Application of Improved Automated Text Mining to Transcriptome Datasets

by

Hui Sun Leong

A thesis submitted to Cardiff University
for the degree of

Doctor of Philosophy

Department of Pathology
School of Medicine
Cardiff University

September 2009

Declaration and statements

Declaration

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

Signed ....

Date 19/11/2009

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Signed ....

Date 19/11/2009

Statement 2

This thesis is the result of my own independent investigations, except where otherwise stated. References are given where other sources are acknowledged. A bibliography is appended.

Signed ....

Date 19/11/2009

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 ....

Date 19/11/2009

Abstract

A major challenge in microarray data analysis is the functional interpretation of gene lists. A common approach to address this is over-representation analysis (ORA), which uses the hypergeometric test (or its variants) to evaluate whether a particular functionally-defined group of genes is represented more than expected by chance within a gene list. Existing applications of ORA have been largely limited to controlled vocabularies such as Gene Ontology (GO) terms and KEGG pathways. Therefore, this work aims at determining whether ORA can be applied to a wider mining of free-text. Initial explorations using the classical hypergeometric distribution to analyse tokens from PubMed abstracts revealed a hitherto unexpected feature: gene lists derived from a typical microarray experiment tend to have more annotation (PubMed abstracts) associated with them than would be expected by chance. This bias, a result of patterns of research activity within the biomedical community, is a major problem for the classical hypergeometric test-based ORA approach, as it cannot account for such bias. The negative effect of annotation bias is a marked over-representation of many common (and likely uninformative) terms, interspersed with terms that appear to convey real biological insight. Several solutions have been developed to address this issue. The first is based on the use of a permutation test, but this nonparametric approach is hampered by being computationally intensive. Two computationally tractable approaches were subsequently developed, which are based on the detection of outliers and the extended hypergeometric distribution. The performances of the proposed text-based ORA approaches were demonstrated on a wide range of published datasets covering different species. A comparison with existing tools that use GO terms suggests that mining PubMed abstracts can reveal additional biological insight that may not be possible by mining pre-defined ontologies alone.

Acknowledgements

I would like to express my appreciation and heartfelt thanks to my supervisor Professor David Kipling for giving me the opportunity to undertake this interesting piece of work as well as providing me help and support throughout the course of this project. The work presented in this thesis owes much to his enthusiasm and expert guidance.

This work was supported by Cancer Research UK (CRUK) under grant number C8731/A5579. Without the funding from CRUK, this work would not have been possible. My sincere gratitude also goes to the Overseas Research Student Awards Scheme, who provides funding to cover most of the tuition fees of this PhD study.

Special thanks go to Dr. Peter Giles for his help, ideas and enlightening discussions. The web server PAKORA presented in Chapter 9 would not have existed without the technical support and help from him.

I am very grateful to several other people for their useful comments, inspirations and insightful discussions, including Professor Peter Holmans, Dr. Lesley Jones and Suraj Menon. Many thanks to Kaye Smith for kindly sharing with me her interferon gene list, which was the primary testing dataset during the early stages of this work. My gratitude is extended to my second supervisor Dr. Ian Brewis for his encouragement and support. Also, I thank Bethan Moore and Sally Ellis for their assistance in sorting out all sorts of paperwork, ordering and requests.

I would also like to thank all the people who have made the time I spent in Cardiff pleasant and interesting. I am very fortunate to have a group of friends with whom I enjoy some wonderful games of badminton and ping pong; their accompaniment and the sports have kept me sane and provided an outlet for frustration while writing up.

