ath1121501	*Arabidopsis thaliana*	20
Total: 7 array-types	**6 species**	**139 genelists**

Table 3.1 Summary of a database of genelists manually extracted from published literature.

This chapter will describe the results of comparing all the genelists within the database with themselves. It was expected from the outset that judging the utility of such an approach solely from the results would be problematic as there is no biological 'truth' with which to measure sensitivity and specificity. However, one potential indicator of accuracy is if the similarities found between experiments are biologically plausible.

Another issue that it was hoped this analysis would address was how far across species could significant links be found. It might be expected that increase in evolutionary distance between any two species would be accompanied by both a decrease in the number of homologous genes shared between the species and greater differences in transcriptional regulation programmes. As the analysis involved comparison of all genelists in the database, across all the species represented, it provided an opportunity to observe if biological links could be found even between microarray experiments carried out on evolutionarily distant species, for example between a human and *Arabidopsis* experiments.

3.2 Technical Methodology

Development of methodologies for the comparison of genelists was based on the explorations of biological annotation and studies of the behaviour of similarity metrics, as described in Chapter 2.

The genelists collected within the database comprised of Affymetrix probeset identifiers. For comparisons between genelists derived from experiments carried out on the same species (Figure 3.1(a)), regardless of whether they were carried out on the same array-type or not, the original lists of probesets were converted into non-redundant lists of the Entrez Gene IDs (EGIDs) with which the probesets were annotated. If the genelists were derived from experiments carried out on the same array-type, the gene universe comprised of all the EGIDs present on that array-type. For comparisons between genelists from different array-types, the gene universe comprised of those EGIDs that were present on both the array-types. To reflect this, the genelists were then filtered to remove any EGIDs that were not present in that "common" gene universe. For comparisons between genelists from experiments carried out on different species (Figure 3.1(b)), lists of probeset identifiers were first converted to lists of non-redundant species-specific EGIDs with which the probesets were annotated. The EGIDs of one genelist were then converted into EGIDs representing the species of the other genelist through the homologous relationships between EGIDs stored in the Homologene database (see Materials and Methods). Here, the gene universe comprised of those (homologous) EGIDs that were present on both the array-types. The genelists were then filtered to remove any EGIDs that were not present in that "common" gene universe.

The size of the overlap between any two pairs of genelists, as well as the gene universe size, was the used to calculate Z-scores and p-values, using the hypergeometric statistical test, to assess the similarity between the genelists.

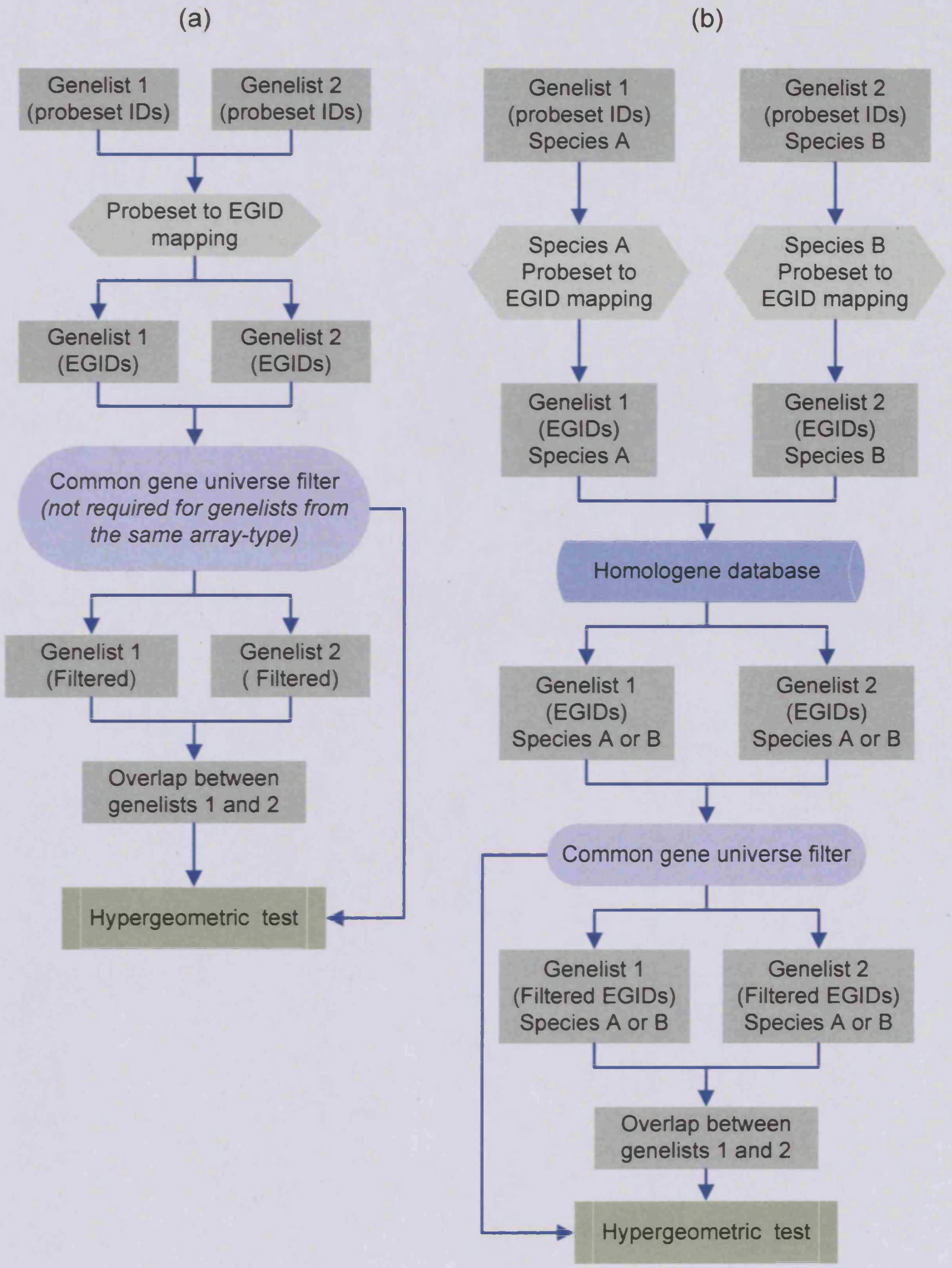

(a)
Genelist 1 (probeset IDs)
Genelist 2 (probeset IDs)
Probeset to EGID mapping
Genelist 1 (EGIDs)
Genelist 2 (EGIDs)
Common gene universe filter *(not required for genelists from the same array-type)*
Genelist 1 (Filtered)
Genelist 2 (Filtered)
Overlap between genelists 1 and 2
Hypergeometric test
(b)
Genelist 1 (probeset IDs) Species A
Genelist 2 (probeset IDs) Species B
Species A Probeset to EGID mapping
Species B Probeset to EGID mapping
Genelist 1 (EGIDs) Species A
Genelist 2 (EGIDs) Species B
Homologene database
Genelist 1 (EGIDs) Species A or B
Genelist 2 (EGIDs) Species A or B
Common gene universe filter
Genelist 1 (Filtered EGIDs) Species A or B
Genelist 2 (Filtered EGIDs) Species A or B
Overlap between genelists 1 and 2
Hypergeometric test

Figure 3.1 Strategies for comparison of genelists (a) within and (b) across species.

3.3 Explorations and Results

3.3.1 An overview of all comparisons

Regarding the question of how far across evolutionary distance could biological similarity be found between genelists, partial insight can be gained prior to the comparison of genelists, by exploring the numbers of homologous genes (if any) that are shared across chips from different species. Table 3.2 shows the number of genes shared between the chips, represented as the percentages of the total number of genes on each chip. The numbers can fall quite low, for example, comparisons between genelists from the *C. elegans* celegans array with those from the *A. thaliana* ath1121501 array would involve a gene universe that consists of only 7.1% of genes on the ath1121501 array and 9.9% of genes on the celegans array.

In Figure 3.2, these values are plotted against the evolutionary distance between the species (see Materials and Methods). It is observed that there is a general trend, as might be expected, of the number of shared genes falling with increase in evolutionary distance.

All possible pair-wise comparisons were then carried out between all 139 genelists in the database. Hypergeometric p-values and Z-scores were calculated for each of the comparisons. Following correction for multiple-hypothesis testing using the Benjamini-Hochberg method (see Materials and Methods), a p-value cutoff value of 0.05 was used to indicate statistically significant similarity between genelists. The number of comparisons found to have significant similarity under this criterion have been represented in Table 3.3 as the percentage of all comparisons between all genelists from any pair of array-types.

	hgu133a	hgu133plus2	mouse4302	rat2302	drosgenome1	celegans	ath1121501
hgu133a	100.0	97.2	79.5	64.4	28.2	18.7	17.9
hgu133plus2	64.8	100.0	70.1	54.5	22.8	14.7	13.9
mouse4302	50.7	67.1	100.0	52.9	21.3	13.9	13.1
rat2302	62.6	79.3	80.5	100.0	27.0	17.6	16.8
drosgenome1	31.3	38.0	37.0	30.9	100.0	20.6	17.2
celegans	15.5	18.3	18.1	15.0	15.4	100.0	9.9
ath1121501	10.6	12.4	12.2	10.3	9.2	7.1	100.0

Table 3.2 – Number of genes shared across different array-types and species. The numbers represent the percentage of genes of the row-wise chips that are present in the column-wise chips. Between array-types from the same species, they represent the number of shared genes (i.e. EGIDs). Between array-types from different species, they represent homologous genes (EGIDs after conversion across species by homology). Grey cells represent comparisons between arrays from the same species.

The proportion of comparisons found to have significant overlaps were highest when comparisons were carried out within a species. Interestingly, links between human genelists can be found with genelists from evolutionarily distant species such as the invertebrates *Drosophila melanogaster* and *Caenorhabditis elegans*, and even the plant *Arabidopsis thaliana*. In Figure 3.3, these values are plotted against the evolutionary distance between the species, and these exhibit a trend of decreasing with increase of evolutionary distance, similar to the behaviour of shared universe sizes in Figure 3.2.

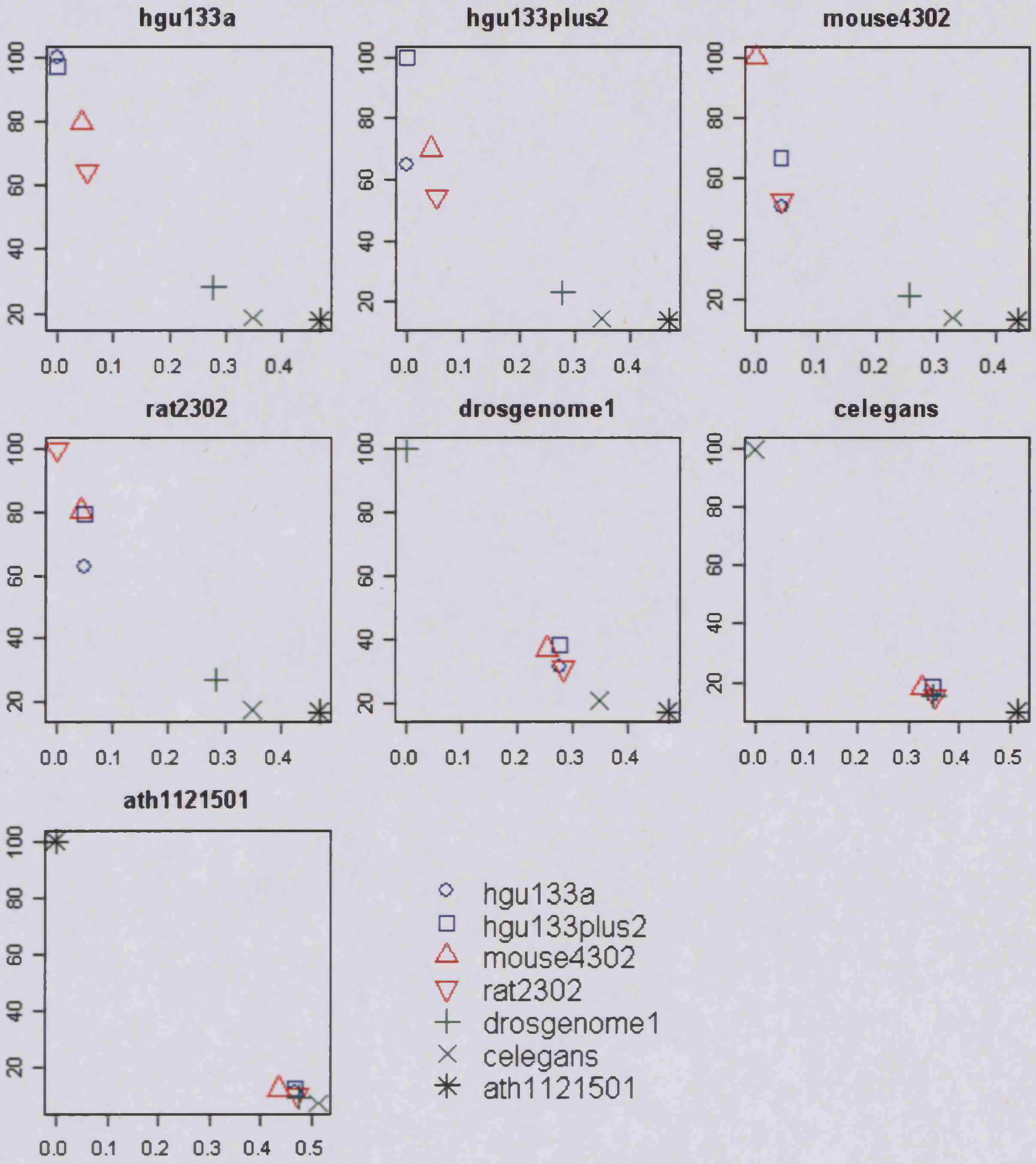


Figure 3.2 Decrease in the number of genes shared between species with evolutionary distance. Data points represent values displayed in Table 3.2. Y-axes represent the percentage of genes for each chip which is shared with the others. X-axes represent evolutionary distances scaled in units of expected fraction of amino acids changed, as calculated using the Dayhoff PAM matrix (see Materials and Methods).

	hgu133a	hgu133plus2	mouse4302	rat2302	drosgenome1	celegans	ath1121501
hgu133a	51.4	53.7	28.4	23.9	5.1	0.2	0.8
hgu133plus2	53.7	69.6	44.7	38.8	3.4	0.0	1.3
mouse4302	28.4	44.7	64.7	35.5	4.0	0.0	0.3
rat2302	23.9	38.8	35.5	52.7	5.0	0.0	0.0
drosgenome1	5.1	3.4	4.0	5.0	66.3	9.1	7.0
celegans	0.2	0.0	0.0	0.0	9.1	80.0	3.6
ath1121501	0.8	1.3	0.3	0.0	7.0	3.6	72.6

Table 3.3 – Comparisons found to show statistically significant similarity. Numbers represent the percentage of all comparisons between genelists that showed significant similarity. Significance was detected using a hypergeometric p-value cut-off of 0.05 following Benjamini-Hochberg corrections (see Materials and Methods). Grey cells indicate comparisons between the same species.

These results appear to be biologically plausible: the decrease in the number of shared homologous genes with increase of evolutionary distance could be a consequence of the evolutionary changes to genome sequences. This factor, along with possible evolutionary changes to transcriptional regulation of biological pathways and networks could explain why the proportion of comparisons found to have significant similarity decreases as the evolutionary distance between the species being compared increases. However, a striking feature observed in the results is the very high proportion of comparisons that appear to have statistically significant overlap sizes when comparisons are carried out between genelists from the same species (or array-type). As can be seen in Table 3.3 (grey cells), 50-80% of all intra-species comparisons are classified as being significant.

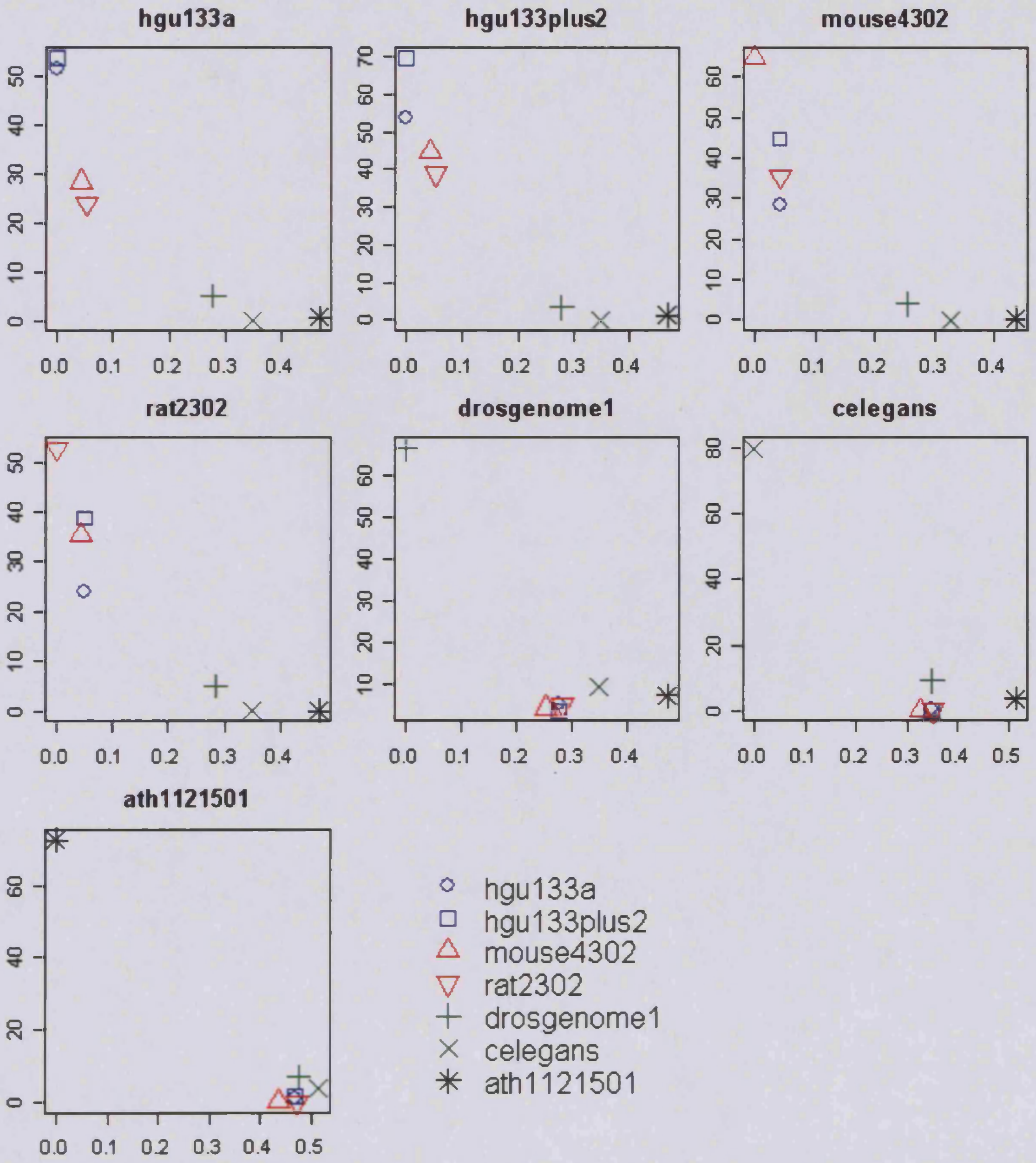


Figure 3.3 Decrease in the number of comparisons showing significant similarity with evolutionary distance. Data points represent values displayed in Table 3.2. Y-axes represent the percentage of comparisons between genelists. X-axes represent evolutionary distances scaled in units of expected fraction of amino acids changed, as calculated using the Dayhoff PAM matrix (see Materials and Methods).

3.3.2 Excess similarity found between genelists from experiments carried out on the same array-type

It might reasonably be expected that lower proportions of cross-species comparisons would yield statistically significant similarity as compared to comparisons carried out between genelists from the same species (or array-type). This is likely to be caused by two effects of evolutionary changes across species. The first effect involves evolutionary changes to genome sequences. This limits the gene universe to only homologous genes that are present on arrays representing the both the species being compared. Prior to comparison, the genelists are required to be filtered to reflect this; very few genes may pass this filter if the shared universe between the species is small (for example, as previously pointed out, between the arrays representing *C. elegans* and *A. thaliana*). The second effect is more biological in nature, involving evolutionary changes to transcriptional regulatory mechanisms, which could cause divergence in biological pathways and processes between species.

While the results described in the previous section are in concordance with these expectations, the very high levels of similarity observed in comparisons of genelists from the same array-type (50-80% of comparisons are assigned significance when using the hypergeometric statistical test) are a cause for concern as such levels of similarity seem biologically implausible.

Hypergeometric Z-score distributions from all possible pair-wise comparisons carried out between genelists from experiments carried out on the same array-type are displayed in Figure 3.4. Control experiments, in the form of simulations, were carried out in parallel as follows: artificial arrays, represented as a set of unique and arbitrary identifiers, were created, of sizes equal to the number of unique Entrez Gene IDs (EGIDs) present on each real array. Genelists were then selected at random and without replacement from these artificial arrays, of the same lengths as those of the experimentally-derived genelists. These randomly created genelists were then compared, thus simulating each comparison of experimentally derived genelists. The grey lines represent the Z-scores distributions from these simulations. As expected, these are centred on medians of zero, implying that on average, there is no similarity between the genelists being compared.

It was expected that the majority of comparisons carried out between a diverse collection of genelists, such as those collected in the database, would yield no significant similarity between genelists, and that, like the simulations, these would yield Z-score distributions centred on medians of zero. However, as can be seen in Figure 3.4, the Z-score distributions from comparisons of the experimentally derived genelists (black lines) are markedly shifted away from those obtained from the simulations, and are centred on medians of 2 or more.

There are two possible interpretations of this observation. The first is that there is indeed some biology common to most, if not all, the experimentally-derived genelists; this seems somewhat biologically implausible when considering the diversity of experiments from which the genelists were derived. The second is that these could be artefacts caused by the violation of some assumption(s) of the hypergeometric statistical test.

To investigate whether this effect is prevalent in other collections of genelists, sets of genelists were downloaded from the L2L database (Newman and Weiner 2005). These are summarised in Table 3.4. All pair-wise comparisons were carried out between genelists from the same array-type and the resultant Z-score distributions are shown in Figure 3.5.

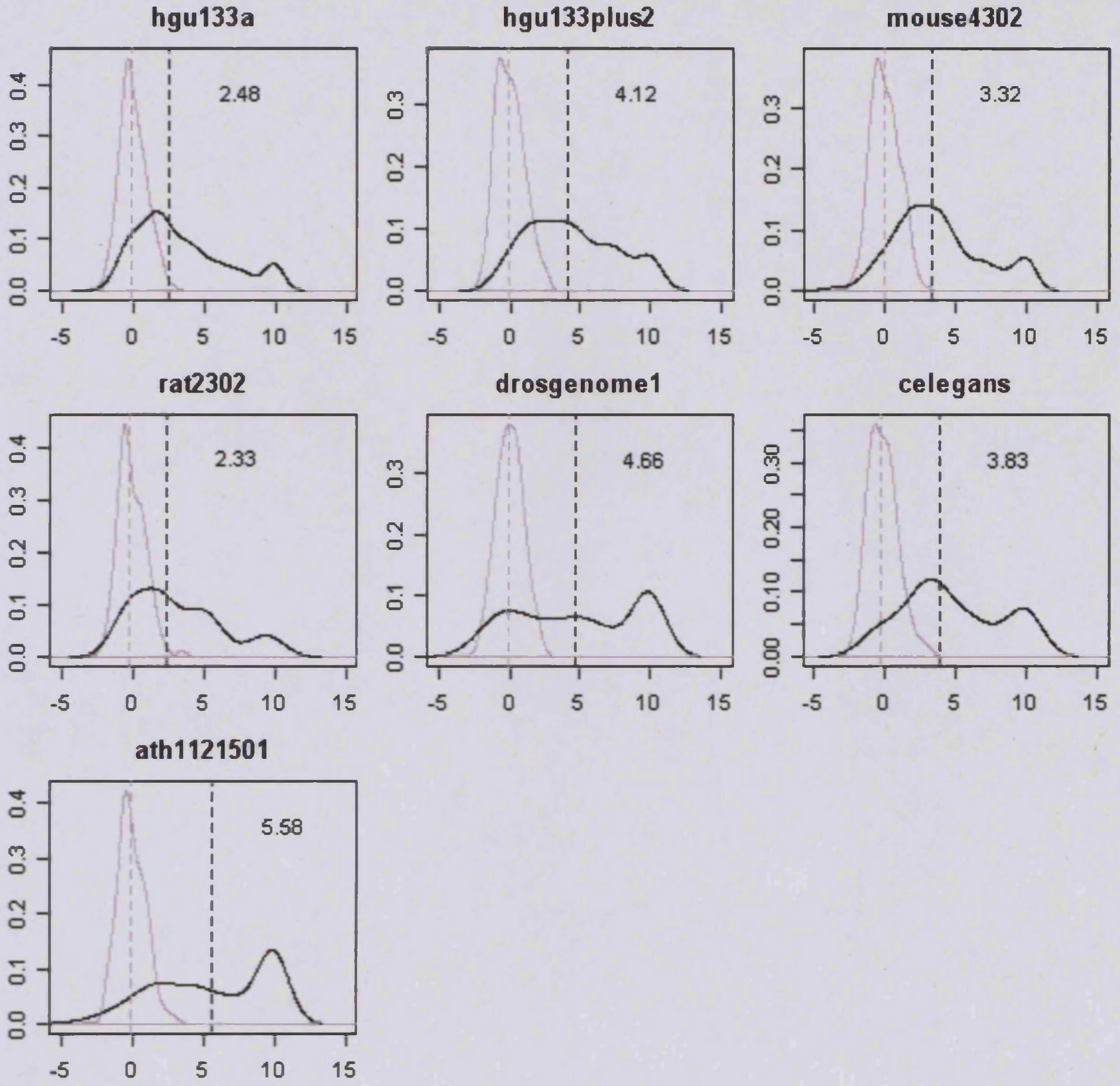


Figure 3.4 Hypergeometric Z-score distributions of comparisons of genelists from the same array-type. X-axes represent Z-scores;Y-axes denote distribution frequencies. Black lines represent comparisons of experimentally derived genelists. Grey lines represent simulations using random genelists using the same genelist and gene universe sizes as comparisons of real-world genelists. Broken lines represent medians of these distributions. Numbers within the plots signify the medians of the Z-score distributions from comparisons of experimentally derived genelists. For ease of visualization, Z-scores were capped at 10, such that all Z-scores >10 were set to 10. This causes some distributions to appear biphasic.

Array-type	**Species**	**Number of genelists**
hgu133a	*Homo sapiens*	37
hgu95a/av2	*Homo sapiens*	105
mgu74a/av2	*Mus musculus*	54
moe430a	*Mus musculus*	13
Total: 4 array-types	**2 species**	**209 genelists**

Table 3.4 Summary of genelists downloaded from the L2L database

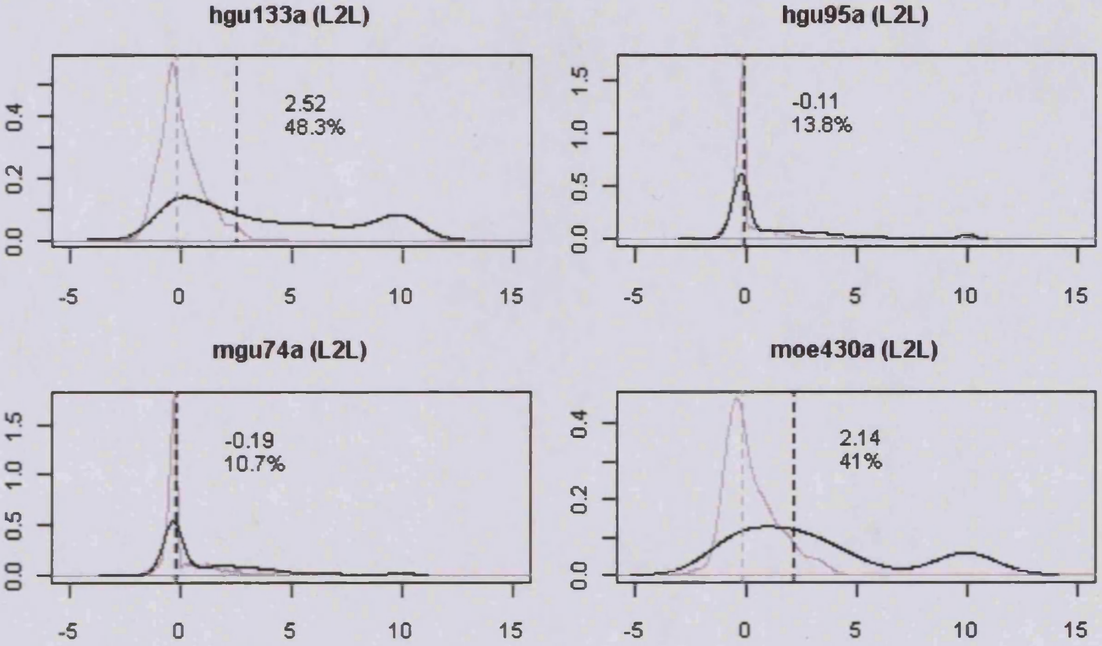


Figure 3.5 Hypergeometric Z-score distributions of comparisons of genelists from the same array-type. X-axes represent Z-scores; Y-axes denote distribution frequencies. Black lines represent comparisons of experimentally derived genelists. Grey lines represent simulations using random genelists using the same genelist and gene universe size as comparison of real-world genelists. Broken lines represent medians of these distributions. Upper numbers in the plots signify the medians of the Z-score distributions from comparisons of experimentally derived genelists; lower numbers denote the percentage of comparisons found to be significant at $p<0.05$ (after FDR correction). For ease of visualization, Z-scores >10 were set to 10. This may cause some distributions to appear biphasic.

As can be seen, Z-score distributions are again positively shifted away from medians of zero in comparisons from genelists of experiments carried out on the hgu133a and moe430a arrays. However, in comparisons of genelists from experiments carried out on the hgu95a and mgu74a arrays, no such shift is observed. In this case the Z-score distributions, like those derived from the simulations, are centred on medians of zero.

These findings are in concordance with the proportions of comparisons which were found to have statistically significant overlaps between genelists (using a p-value cut-off of 0.05 after Benjamini-Hochberg correction): comparisons of genelists from experiments carried out on the hgu133a and moe430a arrays yielded significance for excessive proportions of comparisons (48% and 41% respectively), while the proportions for comparisons of genelists from experiments carried out on the hgu95a and mgu74a arrays that were found to have statistically significant overlap sizes were much more lower (14% and 11% respectively).

These results might suggest that one or more as-yet unidentified effects, which result in the apparent excess similarity amongst genelists from most of the array-types tested, are not prevalent in the hgu95a and mgu74a arrays. However, as is described in the next section, this does not appear to be the case.

3.3.3 Link between significance of similarity and genelist length

Excess similarity between genelists from experiments carried out on the same array-type is observed in most of the array-types tested: the near-ubiquity of this phenomenon is consistent with this being a reflection of an unidentified systematic bias.

One observation in support of this hypothesis is the strong correlation found between the levels of excess similarity found within each array-type and the length of genelists. In Figure 3.6, the median Z-scores of comparisons of genelists within each array-type are plotted against the square root of median genelist lengths for each array-type. These values have a strong correlation to each other, having a correlation coefficient (r) of 0.97. This correlation was also found to be highly significant: a Pearson's product-moment test (see Materials and Methods) yielded a p-value of 1.13 x 10^{-6}.

This apparent relationship then warranted investigation of the question of whether longer genelists tend to find more significant similarities with other genelists than shorter genelists. For this purpose, the median Z-scores of all comparisons of each genelist with all other genelists from the same array-type was recorded and plotted against the square root of the length of that genelist (Figure 3.7).

As can be seen in the figure, the majority of chip types appear to exhibit correlation between the median Z-scores from a genelist and the length of that genelist, though the strength of this correlation is highly variable. It should be noted that amongst the genelists from the L2L database that are derived from experiments carried out on the hgu95a and the mgu74a arrays, which had not shown the excess levels of similarity observed in the other sets of genelists, the correlation is highly significant.

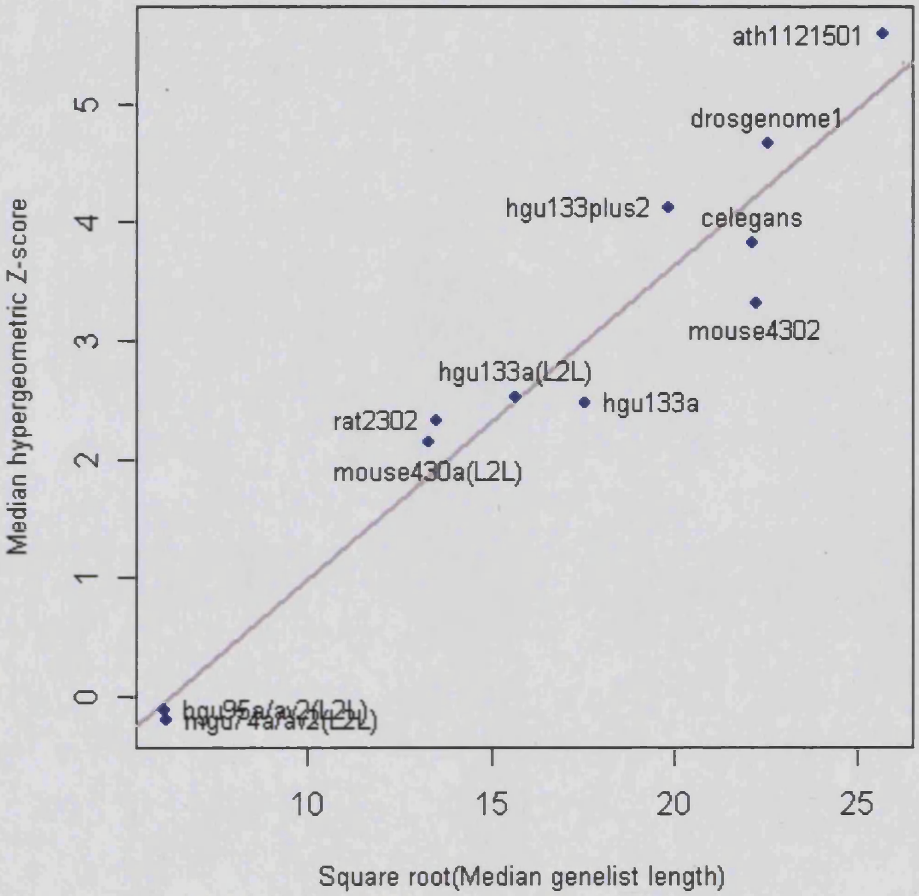


Figure 3.6 Median Z-scores of intra-chip comparisons plotted against the square root of the median length of all genelists from a particular array-type. To avoid possible duplication of genelists, comparisons between genelists from the in-house database and the L2L database were kept separate. Unless indicated otherwise, points represent comparisons made between genelists from the in-house database. The grey line is the line of best fit created by linear modelling of the data.

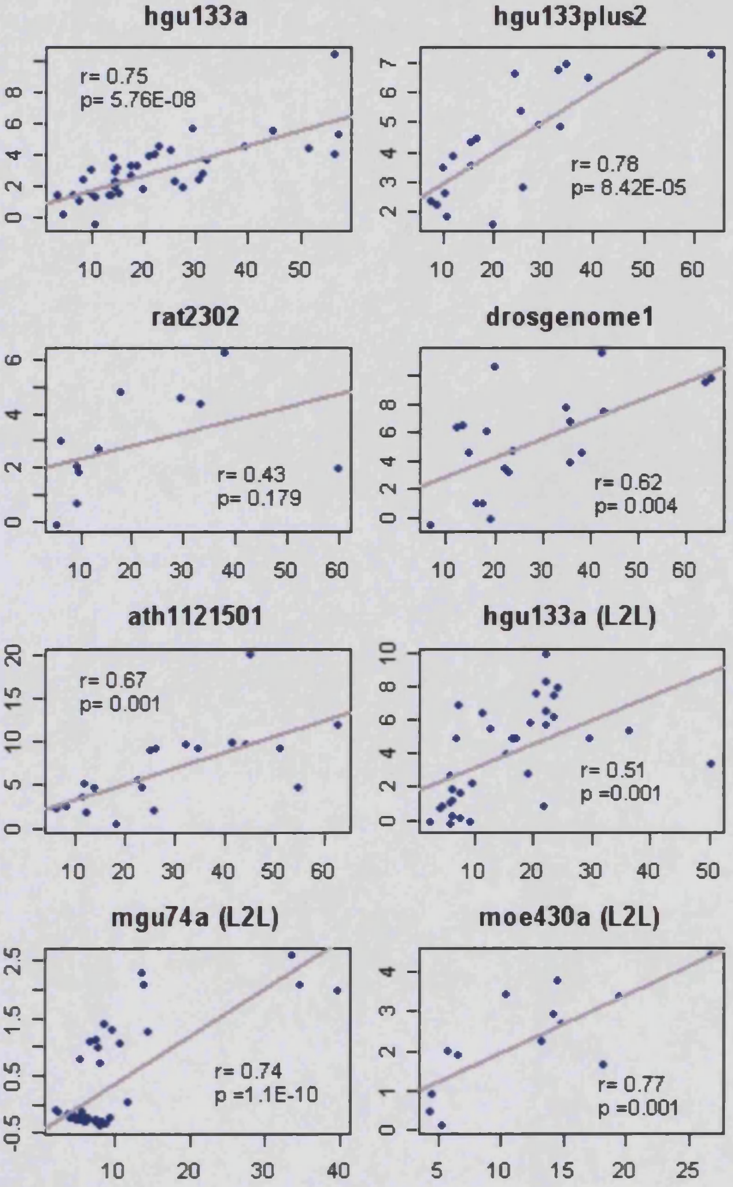


Figure 3.7 Correlation between genelist length and significance. Data points represent the median hypergeometric Z-score of all comparisons carried out between any one genelist and all other genelists from the same array-type (Y-axes), plotted against the square root of the length of that genelist. Grey lines represent lines of best fit created by linear modelling of data points. In the plot, the upper number represents the correlation coefficient while the lower number is the p-value from the Pearson's product-moment test for correlation.

This finding could explain why the shifts in Z-score distributions and excess similarity is not observed amongst the genelists from the hgu95a and mgu74a arrays that were downloaded from the L2L database but are seen in all sets of genelists from the in-house database and all other genelists downloaded from the L2L database: with median genelist lengths of 37 and 37.5 genes for the hgu95a and mgu74a sets respectively (whilst all other sets of genelists had median genelist lengths of 176-660 genes), most genelists in these sets were too short for any shift of Z-score distribution to become noticeable.

To test this, comparisons were carried out for these two sets again, but only using genelists of length greater than 50 genes (an arbitrary cut-off). The resultant hypergeometric Z-score distributions (Figure 3.8) now show shifts away from a median of zero. The proportion of comparisons found to have significant similarity (selected as having $p<0.05$ after Benjamini-Hochberg correction) are also much higher: 45% and 32% for the hgu95a and mgu74a sets respectively (as opposed to 14% and 11% respectively observed prior to filtration of short genelists).

This relationship between genelist length and significance of overlaps between genelists was unexpected, because the hypergeometric test (see simulations described in Chapter 2) is insensitive to genelist length. This length-dependency provides additional support for the hypothesis that the excess similarity observed between genelists from the same array-type reflects some systematic bias, for example, caused by the violation of some assumption(s) of the hypergeometric distribution.

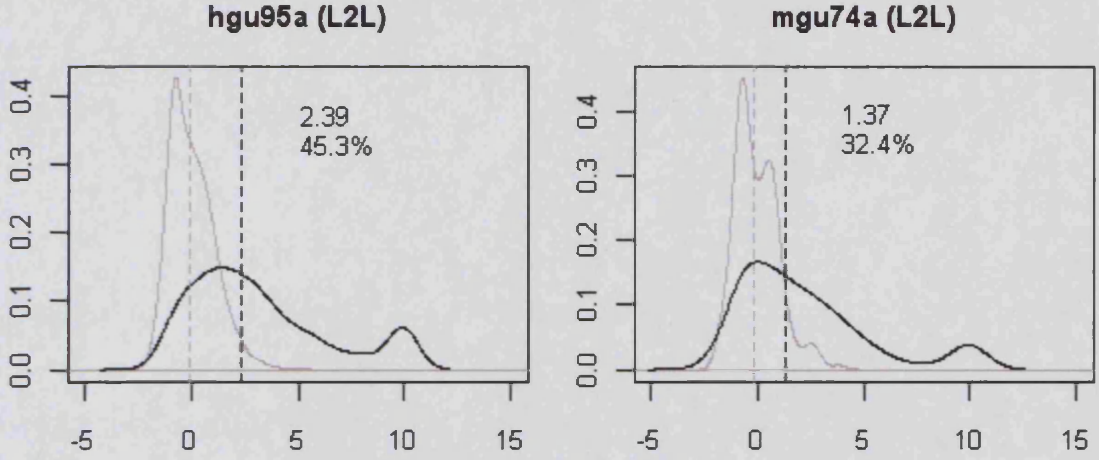


Figure 3.8 – Hypergeometric Z-score distributions of comparisons of genelists from the same array-type. X-axes represent Z-scores; Y-axes denote distribution frequencies. Black lines represent comparisons of experimentally derived genelists. Grey lines represent simulations using random genelists using the same genelist and gene universe size as comparison of real-world genelists. Broken lines represent medians of these distributions. Upper numbers in the plots signify the medians of the Z-score distributions from comparisons of experimentally derived genelists; lower numbers denote the percentage of comparisons found to be significant at $p<0.05$ (after FDR correction). For ease of visualization, Z-scores >10 were set to 10. This may cause some distributions to appear biphasic.

The link between genelist length and significance also makes compensating for the excess similarity observed between genelists more difficult. If the effect of the observed systematic bias could be shown to be uniform, causing the shift of all Z-scores by the same amount or proportion, then the rank-order of comparisons on the basis of Z-scores would remain the same as when there was no bias, and the top-ranked comparisons could be selected for further analysis.

The following chapter describes explorations into the possible causes for the observed relationship between significance and genelist length.

3.4 Discussion

This chapter explored the results of applying the hypergeometric statistical test to assign significance to overlaps between genelists, having carried out all possible pair-wise comparisons between all genelists from a database of genelists extracted manually from the published literature of microarray experiments carried out on a range of species. It was initially found that significant overlaps could be found even between genelists from experiments carried out on evolutionarily distance species, for example between genelists from the human hgu133a and the *Arabidopsis* ath1121501 Affymetrix platforms. It was also found that the proportions of comparisons found to be significant decreased with the evolutionary distances between the species from which genelists were being compared. These could reflect both the decrease in the number of homologous genes (due to evolutionary changes of genetic sequences), and evolutionary changes to transcriptional regulation of biological pathways.

However, very high levels of similarity were found for comparisons between genelists from the same chip type. This was reflected in both the high proportions of comparisons that exceeded thresholds of statistical significance and in hypergeometric Z-score distributions that were positively shifted away from an expected median of zero. These findings have several implications. The first is regarding practicality; for this approach to be of use it should highlight a relatively few interesting overlaps between genelists, which would be seen as occasional outliers (and detected as such). However, as the method finds the majority of comparisons to have significant similarity, it has limited practical utility. The second implication concerns the interpretation of these findings; while it is possible that the results reflect true biological similarities between the genelists, this seems somewhat implausible given the diversity of experiments from which the genelists were derived. A perhaps more likely scenario is that these are the results of some unidentified systematic effect, such as an erroneous assumption within the underlying statistical model.

Subsequently, it was found that significance levels shared strong correlation with genelist length; longer genelists tended to be involved in more comparisons that showed significant similarity between genelists than did shorter genelists. When enough genelists were present in the set, comparisons between them would result in a Z-score distribution shifted well beyond a median value of zero. This is further evidence of some systematic effect, as the hypergeometric statistical test was chosen for the purpose of genelist comparison because the metrics involved are not sensitive to genelist length (in simulations using randomly created genelists – See Section 2.2.3.3).

While it would be possible to quantify and compensate for this relationship between significance and genelist length, which then could result in fewer, more credible numbers of significant links being found, this was not carried out for the following reasons. Firstly, any modelling of the relationship would have been very *ad hoc*, and completely dependent on the data. Secondly (and more importantly), as the reasons for this bias were unknown, there would have been no theoretical grounds for trusting the results from such manipulation of the data.

For these reasons, further explorations focused on attempts to find the possible statistical errors that could have caused the observed relationship between genelist length and significance. Chapter 4 describes explorations of possible violations of assumptions of the hypergeometric distribution that could explain this relationship. It is shown that under the 'biased urn' model of the hypergeometric distribution, the excess levels of similarity observed in comparisons of experimentally-derived genelists from the same array-type can be simulated using randomly created genelists. Chapter 5 then explains why any *ad hoc* methods that attempt to quantify and control for the observed biases could be unreliable when using a single universe size for all comparisons of genelists from the same array-type.

Chapter 4: Modelling violations of the assumptions of the hypergeometric distribution using the 'biased urn' model

4.1 Introduction

One aim of this thesis is to assess the viability and potential utility of comparing lists of differentially expressed genes as representatives of microarray experiments to find possible biological similarities and links between them. Chapter 2 described the development of strategies to translate genelists across different chips and species, and the selection of the hypergeometric statistical test to assess the significance of overlaps between genelists. Chapter 3 described the application of these methodologies to a local database of genelists manually extracted from the published scientific literature of microarray experiments carried out on Affymetrix GeneChip microarrays covering a range of species.

Excess levels of similarity were found for comparisons carried out between genelists from experiments carried out using the same type of GeneChip. Subsequently it was observed that there was a relationship between significance and genelist length, which is not predicted from a simple application of the hypergeometric distribution. These findings raised the possibility that the excess levels of similarity do not reflect any true underlying biology common to these genelists (which would be a somewhat implausible scenario, given the diversity of experiments represented by the genelists), but rather some inherent flaw within the statistical model used to assess significance of overlaps between genelists.

This chapter describes explorations attempting to identify the cause of this observed bias, whereby longer genelists are involved in more comparisons that are called significant than shorter genelists. One possible cause is a violation of some assumption(s) of the hypergeometric statistical test. Such violations might not be totally unexpected in this system, given that a simple statistical model has been applied to the complex issue of gene expression patterns. One particular assumption is that all genes (in an array) have an equal probability of being selected into a list of differentially expressed genes. There are several gene regulation scenarios that violate this assumption.

One such category involves genes that are less likely to appear in a genelist because they are very rarely or never expressed. This may be due to trivial technical reasons (such as a badly designed probe to which the complementary mRNA does not hybridize), or for biological ones (for example genes that are only expressed in certain tissues and/or environmental conditions), or even a combination of the two (for examples, genes which are differentially regulated but have expression levels too low to be detected by current hybridization-based techniques). The presence of such genes then artificially increases the size of the gene universe, and could lead to erroneous assessment of significance when included in calculations of the hypergeometric statistical test.

To illustrate this, consider two genelists that do not share any underlying biology, and have an overlap of size x, which is the expected size of overlap between two randomly created genelists selected from a gene universe *that does not include genes in any of the categories described above* (i.e. a subset of the full gene universe). Now consider that the hypergeometric test is used to assess the significance of this overlap, and all genes represented on the array from which these genelists are derived are included in the gene universe. The hypergeometric test would then assume an expected overlap size of y, but this would be lesser than x, because, as described in Section 2.2.2, the expected size of overlap between genelists decreases with increased universe size (when genelist lengths are constant). Because significance is a function of the deviation of the observed overlap size from the expected overlap size, this comparison would be assigned greater significance (i.e. higher Z-score and lower p-value) than it ought to be. An increase in numbers of this category of (un-expressed/rarely-expressed) genes would result in a greater difference between the size of the 'real' gene universe (i.e. the set of genes which can be selected into a genelist) and the size of the gene universe used in the hypergeometric test (i.e. all genes on the array), which in turn would result in greater significance being assigned to any comparison of genelists.

A second category would involve those genes whose expression patterns are evolutionarily (or otherwise) constrained to be stable. This might potentially include housekeeping genes such as GAPDH, β-actin, or cyclophilin because the expression patterns of these genes may not be expected to show much change in response to experimental conditions (She et al. 2009). As most tests for differential expression assess the level of variability in expression levels between phenotype classes, such genes may be less likely to be selected into a genelist. The presence of this category of genes on an array would have a similar effect on the hypergeometric test as might be expected from the first category of genes.

A third category consists of those genes that show high levels of variability in their expression patterns across different phenotype classes i.e. those that are commonly subject to differential transcriptional regulation in response to experimental conditions. These genes would then be more likely to be selected by tests for differential expression into genelists than other genes. This then would result in an increased probability of these genes being found in common between genelists. In effect, these genes could artificially increase the overlap size between genelists, and cause the comparison to be assigned greater significance by the hypergeometric test than should be the case.

While the above three categories of genes all violate an assumption of the hypergeometric test, namely that all genes have an equal probability of being selected into a genelist, a fourth category of genes violates another assumption: that genes are selected into genelists independently of each other. This category would include co-regulated genes. For example, consider a group of three co-regulated genes, with such similar expression patterns that (if differentially expressed) they would always be selected together into a genelist. Thus the contribution of these genes to an overlap between two genelists that contain these genes would be a value of three (instead of the desired value of one).

More advanced models of the hypergeometric distribution, such as the Fisher's and Wallenius' non-central hypergeometric distributions (Fog 2007), do not make some of the assumptions as the simple hypergeometric model, such as the assumption that all genes have equal probability of being selected. However, their use in the analysis of microarray data has been limited, because these methods require precise quantifications of gene-selection probabilities that are currently unavailable. Due to this, researchers have used simple models like the hypergeometric distribution in microarray data analysis (for example in Over-Representation Analysis of GO terms or pathways within genelists), in the hope that major biological signals will overcome any deficiencies of the statistical models.

So far, no attempts have been made to quantify these complexities or the effects they could have on the results when using statistical models that ignore them. This chapter describes attempts to quantify the effect of these complexities on the naïve implementation of the hypergeometric test to the comparison of lists of genelists using the 'biased urn' model.

4.2 Explorations and Results

4.2.1 The 'biased urn' model

Advanced models of the hypergeometric distribution such as those mentioned above (Fog 2007), involve selection from what has been termed a 'biased urn'. This is an allusion to the urn-filled-with-balls analogy that is popularly used to describe the hypergeometric distribution (see Section 2.2.3.3), and refers to these methods' use in cases where different balls in the urn have different probabilities of being selected. In order to quantify the net effect of complex gene expression patterns on the use of the simple hypergeometric model to assess similarity between genelists, a statistical model was used that also assumed a 'biased urn'.

This model involves an urn (gene universe) that is filled with two sets of balls (genes): a set that can be selected into genelists, and a set that cannot. Consider that the size of this gene universe is x, and the size of the subset of the universe from which genes can be selected is y, such that $x > y$. Genelists were to be selected in a random fashion from amongst this subset of genes and then compared using the hypergeometric test.

However, the universe size to be used in assessment of significance is x rather than y. In such a scenario, the distribution of overlap sizes observed would vary around the overlap sizes expected when using a gene universe of size y. However the statistical test would assume overlap sizes expected when using a gene universe of size x. As shown in Section 2.2.2, the expected overlap size decreases with increase in the size of the gene universe (when genelist lengths are constant). Thus, on average, any comparison of genelists from this model would involve an observed overlap size which is greater than what the model expects, and assignment of greater significance to this comparison than would be warranted by a comparison of two randomly created genelists. This could then cause excess levels of similarity to be found between the genelists, similar to what was observed in the comparisons of literature-derived genelists described in Chapter 3.

This model represents a somewhat over-simplification of the complex expression patterns of genes (such as those described in the introduction to this chapter), and cannot be expected to accurately quantify these effects separately. However it is hypothesized that this model could allow some quantification of the net effect of these factors on the comparison of genelists.

4.2.1.1 Effect of universe size on significance in the biased urn model

It is expected that for a comparison of randomly created genelists using the biased urn model, the significance assigned to the comparison would increase with an increase of the universe size used in the calculations relative to the size of the 'real' gene universe from which the genelists were sampled.

For example, consider the following biased urn scenario: two genelists of 1000 genes each that have been selected from a gene universe of 10,000 genes. Then consider that the observed overlap size between them is exactly as expected i.e. 100 (as calculated using the formula described in Section 2.2.2). A hypergeometric test using the correct gene universe size of 10,000 (i.e. an unbiased urn) would yield a Z-score of zero for this comparison. However, a hypergeometric test that involves a universe size of 20,000 (i.e. a biased urn which has a further 10,000 genes which cannot be selected into genelists), would then assume an expected overlap size of 50, and assign this comparison a Z-score of 7.4. Similarly, using gene universe sizes of 25,000 and 30,000 would yield Z-scores of 9.9 and 11.9 respectively and so on.

To demonstrate this, simulations were carried out as follows. A set of 10,000 unique and arbitrary identifiers was created, representing the subset of a gene universe from which genes can be selected into a gene universe. Two genelists, each of 1000 genes each were then selected at random and without replacement for the gene universe, and the size of the overlap between them was recorded. This was carried out a thousand times.

Hypergeometric Z-scores were then calculated for this distribution of overlap sizes using gene universe sizes ranging from 10,000 to 40,000. Figures 4.1a and 4.1b display box-plots and density curves, representing the Z-scores distributions calculated using these different universe sizes, respectively.

As expected, the distribution calculated using the correct universe size (which represents an unbiased urn) is centred on a median of zero (broken black lines). The other distributions are increasingly shifted away (from a median of zero) as the universe size used for their calculations increase.

The medians of these distributions (horizontal black lines within box-plots) are equal to Z-scores calculated using an overlap size of 1000, as would be expected in two genelists of 1000 genes each randomly selected from a universe of 10,000 genes (red points in Figure 4.1a).

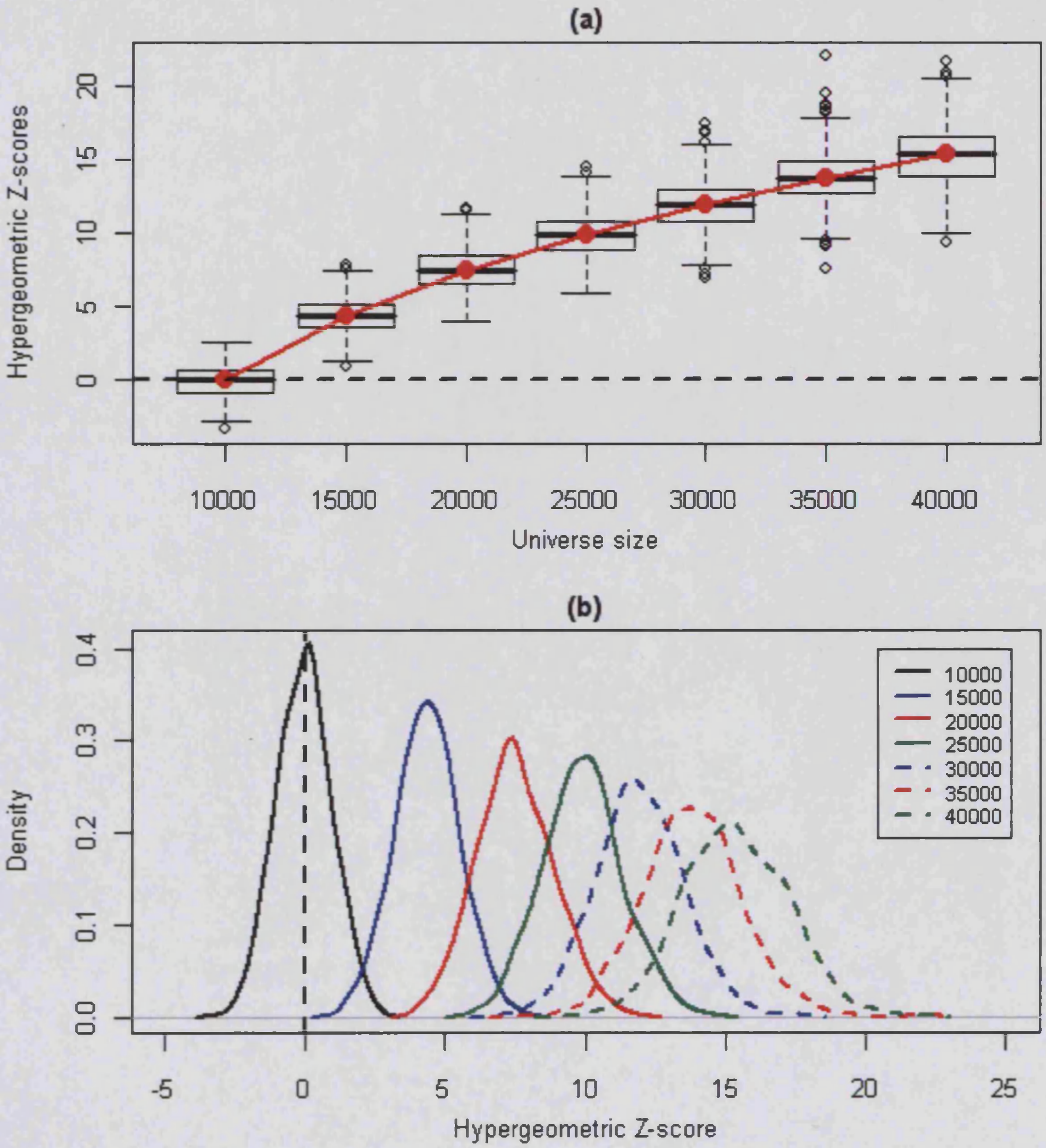


Figure 4.1 Effect of universe size on significance in the biased urn model. (a) Box-plots and (b) density curves of Z-score distributions based on comparisons of two genelists of 1000 genes each selected at randomly from a hypothetical array (a set of 10,000 unique and arbitrary identifiers), iterated a thousand times, created using a range of gene universe sizes (denoted by the X-axis in (a) and the key in (b)). Broken black lines in both figures mark a Z-score of zero. Red lines and points in (a) represent pre-calculated Z-scores expected for their respective distributions.

4.2.1.2 Effect of genelist length on significance in the biased urn model

It is also expected that, in the biased urn model, genelist length is linked to significance. For example, consider the following biased urn model: an array of 20,000 genes of which only half can be selected into genelists. For a comparison between two randomly created lists of 1000 genes each, an overlap size of 100 genes is expected. Under the biased urn model the expected overlap size is 50 and this would yield a Z-score of 7.4. Similarly consider a comparison of two randomly created genelists of 2000 genes each that have an expected overlap size of 400 genes. The biased urn model would then expect an overlap size of 200 and yield an even more significant Z-score of 15.7. A comparison between two lists of 3000 genes (that have an overlap of the expected size) each would yield a Z-score of 24.9, and so on.

To demonstrate this, the following simulations were carried out: a set of 10,000 unique and arbitrary identifiers was created, representing the subset of a gene universe from which genes can be selected into a gene universe. Two genelists, each of 1000 genes each were then selected at random and without replacement for the gene universe, and the size of the overlap between them was recorded. This was carried out a thousand times. Hypergeometric Z-scores were then calculated for this distribution of overlap sizes using a gene universe size of 20,000 genes. This was repeated using a range of genelist lengths varying from 1000 to 2500 genes.

Box-plots and density curves of the resultant Z-score distributions are displayed in Figure 4.2a and 4.2b respectively. A shift of distributions towards increasing significance (Z-scores) with increased genelist size is seen. The medians of these distributions (horizontal black lines within box-plots) are equal to Z-scores calculated using an overlap sizes that are expected in a comparison of randomly created genelists of their respective lengths (red points in Figure 4.2a).

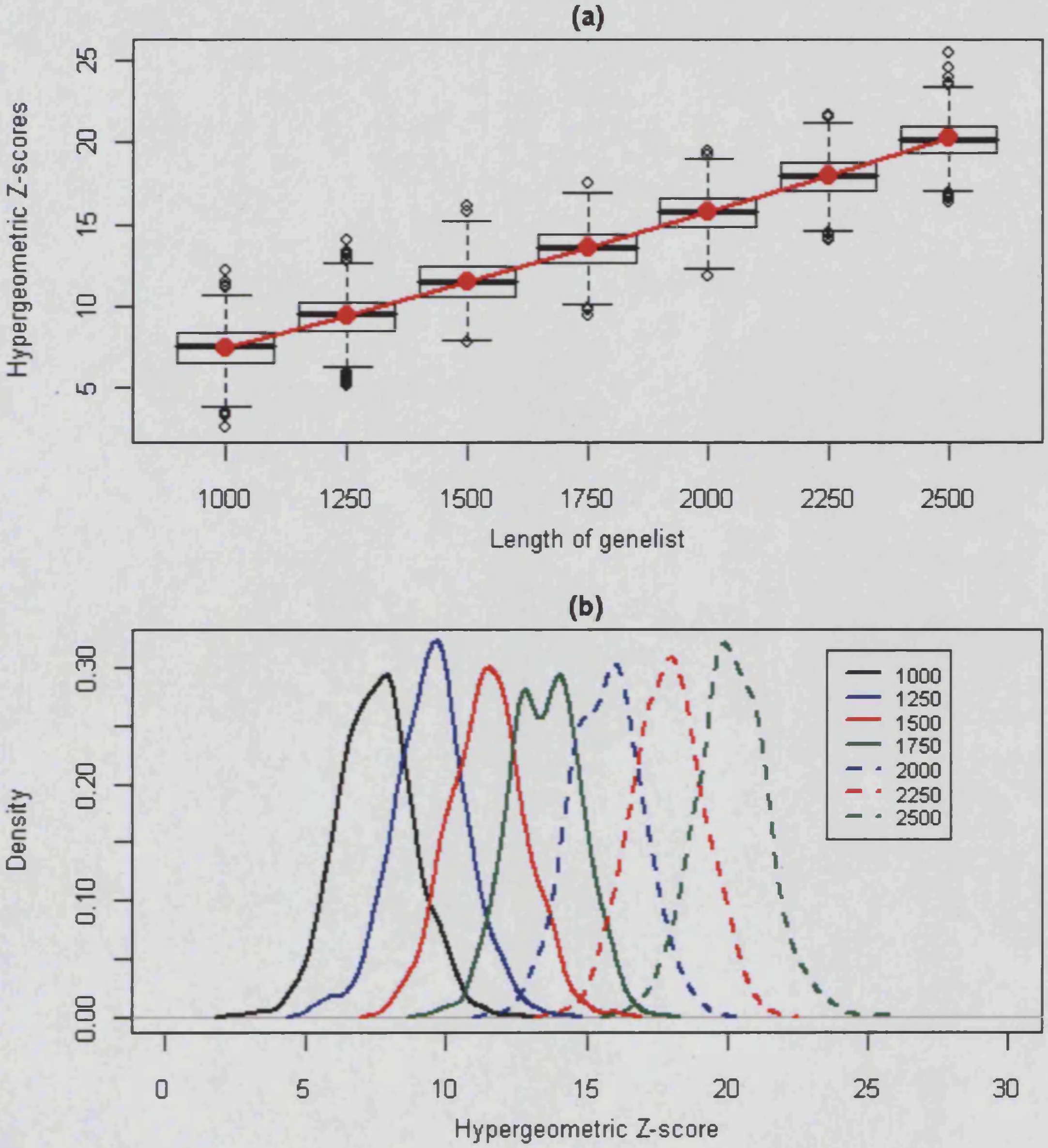


Figure 4.2 Effect of genelist length on significance in the biased urn model. (a) box-plots (b) and density curves of Z-score distributions based on comparisons of two genelists (of the same length), each selected at randomly from a hypothetical array (a set of 10,000 unique and arbitrary identifiers), iterated a thousand times. Calculations used a universe size of 20,000 genes and a range of genelist lengths (denoted by the X-axis in (a) and the legend in (b)). Red lines and points in (a) represent pre-calculated Z-scores expected for the respective distributions.

4.2.1.3 Combined effect of genelist length and universe size on significance in the biased urn model

The explorations described above illustrate how gene universe size and genelist length can independently influence significance when performing the hypergeometric statistical test using the biased urn model.

To investigate if (and how) these factors interact with each other, simulations were carried out as follows: a set of 10,000 unique and arbitrary identifiers, representing a hypothetical array was created. From this, a reference list of 1000 genes was selected randomly and without replacement. Then, a set of genelists of lengths ranging from 1000 to 5000 genes were selected randomly and without replacement.

Each of these was compared to the reference genelist and the observed overlap size was recorded. Hypergeometric Z-scores were then calculated for each comparison using gene universe sizes of 10,000 (representing an unbiased urn), 15,000, 20,000 and 25,000 genes. Figure 4.3a displays the density curves of the resulting Z-score distributions.

As expected, the distribution that was created using a universe size of 10,000 genes (i.e. an unbiased urn) is centred on a median of zero. As seen previously in the analysis described in Section 4.2.1.1, the Z-score distributions that were calculated using the biased urn model are shifted away from a median of zero, and the magnitude of this shift is proportional to the difference between the 'real' gene universe size of 10,000 and the universe size used in the calculations.

There is also a greater change in the shapes of the distributions derived from the biased urn model, as compared to the previous analysis, probably due to a range of different genelist lengths being used (as opposed to the fixed genelists lengths used in Section 4.2.1.1).

In Figure 4.3b the Z-scores are plotted against the square root of the lengths of the test genelists used in the comparisons. As expected, the distribution derived from the unbiased urn shows no influence of genelist length on magnitude. The distributions derived from the biased urn models show an increase of significance with increase of genelist length, which is in concordance with the analysis described in Section 4.2.1.2.

It is also seen that for models involving greater deviations of the gene universe size from the 'real' gene universe size of 10,000 genes, there is a more pronounced effect of genelist length on significance, as evidenced by the increased slopes of the lines of best fit.

Thus, while genelist length and universe size can both influence measures of significance in the biased urn model independently, in cases where both factors are variable, the magnitude of the effect of genelist length on significance depends on how biased the urn is (i.e. the magnitude of the difference in the size of universe used for sampling genelists, and that of the universe size used in the hypergeometric test).

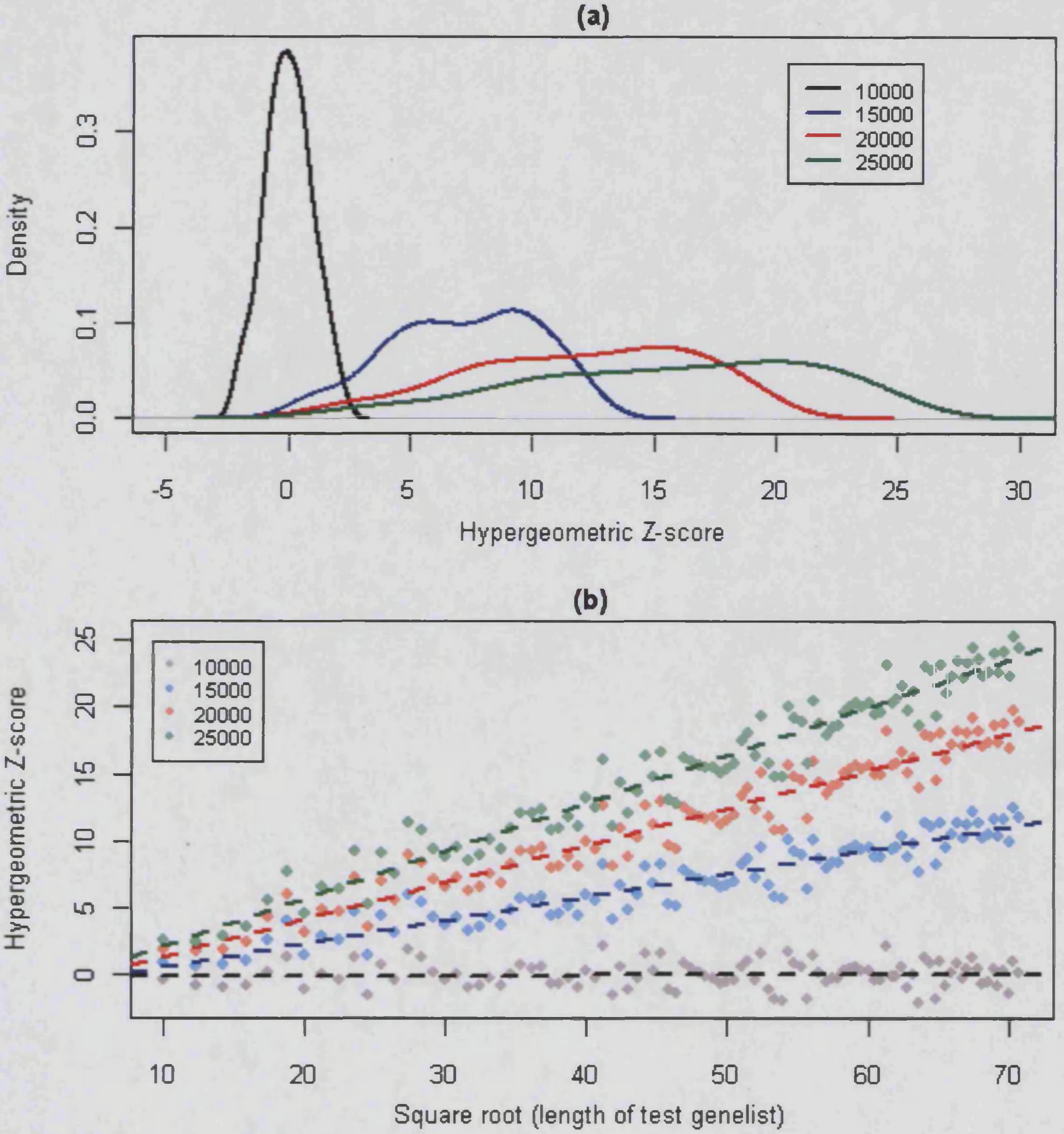


Figure 4.3 Combined effect of genelist length and universe size on significance in the biased urn model. (a) Density curves of Z-score distributions of comparisons of a reference genelist to a set of test genelists of a range of lengths, all of which were selected randomly from hypothetical arrays (sets of unique and arbitrary identifiers) of different sizes. (b) Z-scores plotted against square root of test genelist lengths. Broken lines represent lines of best fit created by linear modelling. The different distributions were created by using different gene universe sizes in Z-score calculations (see figure legends.)

4.2.2 Effective gene universe sizes range from only 35-65% of the genes on an array

The above explorations indicate that the biased urn model is able to simulate both the excess similarity, and the effect of genelist length on significance of comparisons, that is observed in comparisons of real-world genelists. This was achieved simply by changing the universe size used in calculating the hypergeometric Z-scores, thus violating the assumption that all genes in the universe are equally likely to be selected into genelists. In effect, additional genes are added to the universe that cannot be selected into genelists.

Application of this model to a comparison of literature-derived genelists does have some caveats. It involves something of an over-simplification of the complex expression patterns and interactions that occur between genes, in that it assumes only two behaviours: that a gene can be selected into a genelist or it cannot. It thus ignores the continuum of probabilities that quantify the ability of genes to be selected into genelists, and the likelihood that these probabilities may change under different experiments and conditions. The model also ignores the possibility that the gene universes are also likely to be different for different comparisons. For these reasons, it is not reasonable to expect the model to provide resolution and quantification of the many different gene expression behaviours (some of which may cause excess similarity between genelists, and some of which may decrease it). However, it could provide adequate quantification of the net effect of these phenomena in the terms of this model.

The biased urn model was applied to the comparison of literature-derived genelists as follows. Central to the application of this model is the assumption that when the 'true' gene universe size is used in the calculation of Z-scores, the resulting distribution would be centred on a median of zero, thus reflecting the reasonable expectation that in a collection of genelists from a diverse range of experiments, most genelists would not be similar to each other. Decreasing the size of the gene universe used in the Z-score calculation resulted in the shifting of the distribution backwards towards a median of zero. The universe size was iteratively decreased till the resulting distribution was centred on a median of as close to zero as possible. This process was carried out on the comparisons of all sets of genelists (except the L2L hgu95a and mgu74a sets, as their Z-score distributions were already centred on medians of zero), and the results are shown in Table 4.1.

Array-type	Median Z-score	Original gene universe size (number of genes on array)	Estimated 'real' gene universe size	Estimated 'true' gene universe size (% of original)
hgu133a	2.48	13387	7784	58.1%
hgu133plus2	4.12	20080	9256	46.1%
mouse4302	3.32	20981	11794	56.%
rat2302	2.33	13784	7395	53.6%
drosgenome1	4.66	12049	7688	63.8%
celegans	3.83	16107	8650	53.7%
ath1121501	5.58	22568	9635	42.7%
hgu133a (L2L)	2.52	13387	5225	39.%
moe430a (L2L)	2.14	13419	4762	35.5%

Table 4.1 Application of the biased urn model to comparisons of literature-derived genelists. 'True' gene universe sizes were estimated by iteratively reducing the gene universe size used in calculations of Z-scores till the resulting distributions were centred on medians of zero. Array-types represent sets of genelists (derived from those arrays) collected in the local database unless indicated otherwise.

The estimated 'true' gene universe sizes imply that, on average, for any comparison of genelists, the net effect of the complex gene expression patterns that are ignored by the simple hypergeometric test is equivalent to only 35-65% of genes on an array being available for selection into genelists, in terms of the biased urn model.

4.2.2 Randomly selected genelists show excess similarity using the biased urn model

Having estimated the 'true' gene universe sizes, it was then desired to estimate the extent to which the excess similarity observed for comparisons of the literature-derived genelists could be explained by the systematic effects of genelist length and gene universe size on significance, as a result of violations of the assumptions of the simple hypergeometric test. This was attempted by carrying out the following simulations: for each of the sets of genelists, a hypothetical array, represented by a set of unique and arbitrary identifiers, was created. The size of these arrays was that of the estimated 'true' gene universe sizes (e.g. to simulate a comparison of the hgu133a set of genelists, a hypothetical array of 7784 genes was used). Then a set of genelists was selected from this array at random and without replacement, with lengths matched to each of the literature-derived genelists of that set. These genelists were then compared using the hypergeometric test, but now using a gene universe size that was the original number of genes on the array (thus for simulations of comparison of the hgu133a genelists, a universe size of 13387 genes was used).

The resulting distributions are shown in Figure 4.4. It is found that the distributions from the biased urn model are able to simulate the excess similarity observed in the comparisons of literature-derived genelists well. This is reflected both in the similarity of the shapes of the distributions, and also in the similarity in the median Z-scores for these distributions, and the proportions of comparisons found to be significant at $p<0.05$ after Benjamini-Hochberg correction (Table 4.2).

From these observations, it was concluded that much of the excess similarity observed during comparisons of genelists derived from experiments carried out on the same type of GeneChip appears to be an artefact caused by violations of assumptions of the hypergeometric test.

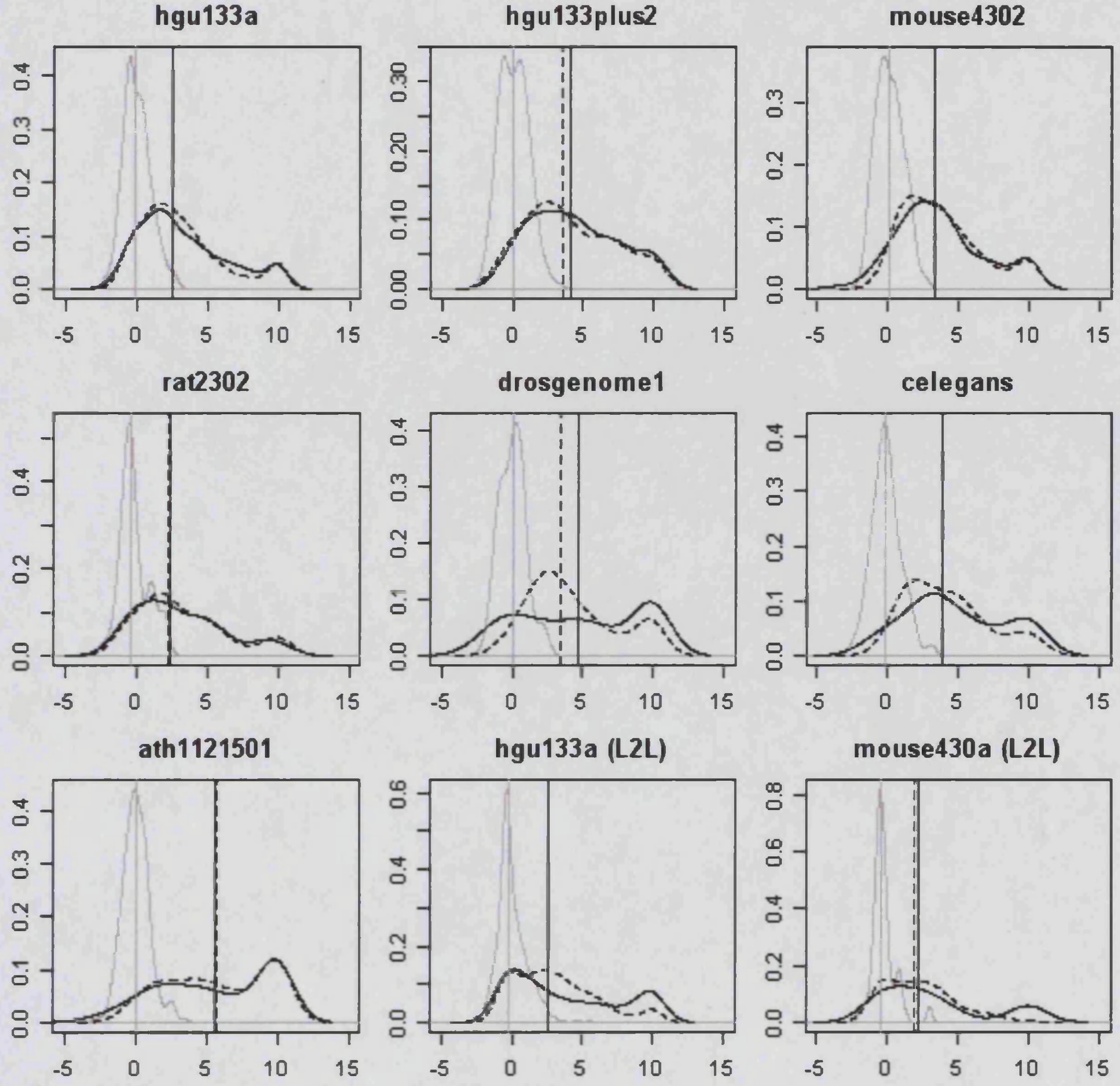


Figure 4.4 Modelling comparisons of literature-derived genelists using the biased urn model. X-axes represent hypergeometric Z-scores; Y-axes denote distribution frequencies. Black curves represent Z-scores from comparisons of literature-derived genelists. Grey curves represent Z-scores from simulation studies using an 'unbiased' urn (i.e. the universe size used for sampling genelists is the same as that used for calculations). Broken curves represent Z-scores simulations using a biased urn. Vertical lines represent medians of their respective distributions. For ease of visualization Z-scores >10 were set to 10. This may cause some distributions to appear biphasic.

Array-type	Comparison of literature-derived genelists		Comparison of randomly created genelists using the biased urn model	
	Median Z-score	% of comparisons significant at p<0.05	Median Z-score	% of comparisons significant at p<0.05
hgu133a	2.48	51.4%	2.51	52.3%
hgu133plus2	4.12	69.6%	3.51	66.1%
mouse4302	3.32	64.7%	3.37	64.7%
rat2302	2.33	52.7%	2.29	43.6%
drosgenome1	4.66	66.3%	3.43	73.2%
celegans	3.83	80%	3.8	74.5%
ath1121501	5.58	72.6%	5.69	78.9%
hgu133a (L2L)	2.52	48.3%	2.59	48.8%
moe430a (L2L)	2.14	41%	1.98	37.2%

Table 4.2 – Modelling comparisons of literature-derived genelists using the biased urn model. Median Z-scores and proportions of comparisons found to be positive at p<0.05 (after Benjamini-Hochberg correction) for comparisons of literature-derived genelists and from simulation studies using the biased urn model. These are derived from the distributions represented as black and broken curves respectively in Figure 4.4. Array-types represent sets of those genelists (derived from those arrays) collected within the in-house database unless indicated otherwise.

4.3 Discussion

As described in Chapter 3, comparisons between genelists derived from experiments carried out on the same GeneChip using the hypergeometric statistical test yielded excess levels of similarity between the genelists. This observation, along with the discovery of a relationship between genelist length and significance, led to the hypothesis that the excess similarity reflected the systematic effects of a flaw in the statistical model (possibly caused by violations of assumptions of the hypergeometric test), rather than any true underlying biology.

This chapter then described investigations of this hypothesis, which involved what has been termed the 'biased urn' model of sampling for the hypergeometric distribution. While the simple hypergeometric test assumes that all genes on an array have an equal probability of being selected into a genelist, the biased urn model forces a bias in the sampling probability, such that a certain proportion of genes on the array can never be selected. This model was used for the investigations because in this model, the significance metrics are functions of genelist length and universe size, and yield excess levels of similarity between randomly created lists of genes, as is seen in comparisons of experimentally-derived genelists.

To simulate the comparisons of experimentally-derived genelists, the biased urn model required three parameters: the genelists lengths (which would be the same as the lengths of the experimentally-derived genelists), the universe size used for calculations of significance (which would be the total number of genes on the array; this is the universe size used in comparisons of experimentally-derived genelists), and the 'true' gene universe size, which is the size of the subset of those genes on the array that can be sampled into genelists. The 'true' gene universe sizes were estimated on an *ad hoc* basis by iteratively re-calculating Z-scores from comparisons of experimentally derived genelists using different gene universe sizes till the resulting distribution was centred on a median of zero. This methodology was based on the assumption that as most genelists from a diverse set of experiments would not be similar to each other, they should yield a Z-score distribution centred on a median of zero, as is seen from comparisons of randomly created genelists.

The magnitude of the difference between these 'true' gene universe sizes and the total number of genes on the array could thus provide a rough quantification of the magnitude of the violation of assumptions of the hypergeometric distribution observed in the data. Strikingly, it was found that the net effect of these violations resulted in the equivalent of only 35-65% of the genes on an array being available for selection into genelists, in terms of the biased urn model.

Simulations carried out using the biased urn model with the parameters described above yielded Z-score distributions that were highly similar to those derived from the original comparisons of experimentally derived genelists. Thus, even though the biased urn model represents an over-simplification of the complex gene expression patterns that violate assumptions of the hypergeometric distribution (some of which are described in Section 4.1), it is able to replicate the excess levels of similarity observed amongst experimentally-derived genelists with a considerable degree of success.

Naïve application of the hypergeometric distribution to the comparison of experimentally-derived genelists is likely, therefore, to yield unreliable results, because of the influence on genelist length on significance metrics. However, as the biased urn model is capable of simulating these effects relatively well, simply by forcing a difference between the sizes of the sampling universe and the size of the universe used in calculation of significance metrics, it can provide some potential solutions. One option may be to simply use the estimated true gene universe size in the hypergeometric test. Another option may be to predict the relationship between genelist length and Z-scores based on the difference between the two universe sizes. The observed Z-scores can be projected on this relationship and significance could then be calculated using the deviation of Z-scores from the predicted relationship. These two methods can only be used in cases where excess similarity is observed amongst genelists (i.e. Z-score distributions are shifted away from medians of zero), as both require estimation of the true gene universe size. Another option which could be of use, particularly in cases like the hgu95a and mgu74a sets of genelists from the L2L database (Newman and Weiner 2005) and thus where the genelists are too short to cause shifts in Z-score distributions, would involve modelling the observed relationship between Z-scores and genelist lengths. For example, a line of best fit could be induced by linear modelling of the data, and significance could then be assigned depending on the deviation of Z-scores from this line.

However, there are issues regarding such solutions. Firstly, they are *ad hoc* and very dependent upon the data, and will thus be potentially prone to biases within the data. Secondly, and more importantly, as these methods use, for all assessments, a single estimated true gene universe size that would actually be the average gene universe size shared between any two experiments, they ignore the possibility that different comparisons are likely to involve different gene universes (which may be of different sizes). For example, it is likely that two experiments carried out on the same tissue type may share a much larger gene universe than those carried out on different tissue types. Estimating the correct universe size is crucial, as (in the biased urn model), it not only influences the significance metrics (see Section 4.2.1.1), but also influences the effects of genelist length on significance metrics (see Section 4.2.1.3).

For this reason, subsequent investigations focussed on estimation of gene universes for experiments. Chapter 5 describes these explorations, which indicated considerable diversity in gene universe sizes, which appear to be affected by both technical and biological (such as tissue-specificity) effects.

Chapter 5: Exploring gene expression patterns with the GNF Expression Atlas

5.1 Introduction

Chapter 3 described explorations involving comparisons of genelists derived from experiments carried out on the same array-type, by using the hypergeometric statistical test to assess significance of overlaps. This revealed what appeared to be excess levels of similarity. Further work revealed a link between genelist length and the significance of comparisons, which is not expected from the hypergeometric test and thus suggested the existence of one or more flaws within the application of the statistical model. Chapter 4 described explorations of possible violations of the hypergeometric test. A statistical model (called the biased urn) was used that could model the excess levels of similarity observed, as well as the relationship between genelist length and significance. It was found that the model, when using randomly created genelists, could provide a reasonable simulation of the significance distributions derived from comparison of experimentally-derived genelists. This was achieved by using a different sized gene universe in the statistical test to that from which the genes were sampled.

While the biased urn model represents an over-simplification of the various gene expression scenarios that could cause violations of the assumptions of the hypergeometric distribution, its ability to at least partially replicate the Z-score distributions derived from comparisons of experimentally-derived genelists does raise the question as to whether it might be possible to model (and thus compensate for) these violations. One concern is that such modelling is likely to be a highly *ad hoc* and very dependent on the data, and would be susceptible to any biases inherent in a particular dataset. Also, such modelling would require assumptions that may be plausible, but somewhat weak (for example, the assumption that comparisons carried out within a set of genelists from a diverse range of experiments would yield a Z-score distribution centred on a median of zero, as is seen from comparisons of randomly created genelists).

One particular concern would be if the same gene universe size (i.e. the subset of genes on an array from which genes can actually be selected into genelists) would need to be used for all comparisons, since from a biological perspective this does not seem justified. This issue is of particular concern because, as simulations described in Chapter 4 show, in the biased urn model, calculations of the significance metrics are sensitive to the size of the gene universe. This also affects the magnitude of the relationship between genelist length and significance. Thus, use of the correct gene universe size is crucial. While the model provides an estimate of the true gene universe size, this only represents the average size of the gene universe that is shared between any two experiments. It is not difficult to conceive of pairs of experiments having very different sizes of shared gene universes (for example due to tissue-specific gene expression). Thus, one might expect that two experiments carried out on similar tissue-types would share a larger gene universe than two experiments carried out on different tissue-types.

To investigate these issues in greater detail the following chapter reports explorations undertaken using the Genomics Institute of the Novartis Research Foundation (GNF) Expression Atlas (Su et al. 2004) dataset. This is a publicly available set of expression profiles of a wide range human and mouse tissues. By providing an opportunity to explore the number of genes expressed in different types of tissues, it allowed some semi-quantitative estimates to be derived as to potential gene universe sizes in different tissue types, and the number of expressed genes shared between the different gene universes.

5.2 Results and Explorations

5.2.1 Explorations of gene universe sizes and gene expression frequencies across 68 different tissue-types

The explorations described below were carried out on a subset of the Expression Atlas (see Materials and Methods) that comprises of the expression profiles of 68 different normal human tissues (having excluded expression profiles derived from foetal tissues and cancerous cell-types) carried out on the Affymetrix human hgu133a array platform. As the experimental procedures carried out to generate the tissue expression profiles that comprise this dataset were all carried out within the same laboratory and on the same microarray platform, analysis of this dataset should minimise problems related to cross-platform and cross-laboratory data integration.

Each of the 68 tissue-types comprising this subset of the GNF Expression Atlas that was selected for further explorations described in this chapter is represented by the expression profiles of two microarrays. To avoid possible biases that might be introduced by instances of several probesets that represent the same gene (see Chapter 2), Entrez Gene IDs were used to represent genes rather than probeset IDs. This was achieved by selecting for each of the Entrez Gene IDs represented on the array, the probeset that shows the greatest median expression levels over all the 68 different tissues (see Materials and Methods).

One estimate for the size of the gene expression universe (that is, the number of genes expressed) for each tissue sample could then be calculated based on the number of genes flagged as 'Present' by the MAS5 algorithm in at least one of the two arrays representing that tissue-type (see Materials and Methods). The sizes of the gene universes based on these criteria are displayed in Table 5.1, and the distribution of these sizes is shown in Figure 5.1.

Tissue-type	Size of gene universe (% of genes on array)	Tissue-type	Size of gene universe (% of genes on array)
Amygdala	46.0	PLACENTA	42.6
CerebellumPeduncles	32.4	Uterus	44.9
CingulateCortex	36.7	UterusCorpus	20.9
Hypothalamus	45.6	Prostate	44.2
MedullaOblongata	32.5	testis	38.5
OccipitalLobe	39.8	TestisSeminiferousTubule	35.2
OlfactoryBulb	37.9	TestisGermCell	44.2
ParietalLobe	31.7	TestisInterstitial	36.6
Pons	24.2	TestisLeydigCell	31.2
PrefrontalCortex	48.3	Heart	29.3
TemporalLobe	27.4	atrioventricularnode	18.3
Thalamus	39.2	Appendix	18.3
TrigeminalGanglion	14.8	721_B_lymphoblasts	52.8
WholeBrain	45.4	PB-CD19+Bcells	44.7
caudatenucleus	35.4	PB-CD4+Tcells	47.2
cerebellum	31.9	PB-CD56+NKCells	47.2
globuspallidus	27.6	PB-CD8+Tcells	48.6
subthalamicnucleus	29.8	PB-CD14+Monocytes	45.1
spinalcord	38.2	lymphnode	40.3
ciliaryganglion	21.4	Lung	40.7
SuperiorCervicalGanglion	13.1	Liver	27.6
PB-BDCA4+Dentritic_Cells	52.8	SkeletalMuscle	13.5
bronchialepithelialcells	41.2	SmoothMuscle	39.6
Pancreas	34.0	CardiacMyocytes	34.8
PancreaticIslets	40.3	BM-CD33+Myeloid	43.9
BM-CD105+Endothelial	43.3	TONGUE	24.7
BM-CD34+	50.0	salivarygland	31.0
BM-CD71+EarlyErythroid	36.1	Pituitary	36.1
bonemarrow	28.4	skin	21.2
WHOLEBLOOD	44.6	thymus	43.9
adrenalgland	33.3	Thyroid	48.3
AdrenalCortex	27.4	Tonsil	35.3
ADIPOCYTE	36.8	trachea	37.3
Ovary	25.7	kidney	32.4

Table 5.1 Numbers of genes expressed in 68 normal human tissues from the GNF Expression Atlas. Numbers indicate percentages of the total number of EGIDs present on the hgu133a array. Expression universes were constructed for each tissue type by selection of genes flagged as 'Present' by the MAS5 algorithm. Tissue-type names were extracted from the names of the CEL files.

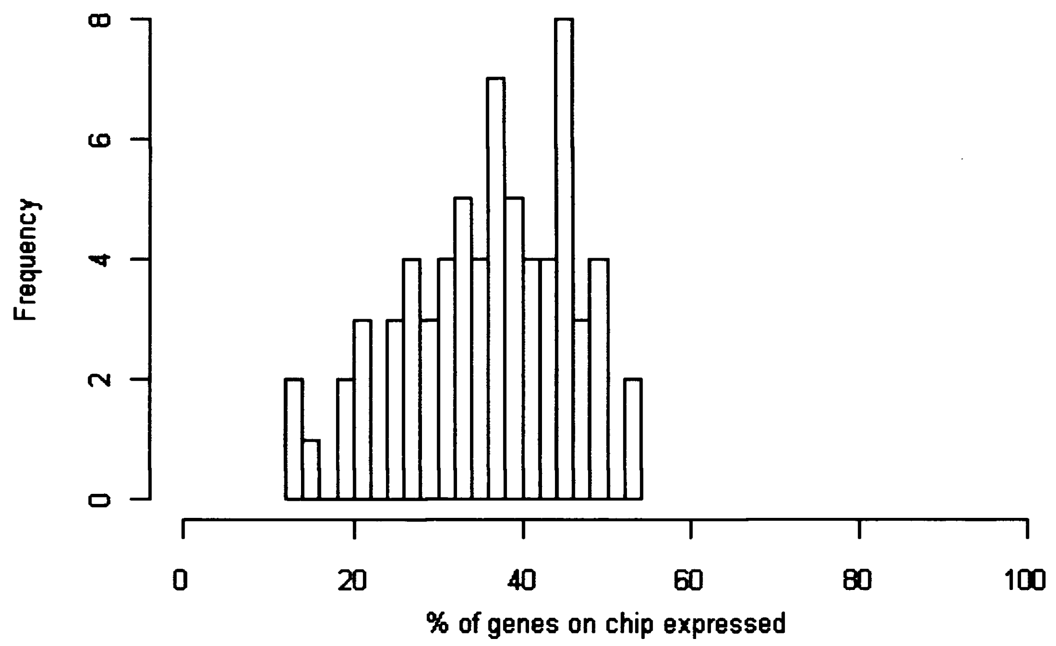


Figure 5.1 Histogram of the number of genes expressed in 68 normal human tissues from the GNF Expression Atlas. These are represented as percentages of the total number of EGIDs present on the Affymetrix hgu133a array. Expression universes were constructed for each tissue type by selection of genes flagged as 'Present' by the MAS5 algorithm.

The median size of the expression universe for these tissues corresponds to 36.6% of the genes (EGIDs) present on the hgu133a array. The number of genes expressed in the different tissues varies considerably, with a standard deviation for this distribution corresponding to approximately 10% of genes on the array. The universe sizes range from 13% of genes on the array, as seen for the superior cervical ganglion tissue, to 53% of genes on the array, as seen for the BDCA4+ dendritic cells.

This analysis also provided the opportunity to investigate gene expression frequencies across these 68 tissue-types. These were calculated as the numbers of tissue-types that each of all the genes present on the hgu133a array were expressed in. To compare this with the distribution of gene expression frequencies expected if genes were selected into expression universes at random, a set of genelists was created, each of the size of one of the tissue gene expression universes, by selecting randomly and without replacement from the total set of genes present on the hgu133a array.

The distribution of gene expression frequencies from the random selection of gene expression universes appears to follow an approximately normal distribution as shown in Figure 5.2b and in the quantile-quantile (Q-Q) normal plot of this distribution in Figure 5.2d. However, the distribution of gene expression frequencies observed in the Expression Atlas (see Figure 5.2a) appears to deviate noticeably from a normal distribution (see Q-Q normal plot in Figure 5.2c), due to the presence of large numbers of genes that are expressed in very few tissues, or in many tissues. Close to 15% of genes on the hgu133a array are not expressed in any of the tissues, and more than 32% of genes are expressed in 5 or less tissues. At the same time, more than 3% of genes on the array are expressed in all 68 tissues, while close to 10% of genes on the array are expressed in 65 or more tissues. In the distribution of gene frequencies of randomly selected genes, no genes cross any of the thresholds described above.

This highly non-normal and non-random selection of genes into expression universes as observed in these 68 tissue types is likely to cause the sort of expression patterns (some of which are described in Section 4.1) that cause violations of assumptions of the hypergeometric distribution when this is used to compare experimentally-derived lists of genes, and in particular, the assumption that all genes are equally likely to be selected into genelists.

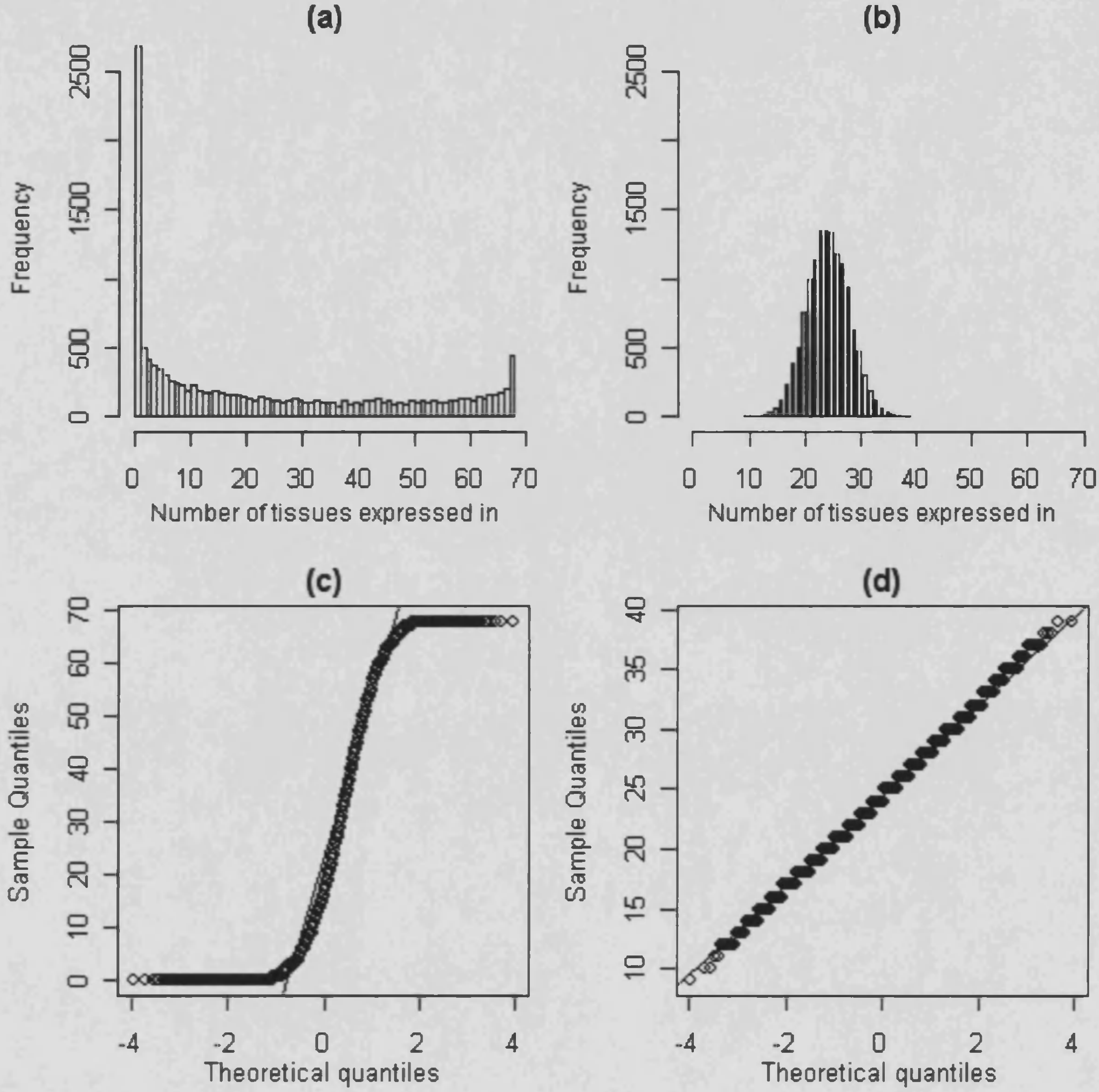


Figure 5.2 Gene expression frequencies observed in 68 different normal human tissue-types. (a) Histogram of gene expression frequencies for all genes on the hgu133a array, i.e. the number of tissues in which each gene is found to be expression (on the basis of being called as Present by the MAS5 algorithm); (c) shows the normal Q-Q plot for this distribution. (b) Histogram of gene frequencies derived from a simulation involving random selection of genes from the hgu133a array into hypothetical expression universes of the same sizes as those derived from the 68 tissues under investigation; (d) shows the normal Q-Q plot for this distribution.

5.2.2 Sizes of genes shared by expression universes of different tissues vary widely

In the previous section it was noted that that the sizes of the expression universes (i.e. the number of genes found to be expressed) for each of the 68 tissue types from the GNF Expression Atlas vary widely. This would then suggest that the number of genes shared between the expression universes of any pair of tissues is highly variable as well, as would be expected from relationship between genelist length and overlap size described in Section 2.2.2.

To investigate this, the number of genes shared between the expression universes (i.e. the sizes of the shared expression universes) was recorded for every possible pair of tissues from amongst the 68 normal human tissues of the GNF Expression Atlas under consideration. A control experiment was then carried out by creating another set of expression universes by randomly selecting from amongst all genes represented on the Affymetrix hgu133a array, sets of genes of the same sizes as the expression universes observed for each of the 68 normal human tissues. Overlap sizes between all pairs of gene sets were also recorded. Box-plots and density distribution curves for both the former ('Observed') and latter ('Control') sets of shared expression universe sizes are displayed in Figure 5.3a and 5.3b respectively.

Firstly, it is observed that distribution of observed overlap sizes is shifted away from that of the simulated overlap sizes: the former is centred on a median value of ~25% of genes on the hgu133A array, which is nearly twice the median of the latter distribution (~12%). A paired t-test comparing these distributions yielded a p-value of $< 2 \times 10^{-16}$, indicating that the difference between the means of these distributions is highly significant. This is most likely to have occurred due to the highly non-random selection of genes into the expression universes of the 68 tissues, as was observed in Figure 5.3.

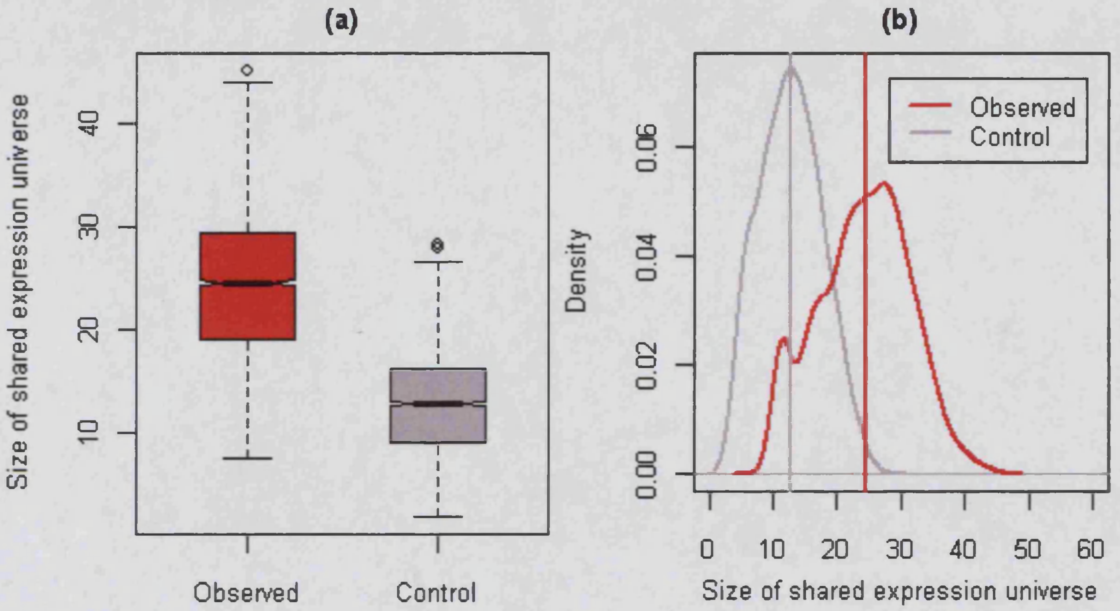


Figure 5.3 Sizes of expression universes shared across 68 normal human tissues. Box-plots (a) and density curves (b) of distributions of the sizes of overlaps between expression universes observed for 68 normal human tissues from the GNF Expression Atlas ('Observed') and from a simulation comparing universes created by selecting randomly from all the genes on the hgu133a array, of the same sizes as the observed expression universes ('Control'). Sizes are represented in terms of percentage of genes on the hgu133a array (Y-axis in (a) and X-axis in (b)). Broken vertical lines in (b) represent medians of the respective distributions.

Secondly, it is observed that the overlap sizes of gene expression universes are variable: the simulated overlap sizes have a standard distribution of 4.9% of genes on the array, while that of the observed overlap sizes is even greater (7.2%).

These observations are as expected because of the sensitivity of overlap sizes to the sizes of the expression universes. To demonstrate this, the overlap sizes were observed in relation to the sizes of the pairs of expression universes being compared. For this purpose, the overlap sizes from both distributions were plotted against the square root of the product of sizes of expression universes (Figure 5.4).

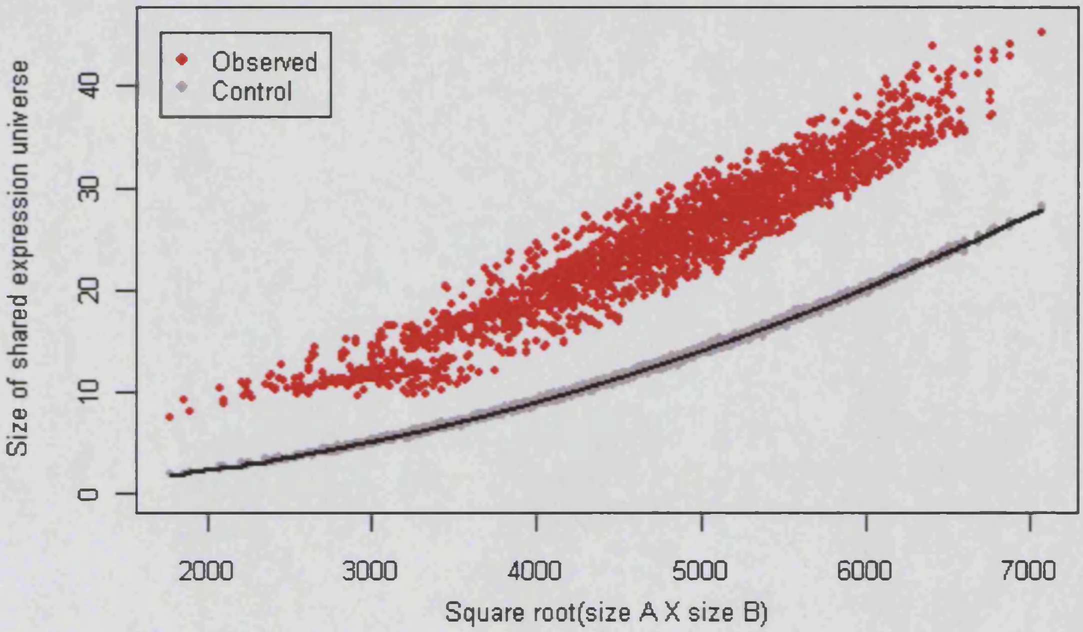


Figure 5.4 Effect of expression universe size on the numbers of genes shared between universes. Points represent observed (red) and simulated (grey) overlap sizes for all pairs of expression universes for 68 tissues of the GNF Expression Atlas. The black line indicates overlap sizes expected between comparisons of universes of those sizes (calculated using the formula described in Section 2.2.2).

Here it is found that both overlap size distributions shown a noticeable positive relationship with expression universe size. The simulated overlap sizes (grey points in Figure 5.4) vary around the overlap sizes that are expected for each comparison of expression universes of those sizes (black line, calculated using the formula described in Section 2.2.2). The entire distribution of observed overlap sizes (red points) is shifted upwards of that of the simulated overlap sizes, reflecting that observed in Figure 5.3. Interestingly, it is also observed that the spread of observed overlap size values is much greater than that for the simulated values. This reflects a greater variability than can be attributed solely to the effects of expression universe size.

5.2.3. Tissue-specificity has a marked effect on gene expression universes

Thus far, it has been observed that there is a considerable level of variability in the number of genes shared by expression universes for each of the 68 tissues in the GNF Expression atlas. Some of this variability can be attributed to the marked variability of the sizes of the expression universes. Furthermore, overlap sizes may also be sensitive to biological effects, such as tissue-specific gene expression; for example, the expression universes from two breast cancer experiments may have more genes in common than those shared between universes from a breast cancer experiment and a colon cancer experiment.

To investigate this requires a methodology that would, in an unbiased manner, find patterns of gene expression to allow creation of groups of tissues that have more similarity within their expression universes than with others. One such method to achieve this could be hierarchical clustering (see Materials and Methods), which would allow unsupervised classification of the 68 tissues under investigation, on the basis of some measure of the similarity of their gene expression universes.

The issue then arises of which measure of similarity could be used for hierarchical clustering. The absolute size of overlaps between expression universes would not be a good choice since, as has been shown in Section 5.2.2, these are strongly influenced by the sizes of expression universes being compared. Hypergeometric Z-scores are another option, as they would represent standardized effect sizes where the systematic effects of expression universe sizes have been accounted for. However, the non-normal and non-random gene expression frequencies described in Section 5.2.1 might well cause violations of the assumptions of the hypergeometric distribution, and result in Z-scores that are sensitive to the effects of expression universe sizes, as has been seen in the comparisons of experimentally-derived genelists (see Section 3.3.3).

To account for the systematic effects of expression universe sizes on the size of their overlaps, an empirical sampling-based strategy was therefore adopted to create a measure of similarity that could be used for the purposes of hierarchical clustering. From each of the 68 different expression universes, which comprise the genes that are expressed in each of the 68 different tissues, a set of 1000 genes was selected randomly and without replacement. All pair-wise comparisons were performed and the sizes of overlap between the sets of genes were then recorded. This was repeated 1000 times, and the median size of overlaps for each pair-wise comparison was then used to perform hierarchical clustering of the 68 tissue-types.

Figure 5.5 shows the results of the clustering. As can be observed there are at least three major clusters that comprise primarily of similar tissue-types. These are a cluster of tissues that are of neuronal origin (including the non-neuronal tissue derived from the pituitary, a neuro-endocrine gland), a cluster of tissues from the testes, and a cluster based on cells from blood (many of which are involved in immunity functions).

Further analyses showed that the distributions of overlap sizes observed between gene-expression universes derived from tissues belonging to any of these clusters were centred on medians that were greater than the medians of distributions of overlap sizes from comparisons of these tissues with all other tissues that did not fall into their respective clusters (Figure 5.6). Unpaired t-tests were then used to assess the statistical significance of the difference of means between the distributions, and all three clusters yielded highly significant p-values (see legends in Figure 5.6).

These findings suggest that the numbers of genes shared between the expression universes of different tissues is affected not just by the systematic effects of variability of expression universe sizes, but also by biological factors such as tissue-specific gene expression patterns.

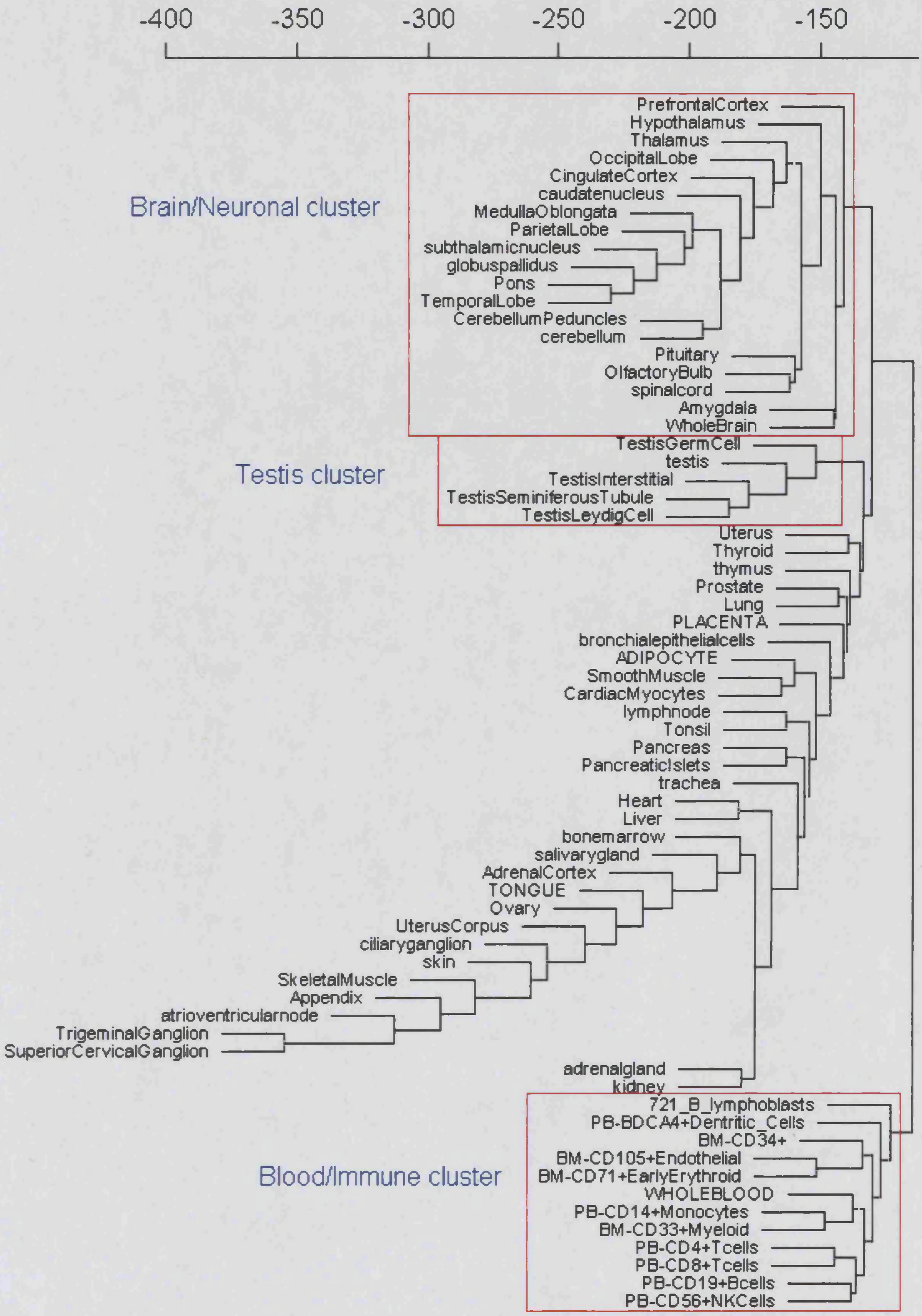

-400
-350
-300
-250
-200
-150
Brain/Neuronal cluster
PrefrontalCortex
Hypothalamus
Thalamus
OccipitalLobe
CingulateCortex
caudatenucleus
MedullaOblongata
ParietalLobe
subthalamicnucleus
globuspallidus
Pons
TemporalLobe
CerebellumPeduncles
cerebellum
Pituitary
OlfactoryBulb
spinalcord
Amygdala
WholeBrain
Testis cluster
TestisGermCell
testis
TestisInterstitial
TestisSeminiferousTubule
TestisLeydigCell
Uterus
Thyroid
thymus
Prostate
Lung
PLACENTA
bronchialepithelialcells
ADIPOCYTE
SmoothMuscle
CardiacMyocytes
lymphnode
Tonsil
Pancreas
PancreaticIslets
trachea
Heart
Liver
bonemarrow
salivarygland
AdrenalCortex
TONGUE
Ovary
UterusCorpus
ciliaryganglion
skin
SkeletalMuscle
Appendix
atrioventricularnode
TrigeminalGanglion
SuperiorCervicalGanglion
adrenalgland
kidney
Blood/Immune cluster
721_B_lymphoblasts
PB-BDCA4+Dentritic_Cells
BM-CD34+
BM-CD105+Endothelial
BM-CD71+EarlyErythroid
WHOLEBLOOD
PB-CD14+Monocytes
BM-CD33+Myeloid
PB-CD4+Tcells
PB-CD8+Tcells
PB-CD19+Bcells
PB-CD56+NKCells

Figure 5.5 Hierarchical clustering 68 different tissues from the GNF Expression Atlas. Clustering was carried out by using, as distances, the negative median overlap size shared between lists of 1000 genes selected randomly from the expression universes of each tissue carried out 1000 times for each possible pair of tissues. Linkage of clusters was performed using the McQuitty method.

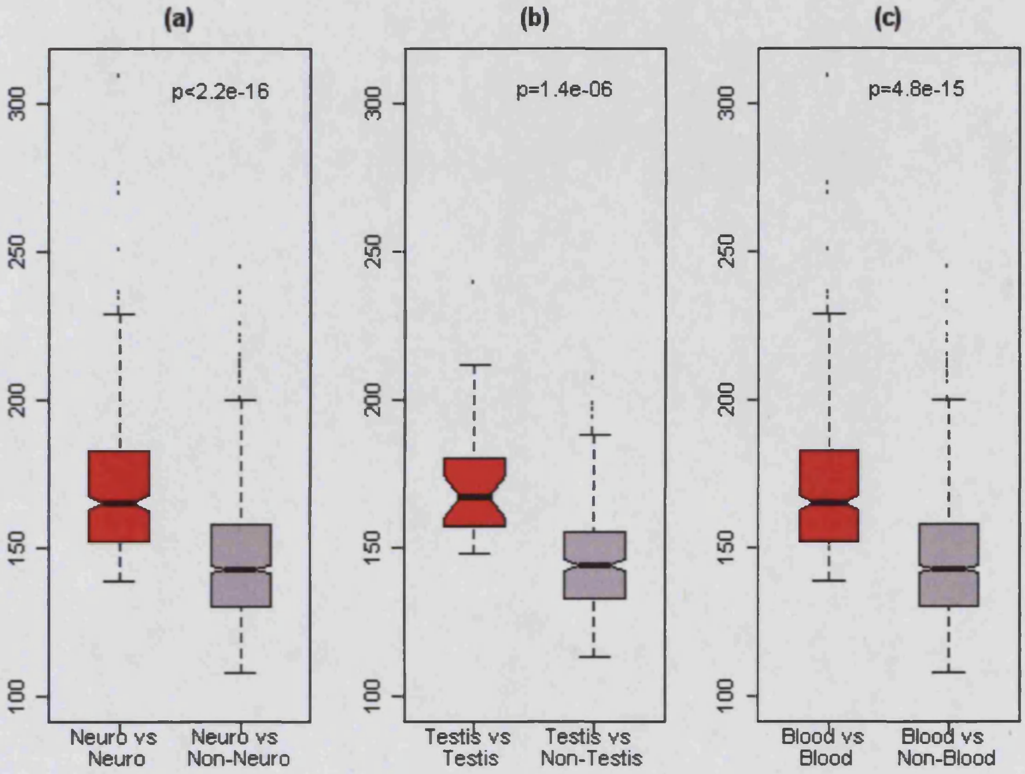


Figure 5.6 Similar tissues share more expressed genes than dissimilar tissues. X-axes represent median overlap sizes shared between lists of 1000 genes selected randomly from the expression universes of each tissue carried out 1000 times for each possible pair of tissues. Box-plots represent overlap sizes derived from comparisons between members of the same cluster (red), and from comparisons between members of the same cluster and all other tissues (grey), for the three clusters of similar tissues derived from hierarchical clustering (see Figure 5.5): (a) neuronal tissues, (b) testis tissues and (c) blood-derived tissues. Legends show p-values from unpaired t-tests comparing each pair of distributions.

5.2.4 Simulating the effects of using an average gene universe size

The biased urn model described in Chapter 4 provides a potential ad hoc solution to the use of all genes represented on an array as the gene universe when using the hypergeometric distribution to assess the similarity between a pair of genelists, by the estimation of a gene universe size representing the 'average' number of genes that is shared by the gene universes of any two experiments. The estimation of the numbers of genes shared between the expression universes of each possible pair of the 68 tissues of the GNF Expression Atlas described in the preceding section of this chapter then provides an opportunity to explore the effects using an estimated average number of genes shared between expression universes to compare genelists.

For this purpose, it was desired to carry out comparisons of simulated genelists where the effect of genelist lengths was controlled for; this would allow for observation of the sole influence of gene universe size on the significance metrics calculated from the hypergeometric distribution (i.e. Z-scores).

One possible strategy is the creation of lists of randomly selected genes from the expression universes of each of the 68 tissues of the GNF Expression Atlas, and comparison of all possible pairs of genelists. However, an accurate assessment of the significance of the overlap between any pair of genelists would require using, as the gene universe, those genes that shared between the expression universes for those tissues that the genelists were sampled from. This would then involve removal, from both genelists, of any genes that are not present within that gene universe. Such filtration of genelists could cause variability in the sizes of the genelists being compared, which is undesirable.

For these reasons, simulations were carried out using the following strategy: the overlap between the expression universes for each possible pair of the 68 tissues in the GNF Expression Atlas were considered separately as gene universes, from which two genelists of equal length (500 genes) were selected randomly and without replacement. The observed as well as the expected overlap size between these genelists was then recorded; the latter value was calculated using the formula described in Section 2.2.2.

Using these overlap sizes, three sets of hypergeometric Z-scores could be calculated, each using a different gene universe size. The first set was calculated by using the 'true' sizes of gene universes from which each pair of genelists was sampled (i.e. the overlap of expression universes for each pair of tissues), which was different for each comparison; this was labelled as Set A. To simulate Z-scores as would be calculated using the biased urn model, the second set (Set B) was calculated using the same gene universe size for each comparison. The size used here was the average size of overlap between the expression universes for all pairs of tissues; this was ~25% of genes on the array. The third set was calculated by using the same gene universe size for each comparison, but this time using the entire number of genes present on the hgu133a array (Set C). Density distributions for all three sets of Z-scores are shown in Figure 5.7.

As can be observed, Z-scores of Set A (black curve) exhibit a distribution that is centred on a median of zero. This is as expected from comparisons of genelists created by random selection of genes. Z-scores of Set C (blue curve) exhibit a distribution that is considerably shifted away from a median of zero. This distribution is centred on a median of 14. This can be thought of as representing the shifted Z-score distributions (indicating high levels of similarity), which were observed for comparisons of experimentally-derived genelists in Chapter 3. Z-scores of Set B (red curve) can be thought to represent Z-scores calculated using the biased urn model described in Chapter 4. This distribution is similar to that obtained by using 'true' gene universe sizes, i.e. Set A in that, it is also centred on a median of zero. However it has a very different shape. To further investigate the effect of gene universe size on these three sets of Z-scores that were all derived using the same set of overlap sizes, all three sets were plotted against the 'true' gene universe sizes for each comparison (see Figure 5.7)

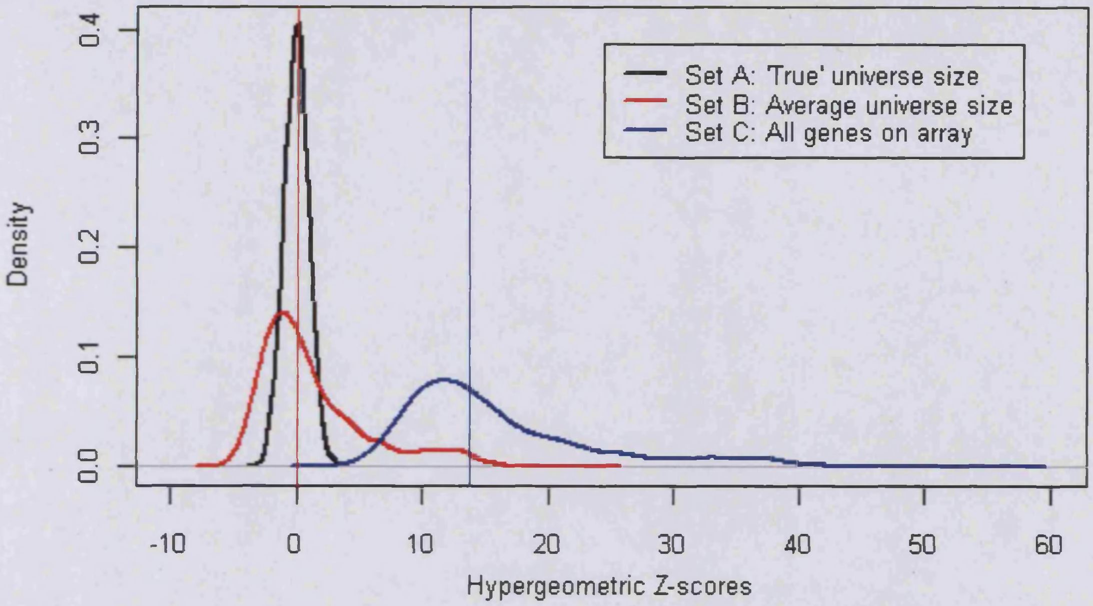


Figure 5.7 Effect of different universe sizes on hypergeometric Z-scores for the same set of comparisons (I). Density curves of Z-scores derived from comparisons of pairs of genelists (of equal lengths) selected randomly and without replacement from each gene-expression universe shared between 68 normal human tissues, calculated using three different universe sizes (see figure key). Vertical lines represent medians of the respective distributions.

As can be observed in Figure 5.7, the expected Z-scores calculated using the 'true' gene universe sizes appear to be unaffected by gene universe size: they are zero for all comparisons (black line), and the observed Z-scores from Set A vary around them (grey points). This is as expected, considering all the comparisons involved genelists that were created by random selection of genes. Z-scores from Set B were, like those from Set A, also found to also be centred on a median on zero in Figure 5.6. However, as can be observed in Figure 5.7, these exhibit a noticeable negative relationship with gene universe size (red points). While Set B Z-scores are similar to Set A Z-scores when the 'true' gene universe size is equal to the estimated 'average' gene universe size (vertical green line), increasing differences between the 'true' gene universe sizes and the estimated 'average' gene universe size appears to cause increasing deviation of Set B Z-scores from the distribution of Set A Z-scores.

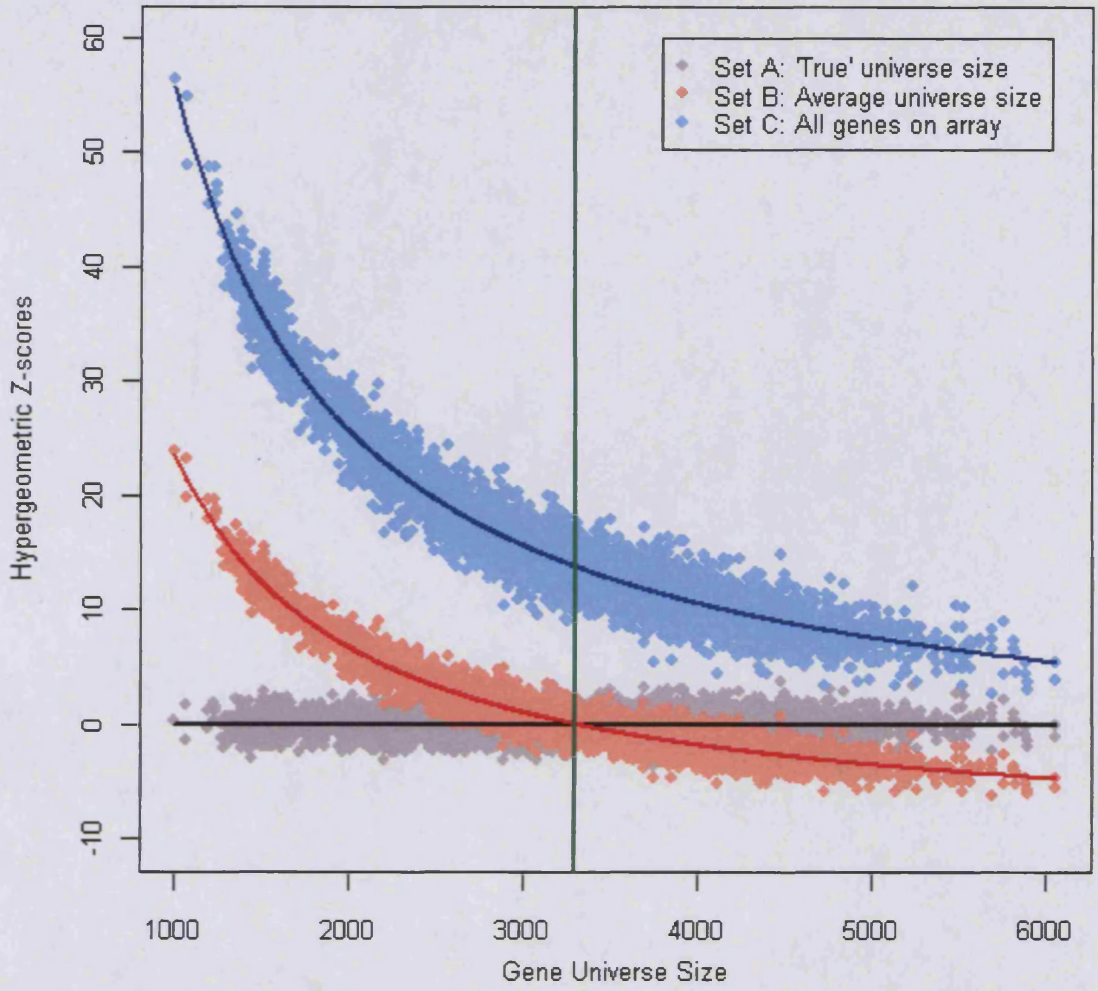


Figure 5.8 Effect of different universe sizes on hypergeometric Z-scores for the same set of comparisons (II). Points represent Z-scores derived from comparisons of pairs of genelists (of equal lengths) selected randomly and without replacement from each set of genes shared between the expression universes for all pairs of 68 normal human tissues, calculated using three different universe sizes (see figure key). Black red and blue lines represent expected Z-scores for the respective distributions. The vertical green line represents the average size of gene-expression universes shared between any pair of the 68 tissues.

The distribution of Set C values is shifted further up from that of the Set A and Set B distributions (reflecting the shift observed in Figure 5.7). It also exhibits a negative relationship with gene universe size, similar to that of the Set B distribution.

5.3 Discussion

The explorations described in Chapter 4 indicated that the 'biased urn' model could be used to control for possible violations of assumptions of the hypergeometric statistical test, when used for comparisons of genelists derived from microarray-based experiments. It involved the ad hoc estimation of an average gene universe size representing the number of genes on an array that could be selected into both genelists being compared. While this model could, in theory, mitigate much of the undesirable effects of using the entire set of genes represented on an array as the gene universe for all comparisons, one particular issue with such a technique is the use of a single gene universe size for all comparisons. This could then lead to erroneous results if the true gene universe sizes vary widely from the estimated average size.

This chapter then described explorations of this issue using a subset of the GNF Expression Atlas dataset (Su et al. 2004), which consisted of the microarray gene expression profiles for a wide range of normal human tissues. This dataset was created in the same laboratory, using the same microarray platform for all samples. Thus, the results of investigations carried out on this dataset would not be subject to the effects of cross-platform and cross-laboratory issues. The expression universes for each tissue (i.e. the genes that were found to be expressed in each tissue) were considered to be estimates of the gene universes for experiments that would be carried out on those tissues. This is because the first criterion in the selection of important genes from an experiment (such as DEGs) is to assess whether those genes are expressed. The overlap between the expression universes for any pair of tissues could then be considered to be the gene universes for the comparison of genelists derived from experiments carried out on those tissues. The findings described are summarized as follows:

- There is considerable variation in the numbers of genes shared between the expression universes of any pair of tissues. This is related to the variability in the number of genes that are expressed for each of the tissues.

- Aside from systematic reasons, such as those mentioned above, the numbers of genes that are shared by the expression universes are also subject to biological factors, particularly tissue-specific gene expression: the gene universe for the comparison of genelists from a pair of experiments carried out on the same tissue-type is likely to be larger than that for two experiments carried out on different tissue-types.
- Simulations were carried out to observe significance values assigned to comparisons of genelists created using a range of differently-sized gene universes, but where the statistical model used for all comparisons the average gene universe size. It was found that such a strategy could produce erroneous results; the magnitude of error would depend on the magnitude of the difference between the true gene universe size for a comparison and the estimated average gene universe size.

Thus while the biased urn model can reduce the excess levels of similarity that may be observed when comparisons of genelists are carried out using a gene universe comprising of all genes on an array, it is still an unsatisfactory methodology due to the use of a single gene universe size for comparisons involving widely ranging universe sizes. In terms of the broader question of whether comparisons of genelists could be carried out as a less resource-intensive alternative to comparison of entire microarray datasets (investigated over Chapters 2-5), it can be concluded that genelists alone do not provide all the information that is required for an accurate comparison of experiments.

The investigations also indicate potential problems in how gene universes are defined for ORA techniques in general (for example, to assess enrichment of GO terms in genelists) (Khatri and Draghici 2005). Here, the presence of genes in the universe for an experiment is a binary concept: they are either present or absent. However, a more realistic and possibly more accurate representation is that of a continuum of probabilities reflecting the likelihood of genes to be selected as interesting (depending on the levels and variability of expression). This is further discussed in Chapter 9.

Part B:

Gene Set Discovery (GSD) - a novel methodology for unsupervised threshold-free discovery of biological themes within microarray datasets

Chapter 6: Gene Set Discovery (GSD): Unsupervised identification of relevant biological themes within microarray datasets

6.1 Introduction

As described in Section 1.4, biological interpretation of data from microarray-based experiments has been aided significantly by the development of gene set analysis (GSA) methodologies. These techniques utilize electronically archived biological knowledge which is available on public databases such as Gene Ontology (Ashburner et al. 2000), KEGG (Kanehisa et al. 2004) and BioCarta (BioCarta 2005). Using these techniques, researchers have been able to identify biological themes (such as pathways or processes) that may be of interest within a particular experiment.

Many of the popular tools to carry out GSA have typically involved over-representation analysis (ORA), which seeks to identify enrichment of biological themes within lists of differentially expressed genes (DEGs) (Huang da et al. 2009; Khatri and Draghici 2005; Rivals et al. 2007). These GSA techniques have been termed 'threshold-based', as they require prior definition of threshold values to identify DEGs. For example, in an experiment to identify genes that are differentially expressed between two classes of samples using a t-test, a p-value threshold of <0.05 may be used. More recently, there has been development of 'threshold-free' GSA techniques that do not require creation of lists of DEGs (Huang da et al. 2009; Nam and Kim 2008). One of the most popular of these, Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005), uses instead a list of all genes on the array, ranked according to their correlation to a pre-selected expression pattern (for example, up-regulation in one class of samples and down-regulation in another class), and then tests for the enrichment of gene-sets towards the top and bottom of the list.

Thus, GSA methods have most often been used after carrying out 'supervised' analyses of microarray datasets, i.e. researchers have knowledge of the sample classes, and select DEGs or rank genes on the basis of pre-selected expression patterns across these known sample classes. However, GSA methods have found little utility in exploratory analyses such as those involving 'class discovery' (see Section 1.3.2). Such analyses are of particular importance in studies of cancer, where sample classes are often not known *a priori*, or where morphology-based classification methods have not successfully resolved sample classes (Alizadeh et al. 2000; Ivshina et al. 2006).

This chapter thus explores the possibility of developing a methodology that could allow identification of biologically relevant themes within a microarray dataset, without requiring prior definition of sample classes. The investigations described focus particularly on the use of hierarchical clustering techniques within 'heatmaps'. This chapter will introduce and explore some of the underlying concepts, and subsequent chapters will describe the application of this methodology to several datasets.

6.2 Results and Explorations

Heatmaps (together with hierarchical clustering) are popular tools to provide visual representations of microarray data and to display gene expression patterns (Eisen et al. 1998). They typically comprise of a matrix where the columns represent samples and the rows represent genes. Each cell is coloured, based on the expression value of that gene in that sample, such that the colour indicates whether the gene is up or down-regulated (for example, compared to the experiment-wide mean), and the intensity of the colour indicates the extent of up- or down-regulation. In analyses of microarray data, usually those involving the identification of a set of genes defined as being of interest by the researcher (e.g. a list of DEGs), it is routine to create a heatmap using those genes as intuitive visual evidence that those genes exhibit the expression pattern of interest.

While clustering and heatmap visualisation have typically been used as 'end stage' tools to represent in visual terms the sets of genes that have been determined to be relevant within any particular experiment, these tools can also be used for *de novo* knowledge discovery. For example, manual inspection of the resultant heatmaps from an unsupervised clustering and visualisation of the data for an experiment for each of the many gene sets that represent biological themes and processes may then allow for visual identification of those gene sets that exhibit expression patterns that are of interest to the researcher. Such a process, although common, is somewhat unsatisfactory due to the element of manual inspection and pattern identification that it involves and the absence of any underpinning statistical methodology.

The investigations described here thus focussed on the identification of a metric that could allow some quantification of the levels of information within a heatmap, and that may therefore be used for identification of gene sets that may be of interest.

6.2.1 Quantification of the information content of a heatmap

6.2.1.1 Hierarchical clustering of gene and samples within heatmaps

Key to the visualization of expression patterns within heatmaps is the process of hierarchical clustering that brings together those genes and samples that exhibit similar expression patterns.

To demonstrate this, a hypothetical gene expression matrix comprising of 20 samples (represented as matrix columns) and 500 genes (represented as matrix rows) was created. To simulate DEGs within the matrix, 25 genes were randomly selected to exhibit up-regulation in 10 randomly selected samples and down-regulation in the other randomly selected 10 samples. Another 25 genes were randomly selected to exhibit the opposite expression pattern in terms of up- and down-regulation in the same samples as the first set of genes.

To represent $\log_2$ median-centred gene expression values, the matrix was initially populated with values sampled randomly from a normal distribution having a zero mean and standard deviation of 0.3 (see Materials and Methods). This provides a very simple data structure to model what is typically seen in a microarray experiment by way of noise. Values in cells representing up-regulated gene expression values were then replaced with values randomly selected from a normal distribution with a mean of 2 (that is, a 4-fold up-regulation) and standard deviation of 0.3. Similarly, values in cells representing down-regulated gene expression values were then replaced with values randomly selected from a normal distribution with a mean of -2 and standard deviation of 0.3. Figure 6.1a represents a heatmap of this hypothetical gene expression matrix without any hierarchical clustering, i.e. the orders of genes and samples are the same as when it was created. Figure 6.1b represents a heatmap of the same matrix after hierarchical clustering of genes and samples. As can be observed, the expression patterns are clearly discernible in Figure 6.1b.

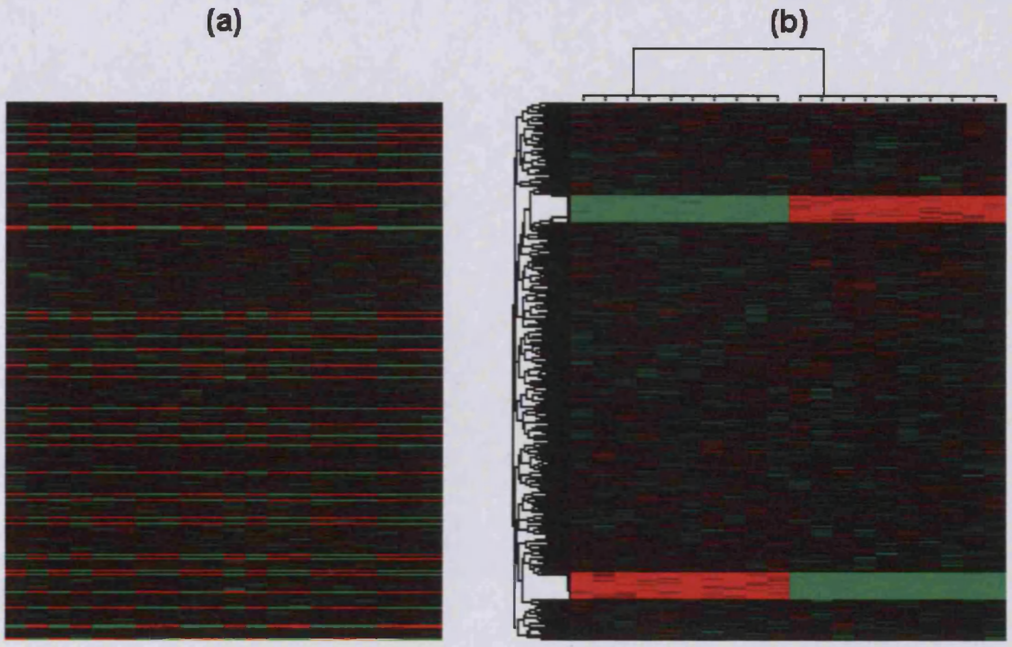


Figure 6.1 Visualization of gene expression patterns through hierarchical clustering. Heatmaps of an artificial gene expression matrix where (a) no hierarchical clustering of genes or samples has been carried out and (b) both genes and samples have been clustered. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Hierarchical clustering was carried out using correlation distance and average linkage.

The presence of expression patterns shared by many genes (which reflects differential expression of genes) within a heatmap indicates that information is contained within the gene set represented in that heatmap. Such an 'informative' gene set (i.e. one that contains DEGs) is likely to be of interest to a researcher. As hierarchical clustering is able to identify expression patterns that are shared between genes and samples (and thereby brings similar genes and samples together), efforts to identify a metric that can quantify information content within the expression matrix for a gene set focussed on this aspect.

6.2.1.2 The effect of information content on gene and sample distance distributions

Hierarchical clustering requires the calculation of 'distances' between each possible pair of genes and each possible pair of samples. These values represent the level of dissimilarity between the expression profiles of each pair of genes or samples (see Materials and Methods). Thus the distance value between a pair of genes or samples that exhibit similar patterns of expression would be lesser in magnitude relative to the distance value for a pair of genes or samples that exhibit dissimilar expression patterns. To investigate the sensitivity of these values to the presence of information content within a gene set, the following simulations were carried out:

First, a hypothetical gene expression matrix comprising of 30 samples (columns) and 500 genes (rows) was created such that it contained no DEGs. Distance distributions for genes and samples were then recorded. A second matrix was created, identical to the first but with one DEG showing an expression pattern of up-regulation in 10 randomly selected samples, and down-regulation in another 10 randomly selected samples. The matrices were populated with gene expression values representing unchanged expression, up- and down-regulation using distributions similar to those used in simulations described in Section 6.2.1.1. A series of matrices was similarly created, by converting increasing numbers of genes into DEGs. Gene and sample distance distributions were recorded for each of these matrices.

Heatmaps of some of these matrices are displayed in Figure 6.2 while the distributions of sample and gene distances for those matrices are displayed in Figure 6.3a and Figure 6.3b respectively. Means of the sample distance matrices (M-SDM) and gene distance matrices (M-GDM) for each of all the hypothetical matrices are plotted against the percentage of DEGs in Figure 6.4a as red and blue points respectively. Standard deviations of the sample distance matrices (SD-SDM) and gene distance matrices (SD-GDM) for each of all the hypothetical matrices are plotted against the percentage of DEGs in Figure 6.4b as red and blue points respectively.

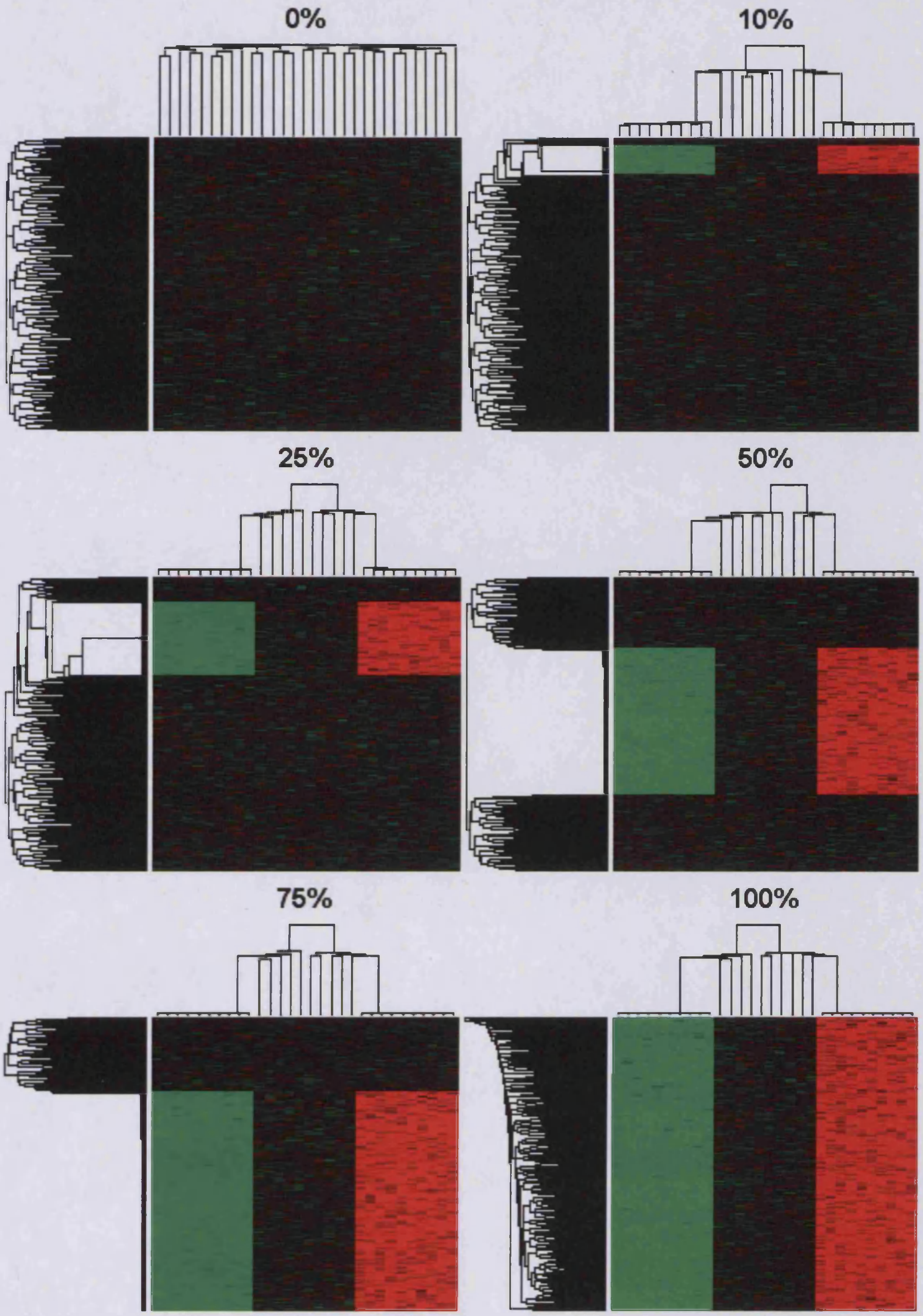

0%
10%
25%
50%
75%
100%

Figure 6.2 Changing levels of information contained within a hypothetical gene expression matrix. Heatmaps represent a series of hypothetical gene expression matrices created with increasing numbers of DEGs. Numbers above the heatmaps indicate the percentage of genes that are DEGs. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Clustering was carried out using correlation distances and average linkage.

As can be observed, in case of the initial matrix with no DEGs (i.e. no information), the distances between all pairs of samples are similar, and the distribution of these distances has a single peak (black line in Figure 6.3a) at ~1. The inclusion of DEGs causes a major change in the shape of the distributions of sample distances. Three peaks are observed: one of the peaks comprises of short distances, and representing distances between samples showing similar expression patterns. This peak approaches a value of 0 (the theoretical minimum correlation distance) as the number of DEGs in the matrices increases. Another peak comprises of long distances, and represents distances between samples showing opposite expression patterns. This peak approaches a value of 2 (the theoretical maximum correlation distance) as the number of DEGs in the matrices increases. The third peak corresponds to the median distance of ~1, and represents distances from samples exhibiting no patterns of differential expression, both to all other samples as well as themselves. These observations demonstrate that increasing the number of DEGs results in the 'strengthening' of sample clusters, in that the intra-cluster distances decrease and inter-cluster distances increase.

As can be seen in Figure 6.4a, increasing the number of DEGs in the expression matrices has no effect on the M-SDM values: these remain stable at ~1. However, the variability of these distributions (SD-SDM) increases continuously, as can be see in Figure 6.4 b.

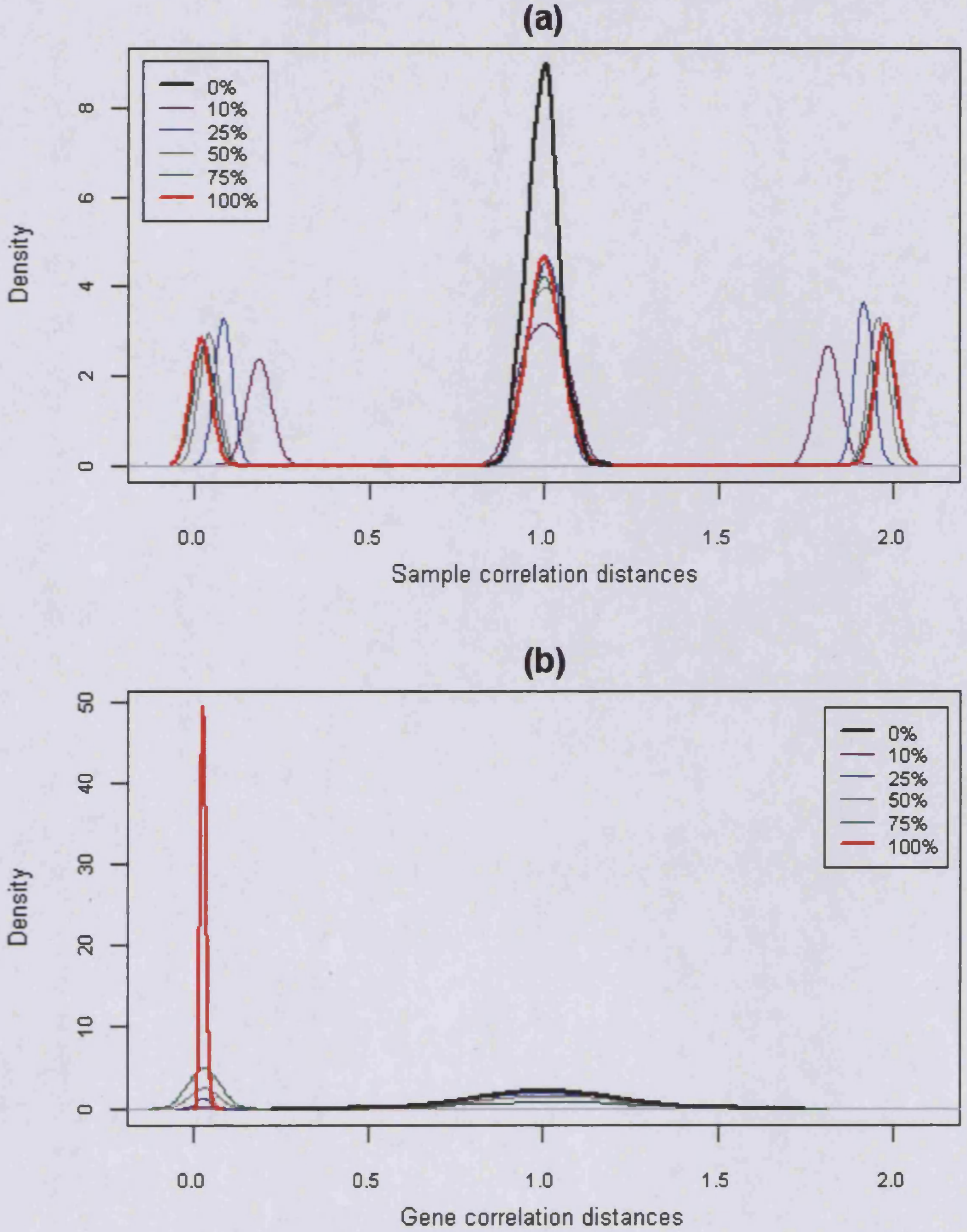


Figure 6.3 Effect of information levels on sample and gene distance distributions (I). Curves represent (a) sample and (b) gene distance distributions derived for each of the hypothetical gene expression matrices represented as heatmaps in Figure 6.2.

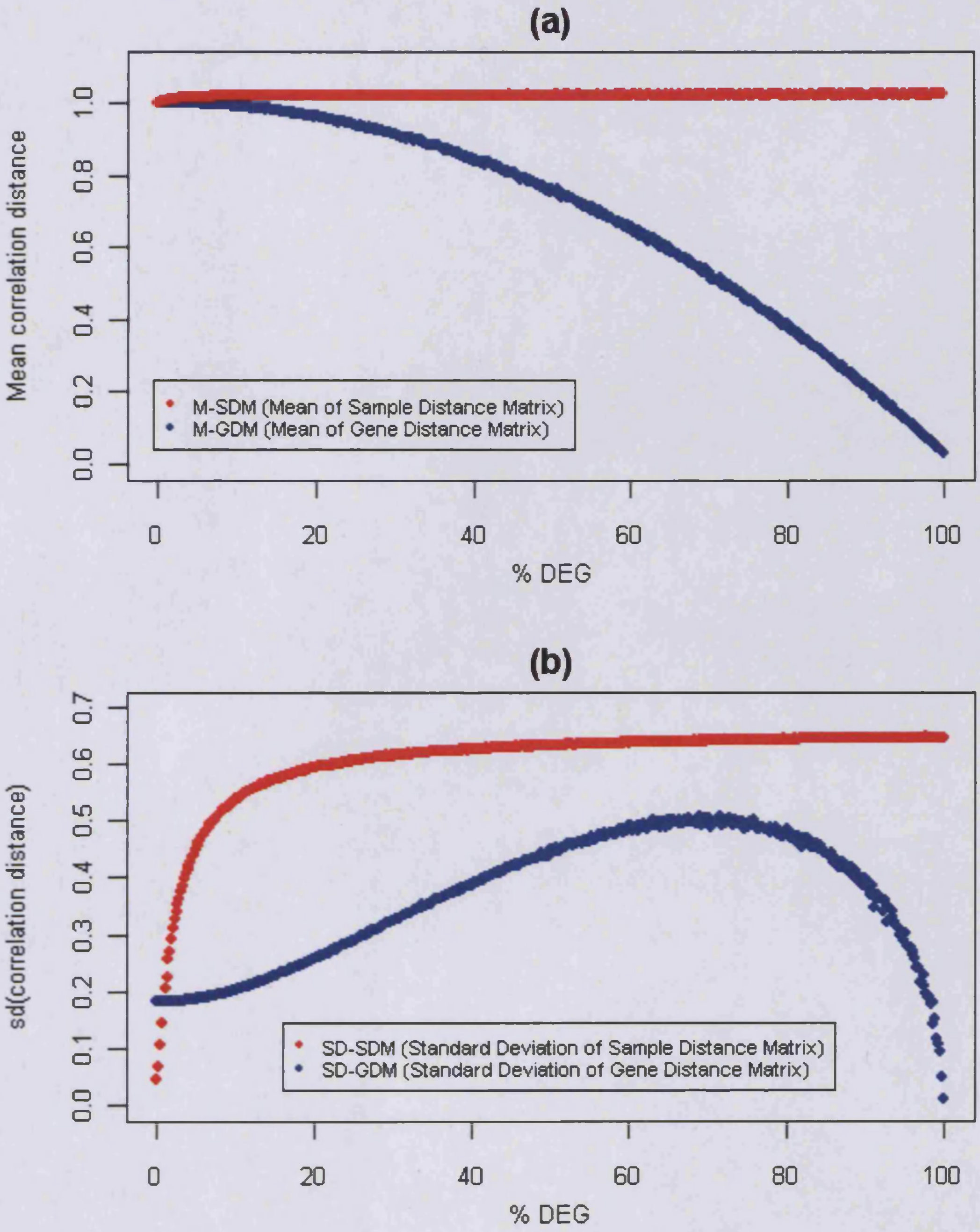


Figure 6.4 Effect of information levels on sample and gene distance distributions (II). Points represent (a) mean and (b) standard distribution values for gene and sample distance distributions of a series of hypothetical gene expression matrices created with increasing numbers of DEGs, some of which are represented as heatmaps in Figure 6.2.

The gene distance distribution in the first matrix (i.e. with no DEGs) is similar to the sample distance distribution for it, having a single peak centred on a mean of ~1, indicating similar levels of similarity/dissimilarity between all genes (black curve in Figure 6.3b). The introduction of DEGs again changes the shape of the distribution: two peaks are now observed. One of the peaks is centred on a distance of ~0, and presumably comprises of distances between the DEGs themselves. The other peak, centred on ~1 presumably comprises of distances from non-DEGs, both between themselves and to the DEGs. As would be expected, as the percentage of DEGs increases, the first peak increases in height, while the second decreases until no non-DEGs remain and all distances are ~0.

As can be observed Figure 6.4a (blue points), the M-GDM values of these distributions start from ~1, when there are no DEGs, and decrease continuously as the proportion of DEGs is increased. The variability of these distributions (SD-GDM) start low, and increase as the proportion of DEGs is increased to a particular point from where they decrease as the proportion of DEGs is increased (blue points in Figure 6.4b).

This observed sensitivity of distance distributions to the introduction of information (in the form of DEGs) within expression matrices raises the possibility that that attributes of these distributions (such as their mean values or variability) could be used to indicate the levels of information (i.e. proportion of DEGs) within these matrices. M-GDM and SD-SDM values exhibited relationships with the proportion of DEGs that were non-linear; but as these were unidirectional, both these metrics could be considered as possible candidates to indicate information levels in an expression matrix. For example, it may be possible to deduce that a gene set contains greater levels of information than another gene set, if the former yields a higher SD-SDM value or a lower M-GDM value. M-SDM values were not investigated further due to their apparent stability to changes in information levels. SD-GDM values were also disregarded in further analyses due to directional changes within their relationship with information levels.

6.2.1.3 Presence of more than one pattern within a gene expression matrix

The explorations described above identified SD-SDM and M-GDM values as metrics that could potentially be used to indicate levels of information with gene expression matrices. These explorations were carried out using simulations to observe the effects of changing the levels of information (i.e. the proportion of DEGs) on gene and sample distance distributions.

Another issue that required exploration was the nature of information within a gene expression matrix, and what effects this could have on the metrics under investigation. In the simulations described in the previous section, all DEGs were created to exhibit the same expression pattern. However, it is possible that there may be two or more groups of DEGs within a gene expression matrix, each of which exhibit different patterns of expression.

To illustrate this, a hypothetical gene expression matrix was created, comprising of 30 samples and 500 genes and labelled as 'HypMat1'. Half of these genes were randomly selected to represent a first group of DEGs, all of which showed the same expression pattern. This pattern was of up-regulation in the first 10 samples and down-regulation in the next 10 samples, and labelled as 'ExPat1'. The matrix was populated with gene expression values representing unchanged expression, up- and down-regulation using distributions similar to those used in simulations described in Section 6.2.1.1. Two identical copies of HypMat1 were then created, called 'HypMat2' and 'HypMat3'.

In HypMat2, the other half of genes were made to represent a second group of DEGs. On this second group, an expression pattern was imposed, that was different to ExPat1, but *corresponded to the same grouping of samples*. This pattern, labelled ExPat2, was of up-regulation in those samples which exhibited down-regulation in ExPat1, and vice-versa.

In HypMat3, these genes were also made to represent a second group of DEGs. However, the pattern of expression imposed on these genes, ExPat3, *corresponded to a different grouping of samples* from what was observed in ExPat1 (and ExPat2). This was achieved by randomly selected 10 samples to exhibit up-regulation, and another 10 randomly selected samples to exhibit down-regulation. Heatmaps of HypMat2 and HypMat3 are displayed in Figure 6.5a and 6.5b respectively.

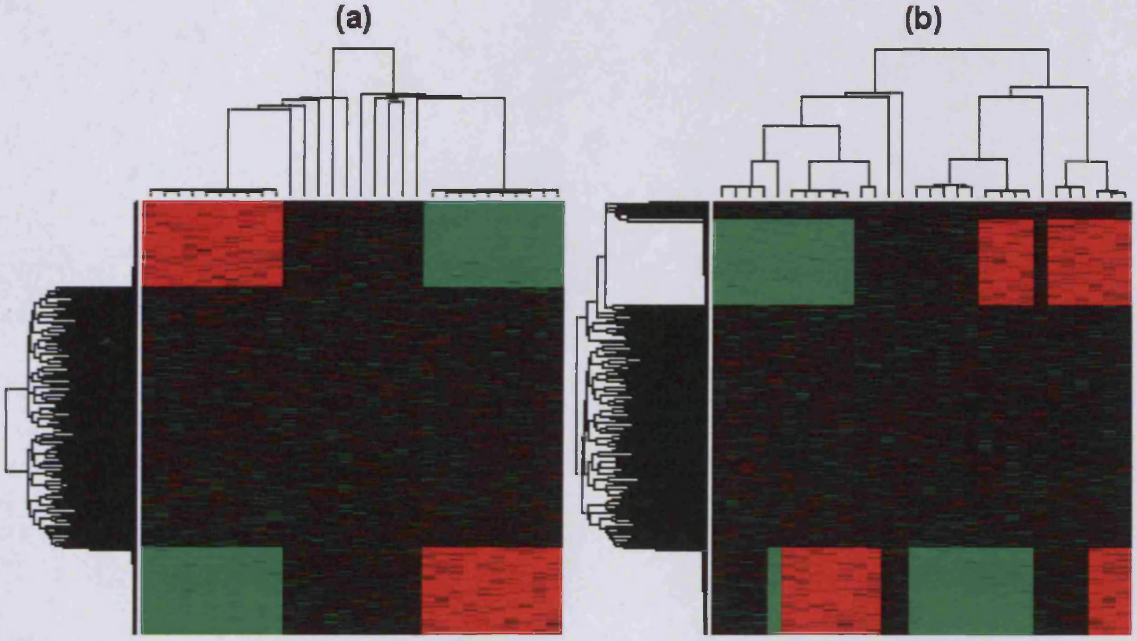


Figure 6.5 Presence of different expression patterns within hypothetical gene expression matrices. Heatmaps represent hypothetical gene expression matrices where all genes are differentially expressed. In both matrices, these comprise two equally sized groups, each exhibiting different expression patterns. In (a) both expression patterns correspond to the same grouping of samples. In (b) the expression patterns correspond to different groupings of samples. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Clustering was carried out using correlation distances and average linkage.

While both HypMat2 (Figure 6.5a) and HypMat3 (Figure 6.5b) have similar levels of information (i.e. the same number of DEGs), it can be argued that a researcher is likely to be more interested in a gene set that can be represented by HypMat2 rather than HypMat3. This is because it represents a single coherent expression profile (in terms of grouping of samples) that can be attributed to the biological theme represented by that gene set.

To observe the effect of changing expression patterns within a gene expression matrix on M-GDM and SD-SDM values, while the total level of information (i.e. number of DEGs) is constant, the following simulations were carried out: first, a hypothetical gene expression matrix comprising of 500 genes and 30 samples, was created such that 50% (i.e. 250) genes were DEGs, and all of which exhibited the expression pattern ExPat1. The matrix was populated with gene expression values representing unchanged expression, up- and down-regulation using distributions similar to those used in simulations described in Section 6.2.1.1. Two series of ten matrices were then created.

One of the series (labelled 'Series A') was created to represent changing expression patterns within a gene expression matrix but where the type of information (i.e. the grouping of samples) remained the same. This was carried out by changing the pattern of expression to ExPat2 of each of the DEGs, 25 genes at a time, till all DEGs exhibited ExPat2. Heatmaps of some of these matrices are displayed in Figure 6.6.

The other series of matrices (labelled 'Series B') was created to represent changing expression patterns within a gene expression matrix, each of which represented a different type of information (i.e. each expression resulted in a different grouping of samples). This was carried out by changing the pattern of each of the DEGs, 25 genes at a time. However for each set of 25 genes, an expression pattern was imposed that did not correspond to that of ExPat1 (and ExPat2).

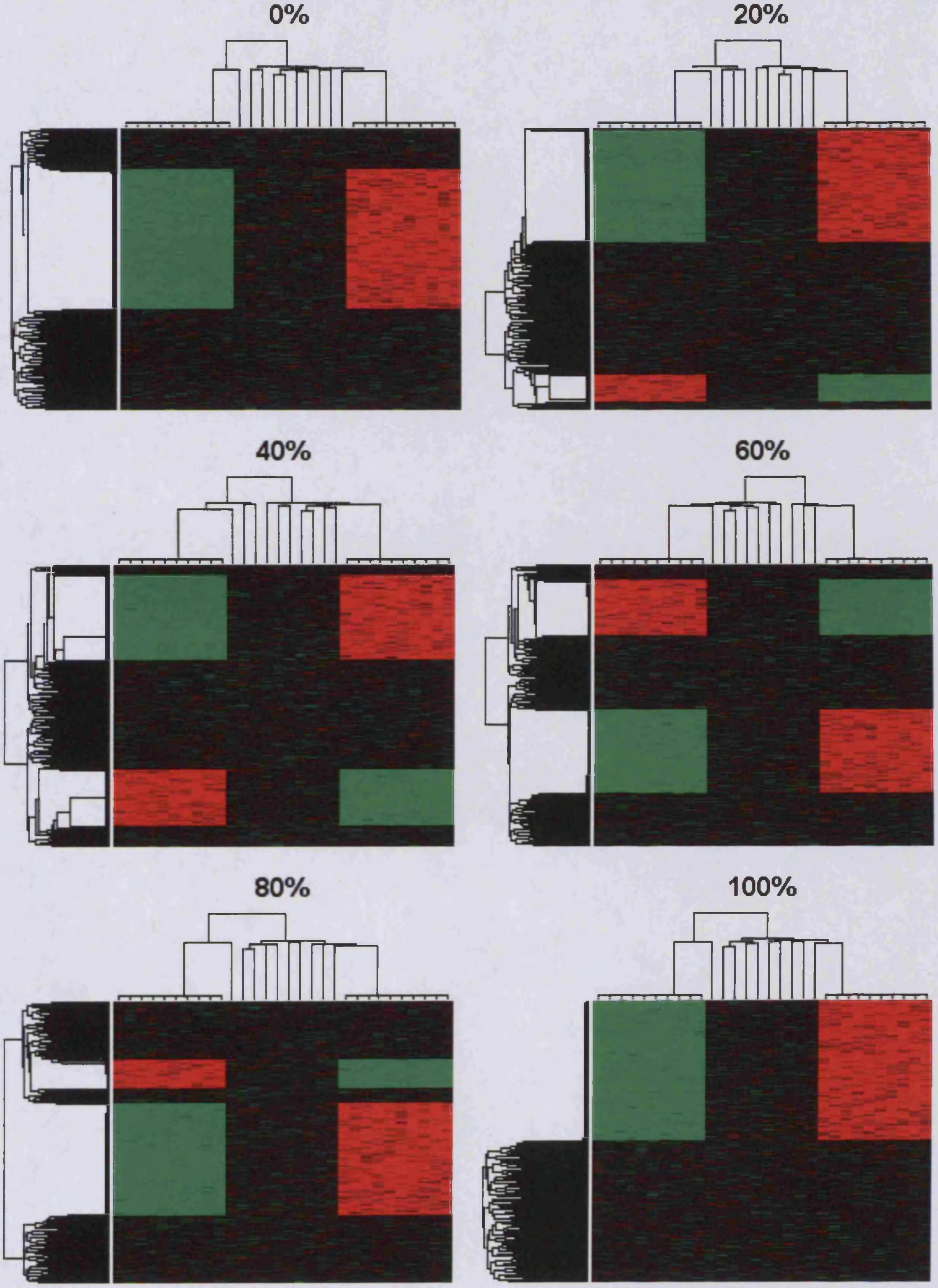

0%
20%
40%
60%
80%
100%

Figure 6.6 Changing expression patterns of DEGs without changing types of information within hypothetical gene expression matrices. Heatmaps represent some of a series of 10 hypothetical gene expression matrices (Series A), starting from one where 250 out of 500 genes are DEGs, all of which exhibit the same pattern of expression. Subsequent matrices were created by changing the expression of 25 DEGs at a time, to show an expression pattern that is different to the first, but one that results in the same clustering of samples. Numbers above heatmaps represent the percentage of DEGs that exhibit the second expression pattern. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Clustering was carried out using correlation distances and average linkage.

This was carried out by randomly selecting 10 samples to exhibit up-regulation, and randomly selecting another 10 samples to exhibit down-regulation, separately for each group of 25 genes. This was continued till a matrix was created where the 250 DEGs consisted of ten equally sized groups (of 25 genes each), each of which exhibited expression patterns that were different from ExPat, and also resulted in different classifications of samples. Heatmaps for some of these matrices are displayed in Figure 6.7.

M-GDM can SD-SDM values were recorded for each of both series of matrices. M-GDM values for Series A and Series B are plotted against the percentage of DEGs with expression patterns that are different from ExPat as red and blue lines respectively in Figure 6.8a. SD-SDM values for Series A and Series B are similarly plotted as red and blue lines respectively in Figure 6.8b.

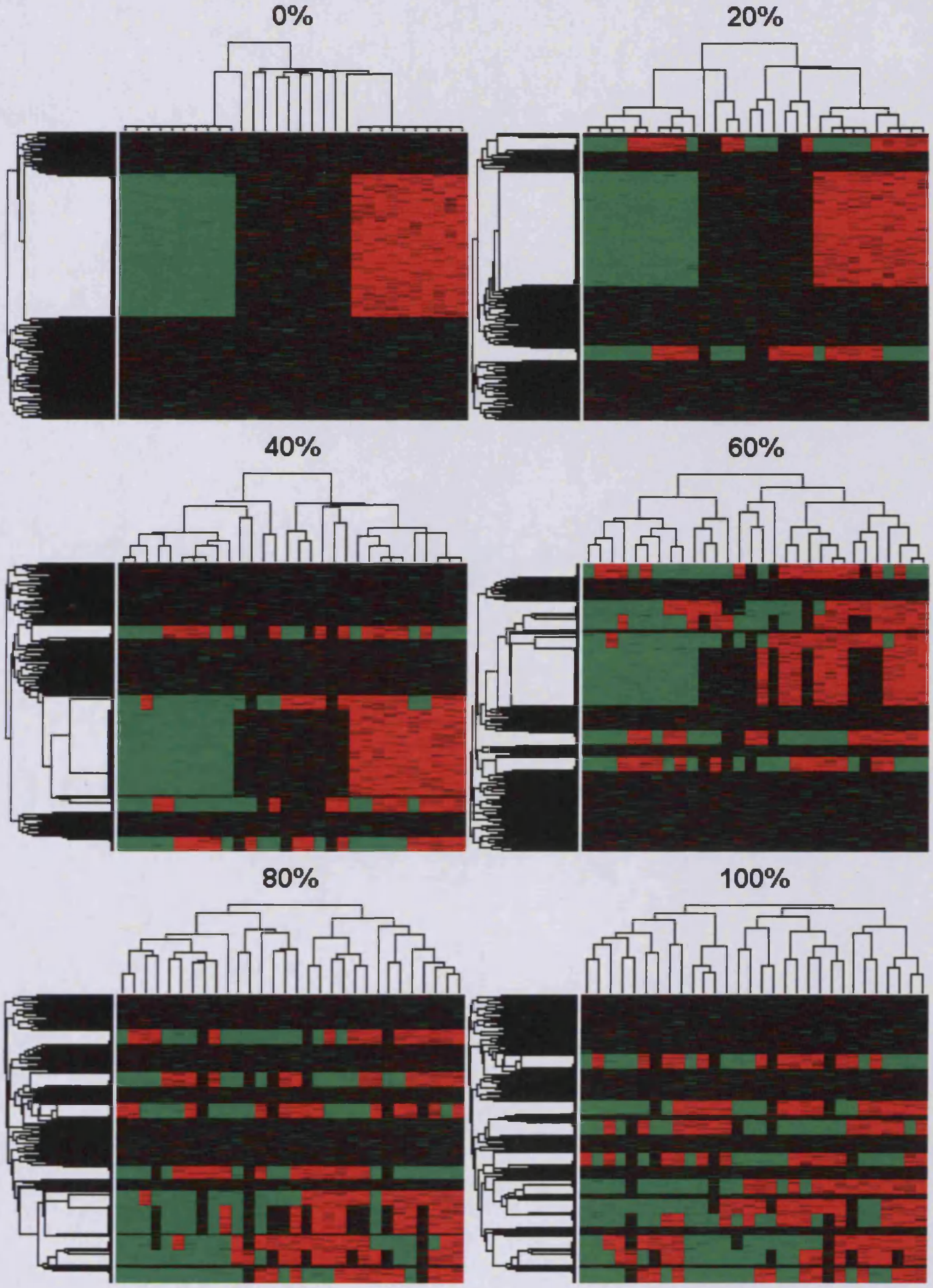

0%
20%
40%
60%
80%
100%

Figure 6.7 Changing expression patterns of DEGs as well as the types of information within hypothetical gene expression matrices. Heatmaps represent some of a series of 10 hypothetical gene expression matrices, starting from one where 250 out of 500 genes are DEGs, all of which exhibit the same pattern of expression. Subsequent matrices were created by changing the expression of 25 DEGs at a time, to show an expression patterns that are different to the first (and each other), and that result in different clustering of samples every time. Numbers above heatmaps represent the percentage of DEGs that exhibit the expression patterns other than the first. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Clustering was carried out using correlation distances and average linkage.

Thus, as is apparent from Figure 6.8, the presence of different expression patterns affects both M-GDM and SD-SDM values even though all matrices contain the same levels of information (all matrices have 250 DEGs). M-GDM values for both series of matrices show increases upon introduction of expression patterns that are different to ExPat (Figure 6.8a). This is can explained as follows: in the starting matrix where all DEGs exhibited ExPat1, distances between all pairs of DEGs were ~0 (the minimum possible correlation distance). In subsequent matrices, the distances between DEGs exhibiting different expression patterns results in replacement of those values with distances >0. The increase is much greater in Series A because ExPat2 represents the diagonally opposite expression pattern of ExPat1, and thus the distances between pairs of genes that exhibit ExPat1 and ExPat2 are ~2 (the maximum possible correlation distance). The M-GDM values for Series A increase till reaching their zenith when equal numbers of genes exhibiting either expression pattern are present. Thereon the values drop in magnitude, as the number of genes exhibiting ExPat2 increases, back down to around the starting value, for the matrix where all DEGs exhibit ExPat2. However, in Series B, the M-GDM values continuously increase, because every successive matrix has an additional expression pattern, reaching their zenith when 10 patterns are present, each of which is exhibited by equal numbers of genes (25 genes each).

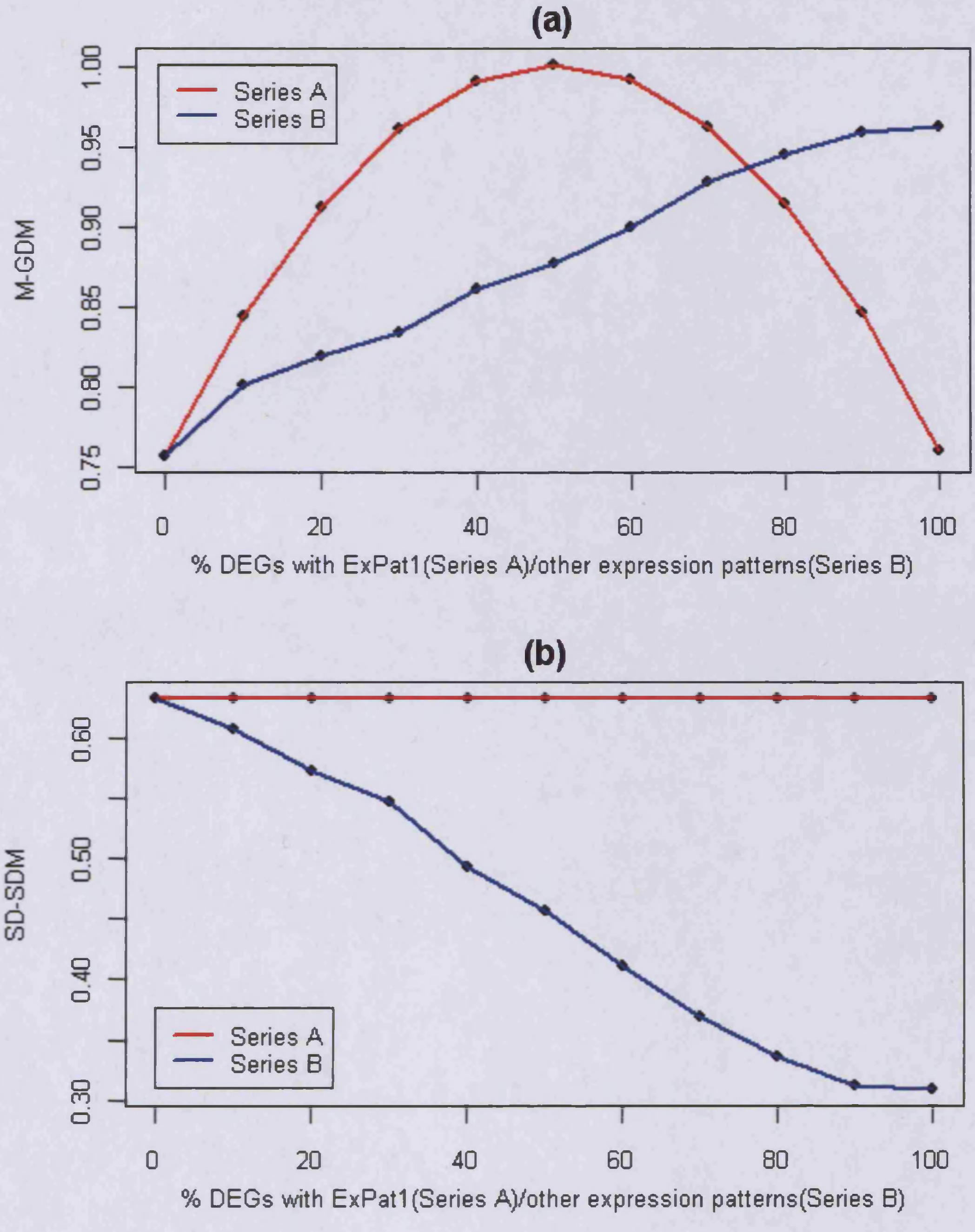


Figure 6.8 Effect of changing expression patterns of DEGs within a gene expression matrix. Points represent (a) M-GDM and (b) SD-SDM values calculated for two series of matrices, some of which are displayed in Figures 6.6 (Series A: same information) and 6.7 (Series B: different information).

As can be observed in Figure 6.8b, SD-SDM values, for Series A appear to be very similar to that of the starting matrix implying that they are not affected by increasing the numbers of genes exhibiting ExPat2. This is presumably because ExPat2 corresponds to the same groupings of samples as ExPat1. However, SD-SDM values show continuous decrease in Series B.

As mentioned earlier, a researcher may not just be interested in those gene sets that contain information, but also the expression patterns represented by that information. In particular, the researcher would be more interested in a gene set where the expression pattern(s) of all DEGs result in the same grouping of samples (i.e. they represent the same type of information) such as those represented in Series A, than in a gene set where different expression patterns can group samples differently (i.e. they represent different types of information) such as those represented in Series B. This is because the former case provides researchers with a simple, coherent relationship between the biological theme represented by that gene set, and the resultant classification of samples. The presence of several types of sample stratification schemes within a gene-set could be of lesser utility and interest to the researcher.

Thus, it is desirable that a metric is not sensitive to the presence of different expression patterns, if they all represent the same type of information (i.e. correspond to the same groups of samples). Similarly the metric should be sensitive to the presence of expression patterns that represent different types of information (i.e. correspond to different groupings of samples). For these reasons, M-GDM values were disregarded for further analyses, which focussed solely on the use of SD-SDM values.

6.2.2. Identification of possible confounding factors

Explorations described in the previous sections identified SD-SDM values as a metric that can potentially identify those gene sets that contain information (i.e. DEGs) particularly if the expression pattern(s) within the expression matrix represent the same type of information (i.e. correspond to the same groups of samples). However, another aspect that must be explored prior to their potential usage as tools for GSA is whether SD-SDM values are also subject to systematic effects of other factors, which may need to be controlled for.

6.2.2.1 Effect of gene set size on the distribution of SD-SDM values

When testing a collection of gene sets for their relevance within any particular microarray experiment, the number of samples involved would be constant for all tests – thus, their influence on SD-SDM values would be uniform across all tests and might not require to be controlled for, as this would not affect the levels of SD-SDM values relative to each other. However, the sizes of gene sets tested could vary greatly. To explore the effect of gene-set size on SD-SDM values, the following simulations were carried out.

A hypothetical gene expression matrix, representing an entire microarray dataset comprising of 30 samples and 10,000 genes was first created. As it was desired to observe the effect of gene-set size alone (i.e. without any possible confounding effects of the presence of information), no genes were made to be DEGs. The matrix was populated with values sampled randomly from a normal distribution with a zero mean and standard deviation of 0.3. A series of gene sets was then created by randomly selecting from amongst those represented in the matrix, ranging in length from 20 to 10,000 genes (i.e. all genes in the matrix), in increments of 20 genes.

SD-SDM values were calculated for each gene set. These values are plotted in Figure 6.9a against the size of the gene sets. In Figure 6.9b, log SD-SDM values are plotted against logs of the gene set sizes.

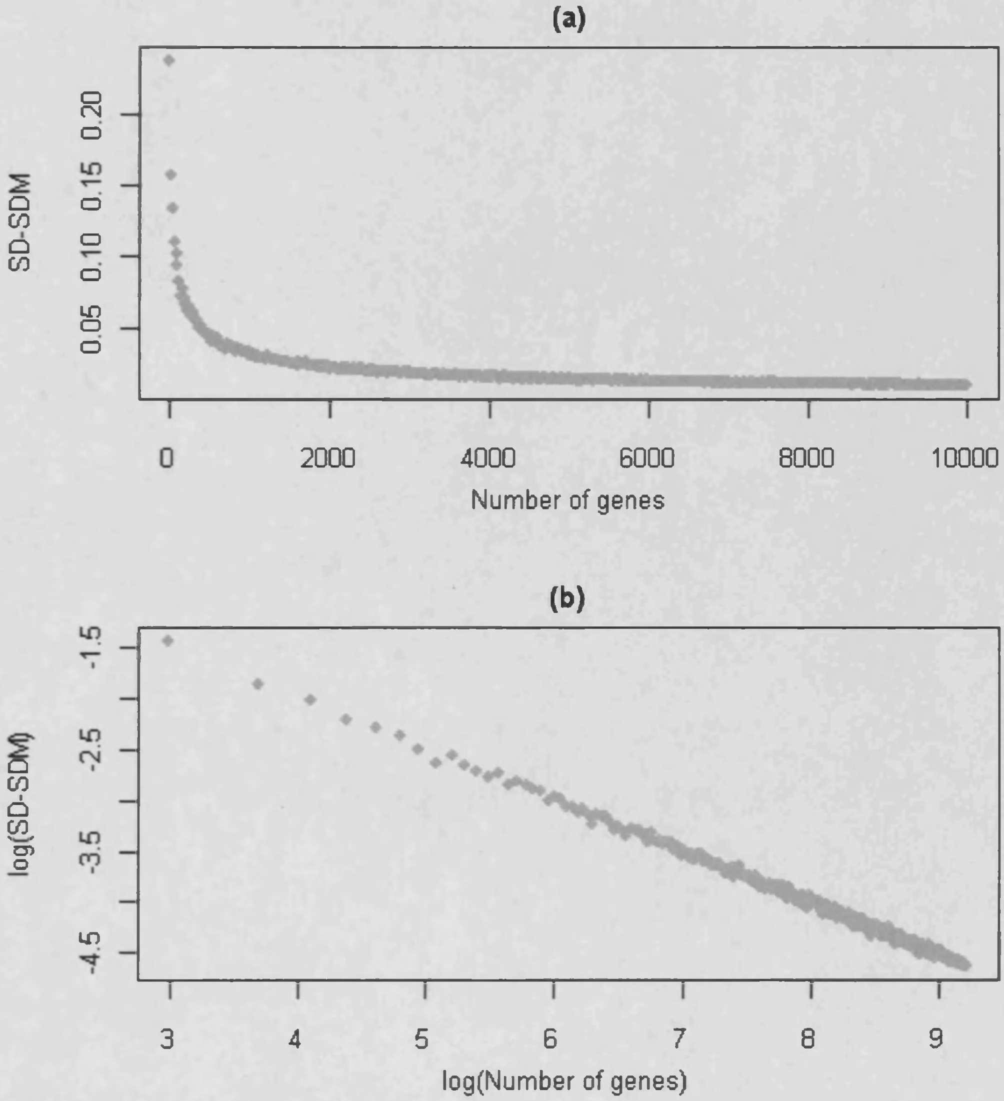


Figure 6.9 Effect of gene-set size on SD-SDM values in the absence of information. Data represented was derived from a series of matrices created by randomly selecting a series of gene-sets of various sizes from a hypothetical gene expression matrix containing no information. In (a) SD-SDM values are plotted against the number of genes. (b) represents the same data as (a) with both axes in log scale.

As can be observed in Figure 6.9a, SD-SDM values decrease with increase of gene set size. This relationship is linear when SD-SDM values and gene set sizes are logged (Figure 6.9b). This systematic sensitivity of SD-SDM values to gene set size thus requires to be controlled for prior to use of SD-SDM values in GSA analysis.

6.2.2.2 Effect of random selection of informative genes on the distributions of SD-SDM values

If one assumes a linear relationship between log SD-SDM values and log gene-set size (as indicated by figure 6.9b), this can then allow for relatively simple and resource-efficient control of the systematic effect of gene-set size on SD-SDM values. However this assumption may not be valid when considering gene expression matrices with information content (i.e. with DEGs). Gene-sets sampled from such matrices could contain informative genes by chance alone (for example, if 10% of genes in the entire expression matrix are DEGs, we would expect 10% of any randomly selected set of genes to be DEGs), and these would exert their own influence on SD-SDM values. This could be further complicated if the information comprises of two or more different expression patterns, each of which could classify samples differently.

To observe the relationship between SD-SDM values and gene set size in the presence of information, first a hypothetical gene expression matrix was created comprising of 30 samples and 10,000 genes, none of which were DEGs. Two identical copies of this matrix were created, labelled ExMat1 and ExMat2. 1000 genes (i.e. 10%) from ExMat1 were randomly selected to represent DEGs. An expression pattern of up-regulation in 10 randomly selected samples and down-regulation in another 10 randomly selected samples was imposed on these genes. In ExMat2, 2000 genes (i.e. 20%) were randomly selected to represent DEGs and the same expression pattern was imposed on them. The matrix was populated with gene expression values representing unchanged expression, up- and down-regulation using distributions similar to those used in simulations described in Section 6.2.1.1.

A series of gene sets was then created by randomly selecting from amongst those represented in the matrix, ranging in length from 20 to 10,000 genes (i.e. all genes in the matrix), in increments of 20 genes. Two SD-SDM values were calculated for each gene set: one each from ExMat1 and ExMat2. These have been plotted in Figure 6.10a, against gene set sizes; SD-SDM values from ExMat1 are coloured red while those from ExMat2 are blue. The same data is displayed in Figure 6.10b with both axes in log space.

As can be observed, the relationships between SD-SDM values and gene-set size in the presence of information (Figure 6.10) are markedly different from their relationships in the absence of information (Figure 6.9). In the absence of information, SD-SDM values showed a continuous decrease in value with increase of gene-set size; this relationship was linear when SD-SDM values and gene set sizes were logged. However, in the presence of information much of this length-dependency of SD-SDM values is lost. This occurs presumably because of random selection of DEGs into the gene sets. The difference in the levels of information between ExMat1 (10% of genes are DEGs) and ExMat2 (20% of genes are DEGs) appears to shift the distributions their SD-SDM values, relative to gene set size, away from each other. This is due to the selection of greater numbers of DEGs into gene sets sampled from ExMat2 as compared to ExMat1.

Thus, we find that the SD-SDM values for any gene set are subject simultaneously to the systematic effects exerted by both gene sets size and any DEGs that may randomly be selected into a gene set. While effects of gene set size alone may be controlled easily through mathematical modelling of these effects, controlling for the effects of randomly selected DEGs is more complicated. Such modelling would require prior knowledge of both the number of DEGs, as well as all the different expression patterns exhibited by those genes. For any given dataset, such knowledge would not be available. For this reason, further investigations focussed on development of *ad hoc* methods that could control for both of these sources systematic effects on SD-SDM values.

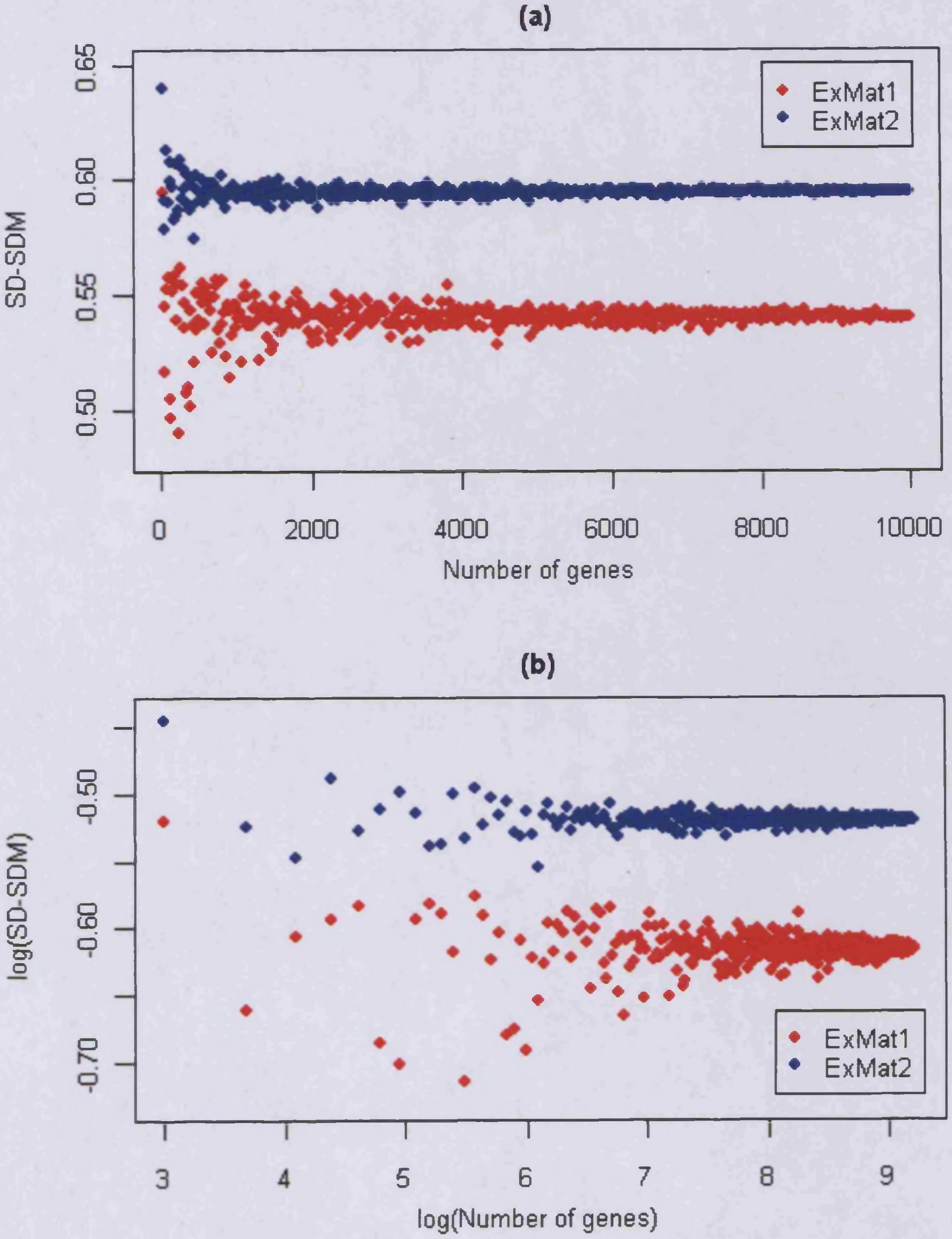


Figure 6.10 Effect of gene-set size on SD-SDM values in the presence of information. Data represented was derived from two series of matrices created by randomly selecting a series of gene-sets of various sizes from two hypothetical gene expression matrices (ExMat1 and ExMat2) containing different levels of information. In (a) SD-SDM values are plotted against the number of genes. (b) represents the same data as (a) with both axes in log scale.

6.2.3. Assessment of significance of SD-SDM values

Explorations described in the previous section showed that SD-SDM values alone may not be satisfactory metrics to identify gene sets that could be of interest to researchers. This is because of the observation that SD-SDM values are sensitive to gene set size and that this relationship is complicated by the possible presence of DEGs in within the gene sets by random chance. Mathematical modelling to control for these effects would require prior knowledge of both the level of information within a microarray dataset (i.e. the number of DEGs), as well as nature of this information (i.e. all possible groupings of samples based on the expression patterns of the DEGs).

Efforts were therefore focussed on an *ad hoc* method that did not require these parameters to be estimated. This involved creation of background distributions of SD-SDM values (i.e. null distributions) with which observed SD-SDM values could be compared, thus allowing for assessment of the significance of the observed SD-SDM values. Two methodologies that could be used for this purpose were identified: the first of these involves a strategy of randomization of values within the expression matrix for any given gene set. Such a strategy can be carried out in three different ways: values can be randomized for each gene (i.e. within each row), for each sample (i.e. within each column), or across both genes and samples. Iteration of this process and recording of the SD-SDM values for each matrix could then create a background distribution of null SD-SDM values with which to compare the SD-SDM value observed for a gene set.

The second strategy is of re-sampling, and involves random selection of gene sets of the same size as the one being tested, from amongst all genes represented in the entire experimental dataset. The background distribution would then comprise of the SD-SDM values for all the expression matrices for these randomly selected gene sets.

Using these background distributions, two measures of significance can be calculated. The first of these is a Z-score, which represents the effect size i.e. the magnitude of the difference between the observed and expected SD-SDM value for gene-sets of a particular size (see Materials and Methods). The second is a p-value, representing the probability that the observed SD-SDM value could have occurred by chance alone.

Both the data randomisation and re-sampling methods for creating background distributions of SD-SDM values have advantages and disadvantages relative to each other. For example, it can be argued that because background distributions derived from the randomization strategy are derived from matrices comprising of the same set of values as that of the expression matrix of the gene set being tested, they may be more comparable to the SD-SDM value observed for that gene-set as compared to those derived from the re-sampling strategy because it involves expression matrices with different sets of values (as they represent different sets of genes). However, as indicated by explorations described in Section 6.2.2.2, DEGs may be present in a gene set simply by chance alone (and not for biological reasons). The randomization strategy may not allow control for their presence (and their influence on observed SD-SDM values) as it involves removal of all structure (brought about by the presence of information) within the expression matrix for a gene set.

To explore the feasibility of using either strategy to create background distributions to assess the significance of observed SD-SDM values, the following simulations were carried out. First, a hypothetical gene expression matrix representing an entire microarray dataset was created, comprising of 30 samples and 10,000 genes, none of which were DEGs. A second matrix was created, identical to the first, except that 2000 genes (i.e. 20% of genes in the matrix) were selected at random to represent DEGs. An expression pattern of up-regulation in 10 randomly selected samples, and down-regulation in another 10 randomly selected samples was imposed on all the DEGs of the second matrix. The matrices were populated with gene expression values representing unchanged expression, up- and down-regulation using distributions similar to those used in simulations described in Section 6.2.1.1.

A series of 100 gene sets, ranging in length from 10 to 1000 genes were then selected at random from both expression matrices. Two SD-SDM values were recorded for each gene-set: one from each parental expression matrix. Background distributions were then created for each gene set using the two candidate strategies identified. Using the re-sampling strategy, for each gene set being tested, 1000 gene sets of the same size were sampled randomly from the entire expression matrix, and their SD-SDM values were recorded for each parental matrix. Three background distributions were created for each test gene set (for each parental matrix) using the randomization strategy, by randomizing expression values only within samples, only within genes and across both genes and samples 1000 times for each gene-set. Z-scores and p-values could then be calculated for each gene-set using each of these distributions.

Table 6.1 shows the number of gene sets assigned significance (i.e. had p-values of <0.05) before and after multiple hypothesis correction using the Benjamini-Hochberg method, for each strategy, and for each parental expression matrix.

Considering that all 100 tested gene sets comprised of randomly selected genes, it is desirable that a useable strategy to assess the significance of the SD-SDM values observed for these gene sets detects little or no significance for them. This appears to be the case for both the re-sampling and randomization strategies when the gene sets were sampled from the matrix that contained no information (i.e. no DEGs): very few gene sets were assigned p-values <0.05, and no gene sets were flagged as significant after multiple hypothesis correction of the p-values using the Benjamini-Hochberg correction. Similar results are obtained from the application of the re-sampling strategy to assess the significance of gene sets sampled from the expression matrix that contained information.

Expression Matrix	**Testing Strategy**		**Number of tests significant at p<0.05 (uncorrected)**	**Number of tests significant at p<0.05 (FDR-corrected)**
Without information	Re-sampling		4	0
	Randomization	Only samples	3	0
		Only genes	2	0
		Genes and samples	2	0
With information	Re-sampling		3	0
	Randomization	Only samples	100	100
		Only genes	100	100
		Genes and samples	100	100

Table 6.1 Numbers of gene sets (out of 100) found to have significant SD-SDM values as assessed by re-sampling and randomization strategies. Gene sets of different sizes were created by random selection of genes from two hypothetical gene expression matrices, one of which contained information (i.e. DEGs), while the other did not.

However, all three versions of the randomization strategy assigned p-values of <0.05 to all test gene sets that were sampled from the expression matrix that contained information. Even multiple-hypothesis correction did not appear to have much impact on these results: all test gene sets were flagged as significant (at $p<0.05$) after FDR-correction of the p-values.

Thus, the re-sampling strategy appears to be insensitive to the presence of information within a gene expression matrix: no significance is detected using this strategy regardless of whether the expression matrix from which gene sets are sampled contained DEGs or not. To explore this, the observed SD-SDM values and background SD-SDM distributions created using the re-sampling strategy for gene sets sampled from the matrix without information (Figure 6.11a) and with information (Figure 6.11 b) were plotted against gene set size.

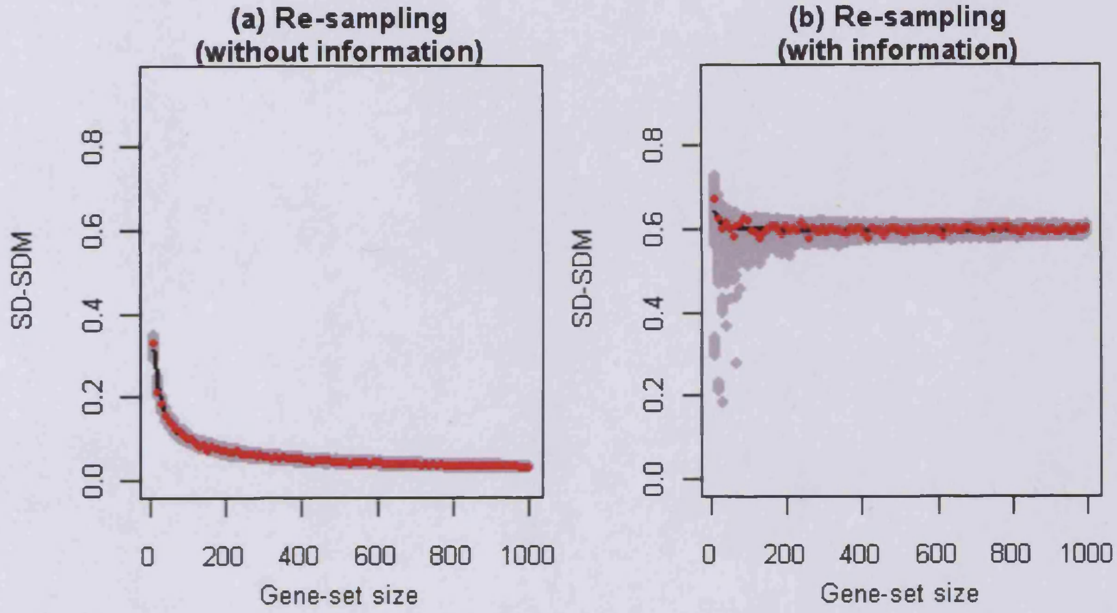


Figure 6.11 Re-sampling based significance testing. Grey points represent background distributions of SD-SDM values derived by re-sampling of gene sets to assess the significance of observed SD-SDM values (red points) for a series test gene sets of various sizes selected randomly from hypothetical matrices (a) without and (b) with information (i.e. DEGs), the results for which are displayed in Table 6.1. Black lines represent the median background SD-SDM value for each size of gene set tested.

As can be observed in Figure 6.11a, the SD-SDM values observed for the test gene sets sampled from the expression matrix with no information content fall well within the background distributions created using the re-sampling strategy. This reflects the absence of any significance detected for these gene-sets. The observed SD-SDM values in Figure 6.11b show a very different distribution to those seen in Figure 6.11a. This is as expected from the explorations described in Section 6.2.2.2, where it was found that when gene sets are randomly sampled from an expression matrix containing information (i.e. DEGs), the DEGs that are randomly selected into the gene sets influence their SD-SDM values. However, the background distribution of SD-SDM values also shows a similar change due to which the observed SD-SDM values again fall within the background distributions. This is because the background distributions were created using gene sets that (like the test gene sets) contained information as well, due to random selection of DEGs into them. The test gene sets were created by random selection of genes and thus contained similar levels of information as those used to create the background distributions. As a result, no significance was assigned to them.

On the other hand, the introduction of information to the expression matrix from which gene sets are sampled randomly changes the results of the randomization strategy radically. To explore this, the observed SD-SDM values for all the test gene sets were plotted against gene set size along with background distributions calculated by randomization of expression values only within genes, only within samples, and across genes and samples in Figures 6.12a, 612b, and 6.12c respectively for gene sets sampled from the matrix with no information, and in Figures 6.12d, 6.12e and 6.12f respectively for gene sets sampled from the matrix containing information.

As can be observed from Figures 6.12a, 6.12b and 6.12c, when gene sets are sampled from expression matrices containing no information, the SD-SDM values observed for them fall well within the background distributions of SD-SDM values: thus, no significance is detected for any of them. However as can be observed in Figures 6.12d, 6.12e and 6.12f, the change in the distributions of observed SD-SDM values for gene sets sampled from the matrix containing information relative to gene set size is not mirrored by changes to the background distributions: these remain similar to as when gene sets were sampled from the expression matrix with no information content.

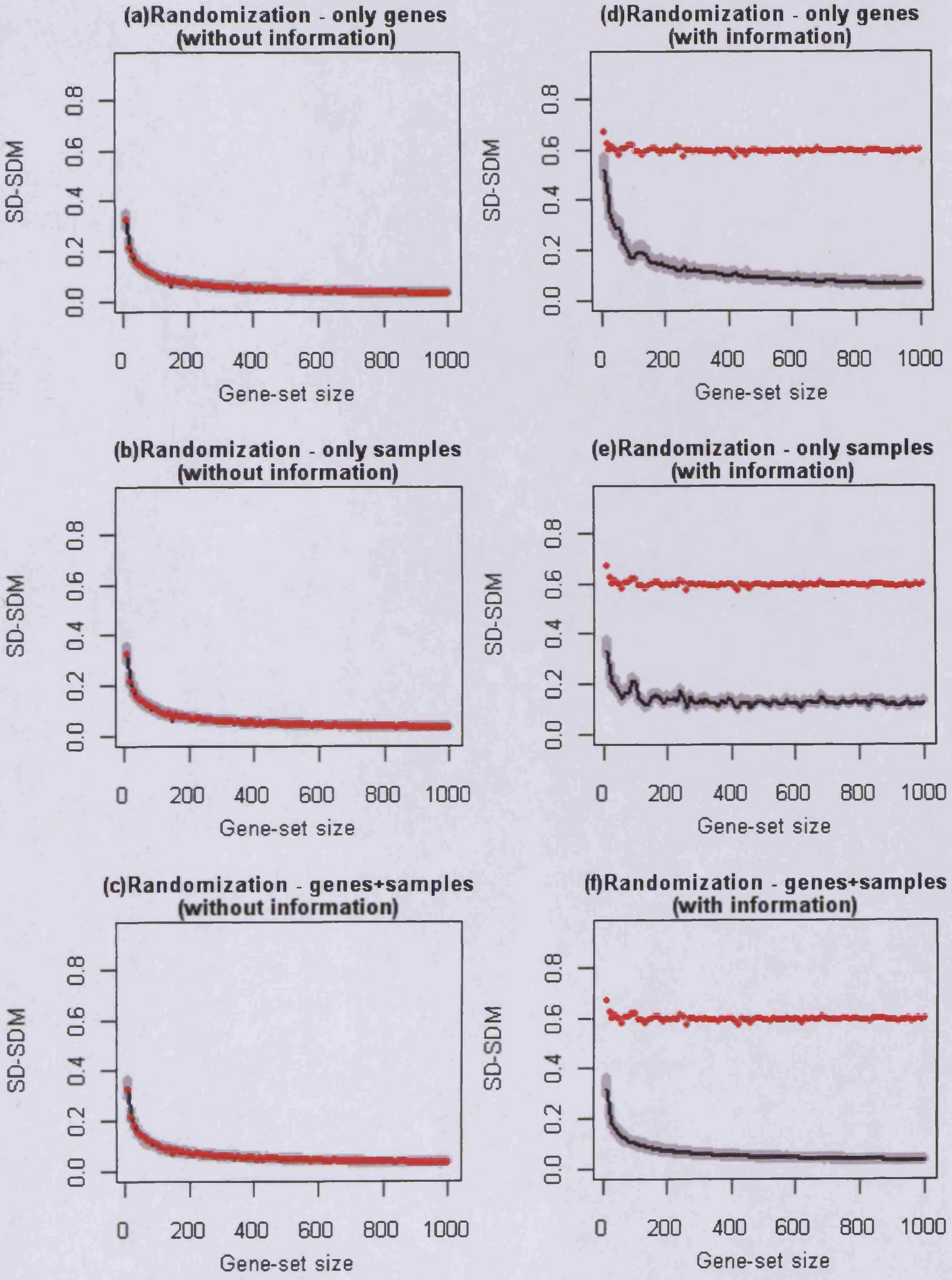

(a)Randomization - only genes
(without information)
(d)Randomization - only genes
(with information)
(b)Randomization - only samples
(without information)
(e)Randomization - only samples
(with information)
(c)Randomization - genes+samples
(without information)
(f)Randomization - genes+samples
(with information)
SD-SDM
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
Gene-set size

Figure 6.12 Randomization based significance testing. Grey points represent background distributions of SD-SDM values derived by randomization of expression values (in three different ways) to assess the significance of observed SD-SDM values (red points) for a series of test gene sets of various sizes selected randomly from hypothetical matrices with and without information (i.e. DEGs), the results for which are displayed in Table 6.1. Black lines represent the median background SD-SDM value for each size of gene set tested.

This is presumably because the randomization strategy to create the background distributions involves the removal of all information, including that represented by randomly selected DEGs. Thus the random selection of DEGs into the test gene sets is not accounted for and as a result all test gene sets are assigned significance.

It is expected that the numbers of DEGs that are randomly selected into gene sets increase linearly in relation to increase of gene set size. Thus it was then desired to investigate if the presence of information within a gene expression matrix could result in the sensitivity of the significance of SD-SDM values (as assigned by the re-sampling and randomization strategies) to gene set size. In Figure 6.13, the Z-scores derived using the re-sampling strategy for test gene sets sampled from the expression matrices with (red points) and without information (grey points) are plotted against gene set size. As can be observed, both distributions are similar: neither set of Z-scores appears to be affected by gene set size. In Figure 6.14, the Z-scores derived using all three types of the randomization strategy for test gene sets sampled from the expression matrices with (red points) and without information (grey points) are plotted against gene set size. As can be observed, for the gene sets sampled from the matrix without information, the Z-scores show no apparent sensitivity to gene set size. However, for gene sets sampled from the matrix with information content, a strongly positive relationship is observed between the Z-scores and gene-set size.

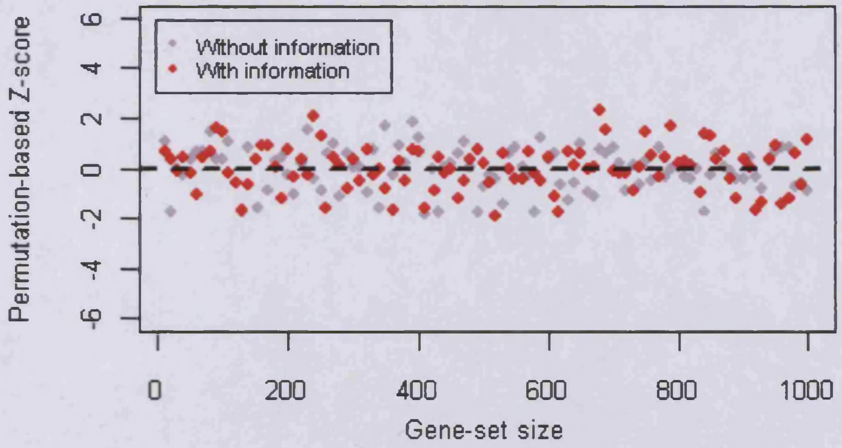


Figure 6.13 Re-sampling based Z-scores. Points represent Z-scores derived by re-sampling of gene sets to assess the significance of observed SD-SDM values for a series test gene sets, the results for which are displayed in Table 6.1. Broken black lines indicate Z-scores of zero.

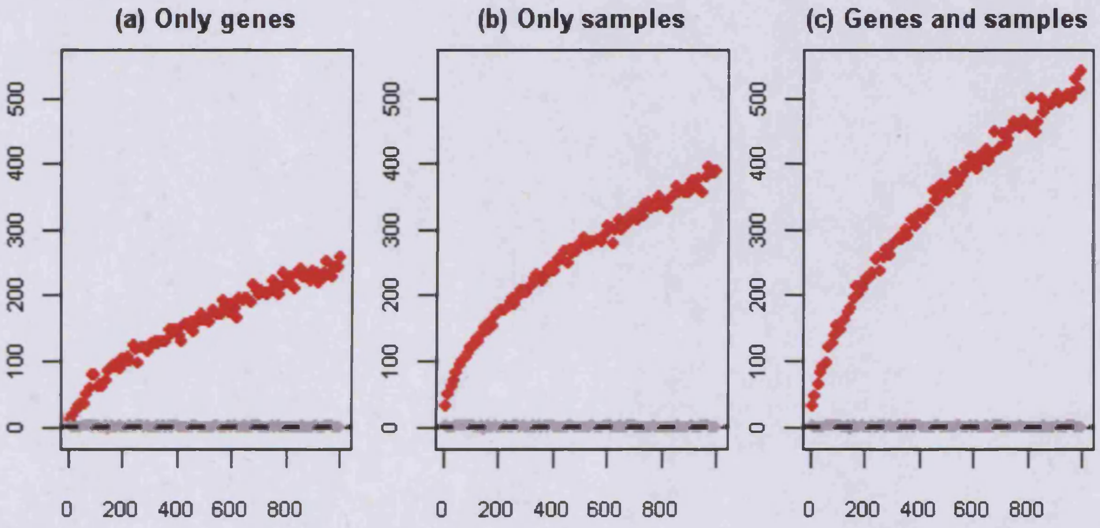


Figure 6.14 Randomization based Z-scores. Points represent Z-scores derived by randomization of expression values (in three ways) to assess the significance of observed SD-SDM values for a series test gene sets selected from hypothetical matrices without (grey points) and with information (red points) , the results for which are displayed in Table 6.1. Broken black lines indicate Z-scores of zero.

These data argue that the randomization strategy is an unsatisfactory methodology with which to assess to significance of SD-SDM values observed for gene sets for two reasons. Firstly, the null hypothesis when using the randomization strategy is that *a gene set contains no information at all.* Thus, the randomization strategy could assign significance to a gene set in which DEGs are present simply by chance alone (and not for biological reasons): this would be a false positive result of no interest to the researcher. Secondly, the presence of information (i.e. DEGs) within an experimental dataset creates an indirect sensitivity of significance levels, as assigned by the randomization strategy, to gene set size (via the random selection of larger numbers of DEGs into larger gene sets).

On the other hand, the re-sampling strategy represents a more satisfactory methodology with which to assess the significance of SD-SDM values observed for gene sets. The null hypothesis in this case is that *a gene set contains no greater level of information than would be expected by chance alone*, and could therefore allow for removal of gene sets containing DEGs simply by chance alone. Also, the levels of significance as assigned by the re-sampling methodology appear to be unaffected by gene set size, regardless of the presence or absence of information within a gene expression matrix.

Thus, the calculation of an SD-SDM value for a gene set, followed by testing of the significance of that value using the re-sampling strategy, comprise a methodology that could potentially be used to identify biological themes that may be of interest to a researcher in an unsupervised way. We term this approach Gene Set Discovery (GSD).

6.3 Discussion

GSA methods, both threshold-based (such as ORA of lists of DEGs) (Khatri and Draghici 2005) and threshold-free (such as GSEA) (Subramanian et al. 2005), are usually carried out following supervised analyses of microarray data that require prior knowledge of sample classes. However, in many studies (particularly those of cancers) researchers may have no *a priori* knowledge of sample classes, or these may be poorly defined (Alizadeh et al. 2000; Golub et al. 1999; Subramanian et al. 2005). This chapter explored the possibility of developing a methodology that could allow GSA analysis of microarray data without requiring prior description of sample classes.

Investigations focussed on heatmaps (and their associated hierarchical clustering) that have typically been used as visually intuitive 'end stage' tools to display expression patterns of genes that are known to be relevant within an experiment (such as lists of DEGs) (Eisen et al. 1998). A particular idea that was explored was that, if manual inspection of a heatmap reveals 'striking' visual patterns indicative of expression patterns shared between many genes, this may imply that the gene set (which may be a biological theme) represented by that heatmap is informative (i.e. contains DEGs) and may thus be of interest to a researcher. However such a method would be somewhat unsatisfactory due to the element of manual inspection and the absence of any underpinning statistical methodology.

To summarise the explorations that were carried out to develop a methodology that could allow unsupervised automated discovery of possibly interesting gene sets:

- A metric, SD-SDM, was found to be suitable for this purpose (see Section 6.2.1) for two reasons. First, it was sensitive to the *levels of information* within a gene set, i.e. the number of DEGs. Second, it was sensitive to the *types of information* within a gene set, i.e. whether the patterns of expression of the DEGs corresponded to the same groups of samples or not.

- SD-SDM values were found to be sensitive to the simultaneous effects of two possible confounding factors: gene set size and the possible random presence of DEGs within a gene set (see Section 6.2.2).
- To control for these factors, two strategies were investigated (see Section 6.2.3). The first of these was a strategy involving randomization of gene expression values. This was found to be unsatisfactory as it involved a null hypothesis that the expression data corresponding to a gene set contained *no information at all*, and thus failed to take into account the possible randomly-selected presence of information within a gene set. The second tested strategy involved re-sampling of gene sets (in a random fashion) from the entire dataset, and had a null hypothesis that a gene set contained *levels of information that were no greater than would be expected by chance alone.* Investigations of this strategy indicated that it could successfully control for the simultaneous effects of both confounding factors.

The calculation of SD-SDM values for a gene set, and the subsequent assessment of the significance of these values using the re-sampling strategy thus constitute a novel methodology called Gene Set Discovery (GSD). As will be explored in subsequent chapters, this method could potentially be used for the discovery of gene sets that may be of interest to researchers in an unsupervised fashion (i.e. without prior definition of sample classes).

While all explorations and investigations described in this chapter were carried out using hypothetical gene expression matrices with artificially introduced information, and gene sets comprising of randomly selected genes, evidence of the utility of this approach in the analysis of 'real-world' microarray datasets requires testing of the GSD methodology on such datasets. This is explored in the next chapter, which describes the implementation of the GSD methodology on four microarray datasets, and analysis of the results.

Chapter 7: Application of the GSD methodology to four microarray datasets

7.1 Introduction

Class discovery based on gene expression signatures in cancer datasets datasets is an important technique (see Section 1.3.2), with numerous published examples of datasets where the samples are morphologically homogenous, but show molecular heterogeneity and varying prognoses (Alizadeh et al. 2000; Bittner et al. 2000; De Cecco et al. 2004; Golub et al. 1999; Ivshina et al. 2006; Perou et al. 2000), and where the expression patterns provide prognostic information beyond what the histological classification is capable of providing.

In the previous chapter, explorations were described outlining the concepts that underpin Gene Set Discovery (GSD), a methodology that can be used simultaneously to identify gene sets (which may represent biological themes such as pathways or GO terms) that could be relevant within an experiment, as well as possibly identify functional classes of samples based on those gene sets. This was illustrated using hypothetical gene expression matrices and simulated patterns of gene expression. Because the GSD methodology does not require prior definition of sample/phenotype classes it was reasoned that it might be of particular use in analysis of cancer datasets, where sample classes may be unknown (or at least, where sample discovery is an aim), or where the classification of samples is problematic. An important feature of GSD is the potential to discover informative gene set signatures in such datasets that have a linking theme between the constituent genes, which in turn may identify opportunities for theme-based drug or prognostic marker development.

This chapter now describes the implementation of the GSD methodology, and its application to four microarray datasets. The first of these is the GNF human tissue expression dataset (Su et al. 2004) which was introduced in Chapter 5. This dataset shows strong sample grouping based on tissue-specific expression patterns. The remaining three datasets that were chosen to illustrate the GSD methodology were cancer datasets; a set of Acute Myeloid Leukaemia (AML) samples from the St Jude Children's Research Hospital (Ross et al. 2004), liposarcoma samples from a collaborator at the Memorial University Medical Centre (see Materials and Methods), and breast cancer samples from Uppsala (Ivshina et al. 2006).

7.2 Technical methodology

The development of the GSD methodology was based on the explorations described in Chapter 6. Figure 7.1 displays how the methodology was applied to four datasets. CEL files, which consisted of the raw expression data, were obtained for each experiment. The MAS5 algorithm was used to extract expression summary values from the CEL files, which was then logged and median-centred to yield the gene expression matrices to which the GSD methodology was applied. The gene set database used for all analyses was that of Gene Ontology Biological Process (GOBP) terms (see Materials and Methods). Only gene sets that consisted of a minimum of 5 genes were utilized, and those gene sets that were greater in size than 10% of all genes represented on array were excluded. SD-SDM values were calculated for each gene set. Background distributions were created using the re-sampling strategy: for each unique size of the GOBP gene sets, 10,000 gene sets of that size were selected at randomly and without replacement from amongst all genes represented on the array. SD-SDM values were calculated for each of these gene sets and these made up the background distributions for each size of gene set tested. By comparing the observed SD-SDM values with the background distributions it was possible to derive Z-scores and empirical p-values for each GOBP term. A p-value cutoff of <0.01 after FDR correction was used for all datasets.

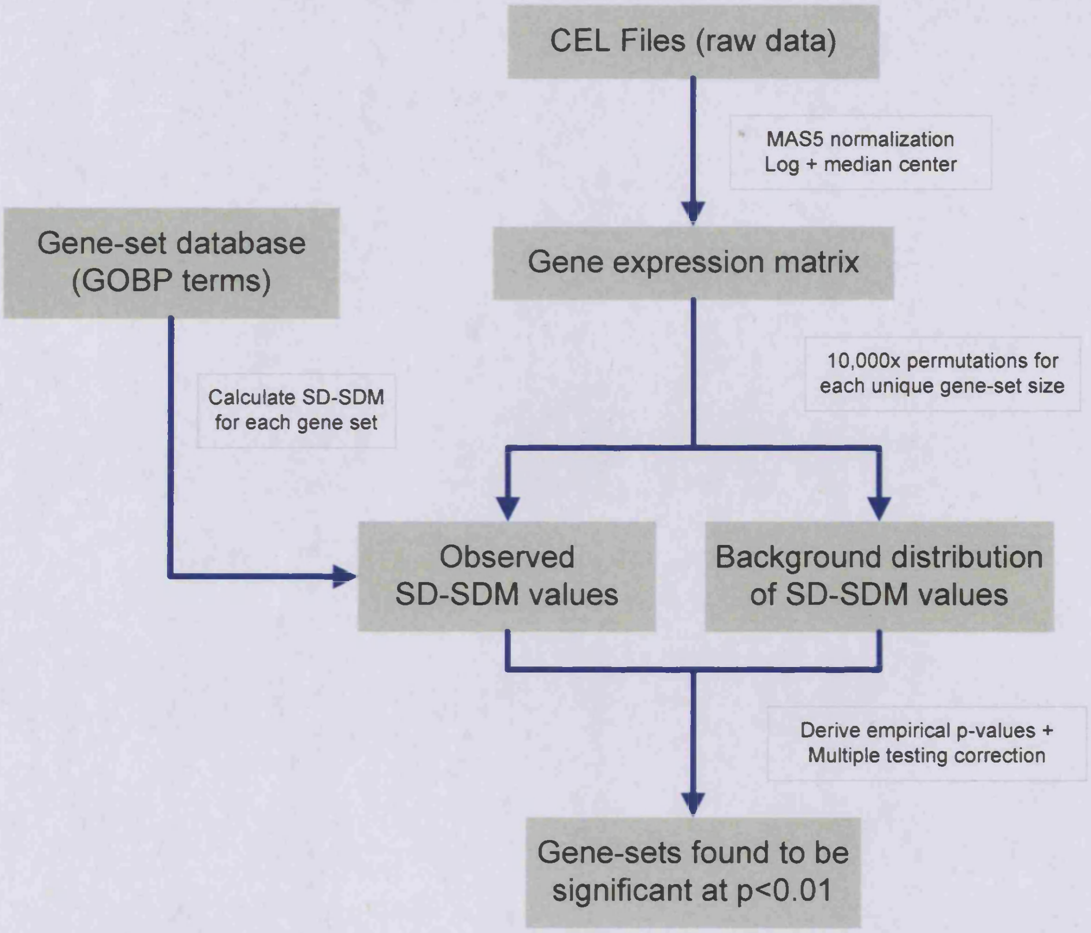


Figure 7.1 Implementation of the GSD methodology.

Testing the efficacy of methods for the analyses of microarray data is problematic due to the absence of 'truth' with which results can be compared. Assessment of the GSD methodology was based on the biological plausibility of the results and/or whether these results are in concord with prior analyses of these datasets.

7.3 Results and Explorations

7.3.1 Analysis of the GNF human tissue expression dataset

The first dataset analyzed using the GSD methodology was from the GNF tissue expression database. The distribution of SD-SDM values observed for all the GOBP terms tested relative to the background distribution of SD-SDM values to which they were compared is displayed in Figure 7.2. A total of 51 GOBP terms (out of 1397) were found to have FDR-corrected p-values less than the significance threshold of 0.01.

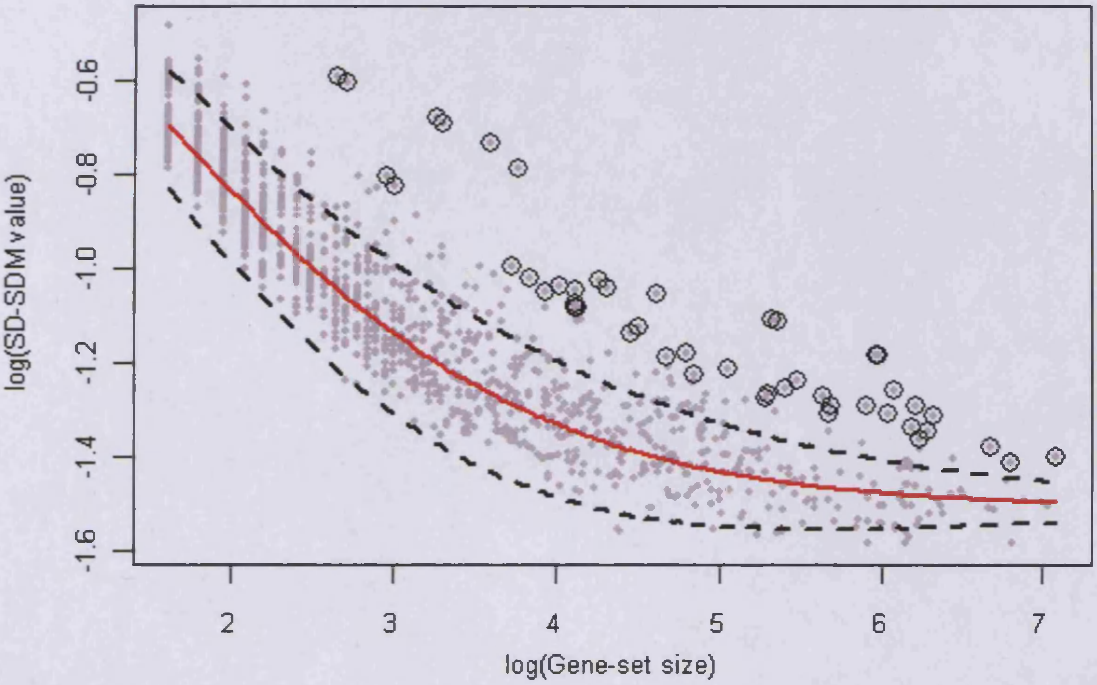


Figure 7.2 Selection of GOBP terms in GSD analysis of the GNF human tissue expression dataset. Grey points represent log SD-SDM values observed for each tested GOBP term. Encircled grey points represent those terms with FDR-corrected p-values of less that 0.01. The red line indicates the median of the background distribution of SD-SDM values for each gene set size. Broken black lines indicate the median ± 2 standard deviations for the background distributions.

While 34 terms had p-values of zero, a more precise ordering of the terms according to their significance could be achieved by using their Z-scores. Table 7.1 displays the top 20 terms selected by GSD analysis of this dataset when ranked by Z-score.

Gene Ontology Biological Process Term	Z-score	p-value (FDR corrected)
GO:0048731_system development	9.3	0
GO:0007399_nervous system development	9.2	0
GO:0030333_antigen processing	8.9	0
GO:0019882_antigen presentation	8.8	0
GO:0019883_antigen presentation, endogenous antigen	8.5	0
GO:0019226_transmission of nerve impulse	8.4	0
GO:0019885_antigen processing, endogenous antigen via MHC class I	8.4	0
GO:0007268_synaptic transmission	8.4	0
GO:0019886_antigen processing, exogenous antigen via MHC class II	7.2	0
GO:0019884_antigen presentation, exogenous antigen	7.1	0
GO:0006412_protein biosynthesis	7.0	0
GO:0007417_central nervous system development	6.7	0
GO:0009059_macromolecule biosynthesis	6..4	0
GO:0050877_neurophysiological process	6.1	0
GO:0030182_neuron differentiation	5.9	0
GO:0048699_neurogenesis	5.8	0
GO:0015672_monovalent inorganic cation transport	5.5	0
GO:0006812_cation transport	5.2	0
GO:0030154_cell differentiation	5.1	0
GO:0030001_metal ion transport	5.0	0

Table 7.1 Top 20 GOBP terms, selected by GSD analysis, of the GNF human tissue expression dataset. Terms are ranked by Z-score values. Terms highlighted in blue represent those processes involved in functions within the nervous system, while those in red are specific to the immune system.

As can be observed, the list is dominated by terms representing processes involved functions of the nervous system and of the immune system. This observation is of interest as it could be a reflection of the fact that samples from brain/neuronal tissues and from blood/immunity-related tissues comprise the two largest groups of similar tissue-types within this dataset. Indeed, terms associated with nervous system processes make up more than a third of all terms selected. On inspection of parent-child relationships between these terms (displayed in Figure 7.3), it is found that nearly all of their ancestral terms that were tested were also selected by GSD analysis. Investigation then focussed on the heatmaps of these selected terms that are associated with nervous system processes. Some of these are displayed in Figure 7.4, together with 'picketplots' to indicate the clusters in which each of the samples was found in the analysis described in Section 5.2.3. Three of the clusters comprised of samples from similar tissues: 'Brain/neuronal', 'Blood/immune' and 'Testis'; all other tissue samples were classified as 'Miscellaneous'. Considerable proportions of genes exhibit strong and consistent patterns of up-regulation in the samples from the Brain/neuronal cluster.

Investigation of heatmaps of terms associated with immune system processes (some of which are displayed in Figure 7.5) also similarly revealed many genes within these terms exhibited higher levels of expression in samples from the Blood/Immune cluster, as well as in samples from other tissues in the Miscellaneous class that are associated with the immune system, (i.e. those labelled as 'tonsil', 'lymph node' and 'thymus'). Heatmaps of all other GOBP terms selected by the GSD analysis (a subset of which are displayed in Figure 7.6) exhibited strong, consistent gene-expression expression patterns that were in concord with the tissue-type based clusters of samples.

Thus, it appears that the GSD methodology is successfully able to detect, without prior information regarding the groups of similar tissues, those biological processes that are specific to the two largest groups of functionally similar tissue-types within the dataset.

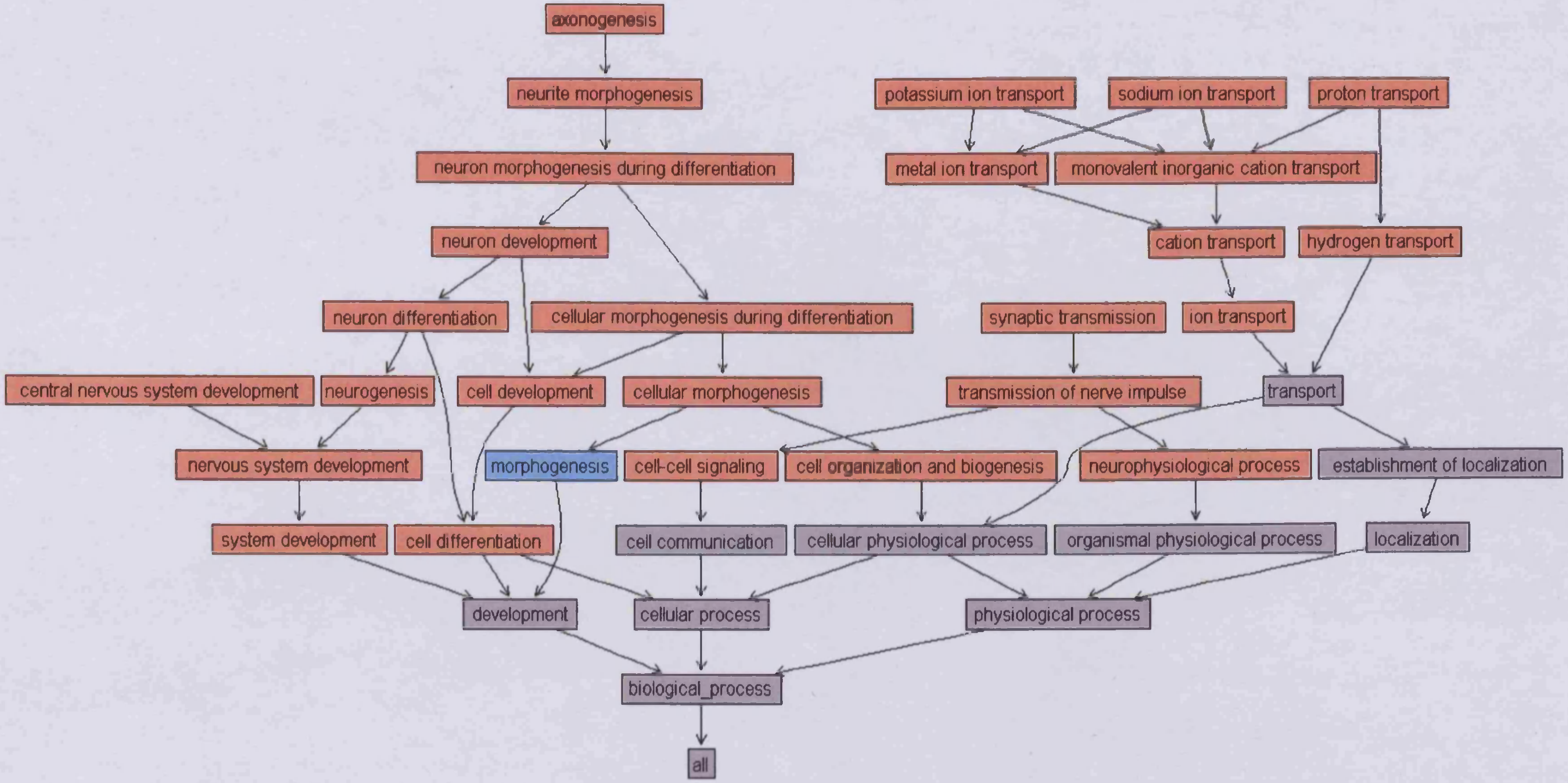
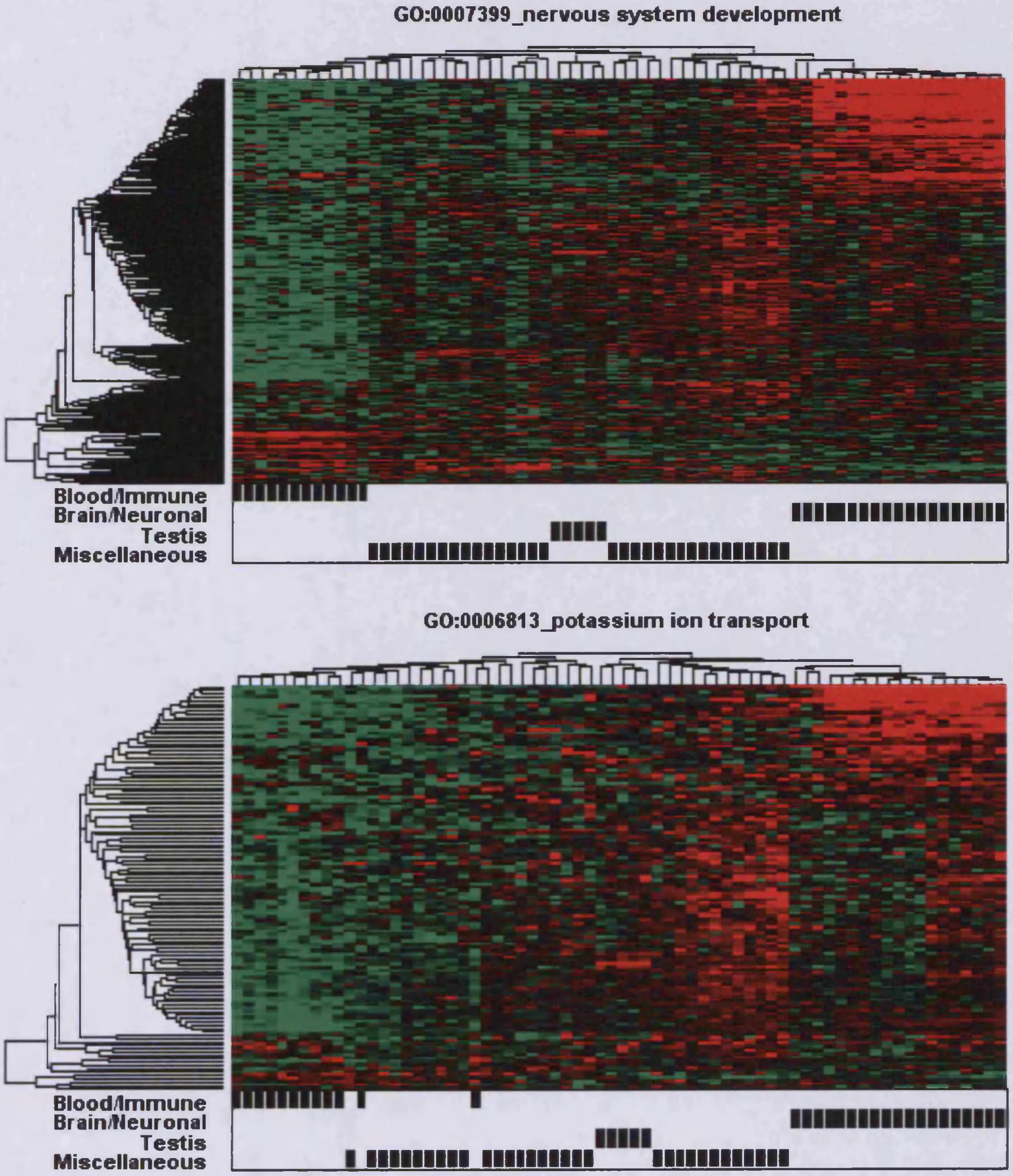

axonogenesis
neurite morphogenesis
potassium ion transport
sodium ion transport
proton transport
neuron morphogenesis during differentiation
metal ion transport
monovalent inorganic cation transport
neuron development
cation transport
hydrogen transport
neuron differentiation
cellular morphogenesis during differentiation
synaptic transmission
ion transport
central nervous system development
neurogenesis
cell development
cellular morphogenesis
transmission of nerve impulse
transport
nervous system development
morphogenesis
cell-cell signaling
cell organization and biogenesis
neurophysiological process
establishment of localization
system development
cell differentiation
cell communication
cellular physiological process
organismal physiological process
localization
development
cellular process
physiological process
biological_process
all

Figure 7.3 GOBP terms selected by GSD analysis of the GNF human tissue expression dataset which are involved in nervous system processes (and their ancestral terms). Nodes represent GOBP terms while the edges represent parent-child relationships between terms. Orange nodes represent terms selected by GSD analysis. Blue nodes represent terms that were tested but not selected. Grey nodes represent untested terms.



GO:0007399_nervous system development

GO:0006813_potassium ion transport

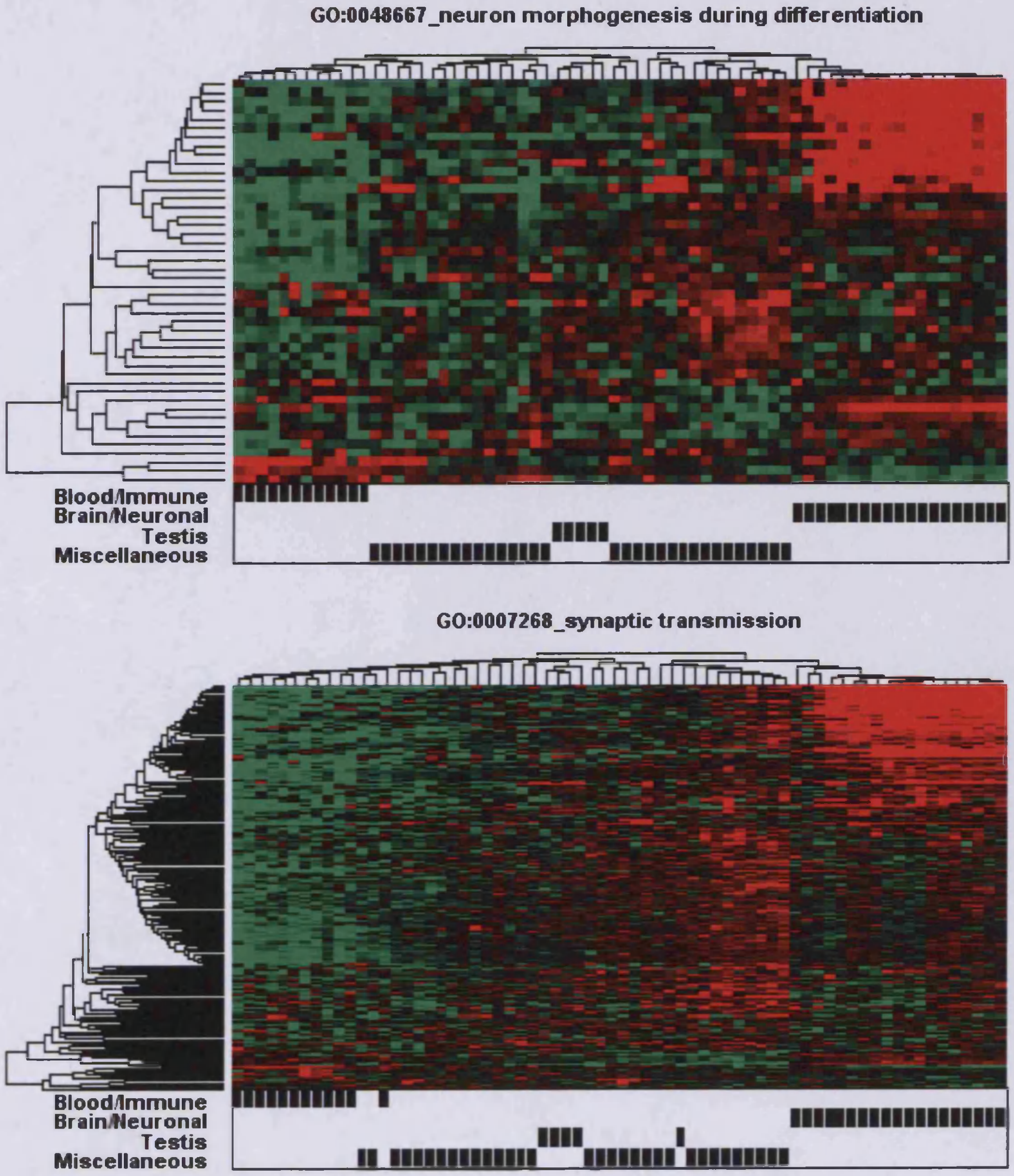
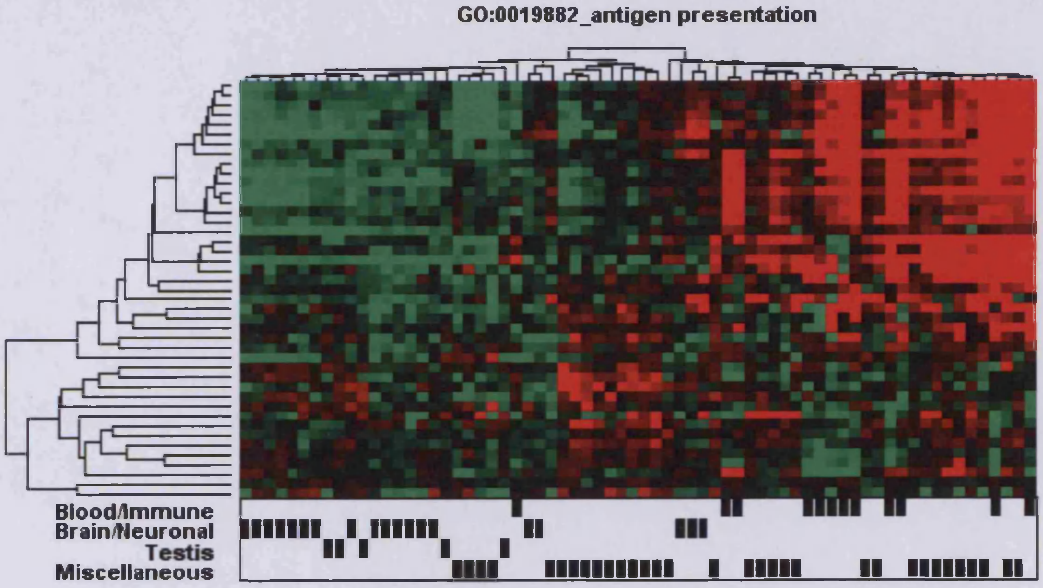


Figure 7.4 Selected GOBP terms specific to nervous system processes detected by GSD analysis of the GNF tissue expression dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 and less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate which tissue-specific clusters the respective samples are grouped into in the analysis described in Section 5.2.3.



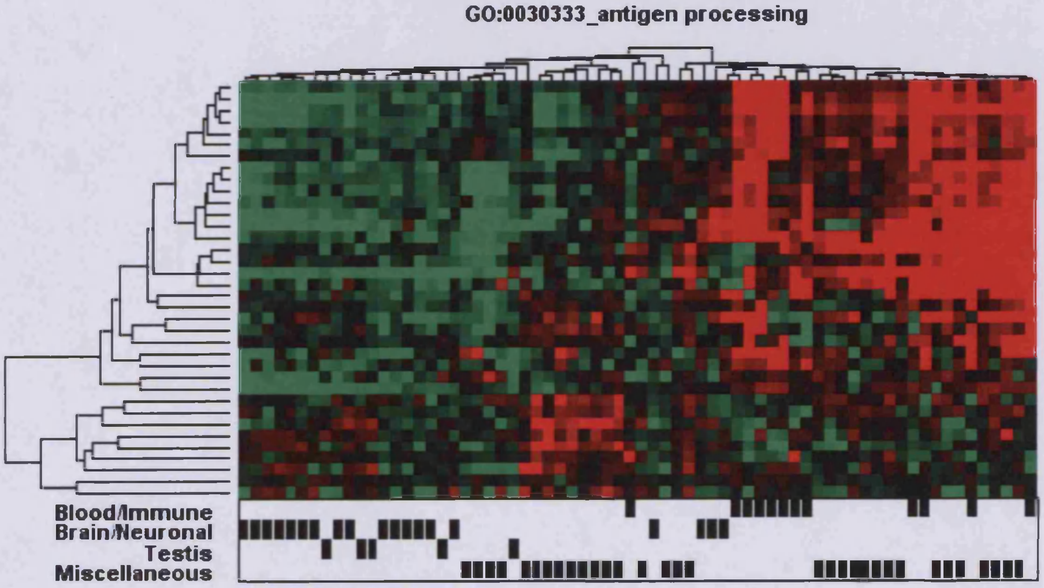
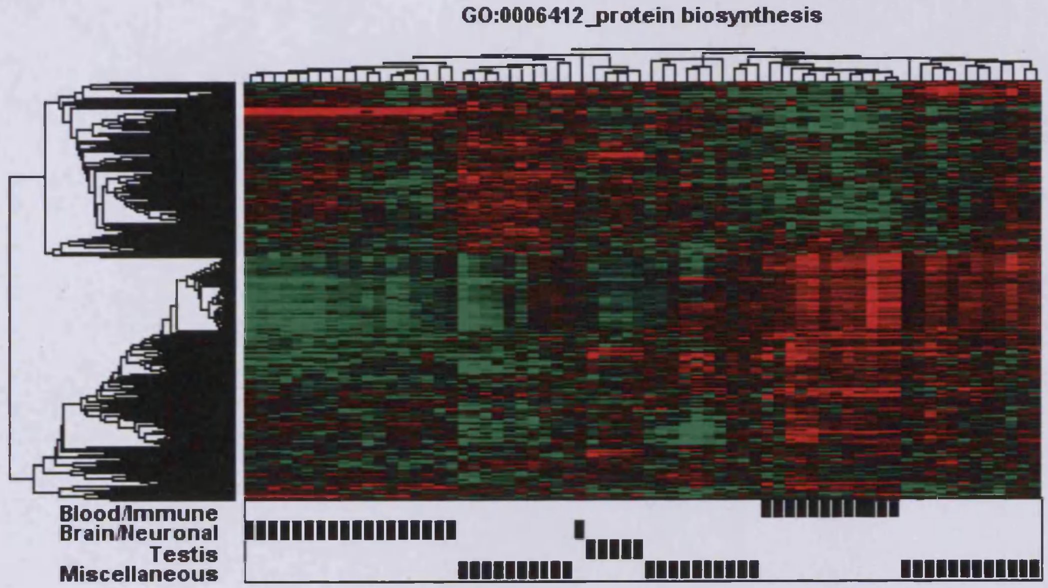
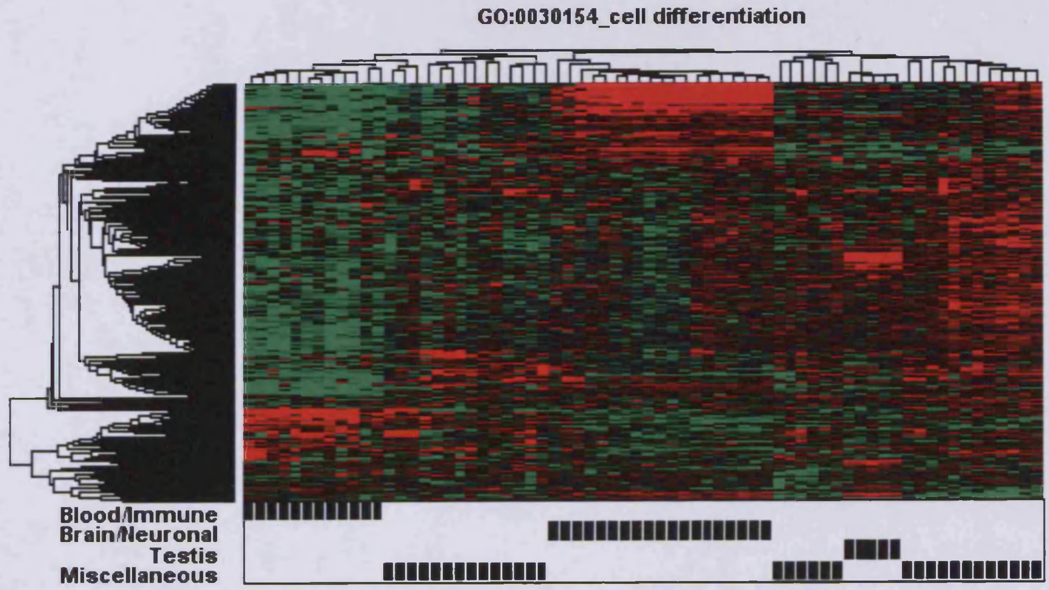


Figure 7.5 Selected GOBP terms specific to immune system processes detected by GSD analysis of the GNF tissue expression dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 and less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate which tissue-specific clusters the respective samples are grouped into in the analysis described in Section 5.2.3.


GO:0006412_protein biosynthesis
Blood/Immune
Brain/Neuronal
Testis
Miscellaneous


GO:0030154_cell differentiation
Blood/Immune
Brain/Neuronal
Testis
Miscellaneous

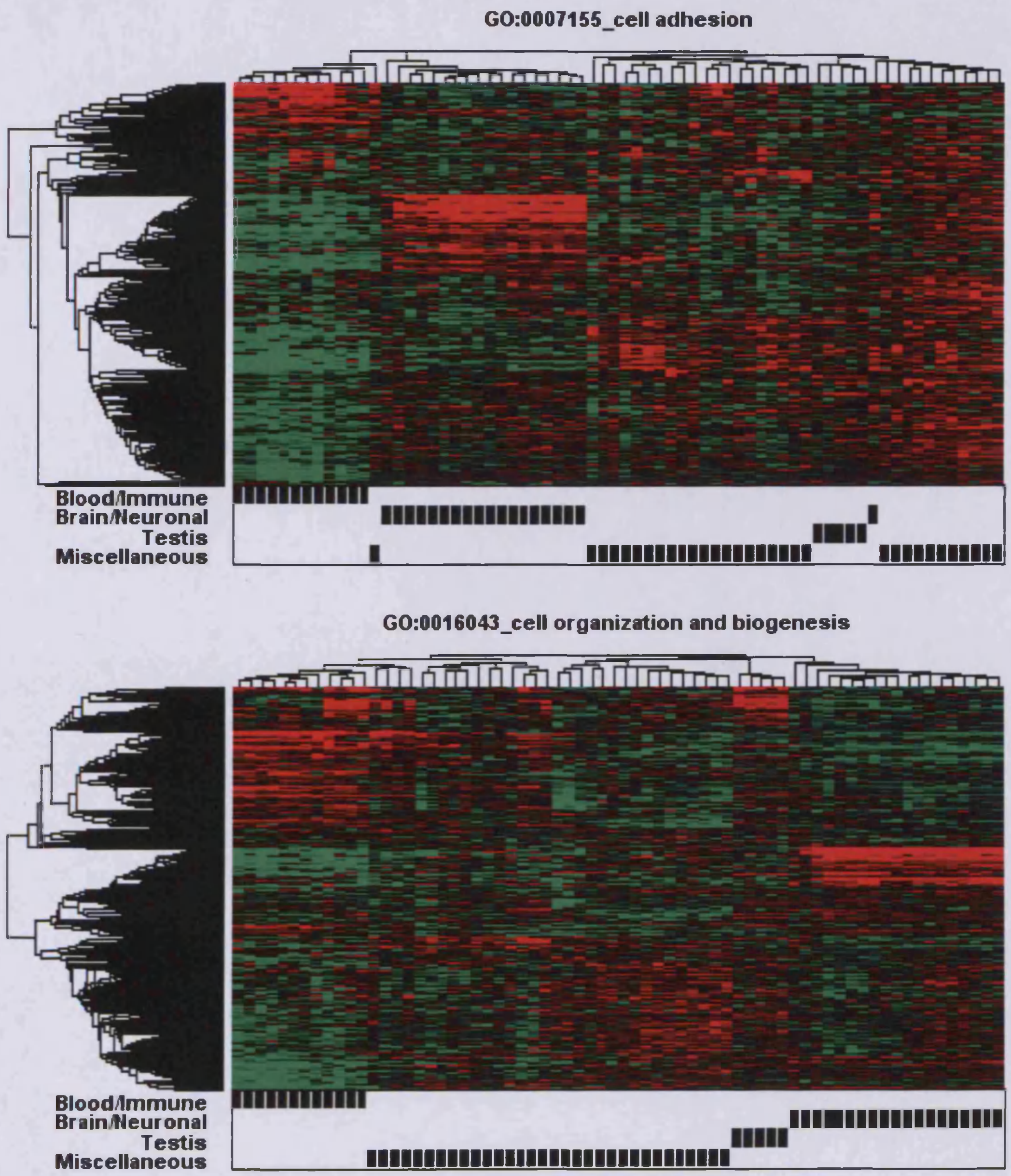


Figure 7.6 Other selected GOBP terms detected by GSD analysis of the GNF tissue expression dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 and less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate which tissue-specific clusters the respective samples are grouped into in the analysis described in Section 5.2.3.

7.3.2 Analysis of the Ross AML dataset

The next dataset analyzed using the GSD methodology comprised 130 samples from paediatric patients with Acute Myeloid Leukemia (AML), created at the St Judes Children's Research Hospital (Ross et al. 2004). Of these, 83 samples could be classified into one of five known genetic sub-types of AML: cases with t(15;17)[*PML-RARα*] (15 samples), t(8;21)[*AML1-ETO*] (21 samples), inv^{16}[*CBFβ-MYH11*] (14 samples), *MLL* chimeric fusion genes (23 samples), and acute megakaryocytic morphology (FAB-M7) (10 samples).

To reduce the complexity of analysis, implementation of the GSD methodology was restricted to these 83 samples, while the other 47 samples that could not be classified in the original study into any genetic subtype of AML were discarded. This therefore allowed GSD to be assessed as to whether it could discover the known classes within this dataset. The distribution of SD-SDM values observed for all the GOBP terms tested relative to the background distribution of SD-SDM values to which they were compared is displayed in Figure 7.7. A total of 12 GOBP terms (out of 1397) were found to have FDR-corrected p-values that were less than the significance threshold of 0.01. These are shown in Table 7.2.

As can be seen, other than the term for "translational initiation", all of the selected GOBP terms can be considered to be involved in processes of immune response – and as is displayed in Figure 7.8, are all descendants of the term "response to stimulus". Given that the samples comprise of mononuclear cells from purified from the bone marrow/ peripheral blood samples from paediatric AML patients; and that these cells represent a critical component of the immune system, it seems plausible that the dominant theme to arise from this dataset relates to the immune response.

Investigations then focussed on the heatmaps derived from these terms, some of which are displayed in Figure 7.9.

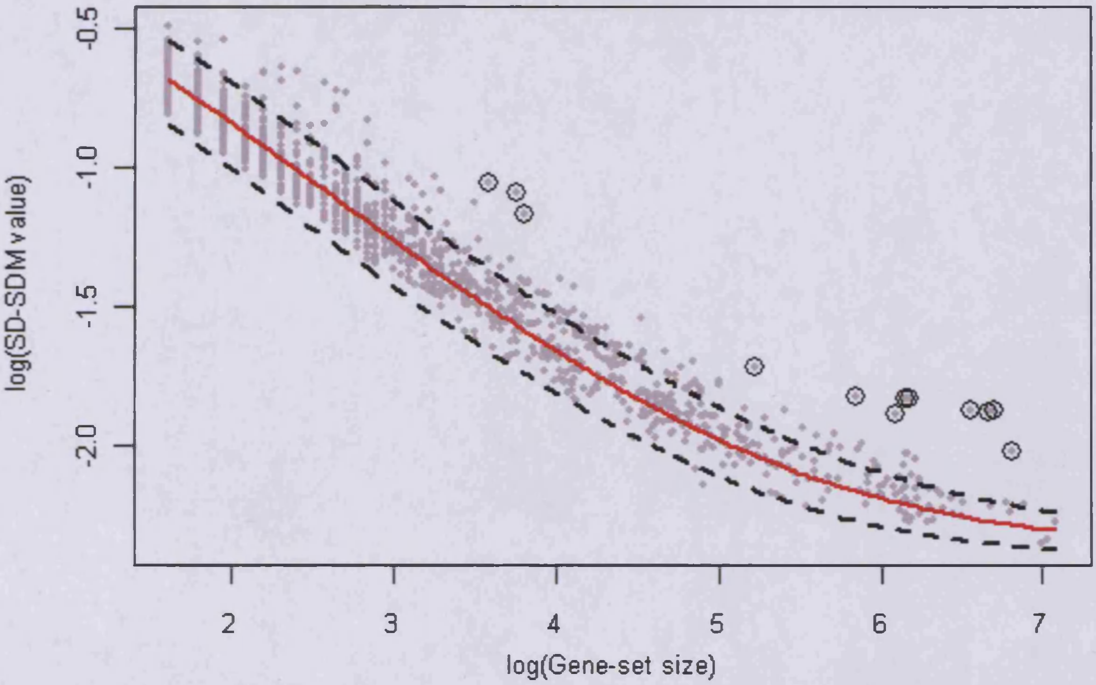


Figure 7.7 Selection of GOBP terms in GSD analysis of the Ross AML dataset. Grey points represent log SD-SDM values observed for each of the GOBP terms tested. Encircled grey points represent those GOBP terms found to exhibit FDR-corrected p-values of less that 0.01. The red line indicates the median of the background distribution of SD-SDM values for each gene set size. Broken black lines indicate the median ± 2 standard deviations for the background distributions.

As most these terms are closely related within the GOBP hierarchy, they share many genes. As a result, most of the heatmaps derived from the terms appear very similar to each other. The patterns of expression observed appear to correlate strongly with the known genotypic classes of AML.

Amongst most of the heatmaps, it is generally observed that the pattern involving the greatest number of genes splits the samples into two major clusters: one which is comprised mostly of samples of the t(15;17)[*PML-RARα*], t(8;21)[*AML1-ETO*], and FAB-M7 subtypes, while the other cluster is dominated by samples of the inv^{16}[*CBFβ-MYH11*] and *MLL* subtypes.

Gene Ontology Biological Process Term	Z-score	p-value (FDR corrected)
GO:0009607_response to biotic stimulus	12.8	0
GO:0006952_defense response	12.4	0
GO:0006955_immune response	11.6	0
GO:0051707_response to other organism	10.2	0
GO:0009613_response to pest, pathogen or parasite	10.1	0
GO:0019882_antigen presentation	8.3	0
GO:0006950_response to stress	8.2	0
GO:0009611_response to wounding	7.8	0
GO:0009605_response to external stimulus	7.6	0
GO:0030333_antigen processing	7.5	0
GO:0006413_translational initiation	7.2	0
GO:0006954_inflammatory response	6.3	0

Table 7.2 GOBP terms selected by GSD analysis of the Ross AML dataset. Terms shown are those found to have FDR-corrected p-values of less than 0.01. Terms highlighted in blue are those that are specific to immune system processes.

Other smaller patterns of expression are also observed that appear to be specific to AML sub-types within the first cluster. As a result, all of the t(15;17)[*PML-RAR*α] samples, all the of the FAB-M7 samples and most of the t(8;21)[*AML1-ETO*] are grouped respectively into three well-differentiated sub-clusters.

No obvious expression patterns that could differentiate amongst the subtypes of samples in the second major cluster (i.e. inv^{16}[*CBF*β-*MYH11*] and *MLL*) were discernible. This is reflected in the grouping of samples within this cluster – there is relatively little discrimination between samples of these sub-types.

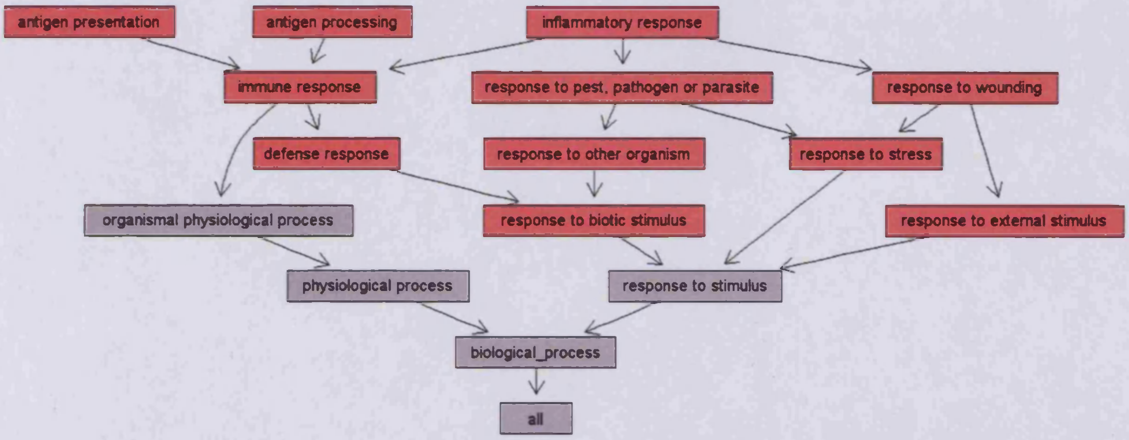
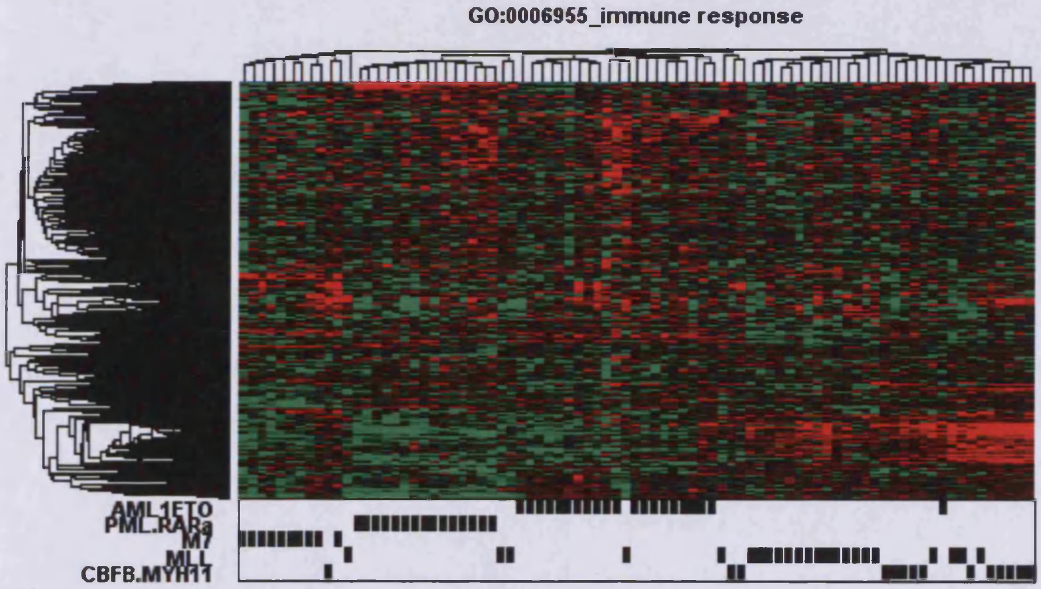


Figure 7.8 GOBP terms selected by GSD analysis of the Ross dataset (and their ancestral terms). Nodes represent GOBP terms while the edges represent parent-child relationships between terms. Orange nodes represent terms selected by GSD analysis. Blue nodes represent terms that were tested but not selected (there are no such terms here). Grey nodes represent untested terms. All terms selected represented with the exception of 'translational initiation'.



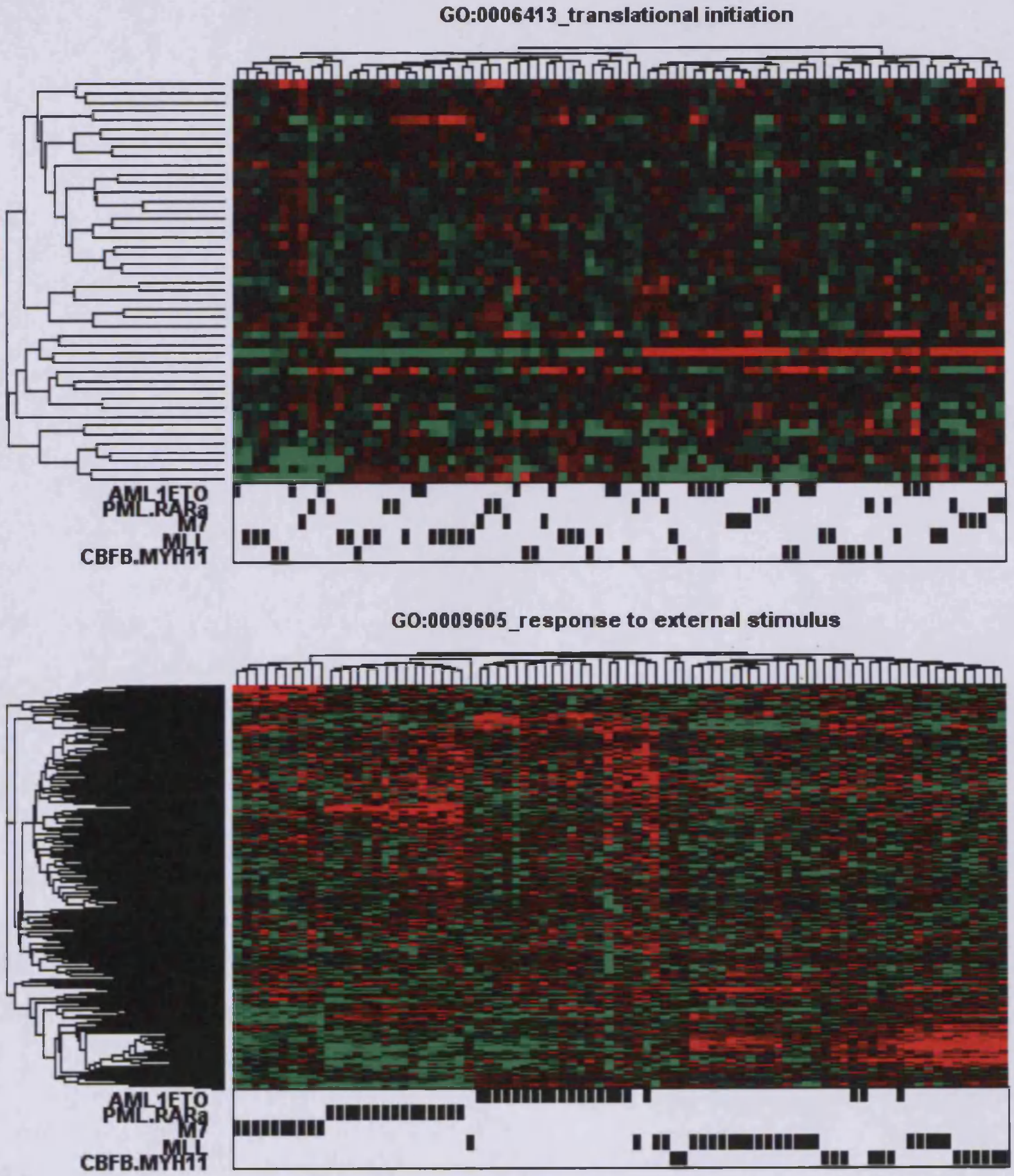


Figure 7.9 Selected GOBP terms detected by GSD analysis of the Ross AML dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 and less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate which of five genetic sub-types of AML that each of the samples was classified as in the original study.

This could be a reflection of the observation made in the original study that there is considerable molecular heterogeneity amongst samples of these AML sub-types, which implied the possibility of molecular sub-groups within these sub-types. The reason for this heterogeneity could also not be explained by the relatively more supervised and focussed analyses carried out in the original study.

Only the GOBP term for translational initiation exhibited a considerably different clustering of samples, which is not unexpected as it comprises a very different biological theme (consisting of a different set of genes) from all the other selected terms. The heatmap for this term (also displayed in Figure 7.9) reveals that there a relatively few informative genes, but strong expression patterns exhibited by these genes results in significantly strong clustering of samples. However the resultant samples clusters do not appear to be in agreement with the known AML genetic sub-types.

In summary, it was found that the majority of gene sets selected by GSD analysis of this dataset represented biological themes that can plausibly be linked to the biology within the dataset. The samples comprised mononuclear cells, which are a critical component of the immune system, and the dominant theme indicate by GSD analysis is that of the immune responses. Furthermore, the information within these gene sets (i.e. the patterns of expression) reflected known phenotypic classes of AML.

7.3.3 Analysis of the Broccoli liposarcoma dataset

The third dataset analyzed using the GSD methodology was a set of 31 liposarcoma samples created using the Affymetrix hgu133plus2 platform. This dataset was provided by a collaborator at the Memorial University Medical Centre (see Materials and Methods). Liposarcomas are a relatively understudied type of cancers, and are of adipocytic origin. Each of the samples could be classified into four sub-types based on histological profiles: well differentiated (WD), myxoid (MYX), de-differentiated (DD) and pleomorphic (PLEO) samples. WD samples represented low grade tumours, while PLEO and DD samples represented high grade tumours.

A total of 1492 GOBP terms were tested on the Broccoli liposarcoma dataset using the GSD methodology. The distribution of SD-SDM values observed for all the GOBP terms tested relative to the background distribution of SD-SDM values to which they were compared is displayed in Figure 7.10. A total of 30 GOBP terms were found to be exhibit FDR-corrected p-values that were less than the significance threshold of 0.01. These terms are displayed in Table 7.3, ranked by their Z-score values. As can be observed, all the terms selected can be grouped into at least three major biological themes: cell division, metabolism (in particular, metabolism of fatty acids and carbohydrates) and the immune response. Relationships between terms for each of these themes are displayed in Figure 7.11.

Investigations then focussed on the heatmaps of these terms. Heatmaps of cell division terms (some of which are displayed in Figure 7.12) showed that most of the information for these terms appear to be limited to sets of genes that share a single expression pattern: of down-regulation in most of the WD samples, and up-regulation in most of the PLEO and DD samples. This is consistent with the tumour behaviour; the high-grade (PLEO and DD) samples exhibit greater levels of proliferative activity than the low-grade WD samples.

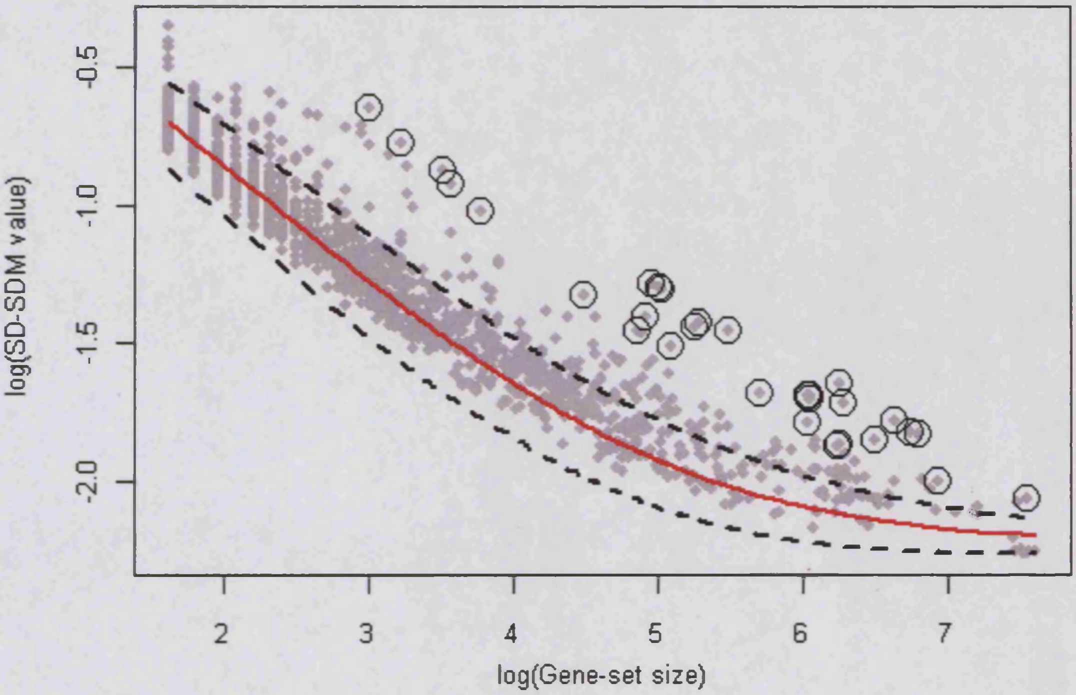


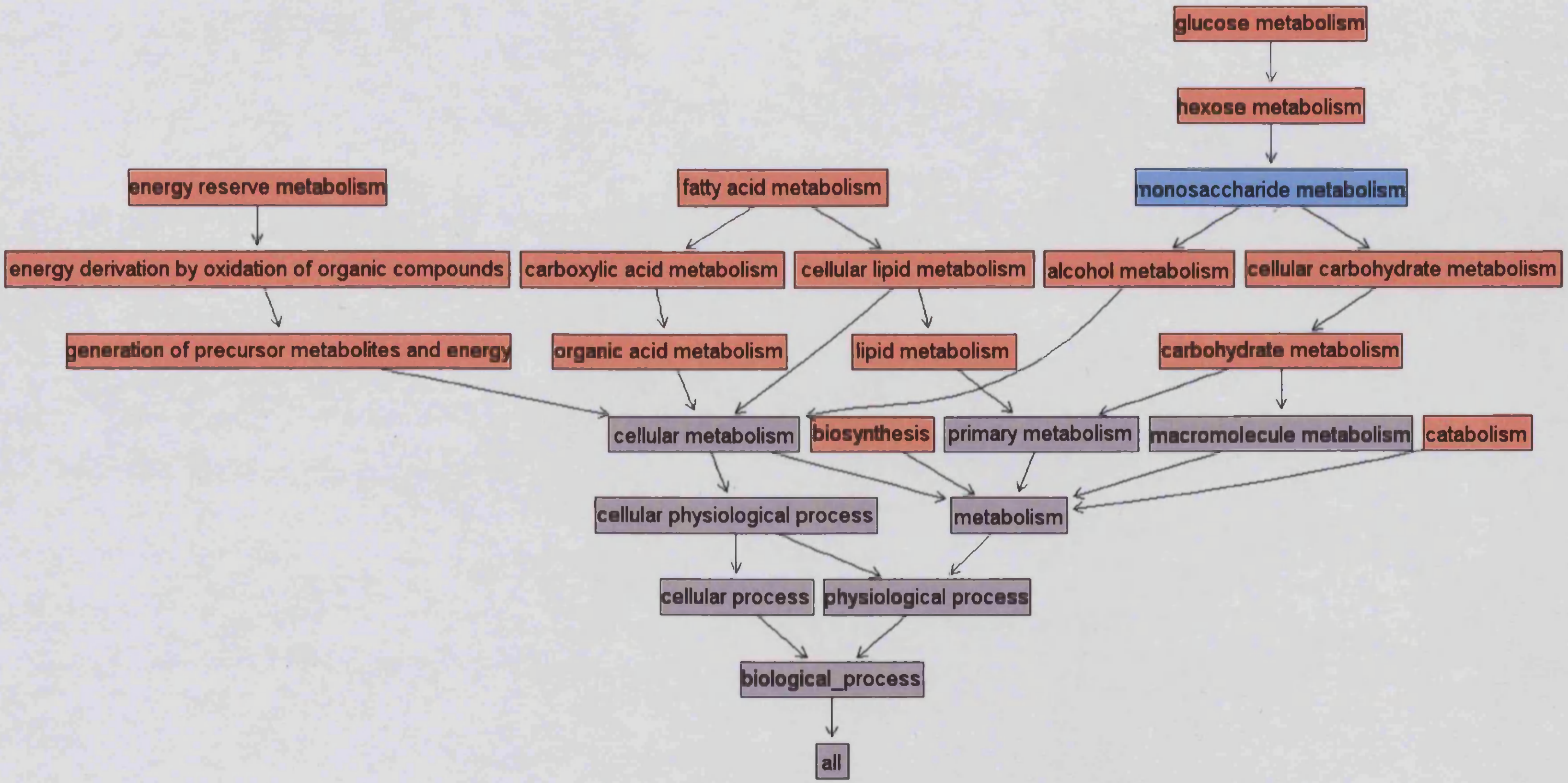
Figure 7.10 Selection of GOBP terms in GSD analysis of the Broccoli liposarcoma dataset. Grey points represent log SD-SDM values observed for each of the GOBP terms tested. Encircled grey points represent those GOBP terms found to exhibit FDR-corrected p-values of less that 0.01. The red line indicates the median of the background distribution of SD-SDM values for each gene-set size. Broken black lines indicate the median ± 2 standard deviations for the background distributions.

Gene Ontology Biological Process Term	Z-score	p-value (FDR corrected)
GO:0000087_M phase of mitotic cell cycle	11.2	0
GO:0006091_generation of precursor metabolites and energy	11	0
GO:0007067_mitosis	10.9	0
GO:0015980_energy derivation by oxidation of organic compounds	10.8	0
GO:0006066_alcohol metabolism	10.7	0
GO:0000278_mitotic cell cycle	9.9	0
GO:0007051_spindle organization and biogenesis	9.5	0
GO:0006955_immune response	9.4	0
GO:0000279_M phase	9.4	0
GO:0006629_lipid metabolism	9.1	0
GO:0009607_response to biotic stimulus	8.8	0
GO:0006112_energy reserve metabolism	8.7	0
GO:0006952_defense response	8.7	0
GO:0019883_antigen presentation, endogenous antigen	8.5	0
GO:0006082_organic acid metabolism	8.5	0
GO:0019752_carboxylic acid metabolism	8.4	0
GO:0030333_antigen processing	8.1	0
GO:0006631_fatty acid metabolism	8	0
GO:0005975_carbohydrate metabolism	8	0
GO:0019882_antigen presentation	7.9	0
GO:0051301_cell division	6.9	0
GO:0019318_hexose metabolism	6.7	0
GO:0044262_cellular carbohydrate metabolism	6.6	0
GO:0007049_cell cycle	6.6	0
GO:0006006_glucose metabolism	6.6	0
GO:0044255_cellular lipid metabolism	6.1	0
GO:0009056_catabolism	5.1	0
GO:0009613_response to pest, pathogen or parasite	4.8	0
GO:0009058_biosynthesis	4.4	0
GO:0050896_response to stimulus	4.3	0

Table 7.3 GOBP terms selected by GSD analysis of the Broccoli liposarcoma dataset. Terms are ranked by Z-score values. Terms highlighted in blue represent those processes involved in metabolism functions, while those in red are specific to cell division processes. The remaining terms are involved in immune response processes.

Heatmaps of the metabolism terms (some of which are displayed in Figure 7.13) also appeared to indicate that much of the information content within these terms was limited to genes that shared a single expression pattern. This pattern was the inverse of what was observed in the heatmaps of cell division terms, i.e. here the expression pattern was of up-regulation in most of the WD samples, and down-regulation in most of the PLEO and DD samples. This is, again, a biologically plausible finding. These liposarcomas are of adipocytic (fatty tissue) origin, and thus the well-differentiated WD samples might be expected to show higher levels of (lipid) metabolism than the poorly differentiated PLEO and DD samples.

Expression patterns observed in heatmaps of the immune response terms (some of which are displayed in Figure 7.14) do not appear to be specific to the known sample classes. There are several possible explanations for this. The samples used were gross biopsies, and no attempt was made to isolate RNA purely from the cancer cells. Thus, if any of the samples contained tumour-infiltrating lymphocytes or showed substantial vascularisation then cells of the immune system would be included in the sample that was analysed. Indeed, several of the samples show strong expression for immunoglobulin light chain (K), a B cell restricted marker that is strongly suggestive of a proportion of the sample being composed of immune cells (see Figure 7.15a). As such, the expression values for immune response pathways may not be strongly correlated with the cancer stage, since it is hypothesised that the signal is derived largely from non-cancer cells.


glucose metabolism
hexose metabolism
energy reserve metabolism
fatty acid metabolism
monosaccharide metabolism
energy derivation by oxidation of organic compounds
carboxylic acid metabolism
cellular lipid metabolism
alcohol metabolism
cellular carbohydrate metabolism
generation of precursor metabolites and energy
organic acid metabolism
lipid metabolism
carbohydrate metabolism
cellular metabolism
biosynthesis
primary metabolism
macromolecule metabolism
catabolism
cellular physiological process
metabolism
cellular process
physiological process
biological_process
all

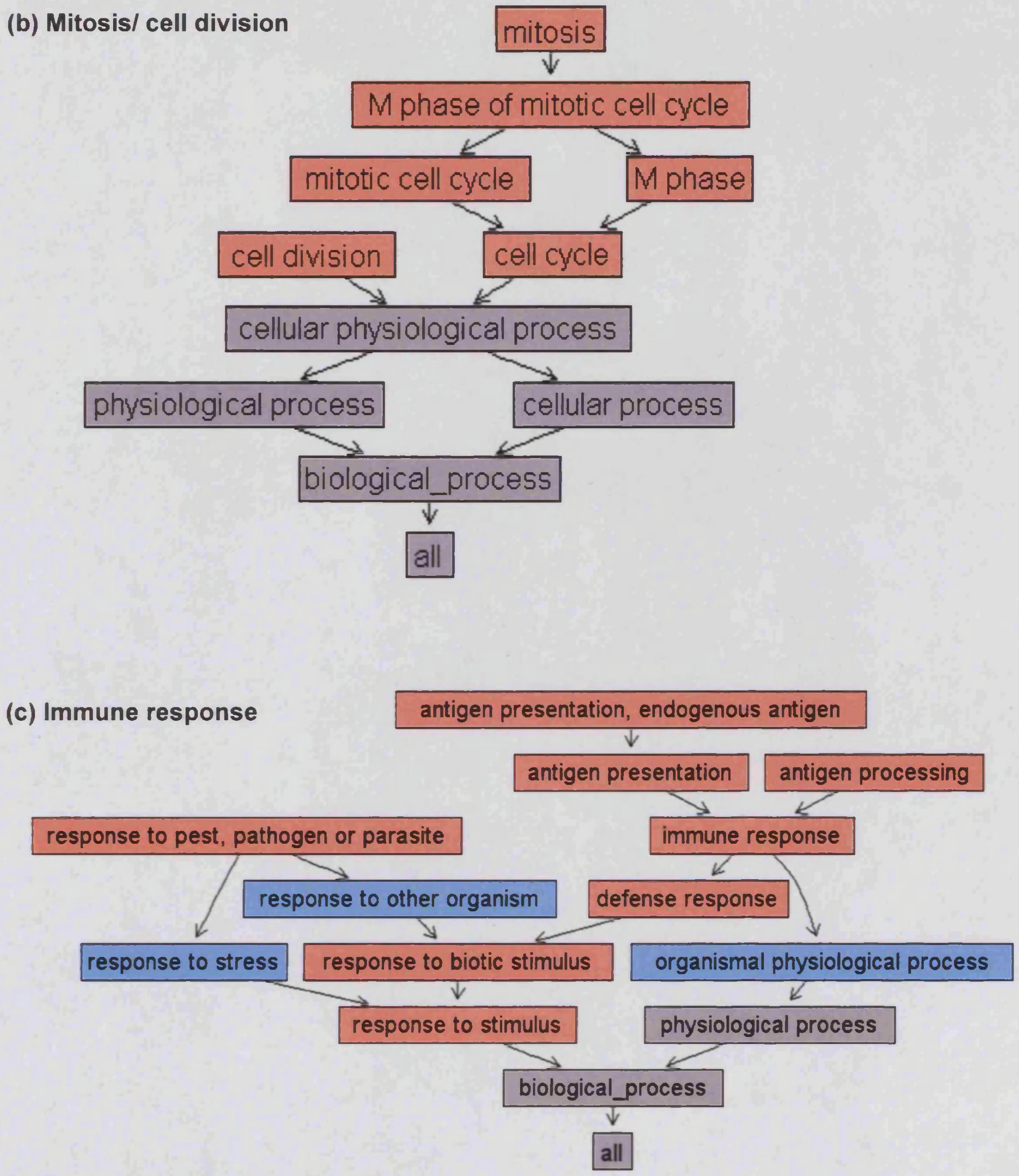
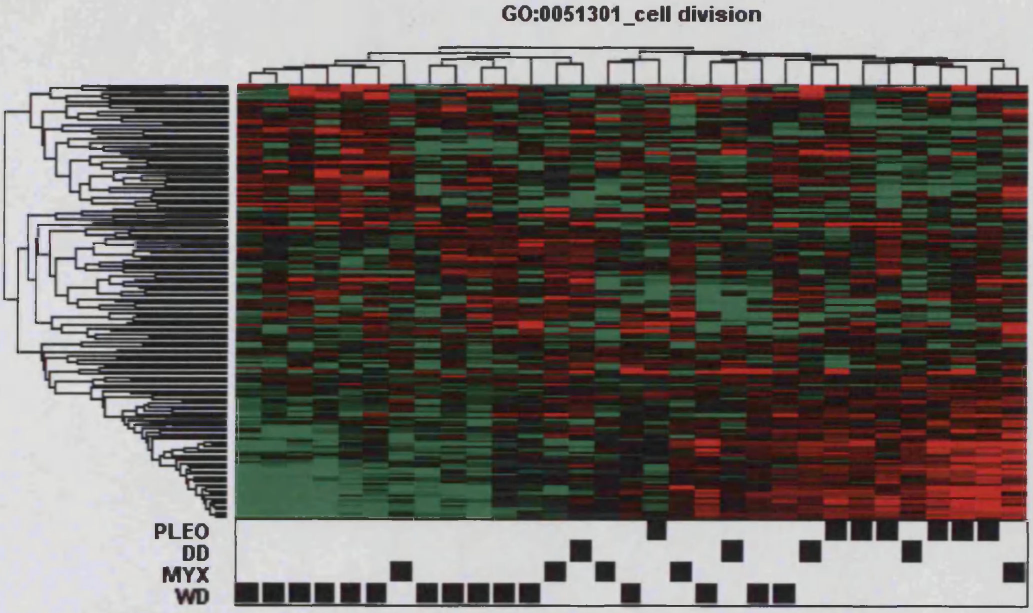
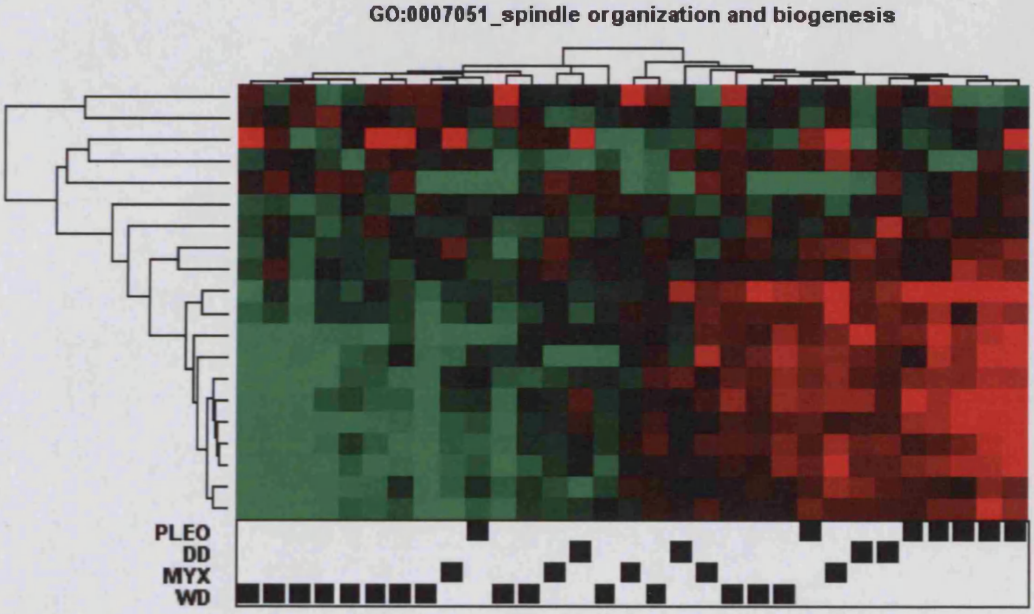


Figure 7.11 Dominant biological themes of GOBP terms selected by GSD analysis of the Broccoli liposarcoma dataset. Nodes represent GOBP terms while the edges represent parent-child relationships between terms. Orange nodes represent terms selected by GSD analysis. Blue nodes represent terms that were tested but not selected. Grey nodes represent untested terms.


GO:0051301_cell division
PLEO
DD
MYX
WD


GO:0007051_spindle organization and biogenesis
PLEO
DD
MYX
WD

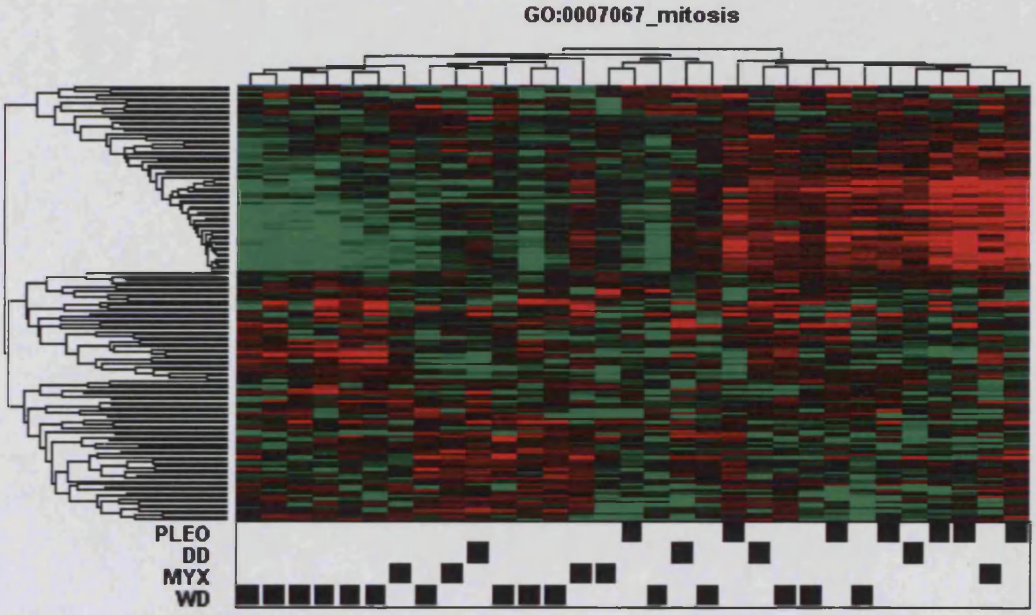
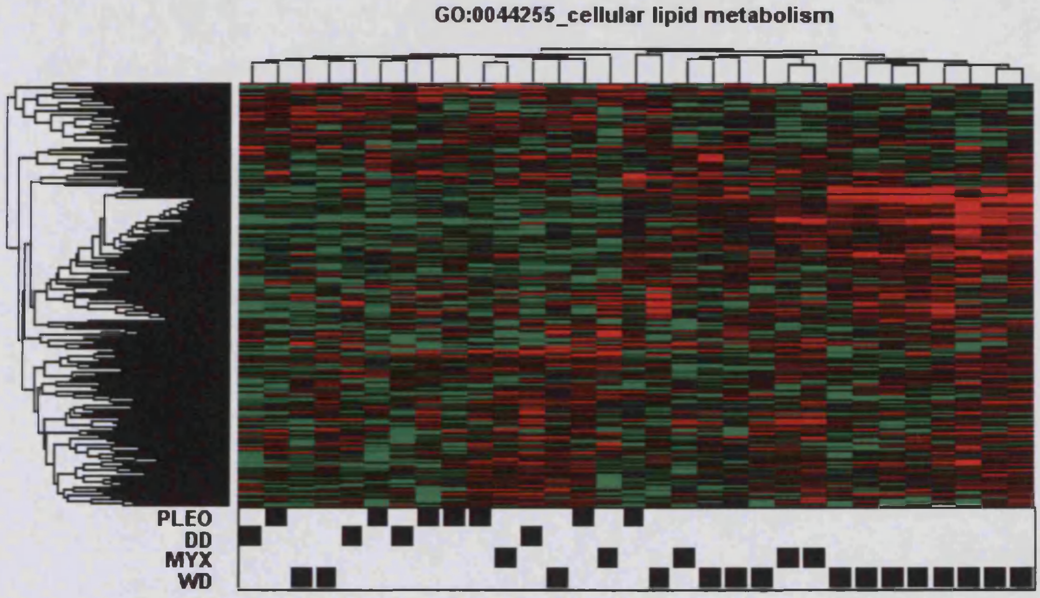
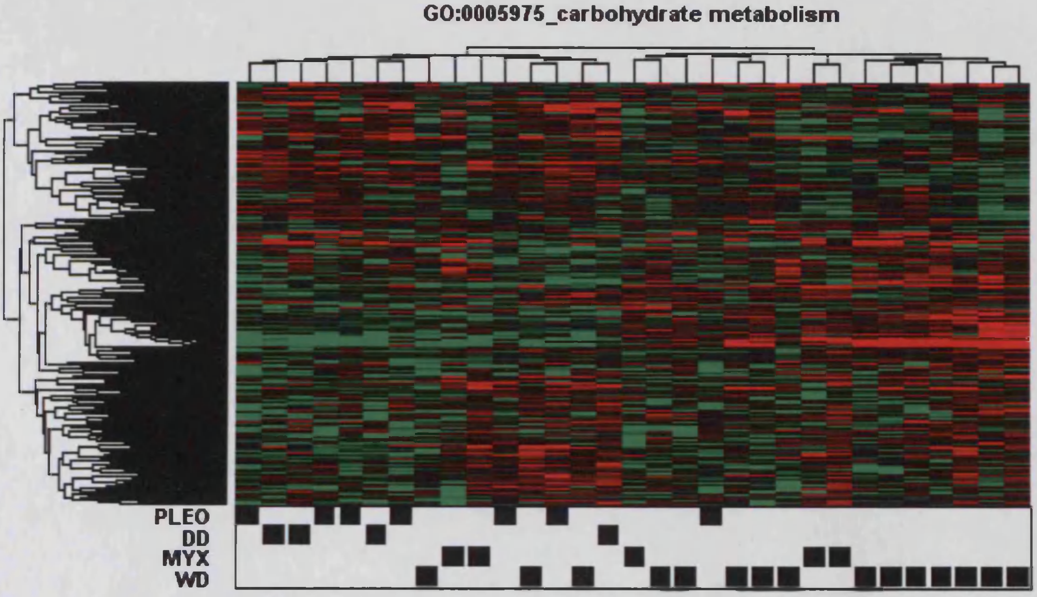
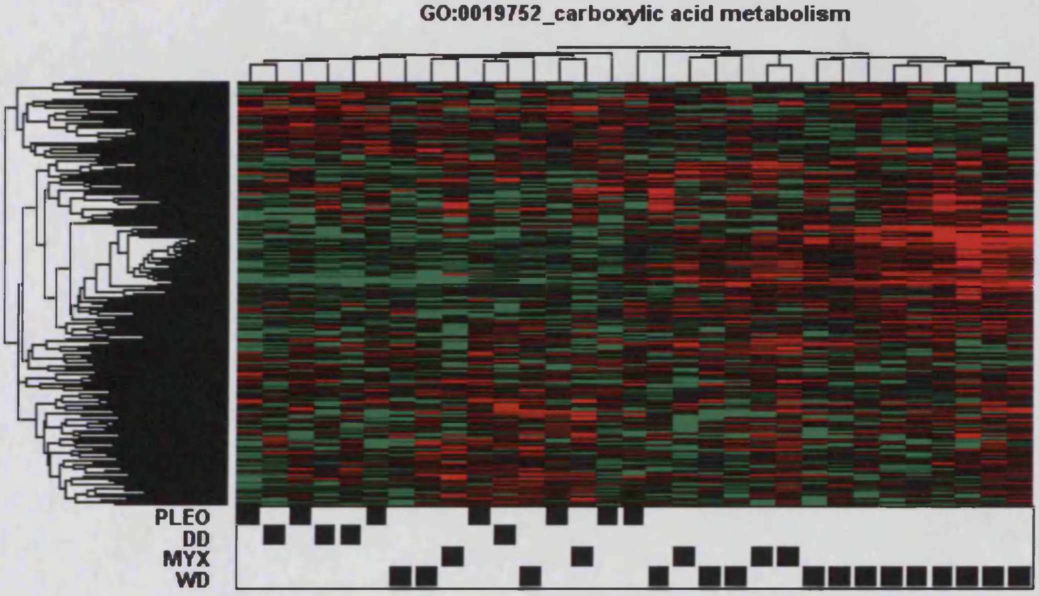


Figure 7.12 Selected GOBP terms specific to cell division processes detected by GSD analysis of the Broccoli dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 or less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate categories of liposarcoma that each of the samples were classified histologically.




GO:0005975_carbohydrate metabolism
PLEO
DD
MYX
WD


GO:0019752_carboxylic acid metabolism
PLEO
DD
MYX
WD

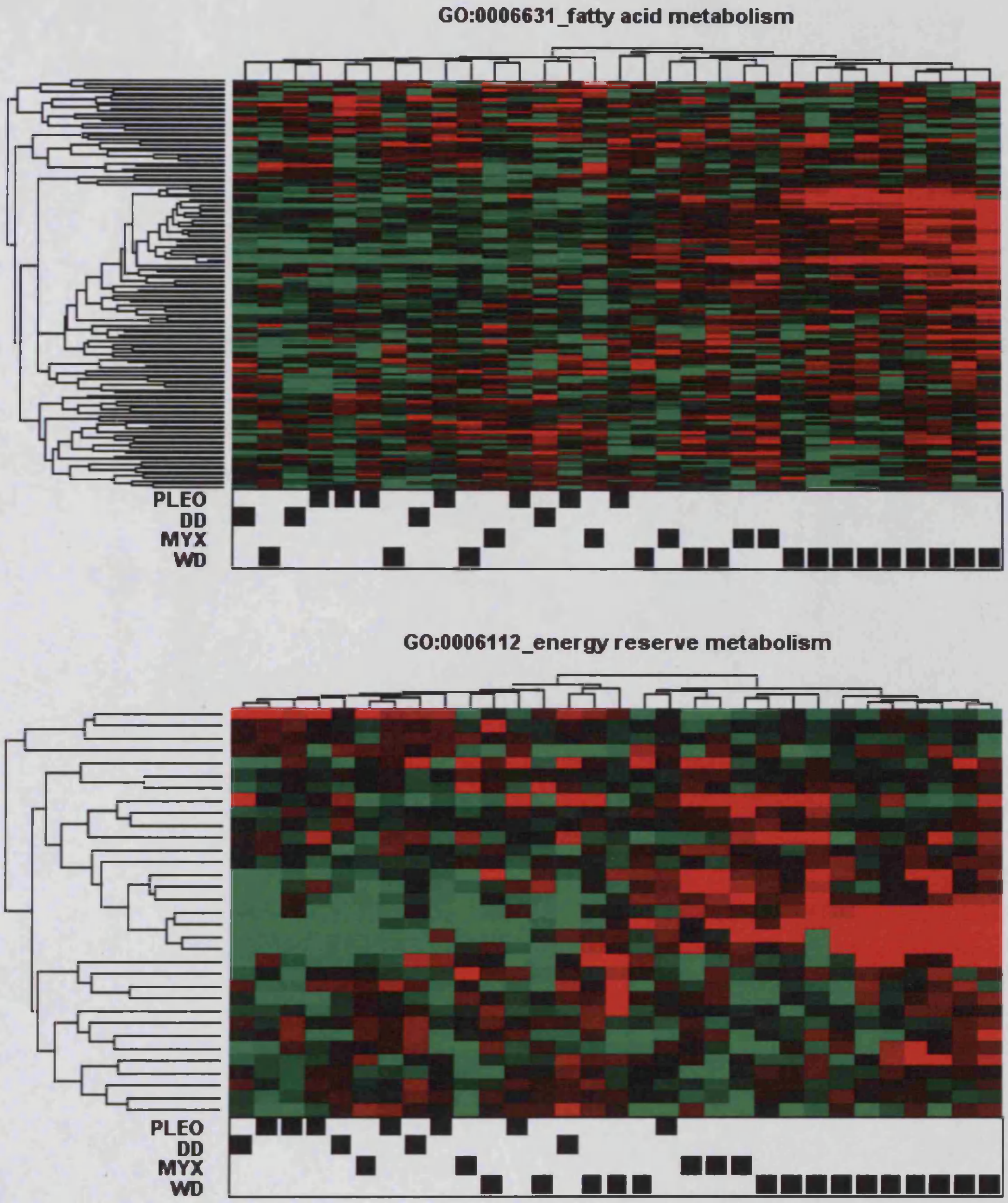
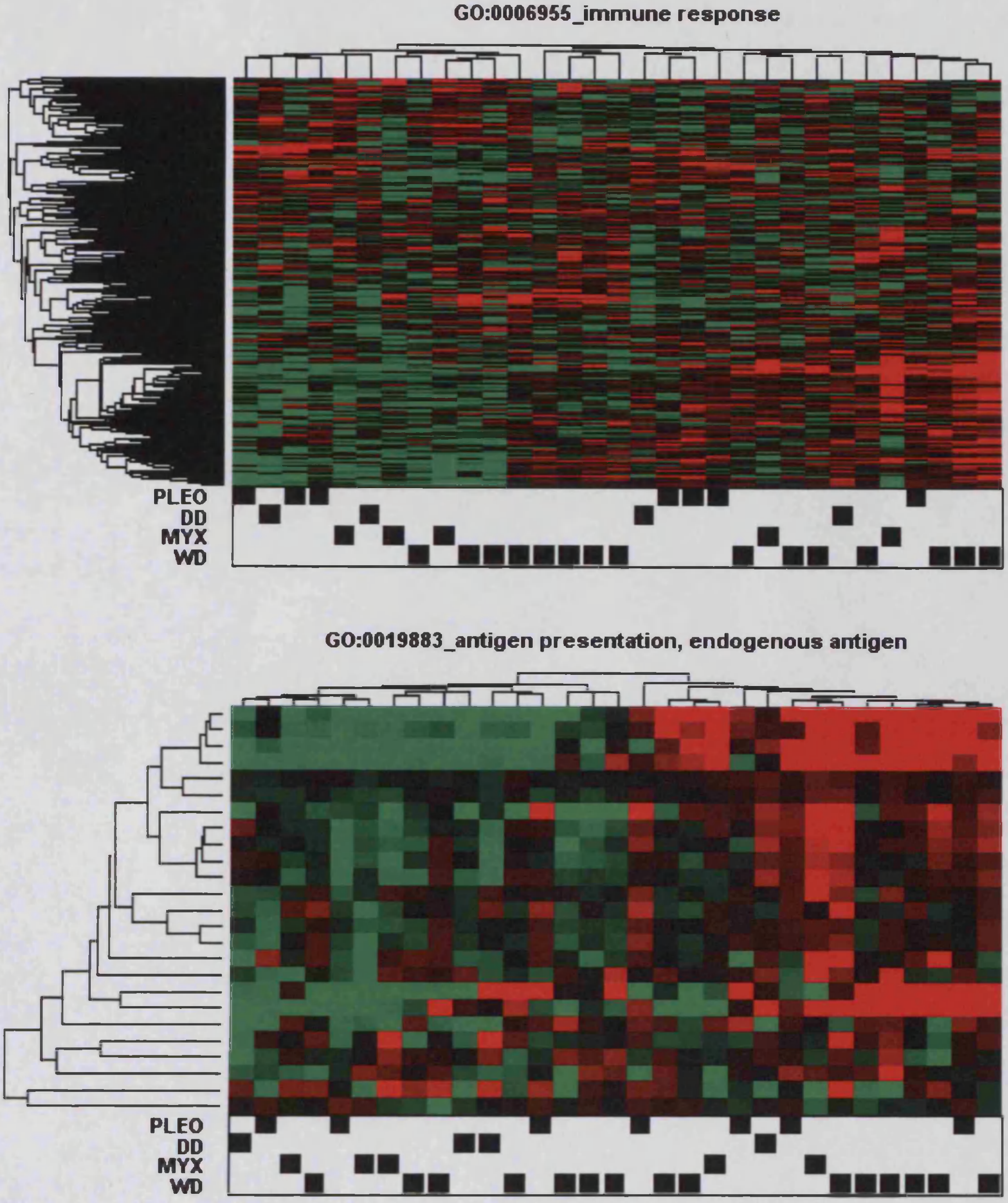


Figure 7.13 Selected GOBP terms specific to metabolism processes detected by GSD analysis of the Broccoli dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 or less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate categories of liposarcoma that each of the samples were classified histologically.



GO:0006955_immune response

GO:0019883_antigen presentation, endogenous antigen

Figure 7.14 Selected GOBP terms specific to immune response processes detected by GSD analysis of the Broccoli dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 or less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate categories of liposarcoma that each of the samples were classified histologically.

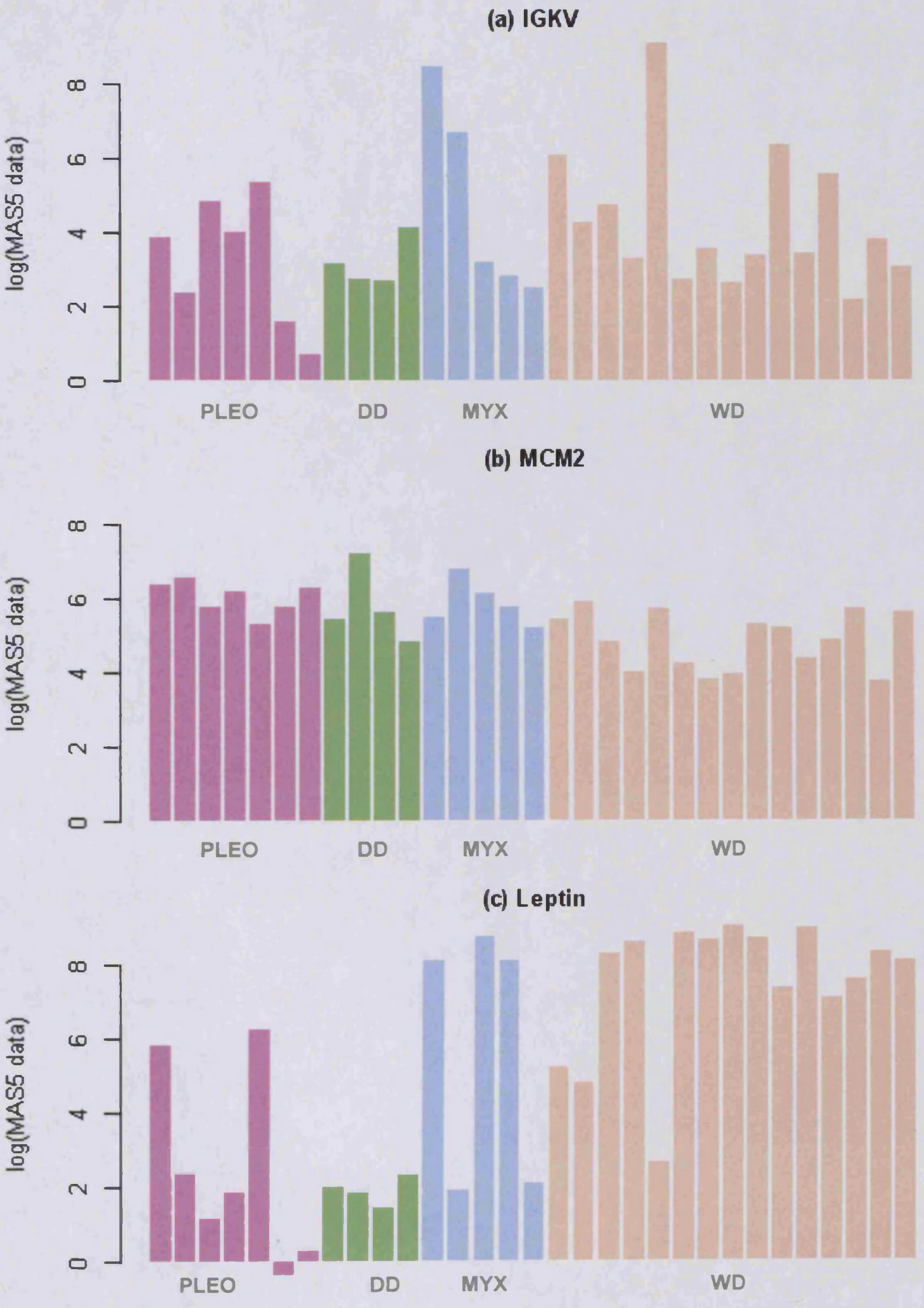


Figure 7.15 Gene expression levels for three biomarker genes. Figures show log MAS5 data for markers for (a) B-Cells, (b) proliferation and (c) adipocytes.

7.3.4 Analysis of the Ivshina breast cancer dataset

The final microarray dataset on which the GSD methodology was tested was the Uppsala breast cancer cohort (Ivshina et al. 2006), which comprised of a total of 249 samples. Using the Nottingham Grading System, which is based on microscopic evaluation of morphological and cytological aspects of tumour cells, each of the samples was classified in the original study into one of three grades. Grade 1 (G1) samples comprised well-differentiated, slow-growing tumours; untreated patients with this grade of tumour have ~95% 5-year survival rates. Grade 2 (G2) samples were moderately differentiated, while grade 3 (G3) samples were poorly differentiated, highly proliferative tumours, and untreated patients with these two grades of tumours have 5-year survival rates of ~75% and ~50% respectively.

The authors showed that G2 tumours could be sub-classified into two categories depending on the similarity of their expression profiles to those of G1 and G3 samples. Using class prediction algorithms, they discovered a set of classifier genes which could accurately discriminate between the 68 G1 and 55 G3 samples. They used this gene set to classify the 126 G2 samples into 83 grade 2a (G2a or 1-like), and 43 grade 2b (G2b or 3-like) samples based on the similarity of their expression profiles to the G1 and G3 samples respectively. Subsequent survival analyses and studies of other clinical variables supported this discrimination, showing significant differences between the two new sub-classes.

GSD analysis was carried out on this dataset. The distribution of SD-SDM values observed for all the GOBP terms tested relative to the background distribution of SD-SDM values to which they were compared is displayed in Figure 7.16. A total of 50 GOBP terms were found to be significant, as they exhibited FDR-corrected p-values of less that 0.01.

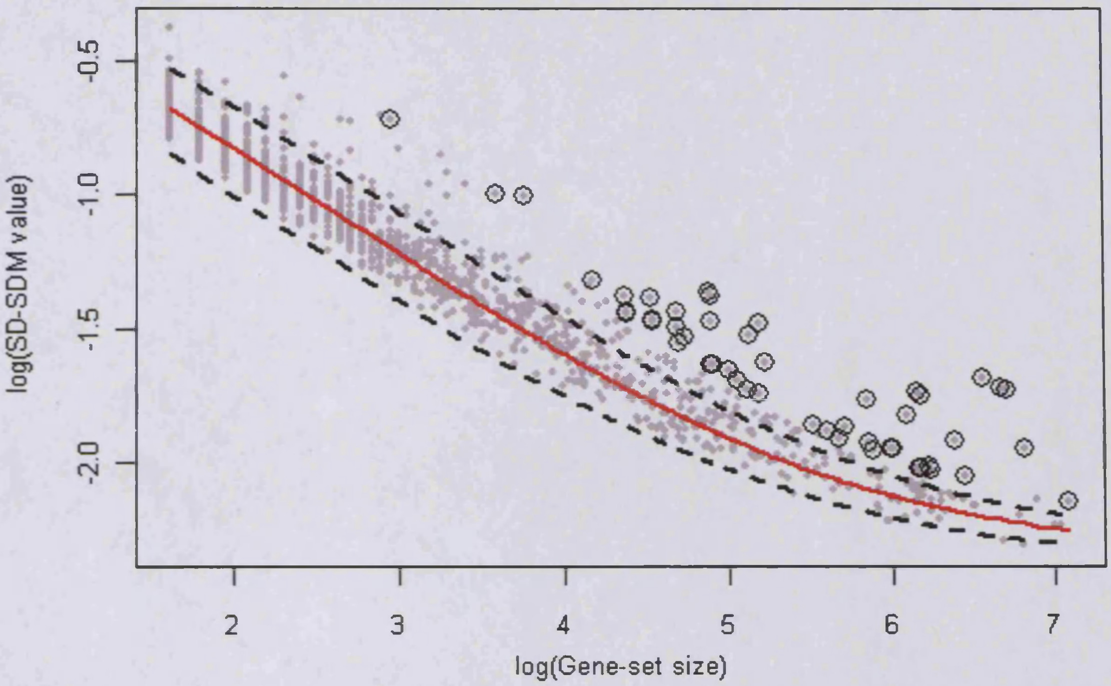


Figure 7.16 Selection of GOBP terms in GSD analysis of the Ivshina breast cancer dataset. Grey points represent log SD-SDM values observed for each of the GOBP terms tested. Encircled grey points represent those GOBP terms found to exhibit FDR-corrected p-values of less that 0.01. The red line indicates the median of the background distribution of SD-SDM values for each gene set size. Broken black lines indicate the median ± 2 standard deviations for the background distributions.

As was the case in the analysis of the GNF human tissue expression dataset, the p-values exhibited by the selected terms were very similar and a more precise ranking of these terms according to their significance could be achieved by using their Z-scores. Table 7.4 shows the top twenty selected terms when ranked by their Z-scores.

As can be seen, the table is comprised entirely of GOBP terms representing two distinct biological themes: immune system and cell division processes. These themes make up the majority of GOBP terms selected, and the relationships between terms from either theme are displayed in Figure 7.17.

Gene Ontology Biological Process Term	Z-score	p-value (FDR corrected)
GO:0009607_response to biotic stimulus	21.1	0
GO:0006952_defense response	21.1	0
GO:0006955_immune response	20.9	0
GO:0051707_response to other organism	13.8	0
GO:0009613_response to pest, pathogen or parasite	13.8	0
GO:0007067_mitosis	11.6	0
GO:0000278_mitotic cell cycle	11.6	0
GO:0000087_M phase of mitotic cell cycle	11.4	0
GO:0006950_response to stress	11.1	0
GO:0000279_M phase	9.8	0
GO:0009605_response to external stimulus	9.7	0
GO:0009611_response to wounding	9.6	0
GO:0051301_cell division	8.8	0
GO:0007049_cell cycle	8.8	0
GO:0019882_antigen presentation	8.5	0
GO:0006954_inflammatory response	8	0
GO:0007017_microtubule-based process	7.7	0
GO:0006968_cellular defense response	7.3	0
GO:0030333_antigen processing	7.2	0
GO:0000819_sister chromatid segregation	7	0.008

Table 7.4 Top 20 GOBP terms selected by GSD analysis of the Ivshina breast cancer dataset. Terms are ranked by Z-score values. Terms highlighted in blue represent those processes involved in functions in immunity processes, while those highlighted in red represent processes that take place during cell division.

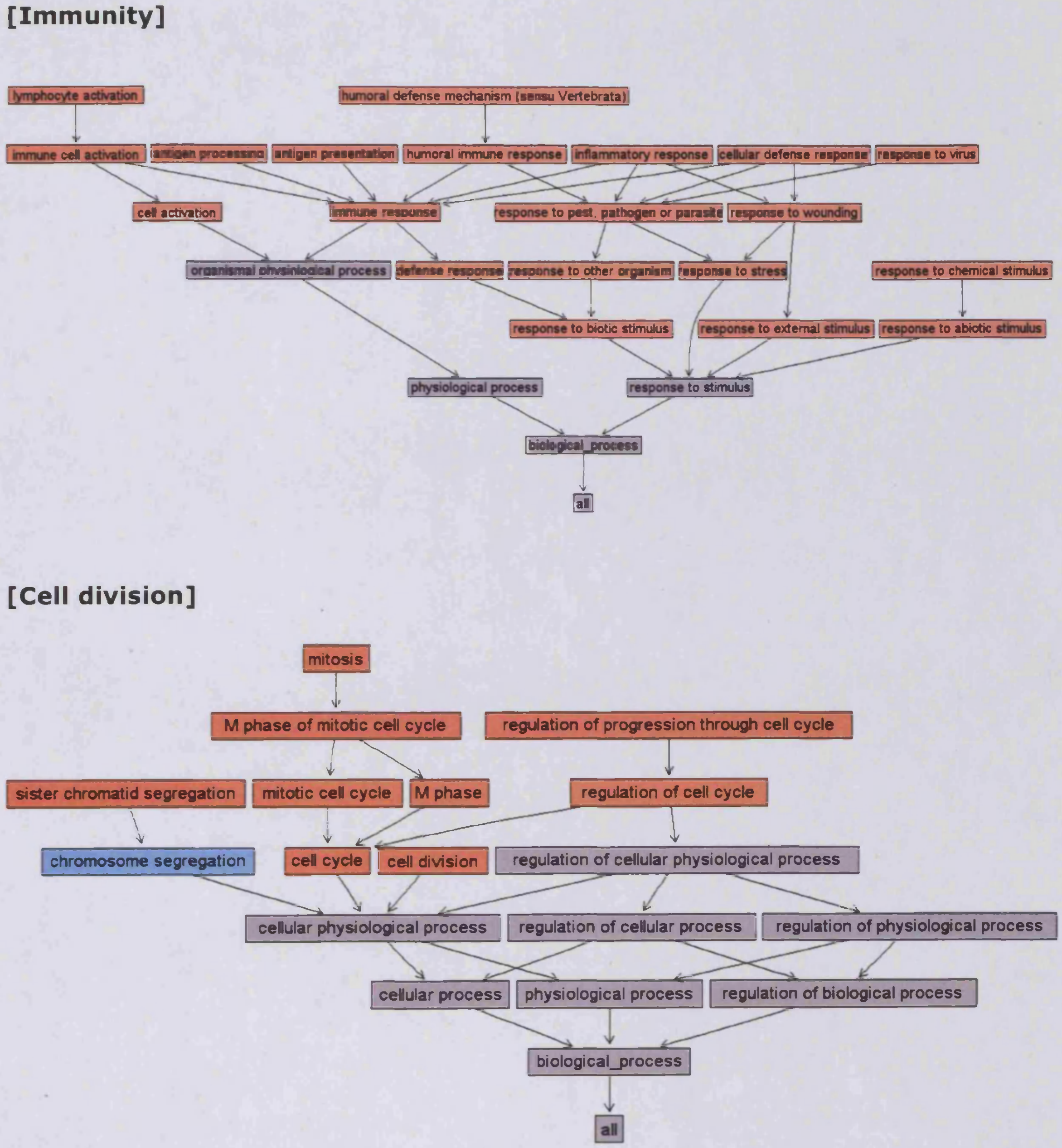

[Immunity]
lymphocyte activation
humoral defense mechanism (sensu Vertebrata)
immune cell activation
antigen processing
antigen presentation
humoral immune response
inflammatory response
cellular defense response
response to virus
cell activation
immune response
response to pest, pathogen or parasite
response to wounding
organismal physiological process
defense response
response to other organism
response to stress
response to chemical stimulus
response to biotic stimulus
response to external stimulus
response to abiotic stimulus
physiological process
response to stimulus
biological_process
all
[Cell division]
mitosis
M phase of mitotic cell cycle
regulation of progression through cell cycle
sister chromatid segregation
mitotic cell cycle
M phase
regulation of cell cycle
chromosome segregation
cell cycle
cell division
regulation of cellular physiological process
cellular physiological process
regulation of cellular process
regulation of physiological process
cellular process
physiological process
regulation of biological process
biological_process
all

Figure 7.17 Dominant biological themes of GOBP terms selected by GSD analysis of the Ivshina dataset. Nodes represent GOBP terms while the edges represent parent-child relationships between terms. Orange nodes represent terms selected by GSD analysis. Blue nodes represent terms that were tested but not selected. Grey nodes represent untested terms.

Indeed, ORA analysis of GOBP terms was carried out on this classifier gene set Materials and Methods), and five terms were found to be significant, exhibiting F corrected p-values of less that 0.01 (see Table 7.6). All five terms represented mi and cell division processes, and had been selected in the GSD analysis.

These findings indicate that this biological theme of mitosis/cell division is likely t relevant to the discrimination between low and high grade breast cancer samples investigate this, inspection of heatmaps of these GOBP terms specific to cell div processes was then carried out. Some of these heatmaps are displayed in Figure along with 'picketplots' that indicate which tumour grade each of the samples classified into in the original study using the Nottingham Grading System.

In the heatmaps of all the selected terms that represented mitosis/cell division proce two major clusters of samples could be observed. In every case, one of the clu appeared to include the majority of G1 samples, and the other appeared to includ majority of G3 samples. It can also be observed that much of the apparent inform within these heatmaps involves genes exhibiting an expression pattern of d regulation in most G1 samples, and up-regulation in most G3. This is biologi plausible, considering that G3 samples represent high-grade tumours which are proliferative than the low-grade tumours represented by G1 samples; thus g involved in cell division processes may be expected to show higher levels of expres in G3 samples than in G1 samples.

EGID	Symbol	Description	GOBP terms
23397	BRRN1	barren homolog 1 (Drosophila)	GO:0016043_cell organization and biogenesis GO:0007049_cell cycle GO:0051301_cell division GO:0007067_mitosis GO:0000087_M phase of mitotic cell cycle GO:0000279_M phase GO:0000278_mitotic cell cycle GO:0006996_organelle organization and biogenesis GO:0051276_chromosome organization and biogenesis GO:0007059_chromosome segregation GO:0000819_sister chromatid segregation GO:0000070_mitotic sister chromatid segregation GO:0007076_mitotic chromosome condensation GO:0030261_chromosome condensation
55143	CDCA8	cell division cycle associated 8	GO:0051301_cell division
4605	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	GO:0006366_transcription from RNA polymerase II promoter GO:0006915_apoptosis GO:0016265_death GO:0012501_programmed cell death GO:0008219_cell death GO:0042981_regulation of apoptosis GO:0043067_regulation of programmed cell death GO:0048519_negative regulation of biological process GO:0043118_negative regulation of physiological process GO:0048523_negative regulation of cellular process GO:0007049_cell cycle GO:0051726_regulation of cell cycle GO:0051243_negative regulation of cellular physiological process GO:0000074_regulation of progression through cell cycle GO:0006916_anti-apoptosis GO:0043066_negative regulation of apoptosis GO:0043069_negative regulation of programmed cell death
2354	FOSB	FBJ murine osteosarcoma viral oncogene homolog B	GO:0006366_transcription from RNA polymerase II promoter GO:0007610_behavior GO:0048519_negative regulation of biological process GO:0043118_negative regulation of physiological process GO:0048523_negative regulation of cellular process GO:0007049_cell cycle

			GO:0051726_regulation of cell cycle GO:0051243_negative regulation of cellular physiological process GO:0000074_regulation of progression through cell cycle GO:0016481_negative regulation of transcription GO:0045934_negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism GO:0009892_negative regulation of metabolism GO:0031324_negative regulation of cellular metabolism GO:0006357_regulation of transcription from RNA polymerase II promoter GO:0045892_negative regulation of transcription, DNA-dependent GO:0000122_negative regulation of transcription from RNA polymerase II promoter
6790	AURKA	aurora kinase A	GO:0006468_protein amino acid phosphorylation GO:0006796_phosphate metabolism GO:0016310_phosphorylation GO:0043412_biopolymer modification GO:0006793_phosphorus metabolism GO:0006464_protein modification GO:0016043_cell organization and biogenesis GO:0007242_intracellular signaling cascade GO:0007049_cell cycle GO:0007067_mitosis GO:0000087_M phase of mitotic cell cycle GO:0000279_M phase GO:0000278_mitotic cell cycle GO:0007010_cytoskeleton organization and biogenesis GO:0006996_organelle organization and biogenesis GO:0019932_second-messenger-mediated signalling GO:0048015_phosphoinositide-mediated signalling GO:0007017_microtubule-based process GO:0007051_spindle organization and biogenesis GO:0000226_microtubule cytoskeleton organization and biogenesis GO:0031647_regulation of protein stability
1062	CENPE	centromere protein E, 312kDa	GO:0016043_cell organization and biogenesis GO:0007049_cell cycle GO:0051641_cellular localization GO:0051649_establishment of cellular localization GO:0046907_intracellular transport GO:0051301_cell division GO:0007067_mitosis GO:0000087_M phase of mitotic cell cycle GO:0000279_M phase GO:0000278_mitotic cell cycle GO:0007010_cytoskeleton organization and biogenesis GO:0006996_organelle organization and biogenesis

			GO:0051276_chromosome organization and biogenesis GO:0006461_protein complex assembly GO:0007059_chromosome segregation GO:0007017_microtubule-based process GO:0051640_organelle localization GO:0051656_establishment of organelle localization GO:0007018_microtubule-based movement GO:0030705_cytoskeleton-dependent intracellular transport GO:0000819_sister chromatid segregation GO:0000070_mitotic sister chromatid segregation
7272	TTK	TTK protein kinase	GO:0006468_protein amino acid phosphorylation GO:0006796_phosphate metabolism GO:0016310_phosphorylation GO:0043412_biopolymer modification GO:0006793_phosphorus metabolism GO:0006464_protein modification GO:0008283_cell proliferation GO:0016043_cell organization and biogenesis GO:0048518_positive regulation of biological process GO:0043119_positive regulation of physiological process GO:0048522_positive regulation of cellular process GO:0051242_positive regulation of cellular physiological process GO:0008284_positive regulation of cell proliferation GO:0042127_regulation of cell proliferation GO:0007049_cell cycle GO:0051726_regulation of cell cycle GO:0000074_regulation of progression through cell cycle GO:0007067_mitosis GO:0000087_M phase of mitotic cell cycle GO:0000279_M phase GO:0000278_mitotic cell cycle GO:0007010_cytoskeleton organization and biogenesis GO:0006996_organelle organization and biogenesis GO:0007017_microtubule-based process GO:0007052_mitotic spindle organization and biogenesis GO:0007051_spindle organization and biogenesis GO:0000226_microtubule cytoskeleton organization and biogenesis GO:0007088_regulation of mitosis GO:0000075_cell cycle checkpoint GO:0007093_mitotic checkpoint
57758	SCUBE2	signal peptide, CUB domain, EGF-like 2	NA
2353	FOS	v-fos FBJ	

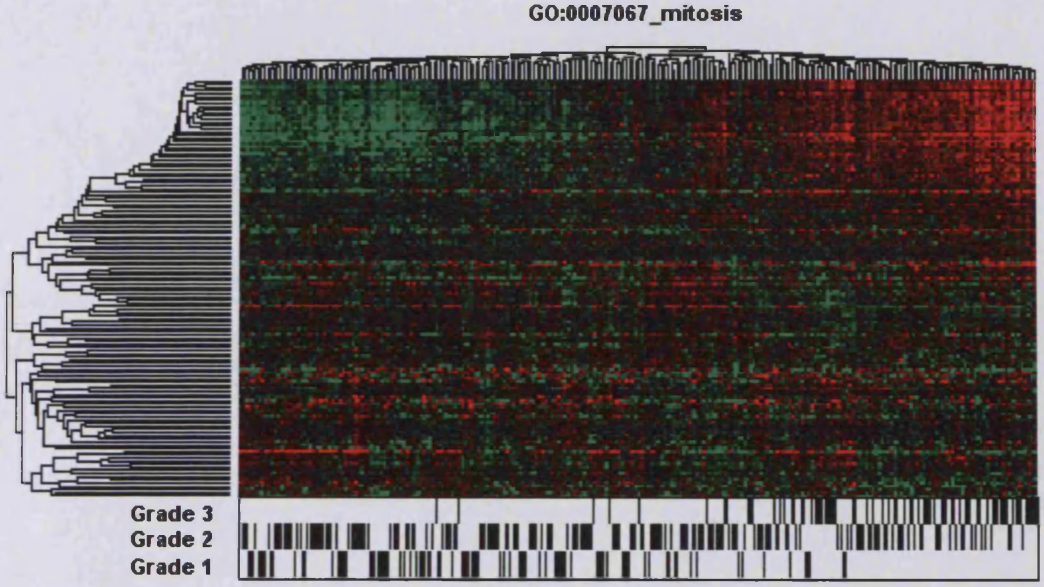
		murine osteosarcoma viral oncogene homolog	GO:0043412_biopolymer modification GO:0006259_DNA metabolism GO:0006950_response to stress GO:0006366_transcription from RNA polymerase II promoter GO:0006954_inflammatory response GO:0009605_response to external stimulus GO:0009611_response to wounding GO:0051707_response to other organism GO:0006952_defense response GO:0006955_immune response GO:0009607_response to biotic stimulus GO:0009613_response to pest, pathogen or parasite GO:0006357_regulation of transcription from RNA polymerase II promoter GO:0043414_biopolymer methylation GO:0006304_DNA modification GO:0006306_DNA methylation GO:0006305_DNA alkylation GO:0040029_regulation of gene expression, epigenetic
22974	TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)	GO:0008283_cell proliferation GO:0007049_cell cycle GO:0007067_mitosis GO:0000087_M phase of mitotic cell cycle GO:0000279_M phase GO:0000278_mitotic cell cycle
2305	FOXM1	forkhead box M1	GO:0042221_response to chemical stimulus GO:0006950_response to stress GO:0009628_response to abiotic stimulus GO:0006366_transcription from RNA polymerase II promoter GO:0006979_response to oxidative stress GO:0006800_oxygen and reactive oxygen species metabolism
9833	MELK	maternal embryonic leucine zipper kinase	GO:0006468_protein amino acid phosphorylation GO:0006796_phosphate metabolism GO:0016310_phosphorylation GO:0043412_biopolymer modification GO:0006793_phosphorus metabolism GO:0006464_protein modification
83461	CDCA3	cell division cycle associated 3	NA

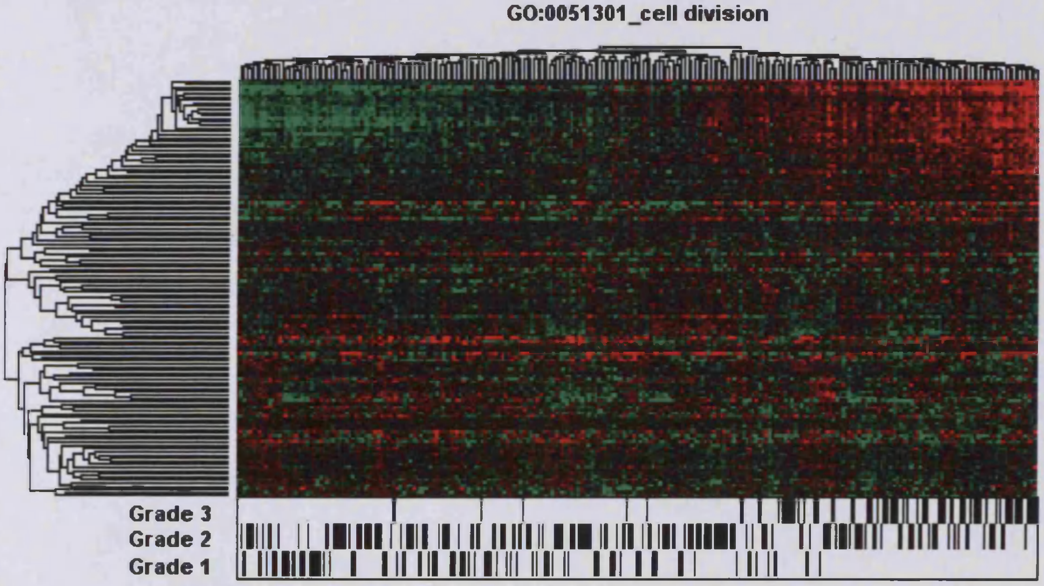
Table 7.5 Classifier genes that differentiate between G1 and G3 breast cancer samples. This was created using the PAM and SWS class prediction algorithm in the original study by Ivshina et al. Of the original 18 genes, only those 13 are shown that are represented on the hgu133a platform. Only those GOBP terms are shown that were tested during the GSD analysis. Terms highlighted in red indicate those selected by GSD analysis of this dataset.

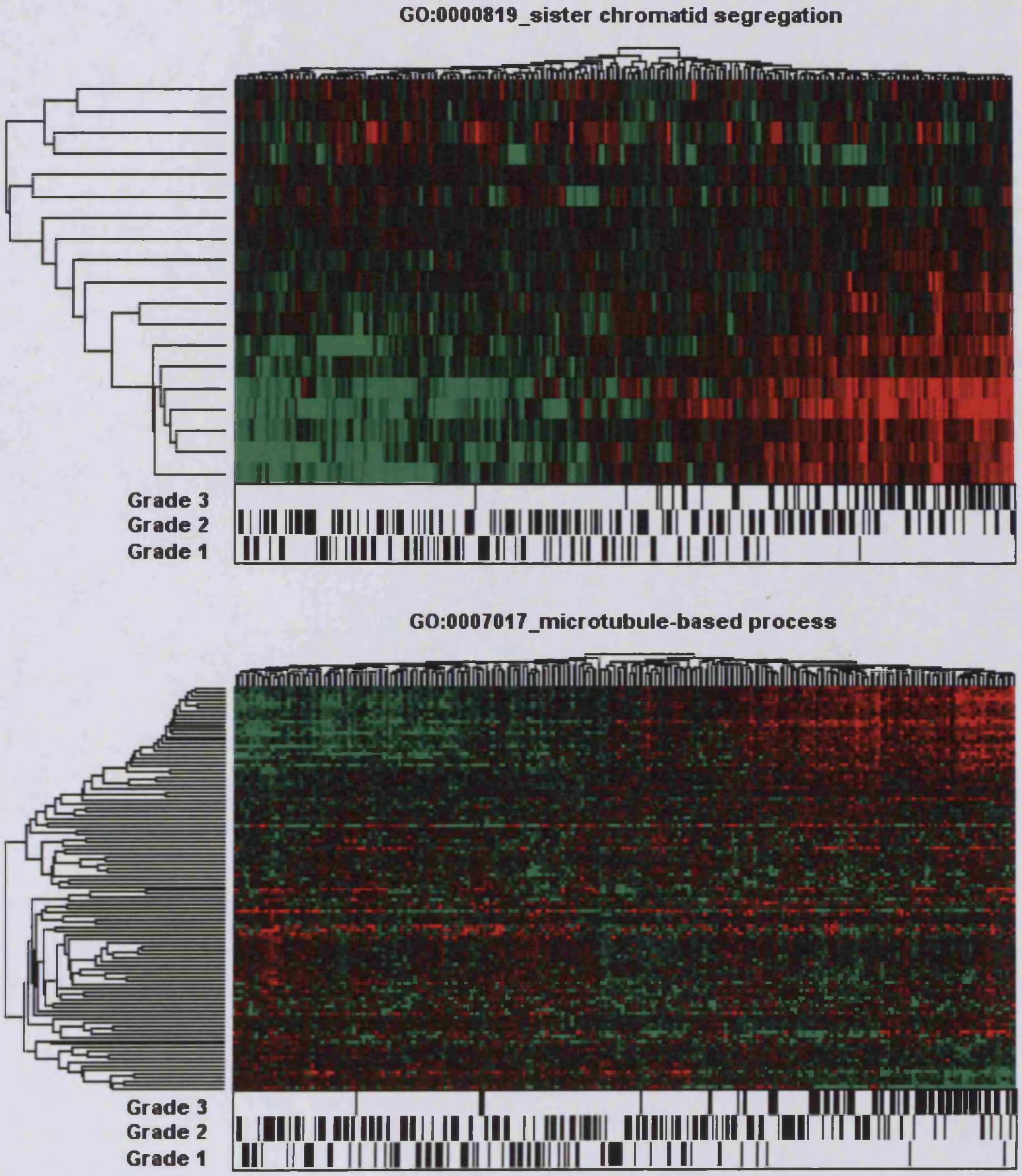
GOBP term	Hypergeometric p-value (FDR corrected)
GO:0007067_mitosis	0.00026
GO:0000087_M phase of mitotic cell cycle	0.00026
GO:0000279_M phase	0.00053
GO:0000278_mitotic cell cycle	0.00053
GO:0007049_cell cycle	0.00054

Table 7.6 GOBP terms found to be enriched in the classifier gene set. ORA analysis of the gene set shown in Table 7.5 was carried out using the hypergeometric test. Terms displayed are those that were found to have FDR-corrected p-values of less than 0.01.

The heatmaps also indicate that the G2 samples do not appear to show any tendency to cluster separately from samples representing other tumour grades: they show apparently random ordering within each of the two major clusters. In the original study, the classifier gene set that could discriminate between G1 and G3 samples was used by the authors to stratify G2 samples into two groups: G2a (G1-like) and G2b (G3-like). Also, in the original study, ORA analysis of genes found to be differentially expressed between G2a and G2b samples detected enrichment of terms specific to cell division and cell cycle processes. It was therefore investigated whether the mitosis/cell division terms selected by GSD analysis could also stratify G2 samples.


GO:0007067_mitosis
Grade 3
Grade 2
Grade 1


GO:0051301_cell division
Grade 3
Grade 2
Grade 1



GO:0000819_sister chromatid segregation

GO:0007017_microtubule-based process

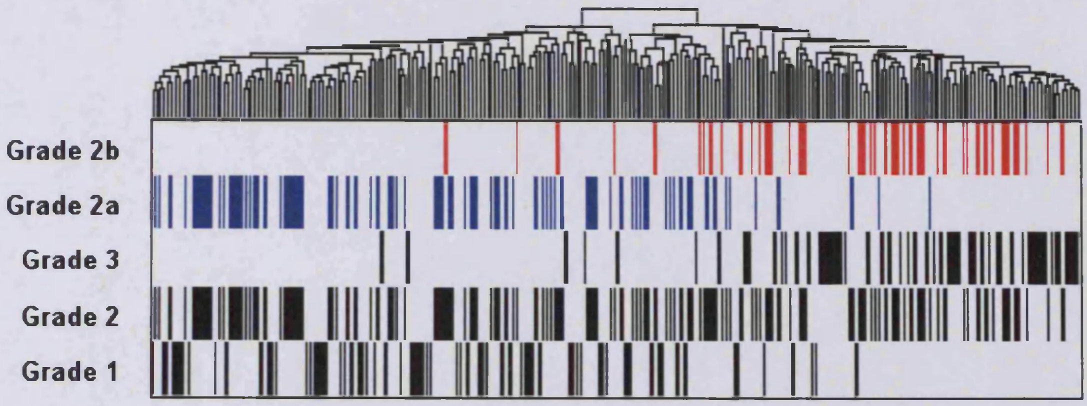
Figure 7.18 Selected GOBP terms specific to cell division processes detected by GSD analysis of the Ivshina Breast Cancer dataset. Heatmaps represent log median-centred MAS5 normalized data ranging in value from -2 (bright green) through 0 (black) to 2 (bright red), and were created using correlation distance and average linkage. Values greater than 2 and less than -2 were set to 2 and -2 respectively. Black bars in the 'picketplots' below the heatmaps indicate the tumour grade in which each of the samples was classified in the original study using the Nottingham Grading System for breast cancers.

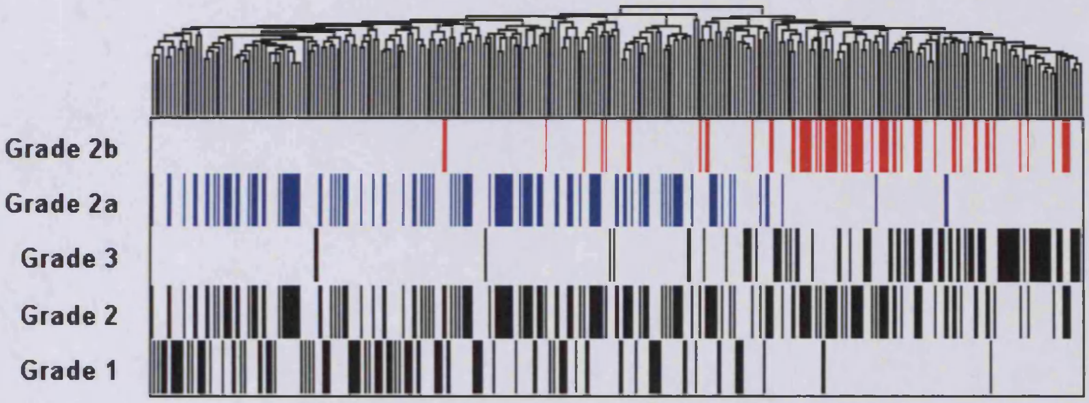
For this purpose, the positions of samples that were classified in the original study as G2a and G2b, within the hierarchical clustering of samples produced for each of the mitosis/cell division terms selected by GSD analysis, were investigated. To illustrate this, displayed in Figure 7.19 are the cluster dendrograms from the heatmaps for some of these terms, along with picketplots supplemented with information regarding which samples were classified as G2a and G2b in the original study.

As can be seen in Figure 7.19, the clustering of G2 samples appears to be in concord with the sub-classes assigned in the original study: most G2a samples appear to be interspersed within the cluster that is also contains the majority of G1 samples. Similarly, most G2b samples are found interspersed within the cluster that includes the majority of G3 samples.

To statistically assess the enrichment of tumour grades within these clusters, the number of samples of each tumour grade within each of the first two hierarchical clusters exhibited by each of the cell division specific GOBP terms was counted. Tests were then carried out, to detect over-representation of each tumour grade within that cluster where the majority of samples for that grade were found, using the hypergeometric statistical test (see Materials and Methods).

These results are displayed in Table 7.7. In every case there is highly significant enrichment of tumour grades within the respective clusters. These findings imply that the grouping of tumour grades, as provided by hierarchical clustering for these terms, is non-random and unlikely to have occurred by chance. This then highlights the importance of the biological theme in the differentiation between low (G1 and G2a) and high (G2b and G3) grade breast cancer samples.


GO:0007067_mitosis


GO:0007049_cell cycle

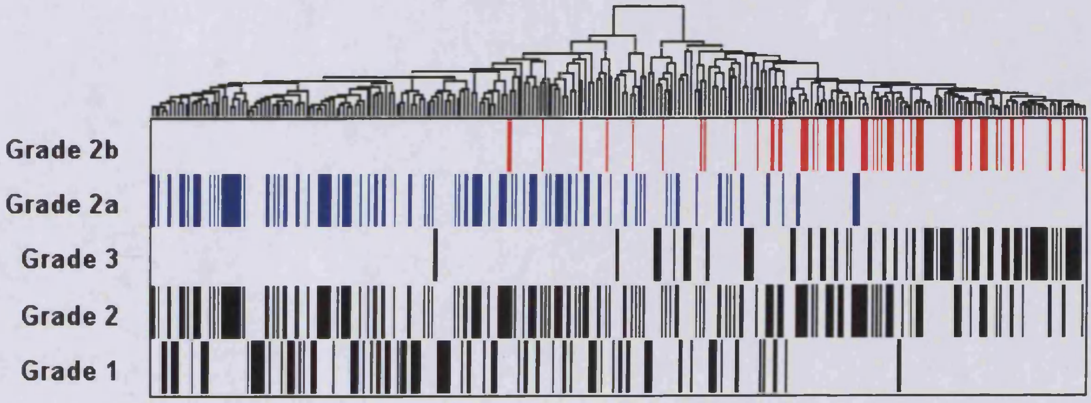

GO:0000819_sister chromatid segregation

Figure 7.19 Stratification of Grade 2 breast cancer samples by hierarchical clustering using GOBP terms specific to mitosis/cell division processes. Figures display dendrograms representing hierarchical clusters of samples classes. Also displayed are picketplots where the tumour grade of each sample is indicated by the presence of coloured bars. Black bars indicate tumour grades G1, G2 and G3, which were assigned to samples in the original study through morphological and cytological assessments using the Nottingham Grading System. Coloured bars indicate sub-classes of G2 samples that were discovered *de novo* in the original study using a gene set created using class discovery techniques.

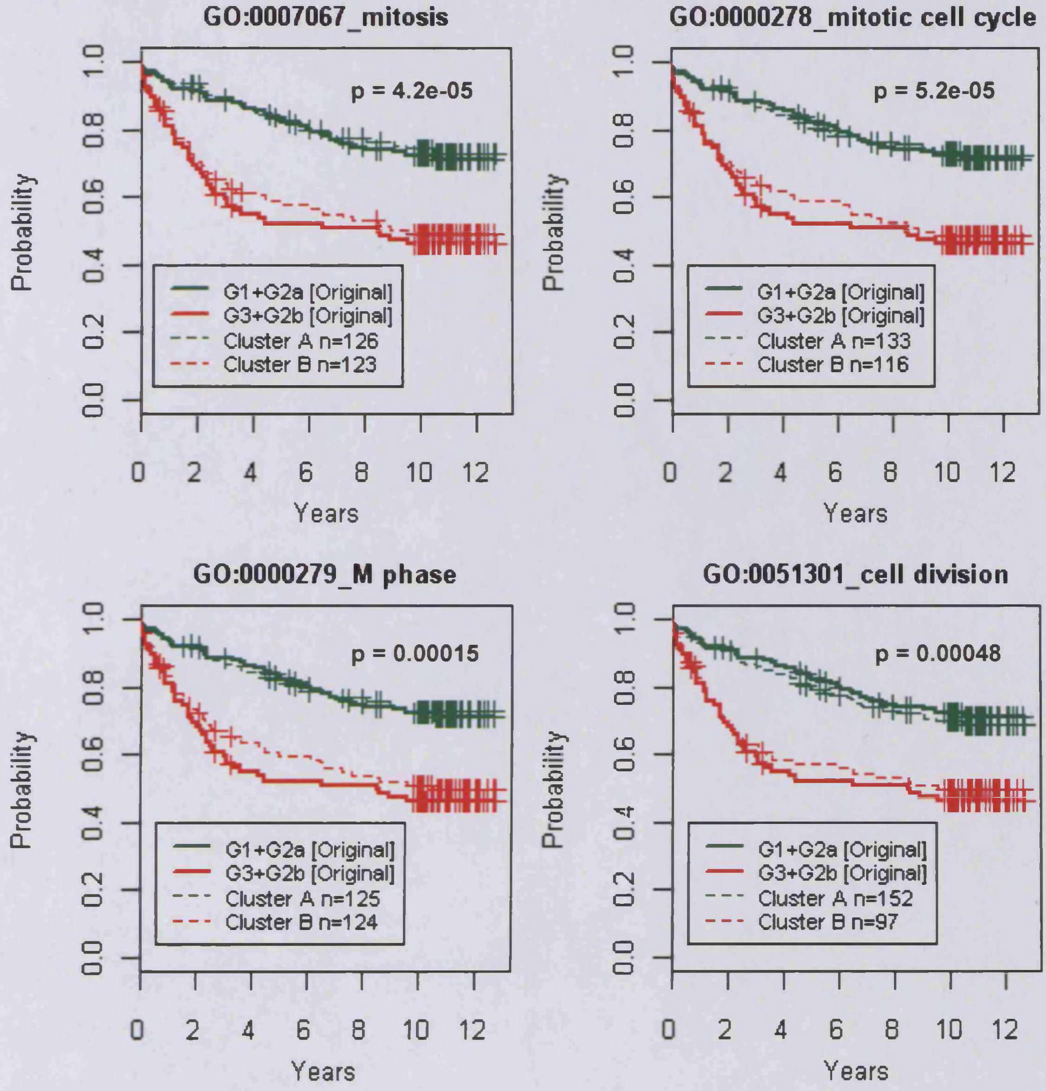
Patient survival data was also available for each sample within the dataset, and in the original study this was used to display the highly significant difference in survival rates between patients with different grades of tumours. Assessment of the difference in survival rates between patients with tumours classified in the original study as either G1 or G2a and those with G3 or G2b tumours using the Cox proportional-hazards regression model (see Materials and Methods) yielded a highly significant p-value of 2.1e-06.

To investigate if the relevance of the biological theme of cell division within this experiment could be reflected in patient survival rates, the following analysis was carried out: for each of the cell division specific GOBP terms selected by the GSD analysis of the dataset, two groups of samples were created by selecting the first two hierarchical clusters (as was done in the explorations displayed in Table 7.7).

The difference in survival rates of patients represented by the samples in each of the two groups was then assessed using the Cox proportional-hazards regression model. Kaplan-Meier survival curves are plotted for the two groups of samples created for each of the GOBP terms in Figure 7.20, along with the survival curves for the groups of samples based on tumour grades assigned in the original study.

Gene Ontology Biological Process Term	Cluster A		Cluster B	
	G1	G2a	G3	G2b
GO:0007067_mitosis	79.4% (1.3e-08)	75.9% (1e-08)	90.9% (3e-13)	90.7% (5e-10)
GO:0000278_mitotic cell cycle	82.4% (8.1e-09)	81.9% (6.4e-11)	94.5% (1.7e-17)	86% (5.5e-09)
GO:0000087_M phase of mitotic cell cycle	77.9% (2e-08)	77.1% (3.6e-10)	92.7% (7.7e-14)	95.3% (5.1e-12)
GO:0000279_M phase	79.4% (8.7e-09)	74.7% (3.1e-08)	90.9% (4.9e-13)	90.7% (7.1e-10)
GO:0051301_cell division	89.7% (1.9e-09)	91.6% (1.4e-13)	90.9% (1e-19)	76.7% (3.5e-08)
GO:0007049_cell cycle	91.2% (1.2e-11)	88% (4.3e-12)	90.9% (5.3e-18)	86% (4.5e-11)
GO:0051726_regulation of cell cycle	82.4% (3.2e-10)	74.7% (5.1e-08)	94.5% (7.8e-16)	88.4% (5.8e-09)
GO:0000074_regulation of progression through cell cycle	85.3% (3.7e-11)	74.7% (3.4e-07)	92.7% (2.2e-15)	86% (1.5e-08)
GO:0000819_sister chromatid segregation	80.9% (3.1e-08)	84.3% (5.6e-13)	96.4% (8.4e-19)	88.4% (7.4e-10)
GO:0007018_microtubule-based movement	79.4% (3.2e-08)	77.1% (5.5e-09)	89.1% (1.7e-12)	90.7% (2.5e-10)

Table 7.7 Enrichment of tumour grades within hierarchical clusters of samples created by GOBP terms specific to cell division processes. Hierarchical clustering of samples was carried out using correlation distance and average linkage for each of the gene sets representing GOBP terms specific to cell division processes that were selected by GSD analysis of the Ivshina breast cancer datasets. Samples were then divided by selecting the first two hierarchical clusters (Clusters A and B) for each gene set separately. Figures indicate the percentage of samples of the dominant tumour grades within those clusters. Also displayed are hypergeometric p-values calculated to assess the enrichment of the tumour grades within those clusters.


GO:0007067_mitosis
p = 4.2e-05
G1+G2a [Original]
G3+G2b [Original]
Cluster A n=126
Cluster B n=123
GO:0000278_mitotic cell cycle
p = 5.2e-05
G1+G2a [Original]
G3+G2b [Original]
Cluster A n=133
Cluster B n=116
GO:0000279_M phase
p = 0.00015
G1+G2a [Original]
G3+G2b [Original]
Cluster A n=125
Cluster B n=124
GO:0051301_cell division
p = 0.00048
G1+G2a [Original]
G3+G2b [Original]
Cluster A n=152
Cluster B n=97
Probability
0.0
0.2
0.4
0.6
0.8
1.0
Years
0
2
4
6
8
10
12

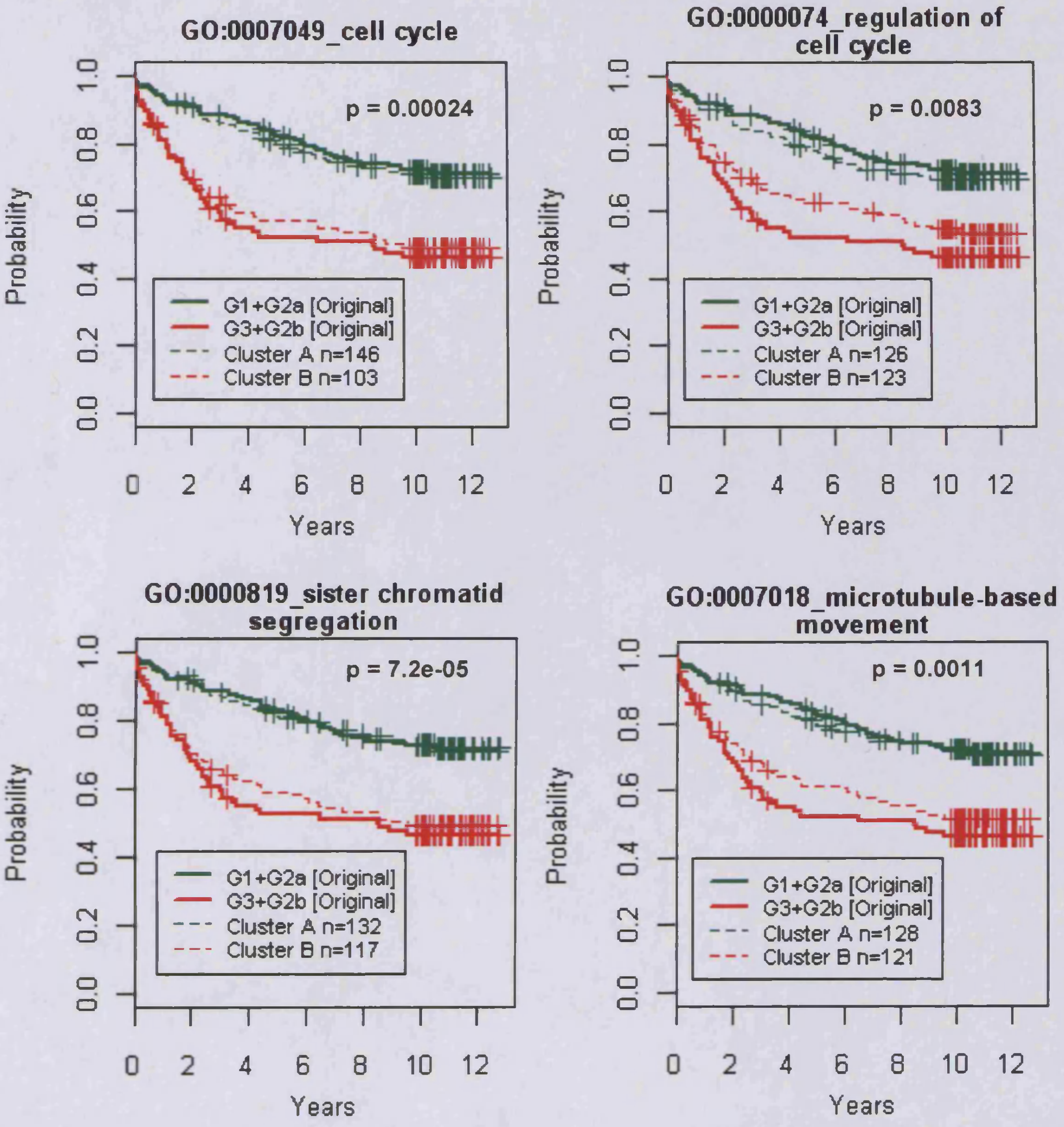


Figure 7.20 Survival curves for patients stratified by hierarchical clustering of samples using gene sets specific to cell division processes. Hierarchical clustering of samples was carried out using correlation distance and average linkage for each of the gene sets representing GOBP terms specific to cell division processes that were selected by GSD analysis of the Ivshina breast cancer datasets. Broken survival curves represent patients grouped in each of the first two hierarchical clusters for each gene set. Solid curves represent groups of patients created according to tumour grades assigned in the original study. P-values in the top right corner indicate the significance of the difference in survival rates of cluster-based groups of patients, using the Cox proportional-hazard regression model.

As can be observed, there is considerable difference in survival rates of patients represented in each of the clusters; assessment of the statistical significance of these differences yields highly significant p-values for all tested gene sets.

Thus, the results of the implementation of the GSD methodology on this dataset are in concord with the findings of the original study on two major aspects: The first is that the GOBP terms selected by the analysis include many that are specific to cell division processes, which was found in the original study to be the most important biological theme with regard to discrimination between high and low grades of tumour.

The second is the finding in the original study that there is no apparent continuum of gene expression levels in G2 samples as they progress from G1 to G3 stages; rather G2 samples appear to comprise at least two distinct molecular sub-classes, G2a and G2b, depending on the similarity of their gene expression profiles to those of G1 and G3 samples respectively. This is supported by the findings that when the hierarchical clustering of samples is carried out using GOBP terms specific to cell division processes, there is a marked tendency for the majority of G1 and G2a samples to cluster together, and for the G2b and G3 samples to cluster together. This is further supported by survival analyses carried out both here and in the original study.

7.4 Discussion

The previous chapter described explorations leading to the development of the GSD methodology using hypothetical gene expression matrices, randomly selected gene sets and artificially imposed information. This chapter then described the results of the application of the methodology to the analysis of four microarray datasets using gene sets based on GOBP terms. Assessment of a methodology like GSD, which is essentially exploratory in nature, is problematic as there is no known 'truth' regarding these datasets. It was thus hoped that some indication of the accuracy and reliability of results could be achieved by investigating whether the results of the analyses were biologically plausible and/or they were in agreement with the results of prior analyses of these datasets using other methodologies.

7.4.1 Overview of results

The GNF human tissue expression dataset (Su et al. 2004), which was used in Chapter 5 to explore tissue-specific gene expression patterns was selected for analysis using GSD because it contains strong tissue-specific expression signatures that lead to clustering of samples based on tissue types. Inspection of the GOBP terms selected by GSD analysis of this dataset revealed considerable numbers of terms specific to immune and nervous system processes. These results most likely reflect the observation that tissues specific to these processes comprise the two largest groups of tissues of similar origin and/or function.

The next three datasets analyzed involved samples from cancer patients: this was because it is expected that the ability of the GSD methodology to identify potentially relevant biological themes within a dataset without requiring prior definition of sample classes could be of utility particularly in the analysis if cancer datasets where sample classes may not be known, or where the samples are difficult to classify.

GSD analysis of the Ross AML dataset (Ross et al. 2004) resulted in selection of many immunity-specific GOBP terms, consistent with the biological nature of the samples (mononuclear cells that involved in immunity functions). Inspection of the heatmaps of the selected GOBP terms revealed gene expression patterns that resulted in groupings of samples that agreed with the five different known genetic sub-types of AML represented within the dataset.

Next, GSD was applied to a liposarcoma dataset provided by a collaborator at the Memorial University Medical Centre. Two out of the three major biological themes selected by GSD for this dataset are biologically plausible in the context of the tumour grades of the samples: genes comprising the mitosis/cell division theme exhibited expression patterns of up-regulation in the high-grade tumour samples and down-regulation in the low-grade samples. The opposite pattern was exhibited by genes comprising the lipid metabolism theme. The third theme, that of immunity, is likely to have been selected as a consequence of the possible presence of cells involved in the immune system within the samples.

Finally, GSD analysis of the Ivshina breast cancer dataset (Ivshina et al. 2006) revealed at least two major biological themes. One of these was cell division; this finding is biologically plausible (as high-grade tumours are typically more proliferative than low-grade tumours) and in agreement with the results of more supervised analyses carried out in the original study that took into account the histological grades of the samples. Further analyses carried out appear to highlight the importance of discovering relevant biological themes within a dataset – simple hierarchical clustering of samples using the GOBP terms specific to cell division could differentiate between patients into groups with statistically significant differences in survival rates.

Thus it is found that the GSD methodology can provide plausible and useable results. The potential utility of this technique is further discussed in Chapter 9.

7.4.2 Genes and expression patterns shared by selected gene sets

Having identified the possibly relevant biological themes within a dataset (which may have been achieved by GSD or any other gene set analysis method), there are at least two issues that need to be addressed in order to help researchers further analyze these results.

The first of these is that, as can be observed in heatmaps of most selected gene sets, much of the information is restricted to only a proportion of genes within a gene set, i.e. only a sub-set of genes are DEGs. This then raises the need for a methodology to identify these informative genes within the selected gene sets.

The second issue is that of comparison of gene expression patterns within the selected gene sets: a researcher would be interested in knowing all the different types of information that are represented by the selected gene sets. Some gene sets may exhibit the same type of information as each other i.e. they could contain expression patterns that result in similar hierarchical clusters of samples. This may be because they represent similar biological themes and share many genes (for example, as would be seen for GO terms that share parent-child relationships). More interestingly, these may represent disparate biological processes (and thus do not share genes) but are affected in the same way by the experimental conditions.

The next chapter describes explorations to resolve both these issues. It first describes a methodology that could allow for ranking (and subsequent selection) of genes based on their information content. It then describes the implementation of a scheme that allows for simultaneous visualization of all gene expression patterns exhibited by all selected gene sets, indication of which patterns are exhibited by which gene sets, and the relationships between the gene sets (based on shared genes).

Chapter 8: Extraction of informative genes and visualization of GSD results

8.1 Introduction

Chapter 6 described the development of GSD, a methodology that could identify potentially relevant biological themes within a microarray dataset. Chapter 7 then described the application of this methodology to four microarray datasets.

Visual inspection of heatmaps of gene sets selected by GSD analysis showed that in many cases much of the information within a gene set appeared to be contained in the expression values for only a subset of the genes. Ostensibly, only these 'informative genes' would be of interest to the researcher and this then raises the need for a methodology for extraction of these genes from within a heatmap.

One of the key differences between the GSD methodology and other gene set analysis methods is that it is unsupervised in terms of the relationship between expression patterns and classes. Typically, supervised gene set analyses of a datasets utilizing prior knowledge of sample classes would involve identification of gene sets that exhibit a single expression pattern (for example, differential expression between tumour and normal tissue). In contrast, GSD analysis attempts to identify those gene sets that exhibit significantly strong expression patterns, regardless of what the expression pattern is, and it is possible that gene sets identified by GSD analysis could exhibit several different expression patterns. This raises the need for identification and comparison of the expression patterns exhibited by gene sets selected by GSD analysis.

This chapter describes the development of a methodology that attempts to identify informative genes within selected gene sets. It then describes the results of the application of this methodology to the gene sets selected by GSD analysis of the four datasets described in Chapter 7.

Also described is the implementation of a strategy for visualization of the results of GSD analysis that can present the results, together with information that is likely to be of interest to researchers, in an integrated format. This allows for simultaneous visualization of the different gene expression patterns exhibited by the selected genes, indication of which gene sets each of the selected genes belongs to, and the relationships between the gene sets (based on shared genes).

8.2 Results and Explorations

8.2.1 Quantifying the prevalence of an expression pattern

The information content of any gene expression matrix is typically limited to groups of genes that exhibit very similar expression patterns. The consistency of these expression patterns leads to strong clustering of samples, which in turn leads to the selection of the gene set by the GSD methodology.

The hierarchical clustering of genes within a heatmap requires the calculation of distances between each possible pair of genes to yield a gene distance matrix (similar to the sample distance matrix used in the hierarchical clustering of samples, which is utilized by the GSD methodology). As these distances quantify the similarity between the expression profiles of any pair of genes, it was reasoned that they could be used to identify informative genes, i.e. those that share expression patterns with many other genes in that gene set.

Each row and column of a gene distance matrix consists of a vector of values of distances of any one gene to all other genes in that gene set. It was hypothesised that the variability of this distribution of distances of one gene to all others could indicate how informative a gene is. The reasoning behind this is similar to that of the use of SD-SDM values in GSD analysis: if the expression pattern of a gene is shared by many other genes, then the distance distribution for that gene would contain higher levels of short distances (to genes with similar expression patterns) and long distances (to genes exhibiting dissimilar or even opposite expression patterns) as compared to the distance distributions of genes that exhibit random expression patterns (i.e. uninformative genes).

To test this hypothesis, a hypothetical gene expression matrix consisting of 30 samples and 500 genes was created by randomly selection from a normal distribution with a zero mean and standard deviation of 0.3 (similar to the scheme used in Chapter 6 – see

Materials and Methods). 200 genes were then randomly selected to exhibit an expression pattern of up-regulation, down-regulation and no change in 10 samples each. Another 50 genes were randomly selected to exhibit a different pattern of expression. Correlation distances were than calculated for every possible pair of genes. The standard deviation of the gene distance vectors (SD-GDV) were then recorded for each of the genes. Figure 8.1 shows the heatmap of this matrix, along with SD-GDV values for each of the genes.

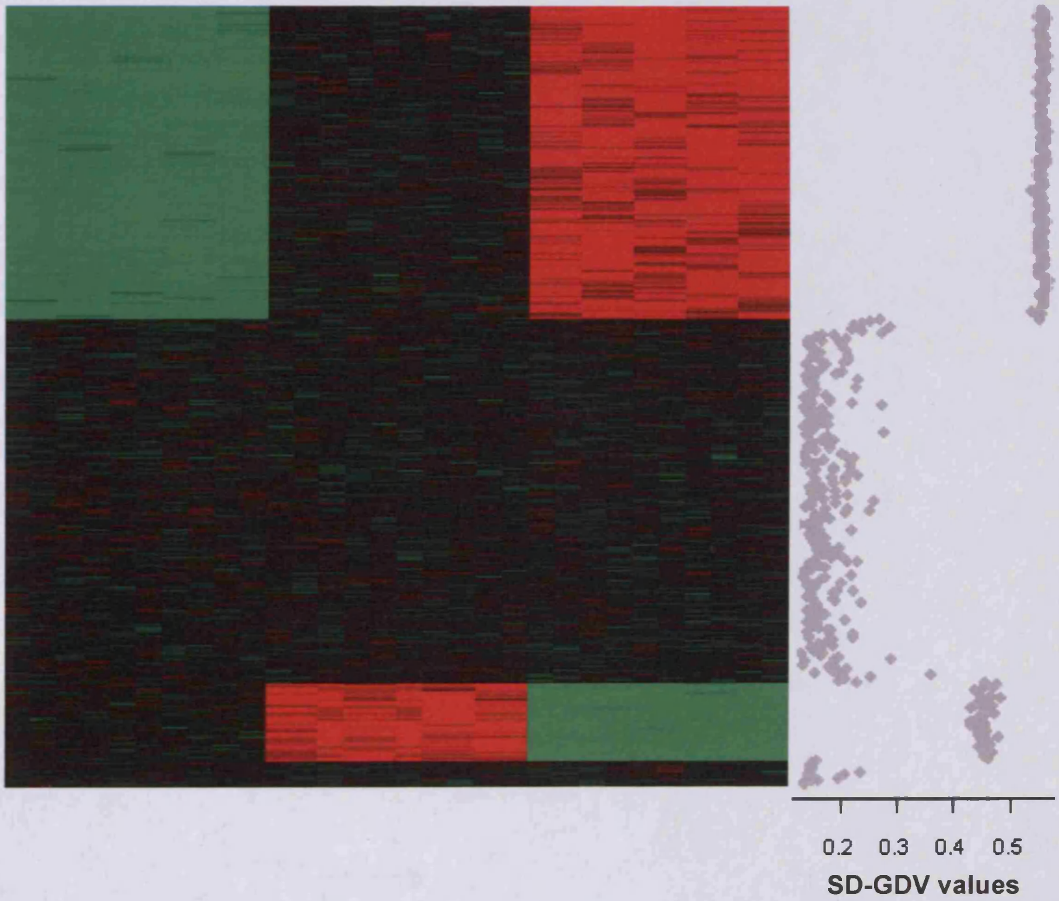


Figure 8.1 Heatmap of a hypothetical gene expression matrix with artificially introduced information content. The heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Hierarchical clustering of genes and samples was carried out using correlation distance and average linkage. Also displayed is a plot of SD-GDV values for each of the genes at their respective positions within the heatmap.

Two observations can be made: firstly, SD-GDV values for genes into which information was experimentally introduced are higher than for the other genes. Secondly, the SD-GDV values for the informative genes are greater when the expression pattern is shared by a greater number of genes: the SD-GDV values for each of the larger set (n=200) of informative genes is 0.55, while that of each of the smaller set (n=15) of informative genes is 0.45. The median SD-GDV for all other genes is 0.16. The distance distributions for each of these sets of genes are displayed in Figure 8.2.

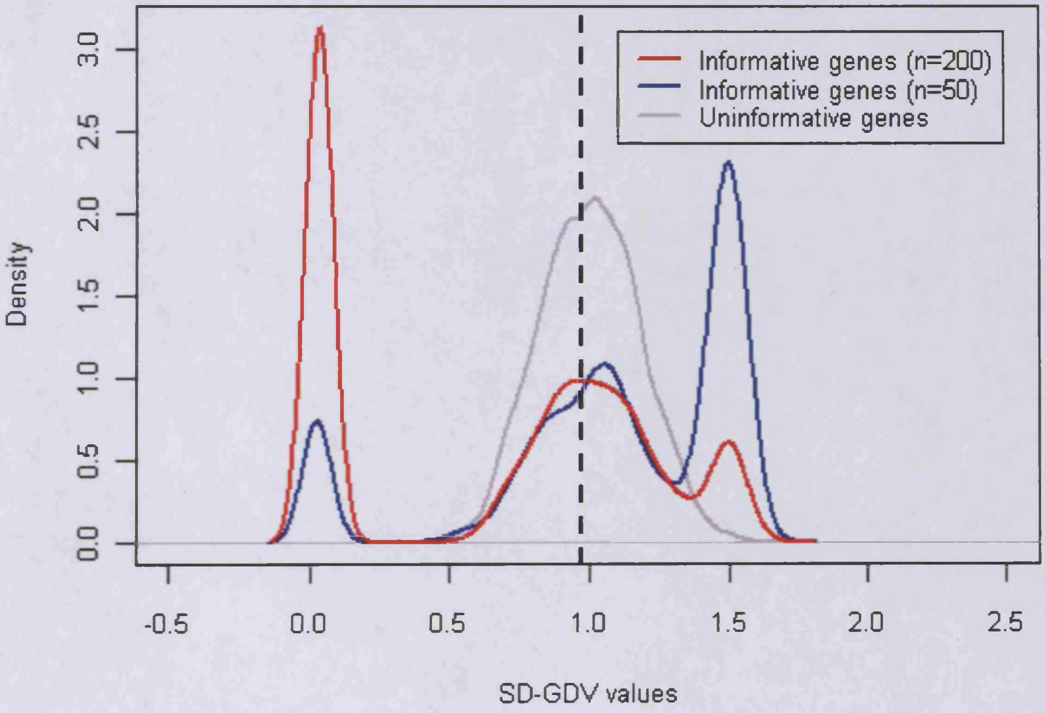


Figure 8.2 Gene distance distributions for informative and uninformative genes. Curves represent gene distance distributions for the gene expression matrix represented in Figure 8.1. Broken vertical line represents the median gene distance for all pairs of genes.

These results suggest that SD-GDV values may be used to quantify how informative a gene is, and thus possibly allow for the extraction of informative genes from within gene sets.

To investigate if similar results could be obtained from 'real-world' data, the GO term for mitosis, which was found to be highly significant in GSD analysis of the Ivshina breast cancer dataset was used. The heatmap for this term is displayed in Figure 8.3, along with the SD-GDV values calculated for each of the genes.

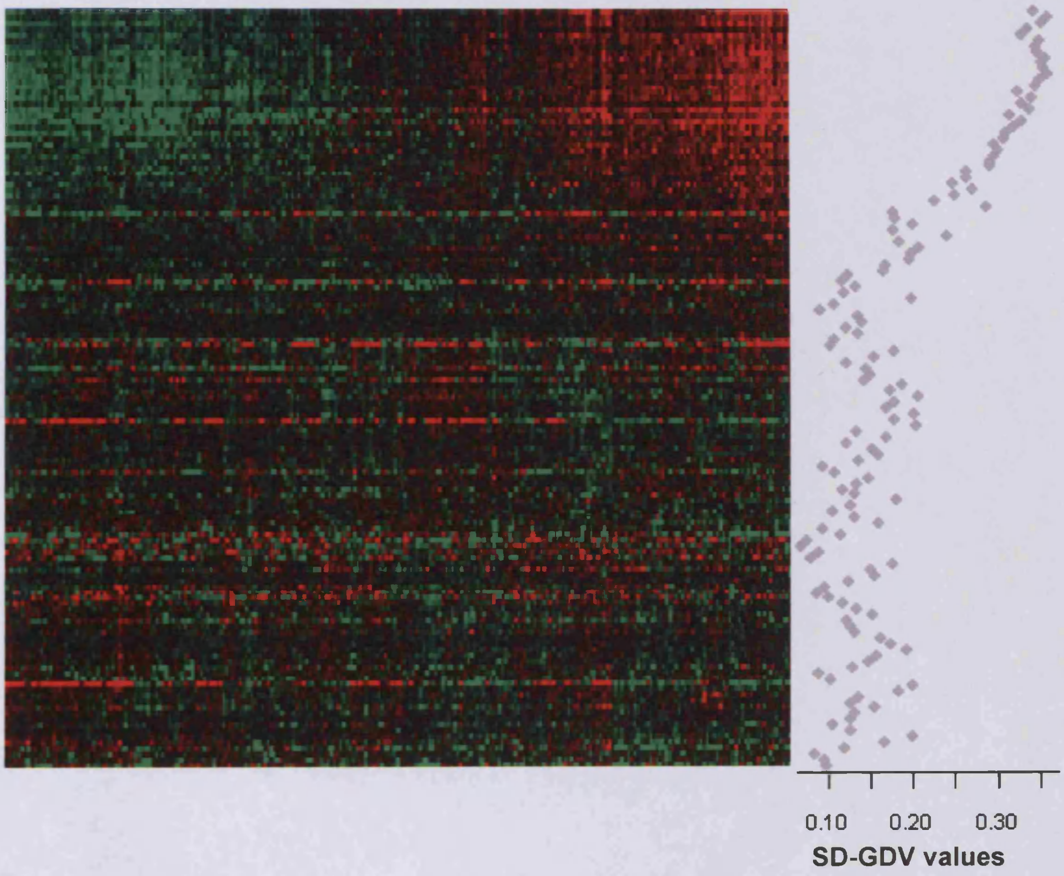


Figure 8.3 Heatmap of the term GO:0007067 (mitosis) on the Ivshina breast cancer dataset. The heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Hierarchical clustering of genes and samples was carried out using correlation distance and average linkage. Also displayed is a plot of SD-GDV values for each of the genes at their respective positions within the heatmap.

Again, it is observed that there is a tendency for SD-GDV values to be higher in those genes that are visually informative, as compared to those that are not. For these reasons, SD-GDV values were presumed to be suitable for the purpose of extracting informative genes from within gene sets; the next section describes the implementation of this technique on the GSD results of datasets analyzed in Chapter 7.

8.2.2 Extraction of informative genes using SD-GDV values

The extraction of informative genes from within gene sets selected by GSD analyses was implemented using a 'global' strategy in order to capture relationships between genes (in terms of their expression patterns) across the different gene sets.

Thus SD-GDV values were calculated for genes of all gene sets selected by GSD analysis of a microarray dataset taken together (that is, the SD-GDV value was calculated for each gene using a vector of distances of that gene to all other genes from all gene sets selected by GSD). This was performed for each of the four microarray datasets described in Chapter 7. In each case, the genes were ranked in decreasing order of their respective SD-GDV values, and then grouped into quintiles (such that the first quintile comprised of 20% of the genes with the highest SD-GDV values). The heatmaps for each of these quintiles is displayed in Figure 8.4.

As can be observed, there appears to be a gradient of information in the quintile heatmaps: the proportion of genes that are informative (i.e. show consistent and strong expression patterns) appears to be highest in the first quintiles. This proportion decreases in successive quintiles till there are few or no apparent informative genes in the fifth quintiles.

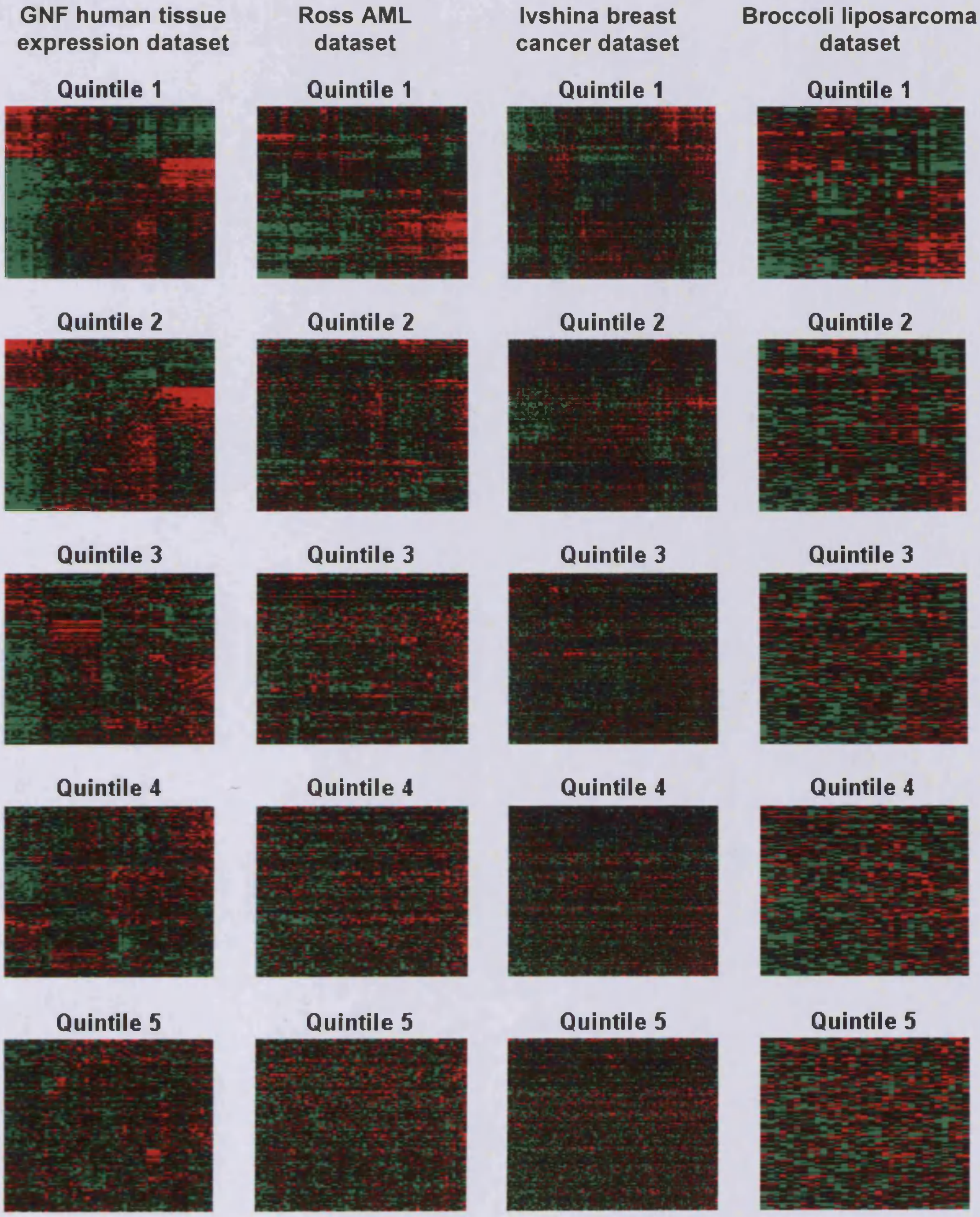
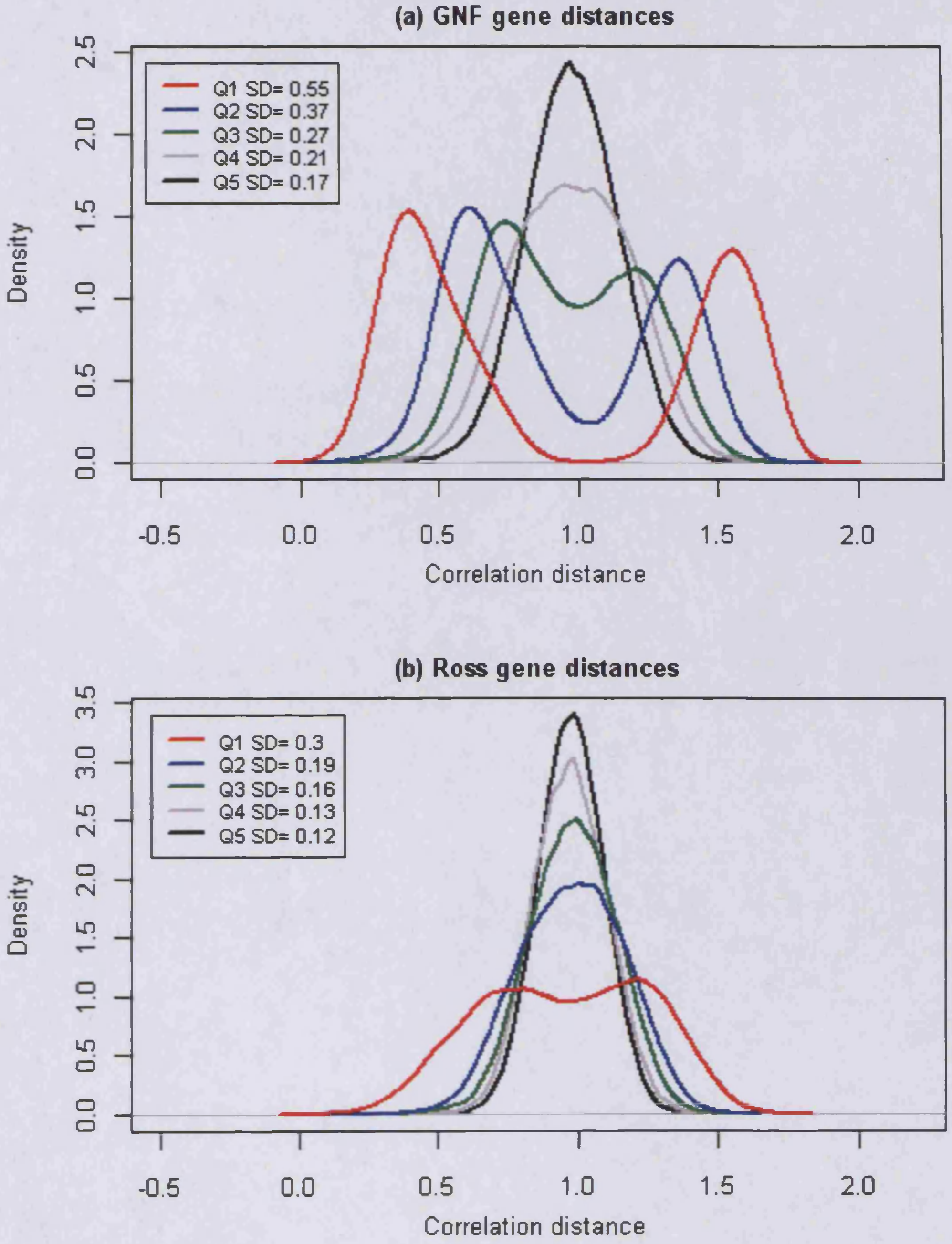


Figure 8.4 Quintile heatmaps of genes from all gene sets selected by GSD analysis ranked by SD-GDV values. Heatmaps represent log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Hierarchical clustering of genes and samples was carried out using correlation distance and average linkage.

Figure 8.5 shows the distributions of distances between all pairs of genes for each of the quintiles. There is a marked decrease in the variability of the distributions moving from the first to the last quintiles implying that the strength of gene clustering is highest in the first quintile, and decreases with each successive quintile. Figure 8.6 shows the distributions of distances between all pairs of samples for each of the quintiles. Here, a noticeable decrease in the variability of the distances can be observed, moving from the first to the fifth quintiles. These indicate that clustering of samples is strongest in the first quintile, and this decreases with each successive quintile.

Thus, ranking of genes according to their SD-GDV values appears to result in the enrichment of informative genes at the top of the order. The issue then arises regarding cut-off SD-GDV values that would need to be set to extract the most informative genes for each dataset. It is apparent (as might be expected) that the number of informative genes varies across datasets. While the vast majority of informative genes appear to be contained within the first two quintiles for the cancer datasets, informative genes can be observed as far as the fourth quintile for the GNF tissue expression dataset (because large numbers of genes exhibit tissue-specific expression patterns).

As 'customized' extraction of informative genes for each of the datasets is subject to possible user bias, an arbitrary selection of the top 20% of genes (i.e. the first quintile) when ranked by SD-GDV values was implemented for each of the datasets in the subsequent explorations.



(a) GNF gene distances

(b) Ross gene distances

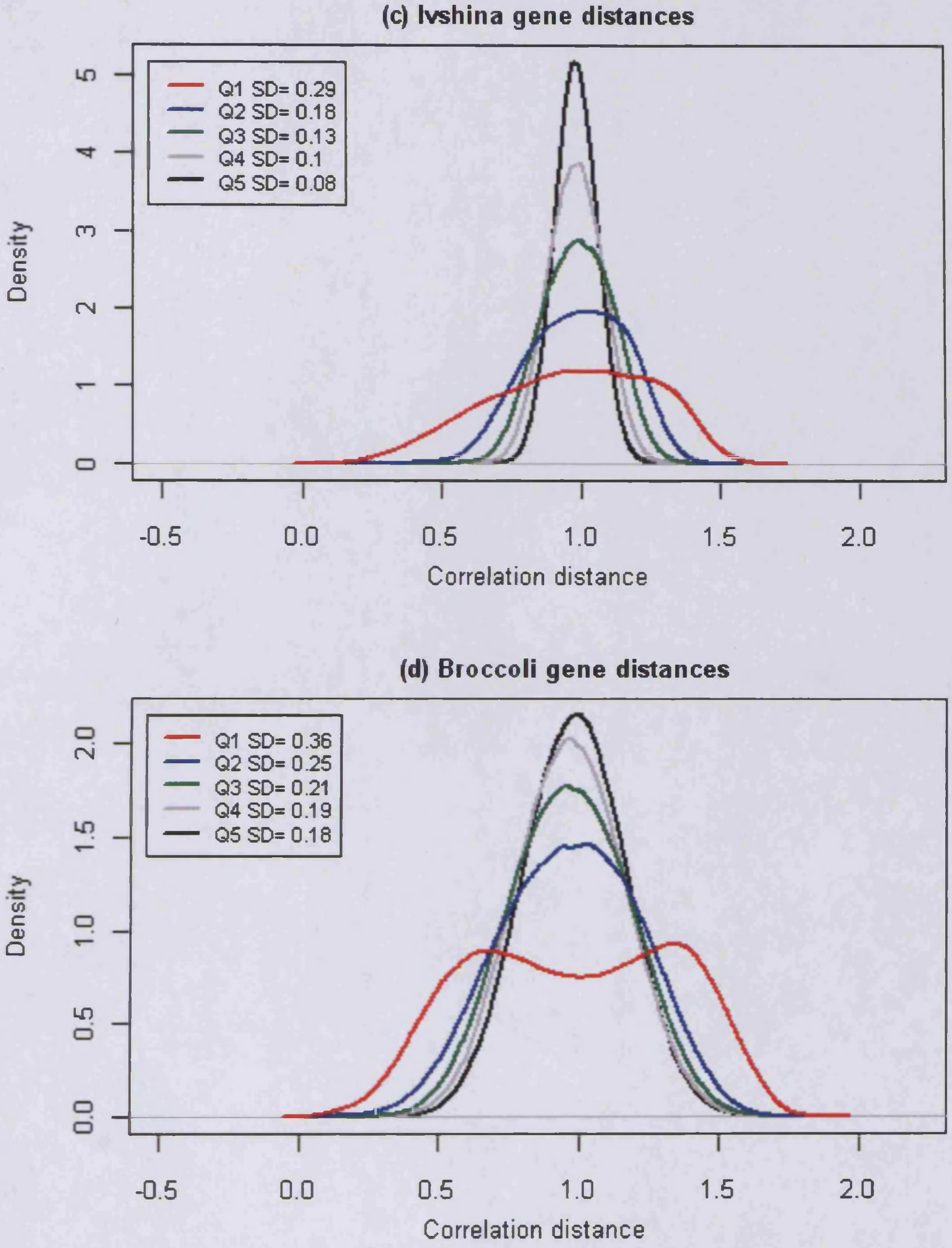
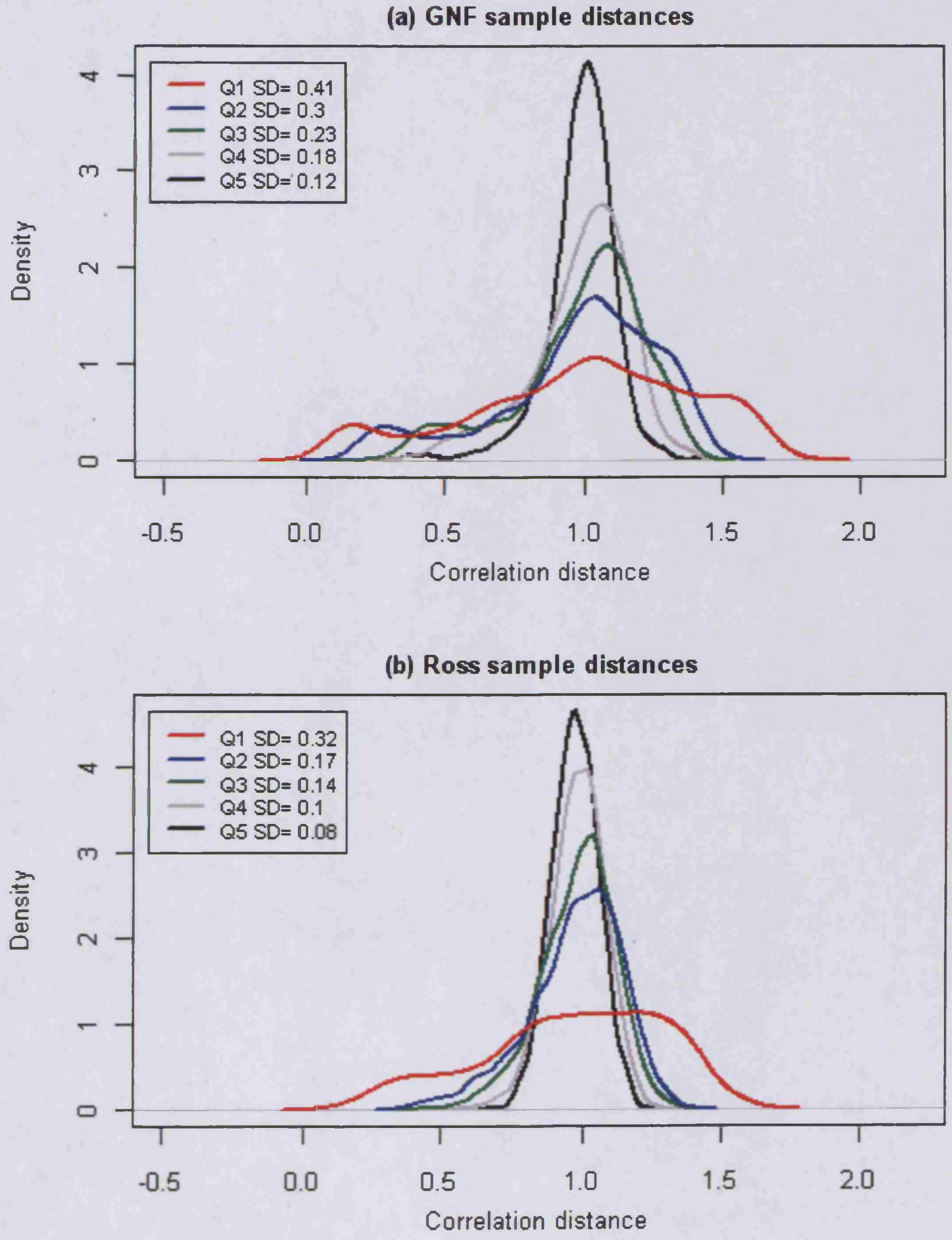
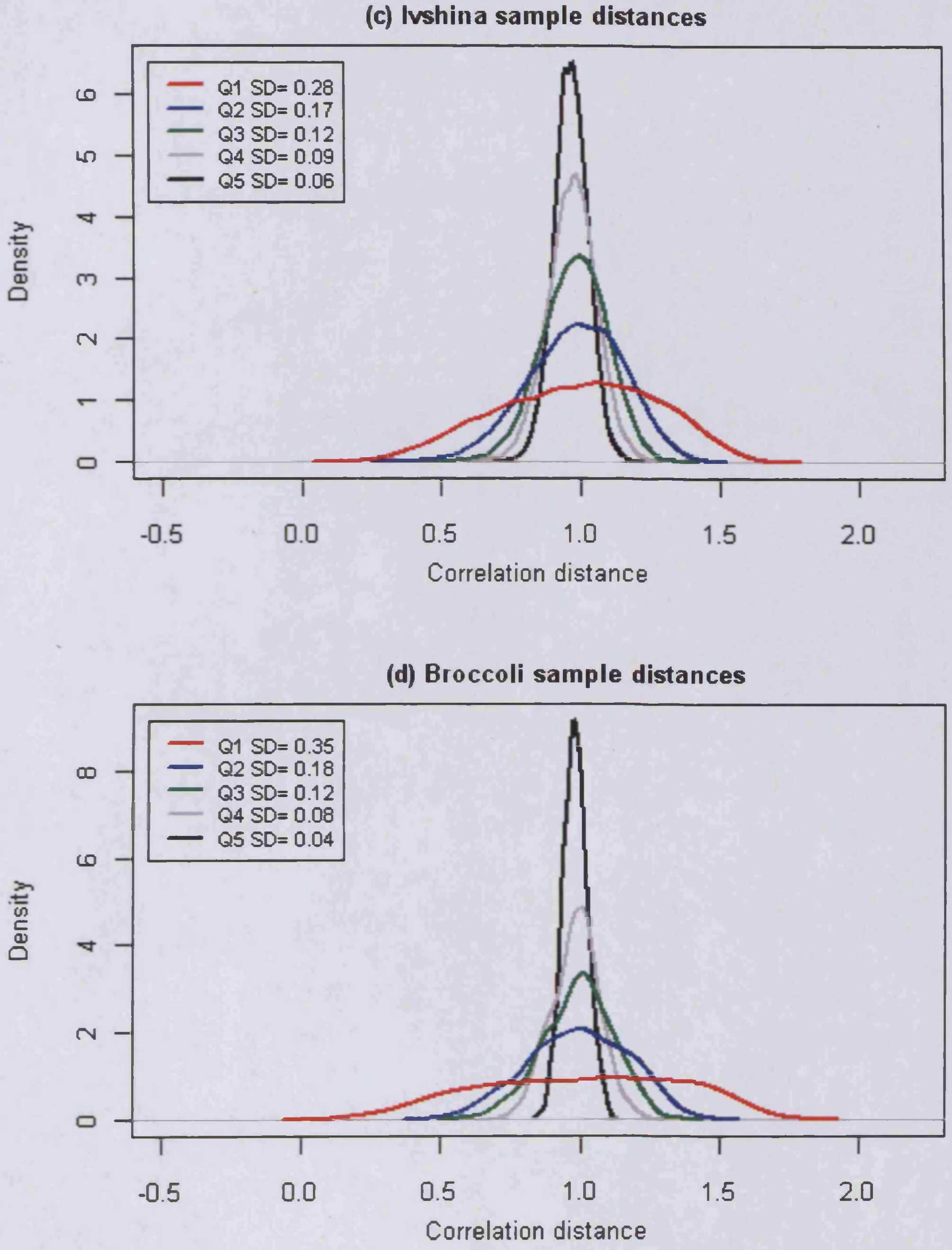


Figure 8.5 Gene distance distributions of gene quintiles. Distributions represent distances between all possible pairs of genes for each of the quintiles displayed in Figure 8.4. Variability of the distributions (measured as standard deviation) is indicated in the figure keys.



(a) GNF sample distances

(b) Ross sample distances



(c) Ivshina sample distances

(d) Broccoli sample distances

Figure 8.6 Sample distance distributions of gene quintiles. Distributions represent distances between all possible pairs of samples for each of the quintiles displayed in Figure 8.4. Variability of the distributions (measured as standard deviation) is indicated in the figure keys.

8.2.3 Visualization of the results of GSD analysis

Having identified possible relevant biological themes and extracted potentially informative genes by implementing the GSD methodology on three public datasets, efforts then focussed on how these results could be presented to a researcher in a format that could integrate all the different types of information.

Visualization of the different expression patterns prevalent in the dataset could be achieved by way of heatmaps of all the informative genes, as these involve hierarchical clustering of genes and samples according to their expression patterns. In a similar approach to the use of picketplots to indicate sample classes (for example, as used with the heatmaps in Chapter 7), a gene-based picketplot was used to show the links between each genes and their corresponding annotation metadata (i.e. GOBP terms). Hierarchical clustering of the terms could be carried out according to the genes shared by them and ordered accordingly in the gene picketplot (see Materials and Methods). The relationships between the terms as identified by the hierarchical clustering could be visualized using dendrograms.

8.2.3.1 Results of GSD analysis of the GNF human tissue expression dataset

The 51 GOBP terms selected by GSD analysis of the GNF human tissue expression dataset comprised of a total of 4123 genes. All possible pair-wise distances between these genes were calculated. An SD-GDV value was derived for each gene using the vector of distances of that gene to all other genes. Genes were then ranked in decreasing order of their SD-GDV values and the first quintile (which consisted of 825 genes) was selected for visualization.

The visualization of these genes based on the scheme described above is shown in Figure 8.7.

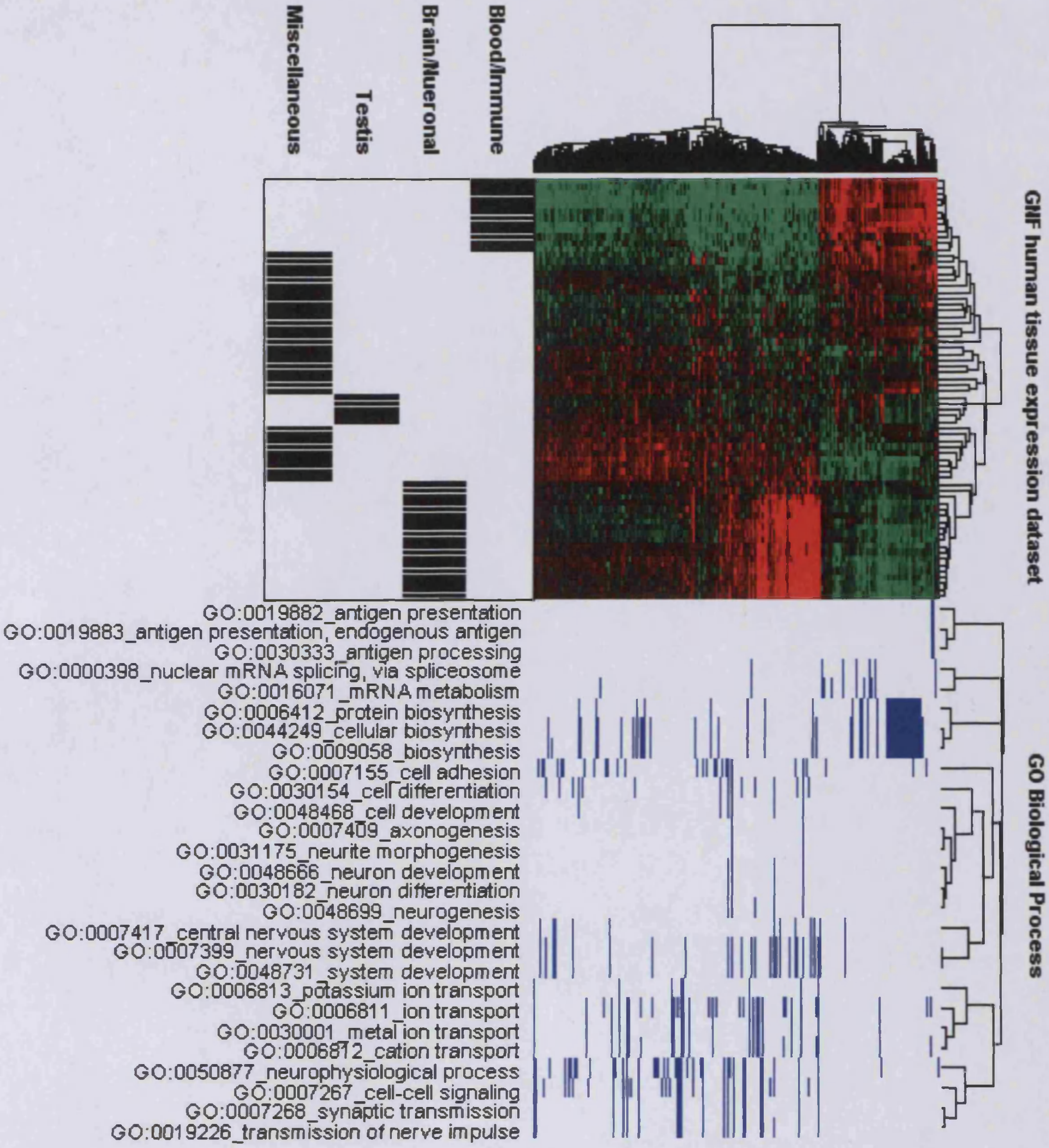


Figure 8.7 GSD analysis of the GNF human tissue expression dataset. Heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Black and blue cells in the picketplots indicate the presence of samples and genes within corresponding sample classes and gene sets respectively, while white cells indicate their absence. All hierarchical clustering was carried out using correlation distance and average linkage. For ease of visualization, some GOBP terms that are very similar to others have not been displayed, as have GOBP terms represented by less than 5 genes in the heatmap.

As can be observed, the heatmap exhibits expression patterns that are primarily based on groups of similar tissues, and the hierarchical clustering of samples has resulted in strong discrimination between these groups. The most visually striking patterns are exhibited in samples that from the Blood/Immune or Brain/Neuronal groups, which are the two largest groups of similar tissues within the dataset.

There appear to be two major clusters of genes: the first of which exhibit up-regulation in the Blood/Immune samples and down-regulation in the Brain/Neuronal samples. This cluster appears to comprise primarily of genes of a few biological themes: immune system processes, mRNA metabolism and cellular biosynthesis. The second cluster comprises of genes that exhibit down-regulation in the Blood/Immune samples. Within this cluster, there a well-defined sub-cluster of genes that exhibit up-regulation within the Brain/Neuronal samples. These appear to be genes that primarily represent biological themes such as nervous system development, synaptic transmission and ion transport.

Thus it is found that the GSD methodology allows for mining of relevant information from within this dataset and visualization scheme can display this information in an integrated format. Consider analysis of the results of implementing the GSD methodology on this dataset without prior knowledge regarding the tissues that these samples represent. Having extracted a set of informative genes and created a heatmap of these, the first discovery would be that there are at least two major groups of samples that show very similar expression patterns (and could thus be from similar tissue-types). The second discovery would involve insight into the nature of these groups of similar samples by inspection of the biological themes represented in their expression patterns: the up-regulation of many genes involved in many nervous system processes in one of the sample groups hints at the neuronal origin of the samples. Similarly, links can be made between the immune system process genes that are up-regulated in the other group of similar samples, and the nature of the tissues they represent.

8.2.3.2 Results of GSD analysis of the Ross AML dataset

The 12 GOBP terms selected by GSD analysis of the Ross AML dataset comprised of a total of 1333 genes. All possible pair-wise distances between these genes were calculated. An SD-GDV value was derived for each gene using the vector of distances of that gene to all other genes. Genes were then ranked in decreasing order of their SD-GDV values and the first quintile (which consisted of 267 genes) was selected for visualization. The visualization of these genes is shown in Figure 8.8.

As can be observed from the heatmap, there are several distinct gene expression patterns, and the combined effects of these is a hierarchical clustering of samples that correlates well with previously known samples classes, i.e. genetic sub-types of AML. The largest of the expression patterns exhibited (in terms of the number of genes involved) causes the first major grouping of genes into the two largest clusters. This expression pattern appears to be strongly influenced by the subtypes of AML represented by the samples, as can be evidenced from the enrichment of samples from certain sub-types within the two clusters: the first cluster (in which most of these genes exhibit down-regulation) comprises of all samples of t(15;17)[*PML-RAR*α] and FAB-M7 subtypes, most of t(8;21)[*AML1-ETO*] samples and relatively few inv^{16}[*CBF*β-*MYH11*] and *MLL* samples. The second cluster (in which most of these genes exhibit up-regulation) consists of the majority of inv^{16}[*CBF*β-*MYH11*] and *MLL* samples, and relatively few t(8;21)[*AML1-ETO*] samples.

The other smaller expression patterns also appear to vary in concordance with AML subtypes. These then allow for well-differentiated sub-clusters (within the first major cluster) of samples representing the t(15;17)[*PML-RAR*α], t(8;21)[*AML1-ETO*], and FAB-M7 subtypes. While no expression patterns were found that could discriminate between the inv^{16}[*CBF*β-*MYH11*] and *MLL* subtypes (in the second major cluster), this is likely to be a reflection of the molecular heterogeneity of these subtypes – which was suggested by the authors in the original study, where highly supervised techniques could not discover adequate differentiation between these two subtypes.

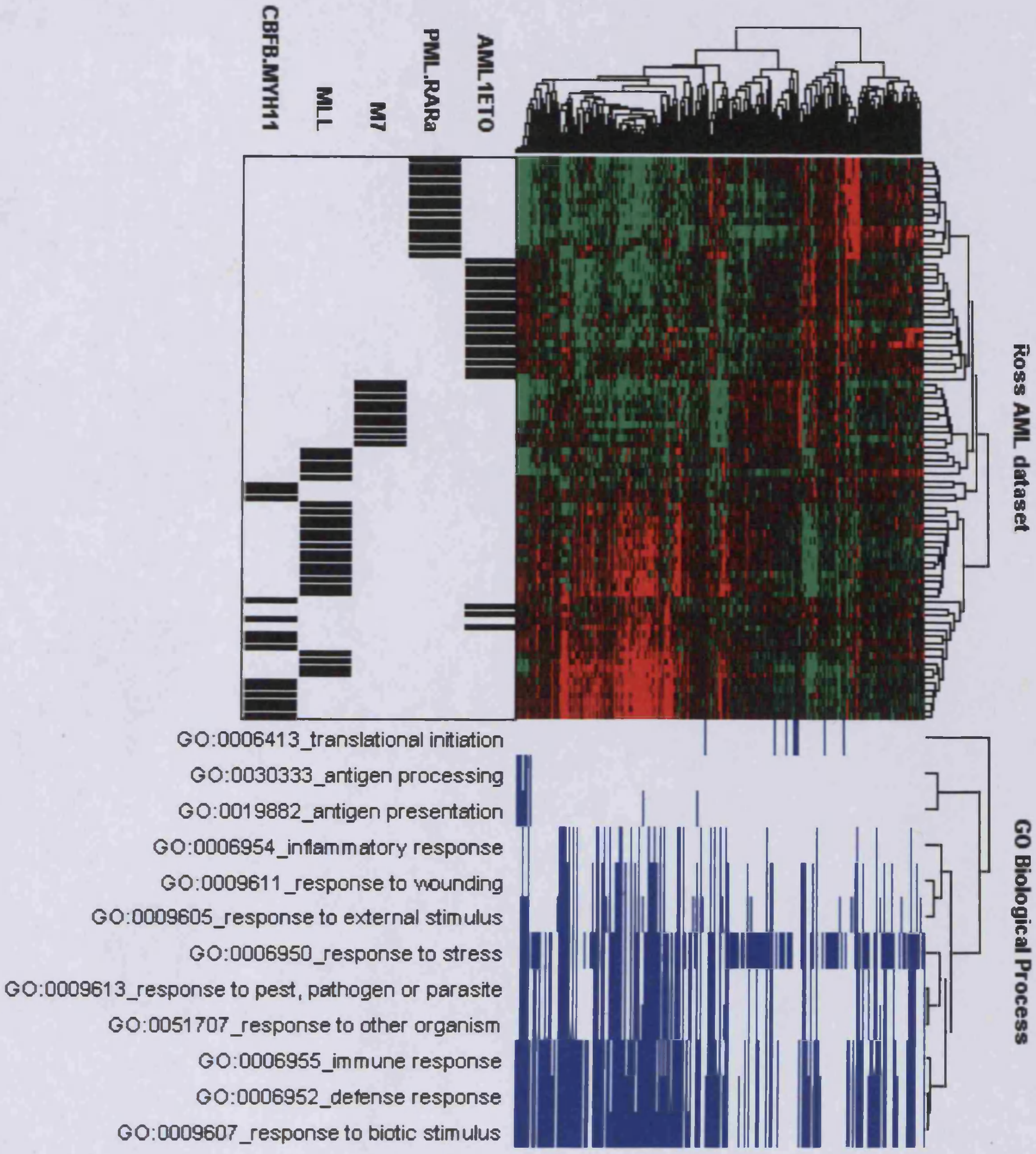


Figure 8.8 GSD analysis of the Ross AML dataset. Heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Black and blue cells in the picketplots indicate the presence of samples and genes within corresponding sample classes and gene sets respectively, while white cells indicate their absence. All hierarchical clustering was carried out using correlation distance and average linkage. For ease of visualization, GOBP terms represented by less than 5 genes in the heatmap have not been displayed.

As noted earlier in Chapter 7, the majority of terms selected by GSD analysis of this dataset all concern a single biological theme: the immune response. The relatedness of these terms is noticeable from the picketplot: many genes are shared between these terms. Most of the terms include genes that exhibit several different expression patterns – this could imply the possible existence of undiscovered functionality-based sub-groups of genes within these terms. At least one expression pattern can be linked with particular GOBP terms: the expression pattern of down-regulation in most t(15;17)[*PML-RAR*α] and FAB-M7 samples, and up-regulation in most of the other subtypes appears to comprise the majority of selected genes that are annotated with GOBP terms for antigen processing and presentation.

8.2.3.3 Results of GSD analysis of the Broccoli liposarcoma dataset

The 30 GOBP terms selected by GSD analysis of the Broccoli liposarcoma dataset comprised of a total of 5042 genes. All possible pair-wise distances between these genes were calculated. An SD-GDV value was derived for each gene using the vector of distances of that gene to all other genes. Genes were then ranked in decreasing order of their SD-GDV values and the first quintile (which consisted of 1009 genes) was selected for visualization. The visualization of these genes based on the scheme described above is shown in Figure 8.9.

As can be observed, the heatmap includes expression patterns that cause a hierarchical clustering of samples which is agreement with the known sample phenotypes. Two major clusters are discernible: one of which includes most of the high-grade samples (all four DD samples and 6 out of 7 PLEO samples), while the other includes most of the low-grade samples (12 out of 15 WD samples).

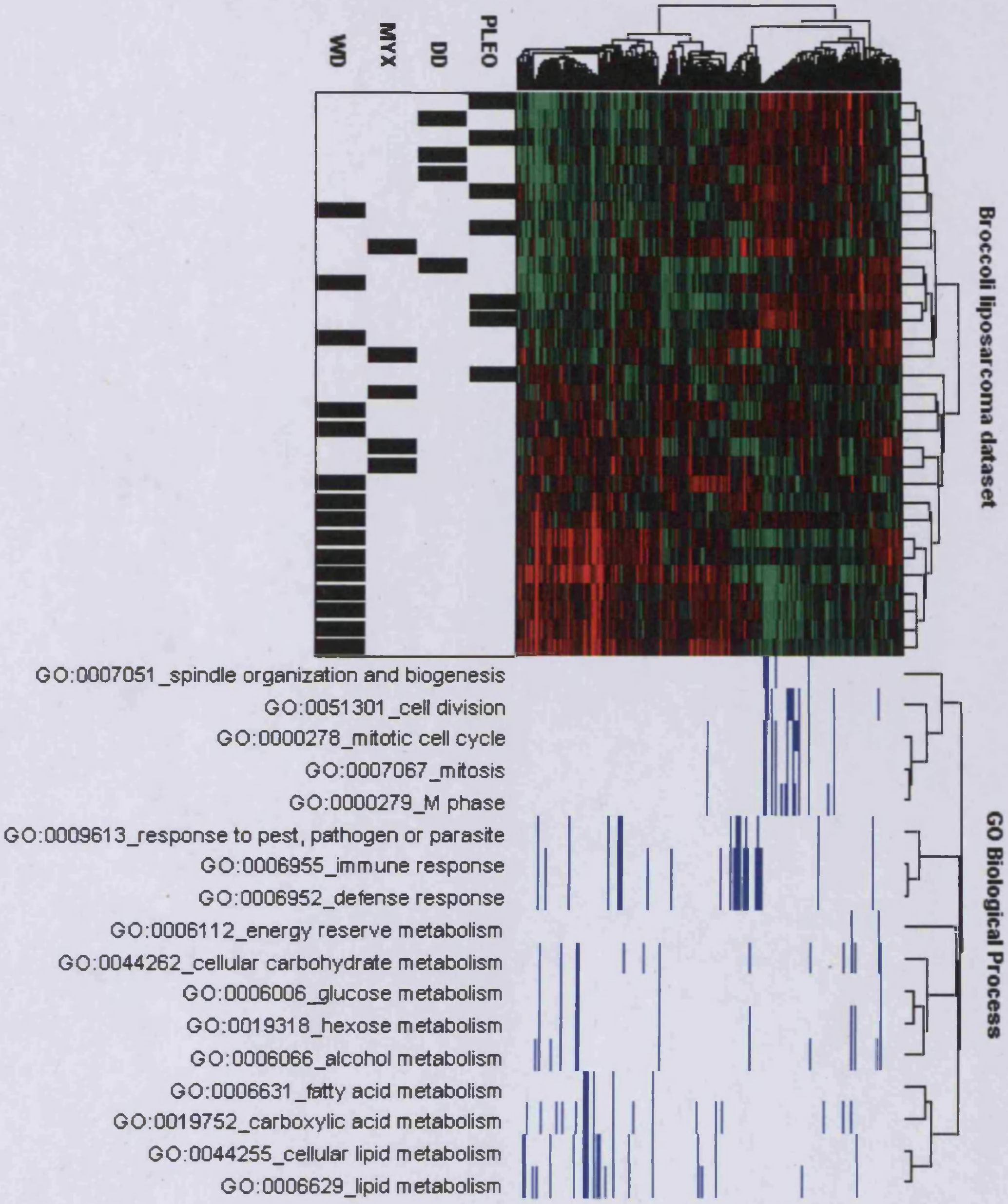


Figure 8.9 GSD analysis of the Broccoli liposarcoma dataset. Heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Black and blue cells in the picketplots indicate the presence of samples and genes within corresponding sample classes and gene-sets respectively, while white cells indicate their absence. All hierarchical clustering was carried out using correlation distance and average linkage. For ease of visualization, some GOBP terms that are very similar to others have not been displayed, as have GOBP terms represented by less than 5 genes in the heatmap.

As was previously described in Chapter 7, the GOBP terms selected by GSD analysis of this dataset could be divided primarily into three major biological themes: mitosis/cell division, metabolism and immunity. The hierarchical clustering of the GOBP terms appears to reflect this – terms comprising these themes fall into three well discriminated clusters. Indeed the hierarchical clustering is able to further sub-divide terms specific to metabolism of lipids and those specific to metabolism of carbohydrates.

Most of the genes annotated with mitosis/cell division terms exhibited expression patterns of up-regulation in the high-grade (DD/PLEO) samples, and down-regulation in the low-grade (WD) samples. The opposite expression pattern was observed for most genes annotated with metabolism terms. As discussed in Chapter 7, these observations are biologically plausible. Genes annotated with immunity terms show a range of different expression patterns.

8.2.3.4 Results of GSD analysis of the Ivshina breast cancer dataset

The 50 GOBP terms selected by GSD analysis of the Ivshina breast cancer dataset comprised of a total of 3771 genes. All possible pair-wise distances between these genes were calculated. An SD-GDV value was derived for each gene using the vector of distances of that gene to all other genes. Genes were then ranked in decreasing order of their SD-GDV values and the first quintile (which consisted of 754 genes) was selected for visualization. The visualization of these genes based on the scheme described above is shown in Figure 8.10.

As can be observed in the heatmap, there are several consistent expression patterns that are exhibited by the selected genes. The clustering of samples is influenced in varying degrees by each of these patterns, and the resultant grouping of clusters does not conform very well to the previously known classes of low and high grade samples.

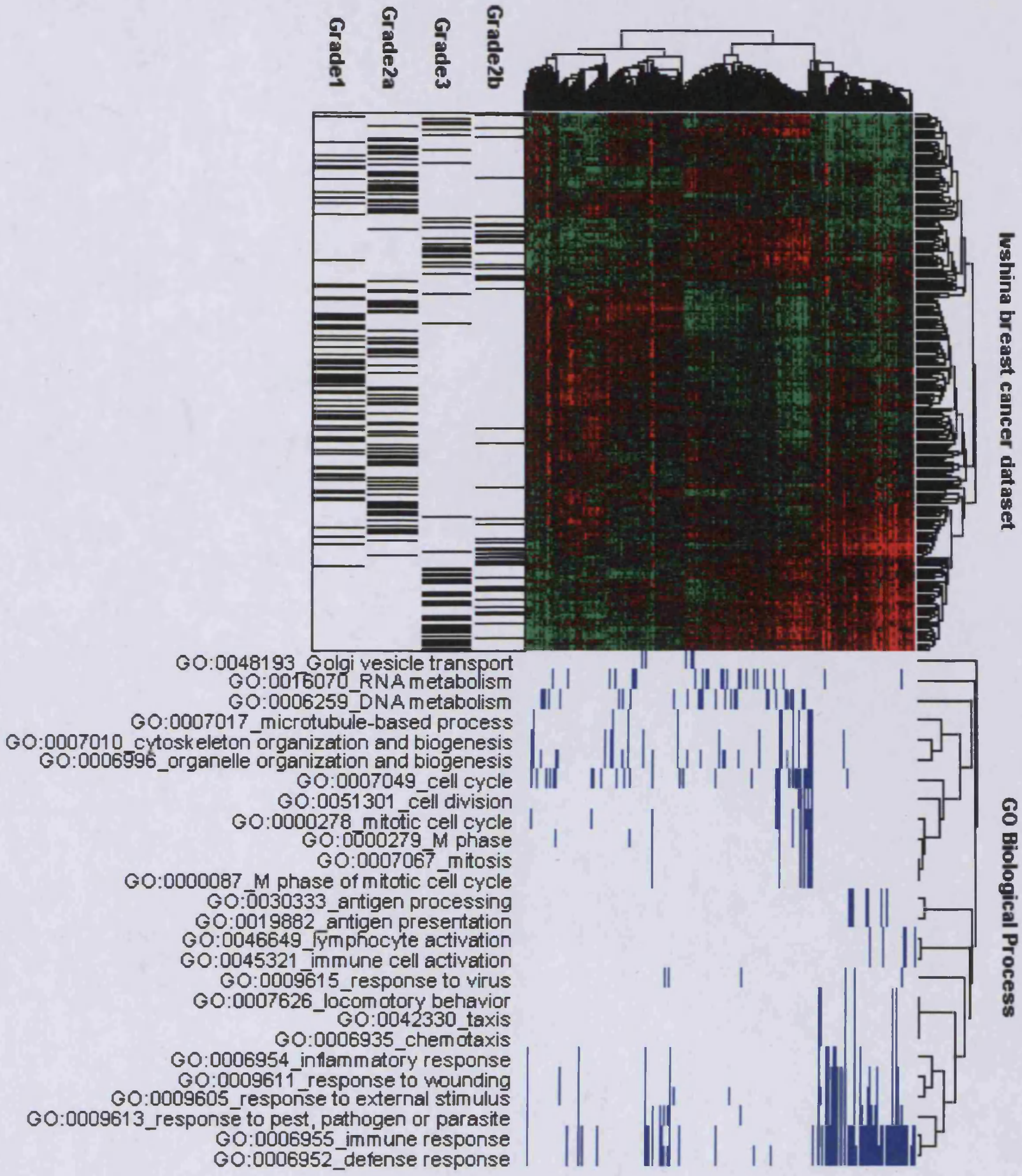


Figure 8.10 GSD analysis of the Ivshina breast cancer dataset. Heatmap represents log median-centred gene expression values ranging from -2 (bright green) through 0 (black) to 2 (bright red). Black and blue cells in the picketplots indicate the presence of samples and genes within corresponding sample classes and gene sets respectively, while white cells indicate their absence. All hierarchical clustering was carried out using correlation distance and average linkage. For ease of visualization, some GOBP terms that are very similar to others have not been displayed, as have GOBP terms represented by less than 5 genes in the heatmap.

The heatmap also indicates at least three major groups of genes. One of these clusters comprises primarily of genes involved in immune system processes – they exhibit a strong expression pattern, and show little concordance with the cancer grades of the samples. The relatedness between these terms is also indicated by the large number of genes shared between them.

The second cluster comprises primarily of genes involved in cell division processes. The expression pattern exhibited by these genes correlates strongly with cancer grades: the majority of these exhibit up-regulation in the high-grade G3 and G2b samples, and down-regulation in the low-grade G1 and G2a samples. The importance of this biological theme of cell division and mitosis in the discrimination between low and high grade breast cancer samples has been evident in both the original study as well as further explorations described in Chapter 7.

The final cluster of genes exhibit a variety of relatively weaker expression patterns, and are annotated to GOBP terms representing a range of biological themes.

The results from this dataset suggest that exploratory analyses such as GSD can discover several biological themes within an experiment, each of which could be affected differently by the experimental conditions, or could be affected by different experimental conditions (and thus show different expression patterns). In this case, one of the biological themes discovered (i.e. cell division and mitosis) appears relevant to the intent of the researchers who conducted the experiment (i.e. discriminating between low and high grade cancer samples on the basis of their gene expression profiles).

8.4 Discussion

This chapter described two further extensions of the framework of GSD analysis of microarray data that could be of use to researchers.

It first described the development of a methodology that attempts to identify the most informative genes from amongst gene sets selected by GSD analysis. It was found that SD-GDV values of genes could allow ranking of the most informative genes, of which an arbitrary number of the highest ranking genes could then be selected. This was implemented by calculating SD-GDV values for genes considering all genes from all GSD-selected gene sets taken together.

Secondly, the chapter described a scheme for visualization of the results that allowed for integration of various types of information. Heatmaps of the selected genes displayed the various expression patterns exhibited by the genes. Picketplots were included to indicate the various gene sets that each of the selected genes was part of. Relationships between the gene sets could also be displayed by hierarchical clustering of the gene sets based on the genes shared between them.

The methodology designed to extract informative genes was applied to gene sets selected by GSD analysis of each of the four microarray datasets described in Chapter 7. While the number and proportion of informative genes varies across datasets, a uniform arbitrary cut-off for selection of genes was implemented in all cases to avoid user bias: in all cases, the top 20% of genes were selected after being ranked in decreasing order of their SD-GDV values. It was observed that this methodology could successfully extract the most informative genes.

Visualization of these results using the scheme developed indicated that it could successfully integrate gene expression data, phenotype data and functional annotation/

meta-data in a format that was visually intuitive: expression patterns of the most informative genes could be observed in the heatmap (Eisen et al. 1998), along with sample class information. Also, genes within the heatmap could be mapped onto the GOBP terms selected by the primary GSD analysis step. Furthermore, relationships between the terms were also displayed and the hierarchical clustering of these terms could indicate larger underlying biological themes.

The methodology for extraction of informative genes is perhaps best suited to analysis of gene sets selected by the GSD methodology, as it is based on similar principles i.e. using the variability of correlation distances based on expression profiles. However, the methodology (as well as the scheme for visualization of results) could be used for further analysis of the results of any other gene set analysis method.

Chapter 9: Summary and General Discussion

This chapter comprises of summarization and discussion of the work described in the preceding chapters of this thesis. These have been divided into two sections representing work described in Parts A and B of this thesis.

9.1 Part A - Integration of microarray based experiments using lists of differentially expressed genes

This section summarizes and discusses Part A of this thesis, which comprises of Chapters 2-5. The primary question explored here was whether large-scale comparisons between microarray experiments could be carried out using genelists derived from those experiments instead of using the entire experimental datasets. Such a strategy could facilitate large-scale comparisons of many different microarray experiments in an unsupervised fashion, and in theory allow for the possibility of finding unexpected links between experiments. Such links could then be further explored using more sophisticated integrative techniques which in turn could lead to novel biological insight.

For this purpose, a database of genelists derived manually from published literature was used to compare genelists using statistical techniques. It was intended to estimate the utility of this strategy by observing the biological plausibility (if any) of the results generated. For example, carrying out unsupervised hierarchical clustering of the genelists on the basis of some metric indicating similarity between them (e.g. hypergeometric Z-scores), and observing if the resultant clusters reflected any underlying biology. Other issues that were desired to be explored included the extent to which biologically meaningful links between experiments could be found across species (in terms of evolutionary distance).

9.1.1 Summary of results and explorations

Chapter 2 described explorations that helped build the experimental framework and strategies with which to carry out comparisons between microarray experiments using their genelists. Firstly, strategies were described regarding how genelists created from different experiments, using different Affymetrix platforms and across different species could be made comparable. Secondly, three different strategies with which the significance of the similarity between any two genelists could be quantified were tested. It was found that the hypergeometric statistical test was suitable for this purpose as it could control for the systematic effects of genelist length and gene universe size.

Chapter 3 then described the results of carrying out comparisons within a database of literature-derived genelists spanning a range of different platforms and species. An unexpected finding was that of implausibly high levels of similarity between genelists derived from the same array-type: hypergeometric Z-scores distributions derived from these comparisons were found to be centred between median values of 2-6. This finding is counter-intuitive as it is expected that most microarray experiments (and thus the genelists derived from them) are dissimilar to each other, and that this would lead to Z-score distributions centred on median values of close to zero, as is observed for comparisons between lists of randomly selected genes. A second related observation was that of a correlation between the size of a genelist and the levels of significance assigned for comparisons of that genelist to all other genelists from the same array-type. This again is of concern, because one of the reasons the hypergeometric statistical test was deemed suitable for these analyses was its ability to control for the systematic effects of genelist length.

Chapter 4 first described speculations regarding the reasons for these unexpected observations, and it was hypothesized that these are more likely to have occurred due to known transcriptional behaviours of genes that violate assumptions of the hypergeometric statistical test, rather than reflecting true biological similarities between genelists. It then described a statistical model called the 'biased urn' which involves the use, in the hypergeometric test, of a gene universe size that is different to that of the gene universe used to sample genelists. It was shown that under this model, comparisons between lists of randomly selected genes could exhibit both the excess levels of similarity and the correlation between genelist size and significance as are observed for the comparisons described in Chapter 3. To simulate these comparisons of literature-derived genelists, the average gene universe size shared between any two experiments was estimated by a re-calculation of the Z-scores with different universe sizes and using that gene universe size for which the Z-score distributions were centred on median values closest to zero. These estimates ranged from only 35-65% of genes represented on an array. It was then found that simulation of comparisons between the literature-derived genelists using the biased urn model with the estimated universe sizes yielded significance patterns that were very similar to those observed in Chapter 3.

This biased urn model and the use of an estimated average gene universe size thus represents a potential solution to control for violations of the assumptions of the hypergeometric test when used to assess comparisons between genelists derived from microarray experiments. However, this strategy has several caveats: firstly it is *ad hoc* and is heavily dependent on the set of genelists used in each analysis. A second issue of greater concern is that this methodology only controls for the observed biases at global levels (i.e. overall significance levels for many different comparisons taken together), and the use of a single estimated average gene universe size for all comparisons could lead to erroneous results if there is wide variability in the true sizes of the gene universes shared between any pair of experiments. Explorations of this aspect were described in Chapter 5 using a subset of the GNF human Gene Expression Atlas comprising the expression profiles of a wide range of normal human tissues carried out in the same laboratory using the same microarray platform (Affymetrix hgu133a). By simply using the number of genes found to be expressed in each tissue-type, it was shown the number of genes expressed in any two tissue-types varies greatly for two reasons. Firstly, systematic effects of the variability in the number of genes expressed in each tissue.

Secondly, for biological reasons (particularly tissues-specific gene expression), similar tissue types share greater numbers of expressed genes than dissimilar tissue types. Finally it was shown that application of an average gene universe size leads to erroneous quantification of significance between lists of genes selected randomly from the expression universe of different tissue-types; the extent of this error is dependent on the magnitude of the difference between the average shared gene universe size used in the statistical test, and the true number of genes expressed in any pair of tissues.

9.1.2 The concept of a 'gene universe' in microarray data analysis

The concept of defining gene universes is relevant to analyses of microarray data for several related reasons. In this thesis, it was first mentioned in Section 1.3.1.1 with respect to multiple hypothesis correction techniques applied to p-values derived from gene-by-gene tests for differential expression. Because the stringency of the correction applied increases with the number of genes tested, it has become convention to remove from the gene universe, for an experiment, those genes that are highly unlikely to be flagged as differentially expressed by a statistical method (Huber et al. 2008; Scholtens and Heydebreck 2005). Such genes include those that are not expressed (due to technical or biological reasons), or those that exhibit low variability in expression levels (for example house-keeping genes that may be evolutionarily constrained to exhibit stable expression levels).

While decreasing the gene universe size could help increase the power (i.e. decrease the number of false negatives) of gene-by-gene testing for differential expression by decreasing the stringency of tests for multiple hypothesis correction, decreasing the gene universe size could help increase the specificity (i.e. decrease the number of false positives) of ORA-based approaches for functional analyses of genelists such as tests for enrichment of biological themes, like GO terms, in genelists. As Falcon and Gentleman point out, the use of an unfiltered gene universe (for example all genes represented on an array) could lead to the erroneous assessments of significance (Falcon and Gentleman 2008). For this reason they include a 'non-specific filtering' step, as part of their GOstats tool for ORA, to filter out low-variability genes from further analyses because these are unlikely to be flagged as differentially expressed (Falcon and Gentleman 2007). The concept is similar to that of the biased urn model described in Chapter 4: statistical tests such as the hypergeometric test assess the significance of the overlap between any two sets of genes by comparing the observed overlap size with that which might be expected from two genelists of similar size comprising of genes selected from the same universe as the genelists being assessed. The inclusion, in the statistical gene universe, of genes that cannot be selected into genelists would decrease the expected overlap size and this would lead to assignment of higher levels of significance to an observed overlap than ought to be.

For this reason, in their review of ORA tools to assess enrichment of GO terms within genelists, Khatri et al criticize those tools that do not filter gene universe sizes; for example some tools even include genes that may not be represented on an array, by using all genes present in the entire genome of an organism. The inclusion of genes that can never be selected as differentially expressed, they argue, "represents a flagrant contradiction of the assumptions of the statistical models used" (Khatri and Draghici 2005). The findings presented in this thesis then extend this criticism to the ORA-based comparisons between genelists derived from microarray experiments. The explorations show that, when using an unfiltered gene universe in the statistical model, not only is there a marked loss of specificity (reflected in the assignment of implausibly high levels of similarity between genelists), but also a loss of control of the systematic effects of genelist size (reflected in the correlation between genelist size and significance levels). These undesirable effects are observed in both the comparisons of genelists derived from real microarray experiments, and in the comparisons of lists of randomly selected genes using the biased urn model.

The possible biological reasons for these effects are speculated on in Chapter 4, but the extent to which these effects could confound analyses is relatively understudied. The finding, in Chapter 4, that the estimated average gene universe size that is shared between any two experiments (which may be used to indicate the net effect of the transcriptional behaviours of genes that violate statistical assumptions) ranges from only 35% to 65% of genes on an array would appear to be cause for concern. It is strongly suggests that a universe size of all genes represented on an array should not be used for comparisons of genelists derived from microarray based experiments.

These effects are particularly an issue for comparisons between experimentally derived genelists as compared to comparisons between an experimentally derived genelist and one derived from manual annotation (for example, as carried in ORA analyses for enrichment of GO terms within a genelist). This is because selection of genes into genelists that are derived from experiments is subject to the biological effects of gene expression patterns that violate statistical assumptions; this may not be the case for those derived from manual annotation (such as gene sets representing pathways or GO terms). Indeed, an experiment was carried out (data not shown) where ORA analysis was carried out on a series of genelists: none of the unexpected trends observed for comparisons between experimentally derived genelists (that is, excess levels of similarity and correlation between genelist length and significance) were observed; there also was very little difference in the results when using a filtered gene universe (only genes annotated with GO terms) as opposed to an unfiltered one (all genes on the array).

9.1.3 Can genelists be used to compare microarray-based experiments?

9.1.3.1 Assessing the significance of the number of shared genes

The comparison of microarray-based experiments using genelists derived from them is based on the rationale that similarity between genelists implies similarity between the experiments. This strategy represents a potentially significant decrease in resource-intensiveness as compared to meta-analytical strategies requiring re-analysis and integration of entire microarray datasets. This would then allow for the possibility of large-scale comparisons between many different microarray experiments in an unsupervised fashion, which could lead to possible discovery of unexpected links between experiments and possible novel biological insights. However, the use of genelists to compare different microarray experiments carried out in different laboratories, on different platforms and across different species has been a controversial issue: there is generally little agreement even between genelists from similar experiments (Cahan et al. 2007; Cheadle et al. 2007; Ein-Dor et al. 2005; Jeffery et al. 2006; Manoli et al. 2006; Tan et al. 2003), and agreement can found only when using very standardized experimental protocols and data analysis strategies (Bammler et al. 2005; Irizarry et al. 2005; Larkin et al. 2005).

For this purpose, it was desired to explore whether carrying out comparisons between genelists derived from a wide range of experiments carried out at different laboratories, with different experimental protocols and data analysis methodologies, and across different species could still yield biologically plausible and useable results, if any. However, as described in Chapters 3 and 4, naïve comparison of genelists using unfiltered gene universes could result in erroneous results: in these circumstances, statistical tests to assess similarity between genelists suffer a considerable loss of specificity, as well as loss of control of the systematic effects of genelist size. The biased urn model described in Chapter 4 can allow for *ad hoc* modelling and control for these statistical biases on a global scale by estimating the average gene universe size for any two microarray experiments. However, as explored in Chapter 5, this model could still yield erroneous results because of the use of a single estimated gene universe size when this size could vary widely due to both systematic effects and biological reasons (such as tissue-specific expression patterns).

A more accurate assessment of the similarity between genelists could be carried out by estimating the gene universe size for each experiment separately, and using for each comparison only those genes that are present in both universes for a pair of genelists. For example, consider two genelists *x* and *y*, with gene universes of *X* and *Y* respectively. These genelists could be compared by using the intersect of *x* and *y* (*xy*) and as the gene universe, the intersect of *X* and *Y* (*XY*). This would also be reflected by filtering *x* and *y* for any genes that are not present in *XY*. However, the estimation of the gene universe for each experiment (for example by removing un-expressed/low expression genes and/or low variability genes) is likely to involve re-analysis of the entire microarray dataset. Such a resource-intensive strategy would then defeat the purpose of comparing genelists as a computationally efficient alternative to more sophisticated meta-analytical strategies.

9.1.3.2 Assessing genelists for shared biological themes

Another strategy that may be used to compare genelists from microarray experiments is to assess whether a pair of genelists share any underlying biological theme (such as GO terms or pathways) (Cheadle et al. 2007; Manoli et al. 2006; Shen et al. 2008). The rationale for this is that if two genelists share a common underlying biology, this may represent a true biological link between the experiments as opposed to a set of overlapping but functionally unconnected genes. Furthermore, this would do away with the requirement for assessing the significance of the size of overlap between genelists (which has been shown in this thesis to be highly problematic), and also provide direct biological insight of the link between the experiments (on the basis of the shared biological themes). There are at least three different methodologies that may be used for this purpose (Figure 9.1).

Firstly, ORA could be carried out for each of the genelists separately to detect biological themes (such as GO terms) that are enriched in each of the genelists (Figure 9.1a). GO terms common to both genelists can then be detected. However, it is possible that the enrichment may have been caused by different subsets of genes annotated with a particular GO term. An extreme example may be one where both genelists exhibit enrichment of a particular GO term, but do not actually share any genes in common.

Secondly, ORA could be carried out on solely the overlapping genes – this would ensure that enrichment is due to the shared genes (Figure 9.1b). However, as would be the case in the first methodology, such analyses would still require estimation of the overlap of gene universe sizes for both experiments being compared; as discussed in Section 9.1.3.1, this would require re-analysis of both experimental datasets. Using the example described in Section 9.1.3.1, the enrichment of GO terms in *xy* would require comparison with the distribution of GO terms in *XY*.

Thirdly, a more sophisticated analysis (Figure 9.1c) could be carried out by testing for enrichment of GO terms in the overlapping genes as compared to the distribution of GO terms in each of the genelists (as opposed to comparison with the distribution of GO terms in the gene universe). However, there is still a need for estimation of the gene universe that is shared between the two experiments because the genelists require to be filtered to reflect this. Using the example in Section 9.1.3.1, *x* and *y* would require a filtering step to remove genes that are not present in *XY*. Furthermore, interpretation of results would be complicated. For example, three sets of GO terms would be derived: terms enriched with respect to both genelists, and those enriched with respect to only one or the other genelist.

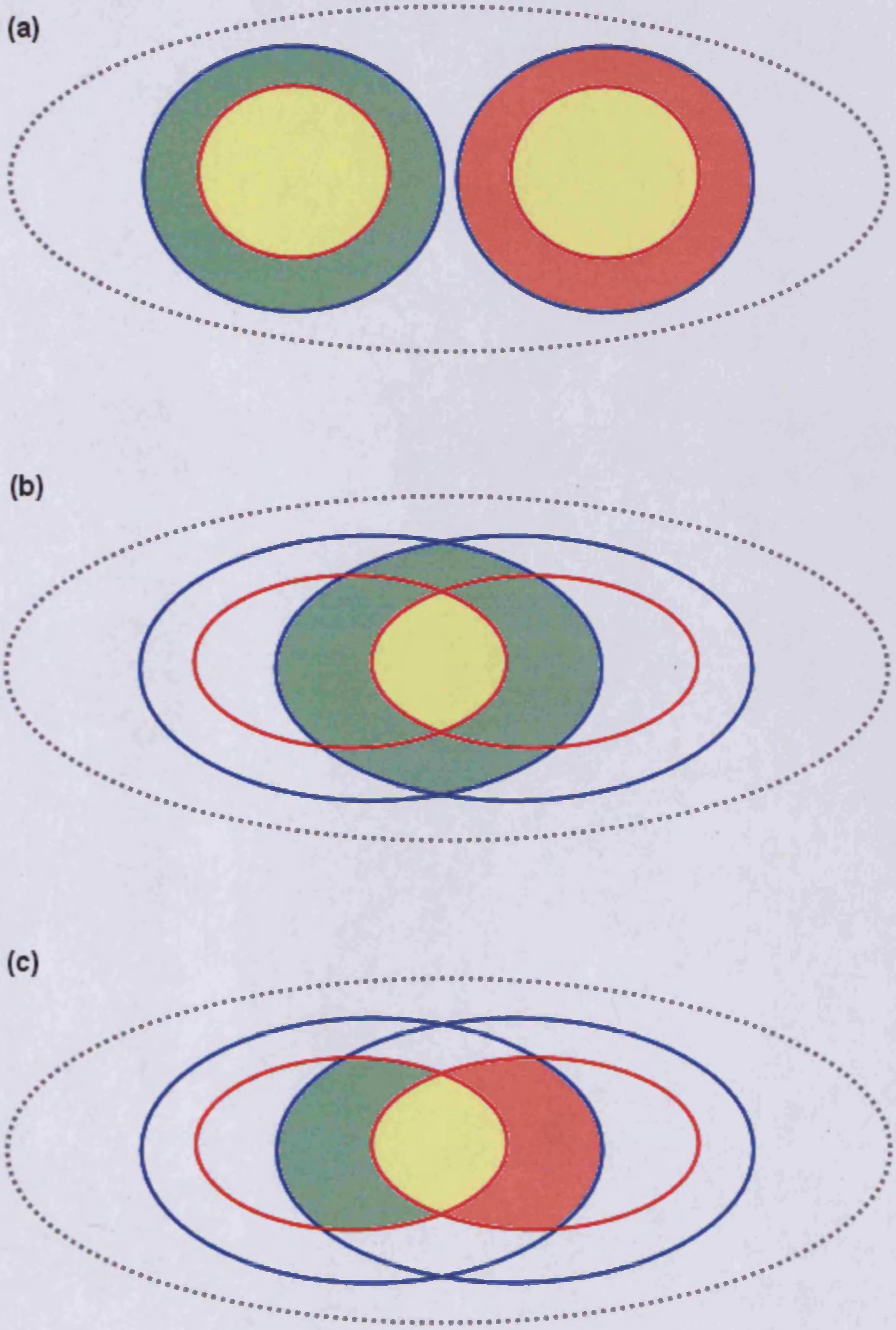

(a)
(b)
(c)

Figure 9.1 Strategies for detection of biological themes shared between genelists. Three different strategies are shown, all using the same scheme: the broken grey oval represents a set of all genes represented on an array; red circles/ovals represent experimentally derived genelists while the blue circles/ovals represent the gene universes for those experiments. The yellow shaded areas indicate those genes that would be tested for enrichment of biological themes (such as GO terms), while the red and green shaded areas represent different gene universes that should be used for those tests.

Other issues of theme-based approaches described above include the inability to rank similarities between the genelists: the number of GO terms shared between genelists may not be an accurate metric because of the parent-child relationships between the terms.

Thus it may be concluded that simply using genelists to compare microarray experiments, even if supplemented with further information (such as the estimated average gene universe size calculated using the biased urn model, or detection of shared underlying biological themes) is problematic and likely to produce erroneous results. While it may be possible to improve the accuracy of these methods by incorporating knowledge about the gene universes for each these experiments, there are several issues regarding such a strategy. Firstly, estimation of the gene universes for each of these experiments is likely to be as resource-intensive as more sophisticated meta-analytical methodologies; this then makes it difficult to justify this strategy as a computationally efficient alternative. Secondly, the definition of a gene universe is problematic in itself.

In ORA-based approaches, the concept of genes in relation to gene universes is essentially binary: a gene is either absent or present in a gene universe. It is more likely that there exists, for each different experiment, a continuum of probabilities for the presence of a gene within the gene universe for that experiment; this is dependent on a range of technical and biological factors unique to each gene (such as discussed in Chapter 4).

9.2 Part B - Gene Set Discovery (GSD): a novel methodology for unsupervised threshold-free discovery of biological themes within microarray datasets

Part B of this thesis comprised of Chapters 6-8, and described explorations that led to the development of the GSD analytical framework for unsupervised theme discovery within microarray datasets. GSD comprises of three stages: selection of relevant gene sets, selection of informative genes within the selected gene sets, and a scheme for integrated visualization of results. Also described are the results of the application of the GSD methodology to the analysis of four different microarray datasets derived from Affymetrix expression profiling platforms.

9.2.1 Summary of results and explorations

Chapter 6 described explorations that led to the development of the first stage of GSD analysis, i.e. the selection of gene sets (that could represent biological themes and functional annotation, such as GO terms and pathways) from a microarray dataset, which could be of interest to a researcher, in an unsupervised fashion. The underlying hypothesis of GSD is that if the gene expression matrix for a particular gene set contains information, in terms of shared patterns of expression, then it is likely to be relevant to the experiment and therefore of interest to a researcher. It was thus desired to derive a metric that could indicate the level of such information within the expression matrix for any gene set. For this purpose, research focussed on the distance matrices that are commonly used for hierarchical clustering of genes and samples (for example, in heatmaps), because these distances represent quantifications of the relationships between genes and between samples based on similarities of their expression profiles.

Simulations using hypothetical gene expression matrices into which known levels of information could be artificially introduced in a controlled manner were used to test several candidate metrics. Of these, the standard deviation of the sample distance matrix (SD-SDM) was selected as a suitable metric for two reasons: firstly, it was sensitive to the presence of information within a matrix, and there was a relationship between the level of information and SD-SDM values. The second desirable property is that, in the presence of more than one type of information (i.e. more than one gene expression pattern) within an expression matrix, SD-SDM values rank an expression matrix where different expression patterns lead to the same grouping of samples over one where different expression patterns each cause different groupings of samples. Thus, it allows for prioritization of gene sets which can be linked to a single stratification scheme for samples.

However, it was also found that SD-SDM values are subject to systematic effects of gene set size and the random presence of informative genes (i.e. genes exhibiting non-random expression patterns). Thus, it was desired to develop a methodology to assess the significance of the SD-SDM value observed for any gene set. Two possible strategies were tested for this purpose using simulation studies: firstly, a re-sampling based competitive strategy that involved creating a null distribution of SD-SDM values using sets of randomly selected genes. Secondly, a randomization based non-competitive strategy which involved creation of a null distribution by randomizing values within the expression matrix for the gene set being tested. The competitive strategy was deemed suitable as it could successfully control for both confounding factors.

While all explorations described in Chapter 6 were carried out using simulated microarray datasets, Chapter 7 then described the results of the application of the GSD methodology to four real-world microarray datasets. To assess how successfully GSD could select relevant gene sets for each of the datasets, assessment was carried out of the results, regarding whether they were biologically plausible and/or they were in concordance with the results of prior analyses of these datasets. In all cases, the GSD methodology was found to successfully identify relevant gene sets and biological themes for each experiment.

Chapter 8 then described the development and implementation of two further extensions of the GSD analytical pipeline. Firstly a metric was developed that could help identify and extract the most informative genes from within a gene set selected by GSD. This metric, the standard deviation of the gene distance vector (SD-GDV), is based on concepts similar to those that led to the implementation of SD-SDM values in the selection of gene sets, and was found to successfully rank genes on the basis of how well their expression patterns were shared. Secondly, a scheme was developed to visualize the results of GSD analyses. It was designed to integrate various types of information, such as the most informative genes from amongst gene sets selected by GSD, the different expression patterns exhibited by these genes, the gene sets that these genes belong to as well as the relationships between selected gene sets, and any phenotypic data that may be available for the samples.

9.2.2 Principles and utility of the GSD methodology

The GSD analytical framework can be considered to be novel in terms of the statistical methodologies underlying its functionality. As far as could be researched, no other technique was found to use SD-SDM values as measures of cluster 'strength', or SD-GDV values as a metric to determine the prevalence of the expression pattern exhibited by a single gene. While the primary utility of the GSD methodology (i.e. the ability to carry out gene set analysis in an unsupervised fashion), is not novel, there very few other options available to researchers for this purpose. Only three other such methodologies could be identified, and of these, two share the same underlying strategy. These methodologies are further discussed in the next section (Section 9.2.3).

The GSD methodology brings together two traditionally disparate modes of analyses of microarray data, which are discussed in the following sections.

9.2.2.1 Functional analysis of microarray data using biological themes

Biological interpretation of microarray data has been aided considerably by the development of databases containing functional annotation of genes (such as GO, KEGG and Biocarta), and of statistical methodologies and tools that allow for analysis of microarray data in terms of the biological themes represented in these databases (typically as sets of genes that are functionally related, or share some common underlying biological theme). The additional biological insight provided by such analyses allows for many more opportunities and greater scope for utility of the results of a microarray based experiment (Bild and Febbo 2005; Curtis et al. 2005).

For example, a list of genes may be identified through a microarray experiment to be predictive of a disease (e.g. cancer). Such results may be sufficient for certain analyses such as the identification of biomarkers. For example, the van't Veer 70-gene breast cancer signature (van 't Veer et al. 2002) has been used in Mammaprint, a molecular diagnostic test to assess the risk of breast cancer metastasis (Slodkowska and Ross 2009). However the utility of such analyses may be limited to such diagnostic tests.

Knowledge of the biological pathways and mechanisms underlying a disease can allow for improved options for diagnosis: for example, if a pathway is found to be differentially regulated in a disease, this knowledge allows for the potential use of antibody-based histo-pathological tests which can be easily carried out in standard clinical laboratories. Such diagnostics tests would represent considerable advantages to gene expression profiling based diagnostic tests using a list of disparate genes, in several aspects: technical (there would be no need for the extraction of RNA from tumour samples, which is problematic [particularly from formalin-fixed samples with degraded RNA]); logistical (samples can be processed on-site rather than sent to other laboratories); and financial (histo-pathological tests would be considerably less expensive than molecular diagnostic tests).

Similarly, knowledge of the higher level biological themes could allow for improved possibilities for intervention using drugs. Consider a pathway that is differentially regulated in the disease state, and several downstream targets of the pathway are selected into a list of differentially expressed genes. Intervention may then be carried out by treatment with drugs targeting each of these genes. However, there may be too many such targets; furthermore, there may be difficulties in designing drugs for them. On the other hand, if the underlying pathway (for example the p38 MAP kinase signalling pathway) is known, this creates the possibility of targeting a few regulatory genes (in this example, p38) rather than many different downstream targets. Furthermore, it would allow for selection of targets for which drugs could be more easily designed.

Some of the earliest techniques to carry out such functional analyses of microarray data were ‘threshold-based’ ORA methods (see Section 1.4.2) that detected enrichment of functional classes of genes within lists of differentially expressed genes. More recently, there has been the development of many ‘threshold-free’ methods for this purpose (see Section 1.4.3), such as GSEA, which do not require prior definition of a list of interesting genes.

9.2.2.2 Class discovery

A key feature of the vast majority of methodologies with which to carry out functional analyses of microarray data (or gene set analyses [GSA]) is that they require supervision in terms of sample phenotype classes: the user needs to define the sub-groups of samples across which differential patterns of expression are expected or desired to be detected. For threshold-based GSA methods, this is carried out by the prior definition of a list of genes that are found by some statistical test to be differentially expressed across two or more sample sub-groups. Threshold free methods directly detect gene sets that are differentially expressed across pre-defined sample sub-groups.

While such analyses are appropriate for many experimental designs, there are also many situations where prior knowledge regarding sample sub-groups may not be known. Indeed, the very purpose of many microarray-based 'class discovery' studies (see Section 1.3.2) is the elucidation of distinct sample sub-groups based on gene expression profiles.

This mode of analysis is particularly used in studies of cancers: quite often it is found that a set of tumour samples are morphologically homogeneous and cannot be differentiated on the basis of histological techniques, but yet show diversity in terms of clinical variables such as survival rates and response to therapeutic drugs. For this purpose class discovery studies are carried out in order to find sub-groups of tumour samples based on their gene expression profiles that may then explain their behaviour with respect to these clinical variables. One particular publication regarding the use of microarrays concerned this very strategy: Alizadeh et al (Alizadeh et al. 2000) studied a set of diffuse large B-cell lymphomas that very morphologically indistinct, but exhibited a wide range of survival rates. Using class discovery techniques, they could identify two distinct sub-classes of these tumours, and found that these new sub-classes exhibited a significant difference in survival rates. The creators of the Ivshina breast cancer dataset (Ivshina et al. 2006) analyzed in Chapter 7 also carried out class discovery: using a classifier gene set, they identified two sub-classes of Grade 2 breast cancers which also showed a significant difference in survival rates (as well as other parameters). Other examples have been cited in Section 1.3.2.

9.2.2.3 Unsupervised theme discovery in microarray datasets using GSD

The GSD methodology described in this work brings together the two modes of analysis described above, such that it can be described as a methodology with which to carry out class discovery using biological themes, or equivalently, unsupervised gene set analysis.

The term 'supervised' has been generally used in the fields of machine learning and microarray data analysis (including this thesis – See Section 1.3.1) to describe methods that require *a priori* knowledge of classes of entities; for example, in order to 'train' methodologies to be able to distinguish between sets of entities where such knowledge of classes may not be known (see Section 1.3.1.2). However, the concept of the level of 'supervision' required by any analytical methodology can be thought to have a more generic meaning: that of the level of user-defined parameters and user-made decisions and input.

In theory, methodologies that require less supervision from users have greater potential for knowledge discovery and are thus more suited for exploratory analyses. For example, threshold-free GSA methods can be considered to be less supervised than threshold-based GSA methods, because they do not require the creation of list of differentially expressed genes. In the latter case, the researcher can choose from a range of statistical threshold levels to define differential expression, each of which could lead to different threshold-based GSA results (Pan et al. 2005). Furthermore, threshold-free methods allow for the possibility of discovering differentially regulated biological themes that may not be enriched within a list of genes exhibiting the greatest changes in gene expression (Ben-Shaul et al. 2005; Breitling et al. 2004; Huang da et al. 2009; Nam and Kim 2008; Subramanian et al. 2005).

GSD can thus be considered to be a threshold-free GSA method, because it seeks to identify relevant gene sets without requiring a list of genes selected by a researcher to be of interest. However, it represents a further decrease in supervision as compared to the vast majority of other threshold-free GSA methods: it does not require prior definition of samples classes across which differential patterns of gene expression are expected or desired to be detected. For example, if the GSD methodology is implemented on a dataset where sample classes are known *a priori*, it has the potential to identify gene sets that contain *unexpected* expression patterns i.e. those that would result in groupings of samples that are different from known sample classes. These may be due to unknown but biologically relevant factors, further analysis of which may lead to novel biological insight. Even if the gene sets detected by GSD exhibit expression patterns that result in clusters of samples that are as expected, GSD represents a more objective method to identify these gene sets.

However, it is surmised that GSD has greater potential utility in class discovery studies in which there may be no *a priori* knowledge regarding sample classes, which as discussed above, is particularly the issue in many studies of cancers. The amalgamation of the concepts of class discovery and theme-based analysis can allow a researcher to ask the question, "which biological theme can stratify this set of cancers?".

The potential utility of the GSD methodology can be illustrated using the Ivshina breast cancer dataset, the analysis of which was described in Section 7.3.3. In the original study, the authors desired to understand the basis for wide range of survival rates of breast cancer patients exhibiting Grade 2 (G2) tumours. Firstly, they used a biological hypothesis: that G2 samples did not represent a continuum of tumour progression stages; rather they comprised of two sub-types. One of these sub-types was similar to the low-grade G1 samples that exhibited good prognosis (high survival rates), while the other was similar to the high grade G3 samples that exhibited bad prognosis (low survival rates). On the basis of this hypothesis, they used highly supervised techniques incorporating tumour grade information to develop a classifier gene set that could discriminate between G1 and G3 samples. They then applied this classifier to the G2 samples and identified two sub-groups based on the similarity of their expression profiles to those of the G1 and G3 samples: G2a (1-like) and G2b (3-like). The subsequent discovery that these newly discovered G2 sub-types exhibited a significant difference in survival rates provided further evidence in support of the authors' original hypothesis.

In contrast, GSD analysis was able to arrive at similar results using a shorter analytical pipeline, and with considerably less supervision: it did not require the authors' hypothesis or knowledge of sample tumour grades. It identified a biological theme (mitosis/cell division) that exhibited a particular gene expression pattern. Simple hierarchical clustering of samples using the genes involved in this theme led to the discovery of two well differentiated classes of samples. It was then found that one cluster contained significant majorities of samples classified in the original study as either G1 or G2a, while the other contained significant majorities of G3 and G2b samples. Furthermore the samples in these clusters exhibited a difference in survival rate that was almost as significant as that achieved in the original study using knowledge of tumour grades and highly supervised techniques.

In addition, GSD could provide direct insight into the biological theme underlying the difference between good prognosis and bad prognosis samples, i.e. mitosis/cell division. As discussed in Section 9.2.2.1, such knowledge regarding biological themes allows for possibilities for improved diagnosis and intervention. For example, using the findings described above, breast cancer diagnoses may be carried out at standard clinical laboratories using simple histo-pathological tests utilizing antibody-based markers for mitosis (such as Ki67 or MCM2), as opposed to extracting RNA and sending off samples to be processed for gene expression profiling at remote locations based on the classifier gene set developed by Ivshina et al or Mammaprint. Furthermore, such diagnoses could identify which patients may require chemotherapeutic drugs, and that they may be treated with drugs specifically designed to target mitosis/cell division pathways.

Most cancer patients show differences in terms of which pathways are deregulated, and as Bild et al point out, knowledge of which pathways are deregulated in which patients allows for the possibility of administration of customized 'cocktails' of drugs to patients that target specifically those pathways that are found to be deregulated in those patients (Bild et al. 2006). The possibility of simultaneous discovery of patient sub-groups based on gene expression profiles along with direct identification of the biological themes underlying these patient stratifications allowed by GSD methodology thus makes it a very suitable potential option for such studies of cancers.

9.2.3 Other methods for unsupervised theme discovery

While other methods have been developed that attempt to evaluate the importance of biological themes (as gene sets) in microarray datasets without requiring prior knowledge of sample/phenotype class information, these are relatively few in number. This section discusses some of these methods, and how they compare to the GSD methodology.

9.2.3.1 Using average pair-wise gene correlation values to evaluate the importance of biological themes

Pavlidis et al (Pavlidis et al. 2002) used the average of all pair-wise Pearson correlation coefficient values for all genes belonging to a gene set as one of three 'functional class scores' with which to evaluate the importance of that gene set within microarray data. The significance of this metric was assessed by comparison with a null distribution created using sets of randomly selected genes of the same length as the gene set being tested. This methodology was also used by Kim et al as the first step of their Gene Set Expression Coherence (GSECA) algorithm (Kim et al. 2007). This metric is similar to the M-GDD metric explored in Section 6.2, because the correlation distance between any pair of genes is calculated as 1-correlation coefficient (see Materials and Methods).

The rationale for the use of this metric is that gene sets in which most genes exhibit co-expression are likely to be of interest to a researcher. However, Pavlidis et al themselves remark that such use of this measure of gene expression profile similarity to evaluate the functional classes of genes may be "too limiting" because "while it may sometimes be true that genes which cluster together have related functions, it is certainly not always the case that genes with related functions cluster together" (Pavlidis et al. 2002).

A related issue regarding the use of this metric is that it requires differentially expressed genes to change in the same direction. For example, consider a microarray dataset with two sample classes, A and B. For any particular gene set, the presence of a sub-group of genes that exhibit up-regulation in samples of class A and down-regulation in samples of class B would increase the magnitude of the average correlation coefficient between genes. However, the presence of another sub-group of genes exhibiting an expression pattern in the opposite direction (i.e. down-regulation in class A samples and up-regulation in class B samples) would cause a decrease in value of this metric. This would then lead to an undesirable decrease in significance assigned to that gene set. Indeed, Breitling et al remark that the ability of their iterative Group Analysis (iGA) method to not be sensitive to differences in expression pattern directions is "a very important feature, because genes that share a functional annotation may include activators as well as inhibitors of a certain process" (Breitling et al. 2004). This issue could become even more acute in datasets involving more than two classes of samples, where there is increased potential for diversity in gene expression patterns. It is for this reason that M-GDD values were rejected as potential metrics for the GSD methodology in Section 6.2.1.3. On the other hand, the SD-SDM values selected as metrics for the GSD methodology were shown to be insensitive to the presence of more than one gene expression pattern, as long as these patterns corresponded to the same sample groupings.

9.2.3.2 Annotation driven clustering of samples using adSplit

Another method with which unsupervised theme discovery with microarray datasets can be carried out is adSplit (Lottaz et al. 2007). The feature that is in common to GSD and adSplit (and distinguishes them from other such methods to identify relevant gene sets without sample class information) is that both methods attempt to select for gene sets on basis of the 'strength' of sample clusters produced by a particular gene set. adSplit achieves this by using the diagonal linear discriminant (DLD) score introduced by von Heydebreck (von Heydebreck et al. 2001) as a measure of cluster strength for samples. This is conceptually similar to the SD-SDM values used in GSD. The significance of DLD scores is determined, similarly to GSD, by using a background distribution of DLD scores calculated for sets of randomly selected genes of the same size as the gene set being tested.

However, there are several differences between adSplit and GSD. The primary difference is that adSplit uses to two-step approach to derive sample clusters: first a hierarchical clustering step is carried out to derive the first two sample clusters. These are then used to calculate cluster centroids for the second step, in which k-means clustering is used to identify two clusters of samples. The DLD score is then calculated for these two clusters. Thus adSplit is limited to the analysis of only two sample clusters at a time. Discovery of more sample clusters can only be achieved by iterative application of adSplit to previously discovered sample sub-clusters. On the other hand, the GSD methodology does not focus on the number of sample classes. Rather, the SD-SDM values utilized by GSD can reflect the strength of sample clusters without actually carrying out clustering (since only distances between samples are used).

adSplit also involves the use of several additional user-defined parameters. For example, for any gene set, only the top 50 genes (by default), when ranked according to the extent to which their expression patterns support the discovered clustering of samples, are used for calculation of DLD values. Also adSplit imposes a default minimum cluster size of 5 samples for clusters derived using a gene set. Interestingly, the authors of adSplit also recommend that for any gene set, the top 5 most differentially expressed genes be ignored (and this is reflected in the default setting for adSplit), as their expression patterns may not be shared by most of the other genes in the gene set. On the other hand, GSD requires only two user defined parameters – the first of which is a p-value cut-off for significant gene sets (which is also required by adSplit) and another to define the most informative genes (for example the top 20%, as used in all examples in this thesis).

It is difficult to choose between GSD and adSplit on theoretical bases. For example, it may be argued that GSD compares favourably because it has fewer user-defined parameters; this implies a decreased scope for user-induced bias (because it is possible that parameters can be changed to achieve results favoured by a researcher). The relatively shorter workflow of the GSD methodology also implies that it may be computationally efficient as compared to adSplit, especially in the analysis of datasets where there may be more than two classes of samples. However, it may also be argued that the additional adjustments made by adSplit (selection of a defined number of genes within a gene set prior to metric calculation, imposition of a minimum cluster size, and rejection of highly variable genes) could lead to biologically more accurate results.

To compare the results of GSD and adSplit, adSplit was used to analyze the four microarray datasets used in Chapter 7 to test the GSD methodology. To ensure that the results were comparable, the parameters for analysis were the same as used for GSD analysis in terms of data normalization (MAS5) and transformation (log and median centring), gene set database (GOBP terms), statistical settings (null distributions created using 10,000 sets of randomly selected gene sets, and a significance cut-off of p<0.01 after FDR correction) (see Materials and Methods). The results of adSplit analyses, as compared to GSD analyses are displayed in Table 9.1

As can be observed, the biological themes selected by adSplit show some degree of biological plausibility and concordance with the results of GSD analyses: the sole term selected for the GNF dataset, GO:0048675_axon extension, could be reflective of the fact that brain/neuronal tissues comprise the largest group of similar tissues in the dataset; this is also thought to have been reflected by the selection of many nervous system-related GOBP terms selected by GSD analysis of the dataset. One of the four terms selected by adSplit analysis of the Ross AML dataset was also selected by GSD (GO:0019882_antigen presentation), while two of the other selected terms are children of it. The only term selected by adSplit analysis of the Broccoli liposarcoma dataset (GO:0009596_detection of pest, pathogen or parasite) is a child term of a one selected by GSD (GO:0009613_response to pest, pathogen or parasite). Nine and two out of the fourteen terms selected by adSplit analysis of the Ivshina breast cancer dataset are associated with the biological themes of cell division and immunity respectively; these are also the two dominant themes shared by GOBP terms selected by GSD analysis of this dataset. However, only one term was selected by both methods (GO:0000819_sister chromatid segregation).

Dataset	Number of selected terms		GOBP terms selected by adSplit analysis
GNF	adSplit	1	GO:0048675_axon extension
	GSD	55	
Ross AML	adSplit	4	**GO:0019882_antigen presentation** GO:0019884_antigen presentation, exogenous antigen GO:0019886_antigen processing, exogenous antigen via MHC class II GO:0006942_regulation of striated muscle contraction
	GSD	13	
Broccoli liposarcoma	adSplit	1	GO:0009596_detection of pest, pathogen or parasite
	GSD	13	
Ivshina Breast Cancer	adSplit	14	GO:0051329_interphase of mitotic cell cycle GO:0006334_nucleosome assembly GO:0050867_positive regulation of cell activation GO:0007051_spindle organization and biogenesis GO:0000226_microtubule cytoskeleton organization and biogenesis GO:0007088_regulation of mitosis GO:0009596_detection of pest, pathogen or parasite GO:0006120_mitochondrial electron transport, NADH to ubiquinone GO:0000075_cell cycle checkpoint **GO:0000819_sister chromatid segregation** GO:0000070_mitotic sister chromatid segregation GO:0042110_T cell activation GO:0050863_regulation of T cell activation GO:0050909_sensory perception of taste
	GSD	50	

Table 9.1 Results of adSplit analyses. The adSplit methodology was used to analyze the four datasets that GSD was tested on (in Chapter 7). Terms highlighted in red were also selected by GSD analysis.

Despite the similarities in the biological themes selected by either methodology, relatively few GOBP terms are common to both, and there are considerable differences between the levels of significance assigned to GOBP terms. Figure 9.2 shows Z-scores derived from each method plotted against each other. As can be observed, there is little or no correlation between the values; many GOBP terms are assigned high levels of significance by GSD but not by adSplit and vice-versa.

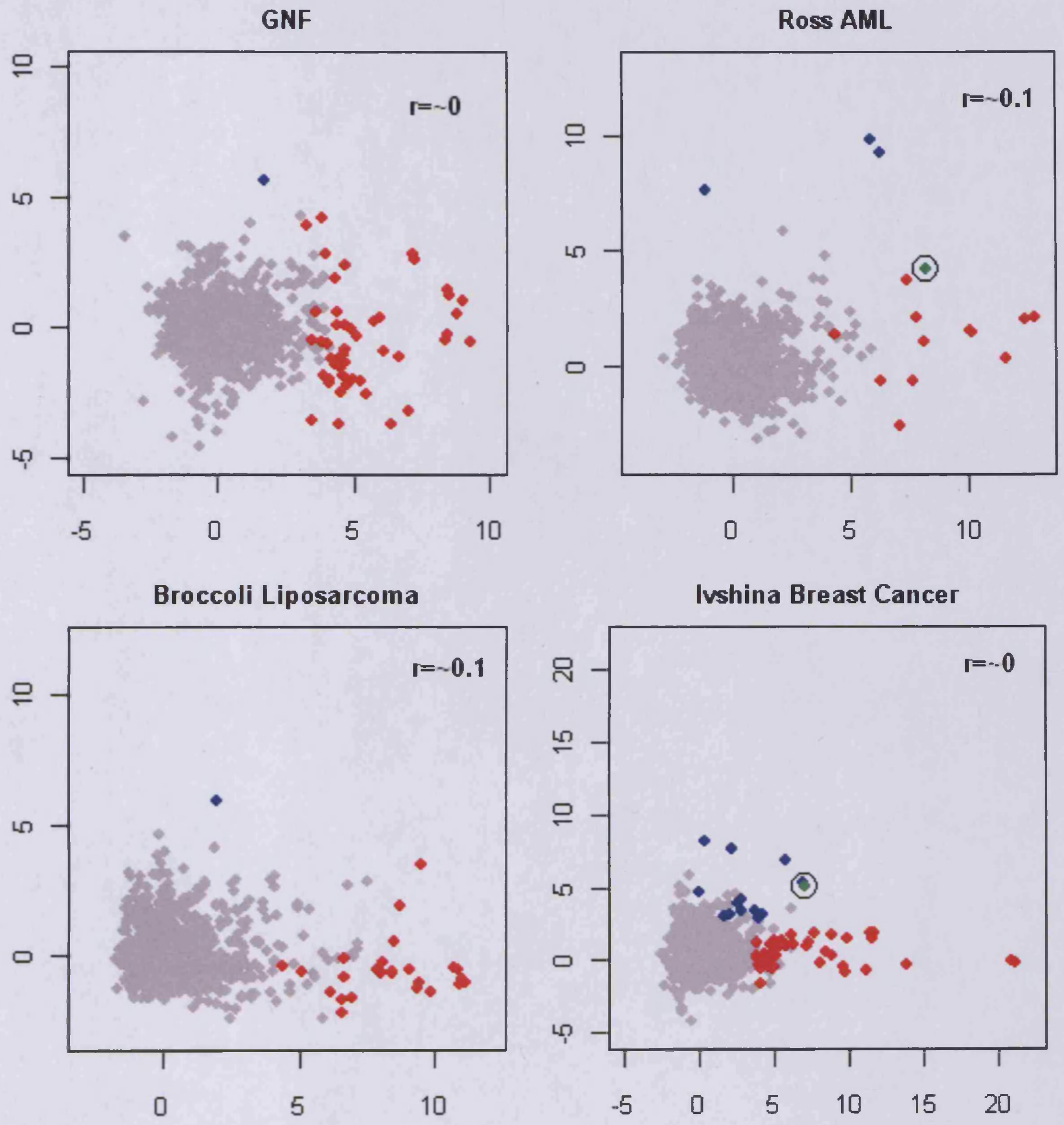


Figure 9.1 Significance levels assigned by adSplit and GSD. Grey points represent Z-scores assigned by adSplit (Y-axes) to GOBP terms, relative to Z-scores assigned by GSD (X-axes) for the same datasets. Blue points represent terms selected by only adSplit (at FDR-corrected p<0.01), while red points represent terms selected only by GSD. Green circled points represent terms selected by both methodologies. Legends show correlation coefficients for both sets of Z-scores.

Based on the comparative analyses described above, it may be concluded that while similar biological themes may be discovered by both methods, GSD exhibits greater sensitivity than adSplit (as it selected significantly more terms). This is particularly advantageous for exploratory studies where sensitivity may be more important to a researcher than specificity.

9.2.4 Development and benchmarking of the GSD methodology

The work described in this thesis regarding the GSD methodology was intended to provide proof-of-principle that it could be used to perform unsupervised theme-based analyses of microarray datasets, and provide useful and biologically meaningful results.

There is scope for further development and refinement of the GSD methodology: for example, other more sophisticated metrics may be used to assess the variability of sample distance matrices, as opposed to the SD-SDM metric described in this thesis. Similarly a more sophisticated metric could be derived to replace the SD-GDV metric used to rank potentially informative genes.

In order to assess any additional utility and advantage provided by the GSD methodology for the purpose of unsupervised GSA of microarray datasets, a comprehensive benchmarking study could be carried out to compare GSD with other methods developed for this purpose (such as the three methods described in the previous section). Such a study would need to include many different datasets, tested using a range of different data pre-processing and transformation schemes, a range of different platforms, and range of different annotation-based sources of gene sets as well (such as KEGG pathways, other GO terms, etc.)

Indeed benchmarking could be carried out to assess whether methods for unsupervised GSA (such as GSD) could provide any advantage over using conventional threshold-based and threshold free GSA methods for datasets where sample phenotype classes are already known. Assessment could be carried out to determine whether unsupervised GSA methods can select similar themes as detected by supervised methods. If unsupervised GSA methods detect additional biological themes, further investigations could be carried out if these additional results could be of interest to a researcher, particularly if these themes represent different groupings of samples than are known.

Chapter 10: Materials and Methods

10.1 Introduction

This chapter provides additional information to support the explorations and results described in all the preceding chapters. Details regarding the all technical methodologies used during this project are provided, as is further information regarding the data analyzed. These are presented on a chapter-wise basis, such that each of the following sections contains supplemental information specific to each of the previous chapters and are named as such.

All analyses were performed using the R (version 2.3.1) statistical programming interface (Ihaka and Gentleman 1996) utilizing Bioconductor (version 1.8) packages (Gentleman et al. 2004).

10.2 Strategies for large-scale integration of microarray data (Chapter 2)

Chapter 2 described a review and analysis of methods with which to compare disparate microarray experiments using lists of differentially expressed genes.

Much of the work described in the chapter concerns two main issues. The first of these was the use of biological annotation to enable comparisons of lists of genes from experiments carried out on different arrays and species. The second issue concerned reviewing the statistical methodologies that could possibly be used to assess the significance of the similarity between any pair of lists.

10.2.1 Conversion of probeset IDs across array-types and species

All information regarding probeset IDs for each of the arrays analysed, as well as the biological annotation available for these, was extracted from Bioconductor libraries for each of the array types. The names of these libraries are displayed in Table 2.1 in Chapter 2.

For all analyses (for example, to observe the number of probeset IDs shared between different chip-types), only the non-control probesets were used. Control probesets were filtered out by removing all probeset IDs with the prefix 'AFFX'.

10.2.1.1 Comparison of list of genes from experiments involving the same species

For reasons described in Chapter 2, comparison of genelists from experiments carried out using the same organism required conversion of probeset IDs into species-specific biological annotation, regardless of whether the lists were derived from the same array or not.

Several different types of annotation are available within Bioconductor libraries for the probesets within each of the different types of arrays, and at least five types of annotation could be used in place of probeset IDs to compare genelists derived from the same species: Entrez Gene IDs (EGIDs), gene names, gene symbols, Unigene IDs and RefSeq IDs.

Annotation is not available for all probesets in any of the arrays analyzed. Also, there are differences in the number of probesets for which each of the different types of annotation is available. The numbers of probesets for which each of the five different types of annotation identified above are displayed in Table 10.1.

	Gene Name	Entrez Gene ID	Refseq ID	Gene symbol	Unigene ID	Total
hgu133a	20056	21803	21139	21435	21284	22215
hgu133plus2	41175	48081	45808	47365	46730	54613
hgu95a	11587	12154	11982	12119	12073	12559
mouse4302	40821	41614	37007	41208	40041	45037
moe430a	22181	22347	21764	22266	22123	22626
mgu74av2	11909	12229	11618	11987	11840	12422
Rat2302	22684	23241	23025	23239	22904	31042
rae230a	13198	13462	13311	13456	13314	15866
rgu34a	7863	7970	7849	7966	7873	8740
zebrafish	11004	12249	8372	12249	11254	15502
drosgenome1	0	13130	13086	13092	12914	13966
drosophila2	0	14232	14175	14181	13973	18769
celegans	0	18480	18473	18473	15099	22548

Table 10.1 Numbers of annotated probesets in Affymetrix microarray platforms. Annotations were derived from Bioconductor meta-data packages. Blue cells indicate the highest number of annotated probe-sets for each array-type.

As can be observed, in all cases, more probesets are annotated with EGIDs than with any other type of annotation. It was also found that across all chips, any probeset that did not have an EGID annotation was also not annotated with anything else.

As the meta-data packages for the *Arabidopsis* arrays did not include EGID annotation, this was derived from the Affymetrix NetAffx Analysis Centre (Liu et al. 2003).

10.2.1.2 Comparison of list of genes from experiments involving different species

Because EGIDs are species-specific, lists of EGIDs from the same species are comparable; however lists of EGIDs from different species are not. In order to facilitate comparison of genelists from different species, the Bioconductor 'homology' packages were utilized. These packages were built using source data from the NCBI Homologene database (Wheeler et al. 2008).

In this format, a set of homologous genes across several different species are linked by a unique Homologene ID. Thus, a human gene can be linked to a homologous mouse gene through the shared Homologene ID. For ease of conversion, a database of homologous genes shared between all the array-types analysed was created by extracting these relationships from the Bioconductor packages.

Inconsistencies were observed in the data contained within these packages. For example, when all EGIDs present on the human hgu133a array were taken, and all their homologous genes on the mouse mouse4302 array were identified using the *hsahomology* package, a total of 10673 pairs of homologous genes were found. However, when changing the order of species, i.e. taking all EGIDs present on the mouse4302 array and identifying all homologous genes present on the hgu133a array using the *mmuhomology* package, a total of 10709 pairs of homologous genes were identified. These two sets shared 10642 gene-pairs in common. This phenomenon was observed in all pairs of species analyzed. To ensure consistency, only those gene-pairs that could be identified using the homology packages for both species were used.

10.2.2 Statistical methodologies to assess similarity between genelists

As described in Chapter 2, the size of overlap between two genelists may not accurately quantify the level of similarity between two genelists, because this measure is sensitive to systematic effects of genelist size and universe size. Three metrics were tested regarding the feasibility of their use in assessment of the significance of observed overlap size.

10.2.2.1 Fold Change

Consider two genelists of lengths $L1$ and $L2$, which have an observed overlap size O. The expected overlap size E can then be calculated as such:

$$E = \frac{(L1 * L2)}{N}$$

Here N represents the size of the gene universe. The fold change (*FC*) can then be calculated as the ratio of the observed overlap size to the expected overlap size, i.e. *O/E*. As these values are asymmetrical in nature, they are usually logged. This was carried out using the R function `log()`, using the default base (exponential of 1, i.e. natural log).

10.2.2.2 Binary similarity

The binary or Jaccard similarity index is the size of the overlap divided by the total number of unique genes present in at least one of the two genelists. Consider two genelists A and B. In set theory terms, binary similarity (*BS*) can be calculated as:

$$BS = \frac{A \cap B}{A \cup B}$$

10.2.2.3 Hypergeometric distribution

The hypergeometric test and its variants have popularly been used in Over-Representation Analysis (ORA) studies, for example, to test for the enrichment of GO terms within a genelist.

Consider two genelists, of sizes $L1$ and $L2$ that come from a gene universe of size N, and have an observed overlap size of O. Using the hypergeometric distribution, at least two metrics can be derived that can help assess the significance of similarities between any pair of genelists. The first of these is a Z-score (Z) which is an 'effect size' that represents a standardization of the observed overlap size, taking into account the sizes of the genelists and of the gene universe. It is calculated as the difference between the observed (O) and expected (E) overlap size, divided by the standard deviation of observed size (sdO). Thus,

$$Z = \frac{O-E}{sdO} = \frac{O - \dfrac{L1 * L2}{N}}{\sqrt{L1\left(\dfrac{L2}{N}\right)\left(1-\dfrac{L2}{N}\right)\left(1-\dfrac{L1-1}{N-1}\right)}}$$

The second is a p-value (p) that is represents the cumulative probability of finding that two genelists of length $L1$ and $L2$ share O or more genes. It is calculated as:

$$p = 1 - \sum_{i=0}^{O} \frac{\binom{L1}{i}\binom{N-L1}{L2-i}}{\binom{N}{L2}}$$

Note that in both equations shown above, the positions of $L1$ and $L2$ are interchangeable (i.e. would lead to the same results). The formula for hypergeometric Z-scores was manually programmed into an R function. Hypergeometric p-values were derived using the R function `phyper()`.

10.3 Comparison of lists of differentially expressed genes using the hypergeometric test (Chapter 3)

Chapter 3 described the application of the hypergeometric distribution to assess the similarity between genelists derived from microarray-based experiments carried out on a range of Affymetrix array-types and species.

10.3.1 A local database of genelists manually extracted from published literature

For this purpose, a local database of genelists was created by manual extraction of genelists from published literature. Details of the numbers of genelists extracted for each Affymetrix array-types are shown in Table 3.1 in Chapter 3.

The genelists that were manually extracted from scientific publications, and any supplementary information provided, were those that were created using statistical tests and algorithms (for example, a t-test to assess differential expression of genes and subsequent multiple-hypothesis correction of p-values), rather than those created using manual curation (for example, a list of the top 20 genes exhibiting the highest levels of differential expression). The number of genelists derived from each of the publications varied widely; when more than one genelist could be derived from a single publication, they were collapsed into a single genelist using the R `union()` function to achieve one genelist per publication.

Details of all genelists are provided in Appendix Ia. All genelists other than those for experiments performed on the Affymetrix hgu133a platform were kindly provided by Miss Hui-Sun Leong of the Department of Pathology at Cardiff University. All genelists are provided in Appendix Ib.

10.3.2 Comparison and statistical assessment of similarity between genelists across array-types and species

An overview of the technical methodology used to make genelists comparable and then assess their similarity is provided in Section 3.2 in Chapter 3.

The hypergeometric test provides a p-value for each test pair of genelists (see Section 10.2.2.3), which represents the probability of the level of similarity (i.e. the number of shared genes) observed for that pair of genelists occurring by chance alone for genelists of that size, when sampled from the universe of that size. Typically a p-value of <0.05 is used as a cut-off level for significance i.e. a pair of genelists are only considered to exhibit statistically significant levels of similarity if that probability of such levels occurring by chance alone is less than 5%.

However, the simultaneous testing of many pairs of genelists creates an issue of multiple hypothesis testing. For example, consider that a genelist is compared to a set of 100 other genelists; under the criterion of selecting those tests that yield p-values of <0.05, it is expected that 5 genelists would show significant similarity with the test genelist just by chance alone. These constitute Type I errors (false positives), and the number of these is expected to increase along with the number of simultaneous tests.

One of the popular methods to deal with this issue is control of the false-discovery rate (FDR) using the Benjamini-Hochberg method (Benjamini and Hochberg 1995). This methodology has been implemented throughout this thesis whenever there has been a need for multiple hypothesis correction, using the R function `p.adjust()`.

10.3.3 Calculation of evolutionary distance between species

To calculate evolutionary distances between the different species from which genelists were compared, source data was extracted from the interactive Tree of Life (iTOL) project of the European Molecular Biology Laboratory (Letunic and Bork 2007). The multiple sequence alignment used to calculate distances within the tree between the 6 species under investigation was downloaded. Using the Protdist program of the Phylip package of programs for phylogenetic analysis (Felsenstein 1993), distances could be calculated between each pair of species. This was carried out using the Dayhoff PAM matrix which comprises of empirically derived probabilities of the change of one amino acid within a protein sequence to another, and the distance computed is in units of the expected fraction of amino acids changed.

10.3.4 Extraction of genelists from the L2L database

As explorations of the local database of genelists indicated excess levels of similarity as assessed by the hypergeometric distribution, it was then desired to investigate if this phenomenon could also be observed amongst genelists collected in external databases. One such database is L2L (Newman and Weiner 2005), and this comprises of genelists that have been manually extracted from published literature. Genelists (where genes were represented as HUGO gene symbols) were downloaded from the L2L databases for experiments carried out on two human and two mouse Affymetrix arrays (see Table 3.4 in Chapter 3) on the 26th of August, 2008. These genelists are provided in Appendix Ic.

While it was difficult to identify which genelists came from the same publication, it was possible to combine lists of genes that were found to be up or down-regulated in the same statistical test for differential expression within an experiment. This was carried out by combining those pairs of genelists whose names had the same prefix (for example, genelists with the names 'XXX_up' and 'XX_dn' were combined).

10.3.5 Assessing the correlation between genelist length and significance

Section 3.3.3 described explorations indicating a link between gene set size and the levels of similarity of that genelist to all others from experiments carried out on the same array-type. Statistical assessment of this relationship was carried out using the Pearson's product-moment test of correlation. Two metrics could be derived from this test: firstly an 'effect size' (called Pearson's r), which varies between -1 and 1. A value of 0 implies no correlation; a positive r value indicates that the dependent variable increases with the causative variable, while a negative value indicates the opposite. The second metric is a p-value to indicate the significance of an observed r value. This was implemented using the R function `cor.test()`.

10.4 Modelling violations of the assumptions of the hypergeometric distribution using the 'biased urn' model (Chapter 4)

Chapter 4 described explorations of the effect of changing the universe size used for the computation of Z-scores when using the hypergeometric distribution to assess the similarity between a pair of genelists. Also described was the implementation of the biased urn model, under which an estimated 'average' gene universe size is used to compute Z-scores instead of the total number of genes present on an array. This average size was estimated as follows: starting from a universe size of all genes on an array, a series of Z-scores distributions representing comparisons of all possible pairs of literature derived genelists was created by successively reducing the gene universe size. The median Z-scores for each of these distributions are plotted in Figure 10.1 against the gene universe size used for calculation of that distribution. The estimated average gene universe size was considered to be that which yielded a Z-score distribution centred on a median value closest to zero (vertical red lines in Figure 10.1).

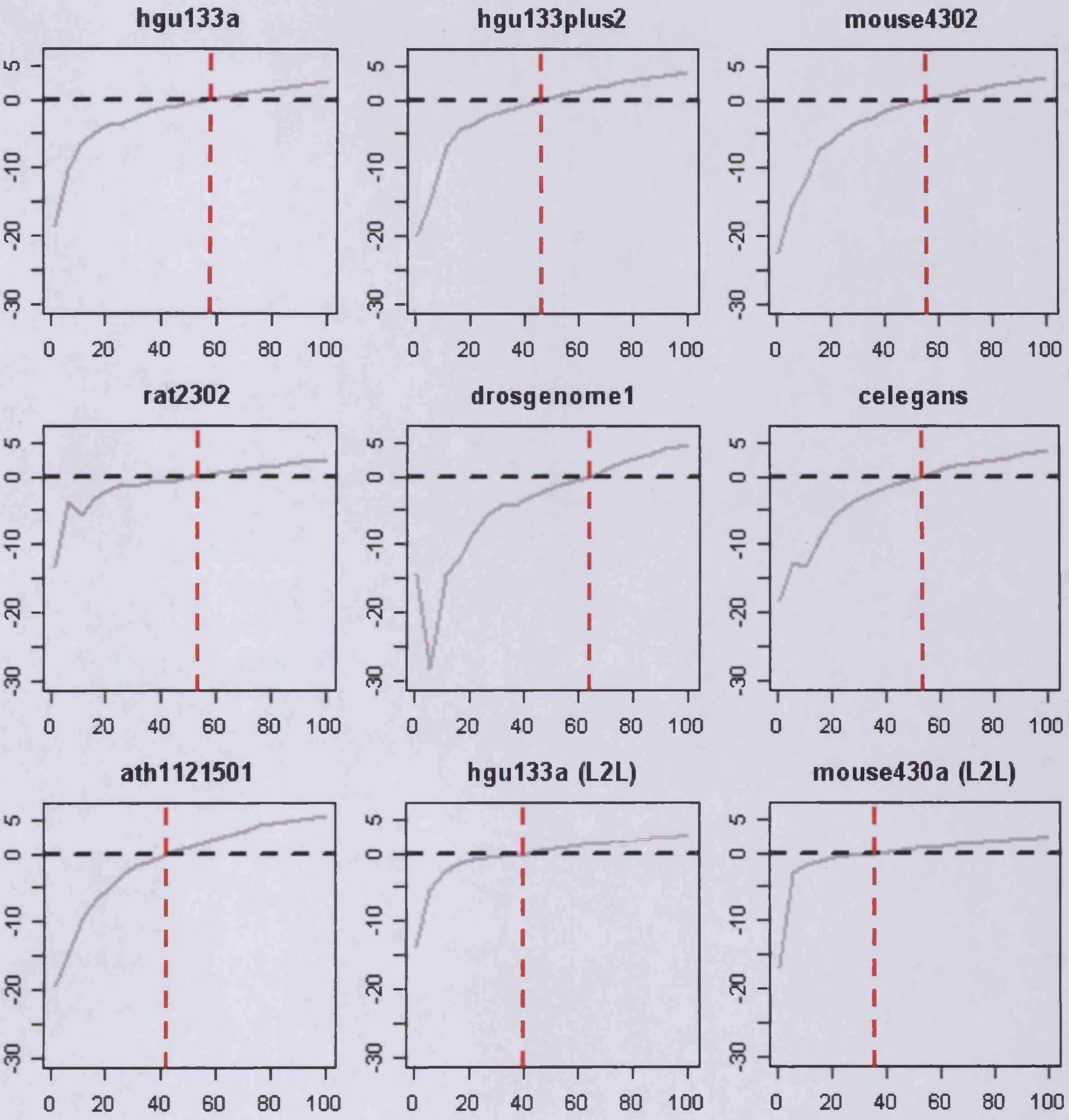


Figure 10.1 ***ad hoc*** **estimation of the average size of the gene universe shared between any two experiments using the biased urn model.** Grey lines represent medians of a series of hypergeometric Z-score distributions for comparisons of all possible pairs of literature-derived genelists, created using a range of different gene universe sizes. X-axes represent gene universe sizes as the percentage of genes represented on each array; Y-axes represent hypergeometric Z-scores. The horizontal black line represents a median Z-score of 0, while the vertical red line represents the gene universe size used as an estimate of the average gene universe size for comparison between any two experiments.

10.5 Exploring gene expression patterns with the GNF Expression Atlas (Chapter 5)

Chapter 5 described investigations of the variability of the sizes of the expression universes for a wide range of human tissue-types using the GNF Expression Atlas.

10.5.1 Data pre-processing and generation of expression universes

Original CEL files representing microarray data for human tissues were downloaded the Gene Expression Omnibus (GEO) database (Edgar et al. 2002), where this data was stored with the series identifier GSE1133. A total of 158 Affymetrix hgu133a samples were downloaded, which represented 79 different human tissues (two samples per tissue). Samples representing cancer cells and foetal tissues were removed, leaving 136 samples representing 68 different normal human tissues.

The probe intensity values that comprise the CEL files were normalized and converted into probeset expression values using the Affymetrix Microarray Suite 5.0 (MAS5) algorithm (Affymetrix 2002). To avoid possible biases caused by the presence of more than one probeset representing the same EGID (as described in Chapter 2), one probeset was selected for each EGID represented on the hgu133a array. For this purpose, the median MAS5-normalized expression values for all probesets were recorded across all 158 samples. For EGIDs having more than one probeset the one exhibiting the greatest median expression value was selected.

MAS5 Present/Marginal/Absent (PMA) calls were then used to identify genes that were expressed in each of the tissues (i.e. their expression universes). A gene was flagged as being expressed in a particular tissue if it was called as 'Present' by the MAS5 algorithm in at least one of the two samples that represented that tissue.

MAS5 normalization of CEL files, and generation of PMA calls was carried out using the `justMAS()` and `detection.p.value()` functions, respectively, both of which are available in the Bioconductor package `simpleaffy` (Wilson and Miller 2005).

10.5.2 Hierarchical clustering of 68 human tissues using overlap sizes of expression universes

Section 5.2.3 described the use of hierarchical clustering for the unsupervised classification the 68 tissues. This was carried out to observe if the resultant groupings of samples could reflect tissue-specific expression patterns. Figure 10.2 displays the scheme used for the creation of a distance matrix for the clustering procedure. The simulations were designed to control for the effects of the variability in the sizes of the expression universes of different tissues on the size of overlap.

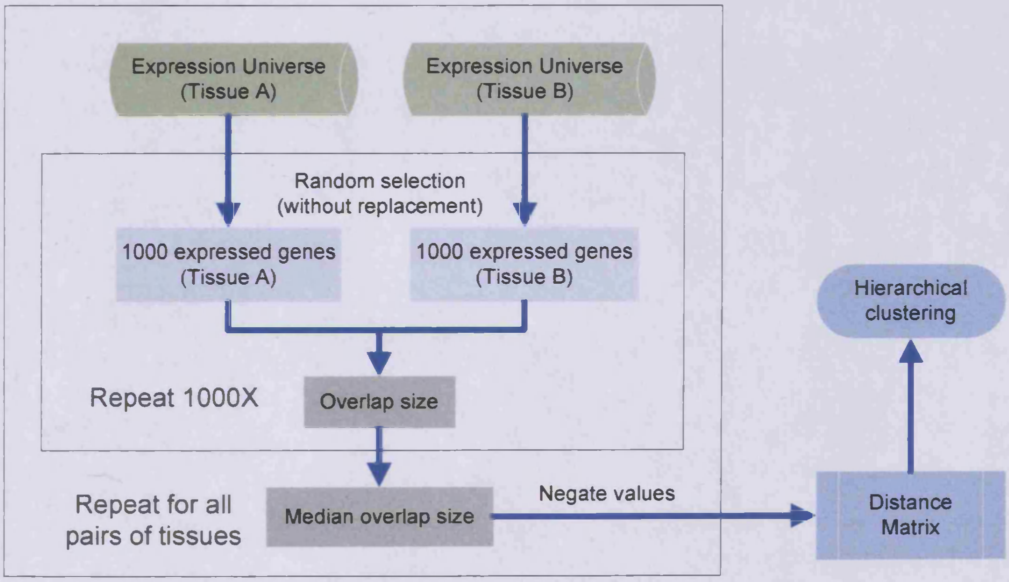

Expression Universe (Tissue A)
Expression Universe (Tissue B)
Random selection (without replacement)
1000 expressed genes (Tissue A)
1000 expressed genes (Tissue B)
Hierarchical clustering
Repeat 1000X
Overlap size
Repeat for all pairs of tissues
Median overlap size
Negate values
Distance Matrix

Figure 10.2 Deriving overlaps sizes for use as distances in hierarchical clustering of tissues.

10.5.3 Simulating the effects of using an estimated 'average' gene universe size.

Section 5.2.4 then described simulations to explore the behaviour of hypergeometric Z-scores, relative to the gene universe size used for their calculation, of comparisons between genelists created by random sampling of genes form universes of various sizes. Figure 10.3 displays the design for these simulations, which was formulated to enable control of the effects of genelist length on overlap size.

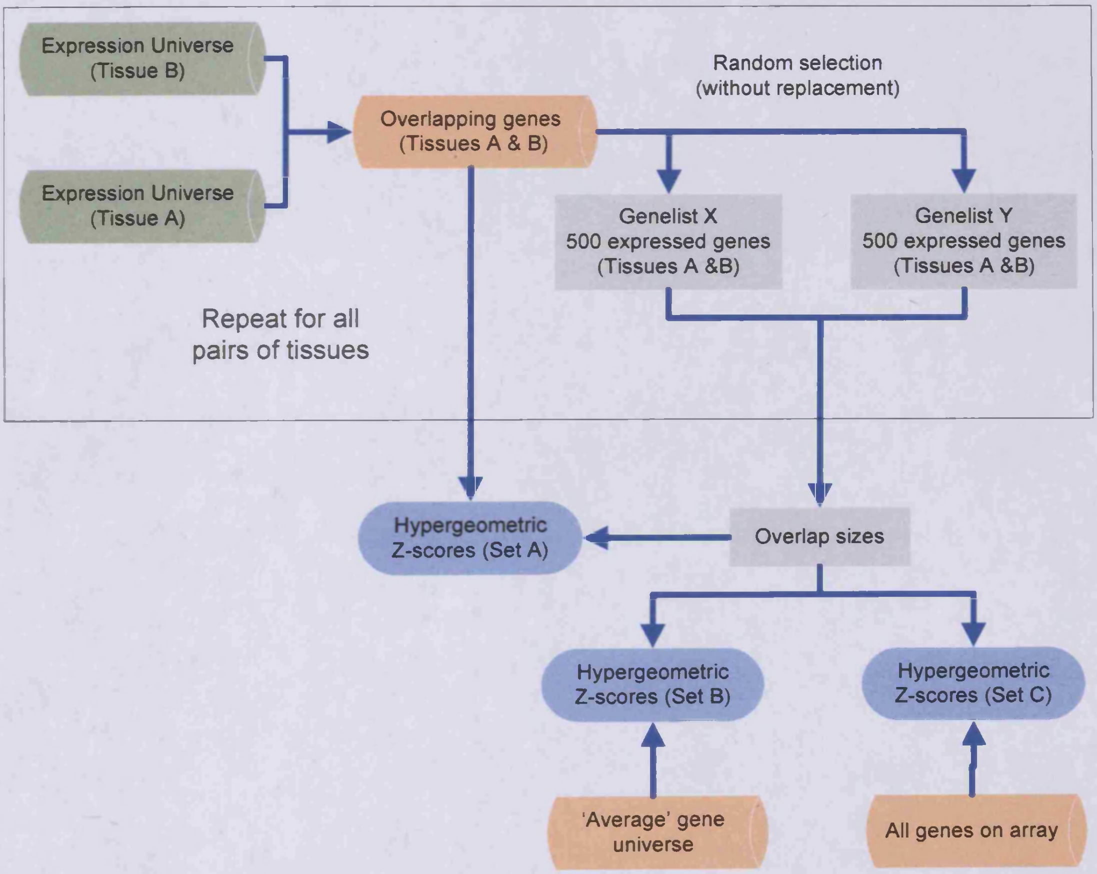

Expression Universe (Tissue B)
Expression Universe (Tissue A)
Overlapping genes (Tissues A & B)
Random selection (without replacement)
Genelist X 500 expressed genes (Tissues A &B)
Genelist Y 500 expressed genes (Tissues A &B)
Repeat for all pairs of tissues
Hypergeometric Z-scores (Set A)
Overlap sizes
Hypergeometric Z-scores (Set B)
Hypergeometric Z-scores (Set C)
'Average' gene universe
All genes on array

Figure 10.3 Calculation of three sets of Z-scores using the same set of overlap sizes.

10.6 Gene Set Discovery (GSD): Unsupervised identification of relevant biological themes within microarray datasets (Chapter 6)

Chapter 6 described explorations using simulated gene expression matrices, and the development of GSD methodology as a tool to carry out unsupervised GSA.

10.6.1 Simulating gene expression matrices

The hypothetical gene expression matrices used in this chapter were intended to simulate log transformed, median-centred expression data. It is expected for genes that are not differentially expressed, expression values would vary around zero. Up-regulation is indicated by values >0; for example, when using log of base 2, a value of 1 represented a two-fold increase in expression as compared to the median value for that gene across all samples. Similarly, a value of -1 represents a two-fold decrease in expression.

Population of the hypothetical gene expression matrices with simulated expression values required estimation of the levels of variability of gene expression values expected in real-world data. For this purpose, 62 Affymetrix hgu133a microarray datasets downloaded by Dr. Peter Giles (Cardiff University) from GEO were used (data not shown). All datasets were logged and median-centred. If more than one probe-set was found to be annotated with the same EGID, the probe-set exhibiting the greatest variability in expression values (measured as IQR) was selected to represent that EGID; this process was carried out for each dataset separately. Standard deviations were calculated for each gene in every dataset, i.e. 62 standard deviation values were derived for each gene represented on the array. The median standard deviations for each gene were then derived. The median of these set of values was ~0.3.

Thus, unchanged gene expression values were simulated by random selection from a normal distribution with a mean value of zero and a standard distribution of 0.3. Similarly, values representing up-regulation and down-regulation were randomly selected from normal distributions with means of 2 and -2 respectively, and a standard distribution of 0.3. This was carried out using the `rnorm()` function in R.

10.6.2 Distances and clustering

One of the most popular distance measures used to quantify the dissimilarity between genes and samples in a microarray dataset is the Pearson Correlation distance (see Section 1.3.2.1). The correlation distances used in this thesis are a variation of this; they differ in the aspect of using the cosine distances of median-centred gene expression data (as opposed to mean-centred data used in Pearson Correlation distance). This is because median values are more robust than mean values; the latter are more sensitive to outlier values. This was carried out by using the R function `Dist()` from the Bioconductor library `amap`. Hierarchical clustering was carried out using the average linkage method through the R function `hclust()`.

10.6.3 Significance testing in the GSD methodology

Chapter 6 described explorations that indicated that SD-SDM values may be suitable metrics to indicate the strength of sample clusters for a given set of genes. The significance of SD-SDM values observed for gene sets could be assessed by comparison with mull distributions of SD-SDM values; two strategies for creating null distributions were tested. A null distribution of 10,000 SD-SDM values was calculated for each separate length of gene set tested. Two metrics to quantify the significance of observed SD-SDM values could be derived: Z-scores and p-values. Z-scores provide 'effect sizes' which are standardized quantifications of the extent of deviation of an observed SD-SDM value from those in the null distribution. This was calculated as:

$$z = \frac{Obs - mean(Null)}{s.d.(Null)},$$

Here, z is the Z-score, *Obs* is the observed SD-SDM value and *Null* is the null distribution for the gene set being tested (based on its length). P-values represent the probability that an observed SD-SDM value could have been observed by chance alone. This was calculated as the number of SD-SDM values, in the corresponding null distribution for an observed SD-SDM value, that are greater than or equal to the observed SD-SDM value, divided by the size of the null distribution (10,000).

10.7 Application of the GSD methodology to four microarray datasets (Chapter 7)

Chapter 7 described the results of the application of the GSD methodology to four microarray datasets.

10.7.1 Data acquisition

The first dataset analyzed using the GSD methodology was a subset of the GNF Expression Atlas (Su et al. 2004). This is the same dataset that was used in Chapter 5. See section 10.5.1 for further details. The second dataset analyzed was the Ross AML dataset (Ross et al. 2004) created using the Affymetrix hgu133a platform. Raw data (CEL files) was downloaded from the website of the St. Jude Children's Research Hospital (http://www.stjuderesearch.org/data/AML1/rawFiles/). Only that subset of the data was used which represented tumour samples that could were categorized in the original study into one of five different AML sub-types. Samples that were classified as 'other' were not considered. The third GSD analysis described was that of the Broccoli liposarcoma dataset, which was created using the Affymetrix hgu133plus2 platform. This was provided by Dr. Dominique Broccoli from the Curtis and Elizabeth Anderson

Cancer Institute at the Memorial University Medical Centre, USA. The final GSD analysis was performed on the Ivshina breast cancer dataset (Ivshina et al. 2006). Both raw data (CEL files) and clinical data (used in survival analyses) were downloaded from the Gene Expression Omnibus (GEO) database (Edgar et al. 2002), where this data was stored with the series identifier GSE4922. Only data from the Uppsala cohort was used. This dataset was created using the Affymetrix hgu133 set platform, which includes both the hgu133a and hgu133b arrays. Only data from the hgu133a arrays was used.

10.7.2 Data pre-processing and transformation

All datasets were normalized using the MAS5 algorithm. This was carried out using the R function `justMas()` which is available in the Bioconductor package `simpleaffy` (Wilson and Miller 2005). Normalized data was then further processed prior to application of the GSD methodology firstly by log transformation and then by median-centring. Log transformation was carried out with the R function `log()`.

10.7.3 Gene Ontology Biological Process terms

The series of gene sets representing biological themes used for GSD analyses of microarray datasets were Gene Ontology Biological Process (GOBP terms). Gene sets were created for each different Affymetrix platform separately. These were derived from the Bioconductor annotation package `GO`. This was carried out by extracting the GOBP term annotation for each EGID represented on any particular array. Gene sets for each GOBP term could then be created by selecting all genes which are annotated with. Because the Bioconductor annotation package provides for each EGID only the most specific GOBP annotation, the gene set for each GOBP term was also made to include all genes annotated with any descendant GOBP terms. For all GSD analyses, those GOBP terms comprising of less than 5 genes, or more than 10% of genes present on an array, were excluded.

10.7.4 Hypergeometric tests

ORA analysis of the classifier developed by Ivshina et al (Ivshina et al. 2006) was carried out using the hypergeometric statistical test (see Section 10.2.2.3). The gene universe used for this purpose comprised of all EGIDs represented on the Affymetrix hgu133a array for which GOBP term annotation was available. Thus, two genes from the classifier (EGIDs 57758 and 83461) were excluded because they were not annotated with any GOBP terms. The hypergeometric test was also used to assess the enrichment of tumour grades in sample clusters derived from mitosis/cell cycle GOBP terms selected by GSD analysis of the Ivshina breast cancer dataset. The universe comprised of all samples in the dataset.

10.7.5 Survival analyses

Kaplan-Meier survival curves and assessment of the difference between survival rates were carried out using the R functions `survfit()` and `coxph()` respectively. Both functions are available as part of the R package `survival`.

10.7.6 Biomarkers

Figure 7.20 shows the expression values for three biomarkers in the Broccoli liposarcoma dataset. The first (7.20a) is that for the IGKV gene which is represented on the hgu133plus2 array by the probeset 214768_x_at. The second (7.20b) is that for the MCM2 which is a biomarker for proliferation, and is represented by the probeset 202107_s_at. The third (7.20c) is that for the adipocytic biomarker leptin, which is represented by the probeset 207175_at. The data represented in the plots is log transformed MAS5 expression data.

10.8 Extraction of informative genes and visualization of GSD results (Chapter 8)

This chapter described two further extensions developed for the GSD framework of analysis. The first of these was a methodology to extract informative genes from within gene sets selected by the GSD methodology. The second was a scheme for integrated visualization of results.

10.8.1 Simulation of expression matrices

To simulate gene expression matrices to test the metric developed to extract informative genes, the same principles were used as for the simulations described in Chapter 6 (see Section 10.6.1)

10.8.2 Clustering of gene sets

Part of the scheme developed to visualize results of GSD analyses was displaying of relationships between GOBP terms. This was carried out by creating a binary matrix using genes selected as informative by GSD. In this matrix, the rows represented the genes and the columns represented GOBP terms. The presence of GOBP annotation for a gene was indicated by a cell value of 1, and absence by 0. Using this correlation distance matrix could be created for the GOBP terms, which was in turn used for hierarchical clustering using the average linkage method. The dendrogram derived from the clustering was used to represent relationships between the GOBP terms in the visualization scheme.

10.9 Summary and General Discussion (Chapter 9)

Chapter 9 comprises primarily of summarization and discussion of the results and explorations described in Chapter 2-8.

10.9.1 Analysis of microarray datasets using *adSplit*

One of the few instances of primary research described in Chapter 9 involves analysis of microarray data using the adSplit methodology (Lottaz et al. 2007). This was carried out as an initial comparison between GSD and adSplit. For this purpose, the same datasets were used as had been used for GSD analyses described in Chapter 7, and the same pre-processing (MAS5) and transformation (log and median-centring) steps were undertaken. The same series of GOBP terms were also used, except without GOBP terms that contained only 5 genes, because with the default settings, adSplit removes the top 5 most variable genes from analyses. For this purpose, the FDR correction was re-applied to p-values derived from GSD, without these excluded GOBP terms.

The adSpit methodology was implemented using the `diana2means()` function available in the Bioconductor package `adSplit`.

Appendices

All appendices are provided in the attached CD-ROM.

Appendix I: Literature-derived genelists

(a) Details of literature-derived genelists

An Excel spreadsheet is provided containing the details of all the genelists that were manually extracted from the literature, and used for the comparisons described in Chapter 3.

(b) Literature-derived genelists

This comprises of a folder containing all the literature-derived genelists that were compared in Chapter 3. These are provided as R objects names after the Affymetrix arrays that the genelists were derived from. Each of the R objects contains a binary matrix called 'EGMat', where the rows represent all genes (as Entrez Gene IDs) on the respective array, and the columns represent experiments. A cell value of 1 indicates the presence of the gene represented by that row in the genelist derived from the experiment represented by that column; a value of 0 indicates its absence.

(c) Genelists from L2L

This comprises of a folder containing genelists that were downloaded from the L2L database and compared in Chapter 3. They are provided in a similar format as the genelists in Appendix Ib, except that genes are represented by HUGO gene symbols rather than Entrez Gene IDs.

Appendix II: R code for GSD analysis

(a) GOBP_lists

R code to derive Gene Ontology Biological Process (GOBP) annotations for all genes that are represented on a particular Affymetrix array.

(b) Transformer

R code for log transformation and median centring of MAS5 gene expression data, as well as the selection of a single probeset per Entrez Gene ID represented on the array.

(c) Null Distribution

R code to create the null distribution of SD-SDM values for comparison with those derived from GOBP terms.

(d) GSD

R code to calculate SD-SDM values for GOBP terms and the comparison of these with the null distribution SD-SDM values to derive p-values and Z-scores.

(e) Informative genes

R code to extract informative genes from within GOBP terms selected by GSD.

(f) Visualization

R code to visualize results of GSD analysis using an integrated scheme.

(g) GSD Usage

R code demonstrating how the above functions can be used, starting from MA5 data.

Bibliography

Affymetrix 2001. *Array Design for the GeneChip Human Genome U133 Set.* Affymetrix Inc. Santa Clara, CA.

Affymetrix 2002. *Statistical Algorithms Description Document.* Affymetrix Inc. Santa Clara, CA.

Aggarwal, A. et al. 2006. Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res* 66(1), pp. 232-241.

Alizadeh, A. A. et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769), pp. 503-511.

Allison, D. B. et al. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7(1), pp. 55-65.

Ashburner, M. et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1), pp. 25-29.

Bammler, T. et al. 2005. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2(5), pp. 351-356.

Ben-Shaul, Y. et al. 2005. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics* 21(7), pp. 1129-1137.

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, Series B (Methodological)* 57(1), pp. 289-300.

Bild, A. and Febbo, P. G. 2005. Application of a priori established gene sets to discover biologically important differential expression in microarray data. *Proc Natl Acad Sci U S A* 102(43), pp. 15278-15279.

Bild, A. H. et al. 2006. Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer* 6(9), pp. 735-741.

BioCarta 2005. www.biocarta.com.

Bittner, M. et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406(6795), pp. 536-540.

Bolstad, B. M. et al. 2005. Pre-processing High-density Oligonucleotide Arrays. In: Gentleman, R. et al. eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. pp. 13-32.

Borozan, I. et al. 2008. MAID : an effect size based model for microarray data integration across laboratories and platforms. *BMC Bioinformatics* 9, p. 305.

Breitling, R. et al. 2004. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* 5, p. 34.

Brown, P. O. and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat Genet* 21(1 Suppl), pp. 33-37.

Butte, A. 2002. The use and analysis of microarray data. *Nat Rev Drug Discov* 1(12), pp. 951-960.

Cahan, P. et al. 2005. List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. *Gene* 360(1), pp. 78-82.

Cahan, P. et al. 2007. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene* 401(1-2), pp. 12-18.

Causton, H. C. et al. 2003. Analysis of gene expression data matrices.*Microarray Gene Expression Data Analysis: A Beginner's Guide*. pp. 71-133.

Chan, M. M. et al. 2005. Gene expression profiling of NMU-induced rat mammary tumors: cross species comparison with human breast cancer. *Carcinogenesis* 26(8), pp. 1343-1353.

Chang, H. Y. et al. 2004. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2(2), p. E7.

Cheadle, C. et al. 2007. A rapid method for microarray cross platform comparisons using gene expression signatures. *Mol Cell Probes* 21(1), pp. 35-46.

Choi, J. K. et al. 2004. Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett* 565(1-3), pp. 93-100.

Choi, J. K. et al. 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19 Suppl 1, pp. i84-90.

Cope, L. M. et al. 2004. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20(3), pp. 323-331.

Curtis, R. K. et al. 2005. Pathways to the analysis of microarray data. *Trends Biotechnol* 23(8), pp. 429-435.

De Cecco, L. et al. 2004. Gene expression profiling of advanced ovarian cancer: characterization of a molecular signature involving fibroblast growth factor 2. *Oncogene* 23(49), pp. 8171-8183.

de Magalhaes, J. P. et al. 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* 25(7), pp. 875-881.

DeMeo, D. et al. 2006. The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Proc Am Thorac Soc* 3(6), p. 502.

Dudiot, S. and Fridyland, J. 2003. Classification in Microarray Experiments. In: Speed, T. ed. *Statistical Analysis of Gene Expression Microarray Data.* pp. 93-130.

Edgar, R. et al. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1), pp. 207-210.

Ein-Dor, L. et al. 2005. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21(2), pp. 171-178.

Eisen, M. B. et al. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95(25), pp. 14863-14868.

Ellwood-Yen, K. et al. 2003. Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell* 4(3), pp. 223-238.

Falcon, S. and Gentleman, R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* 23(2), pp. 257-258.

Falcon, S. and Gentleman, R. 2008. Hypergeometric Testing Used for Gene Set Enrichment Analysis. In: Hahne, F. et al. eds. *Bioconductor Case Studies.* pp. 207-220.

Fang, H. et al. 2005. Bioinformatics approaches for cross-species liver cancer analysis based on microarray gene expression profiling. *BMC Bioinformatics* 6 Suppl 2, p. S6.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.

Finocchiaro, G. et al. 2005. Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control. *BMC Bioinformatics* 6 Suppl 4, p. S14.

Fog, A. 2007. *Biased Urn Theory*. Biconductor Vignette.

Gentleman, R. and Huber, W. 2008. Preprocessing Affymetrix Expression Data. In: Hahne, F. et al. eds. *Bioconductor Case Studies*. pp. 25-26.

Gentleman, R. C. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10), p. R80.

Golub, T. R. et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), pp. 531-537.

Grutzmann, R. et al. 2005. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 24(32), pp. 5079-5088.

Hamid, J. S. et al. 2008. Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics* 2009.

Hu, P. et al. 2005. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics* 6, p. 128.

Huang da, W. et al. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1), pp. 1-13.

Huber, W. et al. 2008. Differential Expression. In: Hahne, F. et al. eds. *Bioconductor Case Studies*. pp. 89-102.

Hwang, K. B. et al. 2004. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics* 5, p. 159.

Ihaka, R. and Gentleman, R. 1996. R: A language for data analysis and graphics. *Journal of Computational Graphics and Statistics* 5(3), pp. 299-314.

Irizarry, R. A. et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2), pp. 249-264.

Irizarry, R. A. et al. 2005. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2(5), pp. 345-350.

Ivshina, A. V. et al. 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* 66(21), pp. 10292-10301.

Jeffery, I. B. et al. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 7, p. 359.

Kanehisa, M. et al. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32(Database issue), pp. D277-280.

Keegan, K. P. et al. 2007. Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Comput Biol* 3(11), p. e208.

Khatri, P. and Draghici, S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21(18), pp. 3587-3595.

Kim, R. D. and Park, P. J. 2004. Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol* 5(9), p. R70.

Kim, T. M. et al. 2007. Inferring biological functions and associated transcriptional regulators using gene set expression coherence analysis. *BMC Bioinformatics* 8, p. 453.

Kuruvilla, F. G. et al. 2002. Vector algebra in the analysis of genome-wide expression data. *Genome Biol* 3(3), p. RESEARCH0011.

Lander, E. S. 1999. Array of hope. *Nat Genet* 21(1 Suppl), pp. 3-4.

Lander, E. S. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822), pp. 860-921.

Larkin, J. E. et al. 2005. Independence and reproducibility across microarray platforms. *Nat Methods* 2(5), pp. 337-344.

Lee, J. S. et al. 2004. Application of comparative functional genomics to identify best-fit mouse models to study human cancer. *Nat Genet* 36(12), pp. 1306-1311.

Lee, J. S. et al. 2005. Comparative functional genomics for identifying models of human cancer. *Carcinogenesis* 26(6), pp. 1013-1020.

Letunic, I. and Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1), pp. 127-128.

Lipshutz, R. J. et al. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* 21(1 Suppl), pp. 20-24.

Liu, G. et al. 2003. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 31(1), pp. 82-86.

Lockhart, D. J. and Winzeler, E. A. 2000. Genomics, gene expression and DNA arrays. *Nature* 405(6788), pp. 827-836.

Lottaz, C. et al. 2007. Annotation-based distance measures for patient subgroup discovery in clinical microarray studies. *Bioinformatics* 23(17), pp. 2256-2264.

Manoli, T. et al. 2006. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* 22(20), pp. 2500-2506.

McCarroll, S. A. et al. 2004. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 36(2), pp. 197-204.

Mootha, V. K. et al. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34(3), pp. 267-273.

Moreau, Y. et al. 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet* 19(10), pp. 570-577.

Mulligan, M. K. et al. 2006. Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc Natl Acad Sci U S A* 103(16), pp. 6368-6373.

Nam, D. and Kim, S. Y. 2008. Gene-set approach for expression pattern analysis. *Brief Bioinform* 9(3), pp. 189-197.

Newman, J. C. and Weiner, A. M. 2005. L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol* 6(9), p. R81.

O'Brien, K. P. et al. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 33(Database issue), pp. D476-480.

Pan, K. H. et al. 2005. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci U S A* 102(25), pp. 8961-8965.

Pan, W. 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18(4), pp. 546-554.

Parkinson, H. et al. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35(Database issue), pp. D747-750.

Parmigiani, G. et al. 2004. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* 10(9), pp. 2922-2927.

Pavlidis, P. et al. 2002. Exploring gene expression data with class scores. *Pac Symp Biocomput*, pp. 474-485.

Pease, A. C. et al. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* 91(11), pp. 5022-5026.

Perou, C. M. et al. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797), pp. 747-752.

Quackenbush, J. 2001. Computational analysis of microarray data. *Nat Rev Genet* 2(6), pp. 418-427.

Ramasamy, A. et al. 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 5(9), p. e184.

Rhodes, D. R. et al. 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 62(15), pp. 4427-4433.

Rhodes, D. R. et al. 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* 101(25), pp. 9309-9314.

Rivals, I. et al. 2007. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 23(4), pp. 401-407.

Ross, M. E. et al. 2004. Gene expression profiling of pediatric acute myelogenous leukemia. *Blood* 104(12), pp. 3679-3687.

Rubin, E. 2005. List mania: interpreting microarray results with the L2L server. *Brief Bioinform* 7(1), pp. 212-122.

Schena, M. et al. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235), pp. 467-470.

Schlicht, M. et al. 2004. Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium. *BMC Genomics* 5(1), p. 58.

Scholtens, D. and Heydebreck, A. 2005. Analysis of Differential Gene Expression Studies. In: Gentleman, R. et al. eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. pp. 229-248.

She, X. et al. 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* 10, p. 269.

Shen, R. et al. 2008. Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med Genomics* 1, p. 28.

Slodkowska, E. A. and Ross, J. S. 2009. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* 9(5), pp. 417-422.

Smith, D. D. et al. 2008. Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics* 9, p. 63.

Stevens, J. R. and Doerge, R. W. 2005. Meta-analysis combines affymetrix microarray results across laboratories. *Comp Funct Genomics* 6(3), pp. 116-122.

Su, A. I. et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101(16), pp. 6062-6067.

Subramanian, A. et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43), pp. 15545-15550.

Sweet-Cordero, A. et al. 2005. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet* 37(1), pp. 48-55.

Tan, P. K. et al. 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 31(19), pp. 5676-5684.

Tarca, A. L. et al. 2006. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol* 195(2), pp. 373-388.

Troyanskaya, O. G. 2005. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform* 6(1), pp. 34-43.

Tsai, J. et al. 2001. RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol* 2(11), p. SOFTWARE0002.

van 't Veer, L. J. et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871), pp. 530-536.

Venter, J. C. et al. 2001. The sequence of the human genome. *Science* 291(5507), pp. 1304-1351.

von Heydebreck, A. et al. 2001. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics* 17 Suppl 1, pp. S107-114.

Vukmirovic, O. G. and Tilghman, S. M. 2000. Exploring genome space. *Nature* 405(6788), pp. 820-822.

Wang, J. et al. 2004. Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20(17), pp. 3166-3178.

Warnat, P. et al. 2005. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 6, p. 265.

Wennmalm, K. et al. 2005. The expression signature of in vitro senescence resembles mouse but not human aging. *Genome Biol* 6(13), p. R109.

Wheeler, D. L. et al. 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36(Database issue), pp. D13-21.

Wilson, C. L. and Miller, C. J. 2005. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 21(18), pp. 3683-3685.

Xu, L. et al. 2005. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics* 21(20), pp. 3905-3911.

Yi, Y. et al. 2007. Strategy for encoding and comparison of gene expression signatures. *Genome Biol* 8(7), p. R133.

Zhou, X. J. and Gibson, G. 2004. Cross-species comparison of genome-wide expression patterns. *Genome Biol* 5(7), p. 232.