Last but not least, I am most grateful and obliged to my parents and my brother. I cannot thank them enough for their encouragement, love and continued support throughout the years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

cDNA	Complementary DNA
DAG	Directed Acyclic Graph
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
EGID	EntrezGene identifer
EST	Expressed Sequence Tag
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GeneRIF	Gene Reference into Function
GENIA	Gene Expression Information System for Human Analysis
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HGNC	HUGO Gene Nomenclature Committee
HUGO	Human Genome Organization
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
ISG	Interferon-Stimulated Genes
IVT	*In vitro* Transcription
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAD	Median Absolute Deviation
MEDLINE	Medical Literature, Analysis, and Retrieval System Online
MeSH	Medical Subject Headings
MLE	Maximum Likelihood
mRNA	Messenger Ribonucleic Acid
MSigDB	Molecular Signatures Database
NCBI	National Center for Biotechnology Information
NER	Name Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
OMIM	Online Mendelian Inheritance in Man

ORA	Over-Representation Analysis
PMID	PubMed unique identifier
SD	Standard Deviation
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System
URL	Uniform Resource Locator

Chapter 1

Introduction

1.1 Motivation and aims

The output of a microarray experiment is typically one or more lists of genes that show an “interesting” change in expression in the context of that experiment. This is often not the end point of the analysis, but the start of a complex process of deriving biological interpretation. Many researchers interpret their results by manually reviewing the function of each gene based on literature or database searches, in an attempt to make judgements about the underlying biology. Given the number and diversity of genes involved, deriving biological interpretation for such lists manually is a time-consuming and unwieldy task. Frequently, the list of candidate genes is too long to develop a parsimonious view of the underlying biological processes.

The need to formalise this interpretation process has led to the development of a range of methods, of which a statistical procedure known as Over-Representation Analysis (ORA) is becoming a common standard for interpreting microarray gene lists. The fundamental question asked by ORA is: what biological terms or functional categories are represented in the gene list more often than expected by chance. The most common approach to test this statistically is by using the hypergeometric test (or variants such as Fisher’s exact test) to calculate the probability of seeing at least a particular number of genes containing the biological term of interest in the gene list. To date, this mode of analysis has been implemented (with minor variations) in more than 40 software tools.

However, existing applications of ORA are largely limited to the mining of pre-defined, controlled vocabularies such as Gene Ontology (GO) or pathway annotations from KEGG. These resources are mostly hand-crafted by experts, with the aim of providing a structured, distilled description of the biological knowledge about genes in the peer-reviewed published literature. Annotating genes with these controlled vocabularies is a labour-intensive task, and therefore these pre-defined functional annotations are inevitably limited in scope and cannot be expected to represent all of the concepts in all areas of biology that might be of interest to biologists. Scientific literature, on the other hand, contains a much greater wealth of information about genes and therefore constitutes a valuable resource for interpreting genome-wide experiments.

The aim of the research described in this thesis is to determine whether the successful applications of ORA can be extended beyond the mining of controlled vocabularies to a wider mining of free-text. Initial effort will focus on applying the classical hypergeometric distribution as the statistical model to mine text-based information encapsulated in the PubMed abstracts.

The objectives are:

- Address to what extent text-based ORA approach can assist in the biological interpretation of microarray gene lists.
- Develop approaches for mining literature-based information associated with a list of differentially expressed genes, and to search within them for terms that are significantly over-represented.
- Test and validate the proposed approaches using publicly available datasets.
- Provide an easily accessible web interface for the proposed methods.

Several aspects of microarray technology, gene expression data analysis and biological text mining will be treated in this Chapter, in order to provide a general background to the current research. Particular emphasis will be placed on the current applications of ORA and related methods (Section 1.3). The concepts of biological text mining and the methods they used will be introduced in Section 1.4. Existing ontology-based ORA methods and text mining tools dedicated to the mining of gene expression data will be surveyed and compared.

1.2 Microarrays

Life depends on the ability of cells to store, retrieve, and translate the genetic instructions required to make and maintain a living organism. This genetic information is stored in deoxyribonucleic acid (DNA)[1]. The "central dogma" of molecular biology is a paradigm of genetic information flow in living organisms which states that the genetic information flows from DNA to mRNA to proteins (Figure 1.1). During transcription, the information contained in the DNA sequence is used to produce mRNA (messenger ribonucleic acid), and then mRNA is used as a template to synthesise proteins during translation. Proteins are the active components of the cells that are responsible for a wide range of intra- and extracellular activities, including enzymatic activity, transport, storage, and providing structural integrity to cells.

A gene is a segment of DNA that encodes specific information for making a protein. In addition to expressing mRNAs that encode proteins, genes could also encode transcripts (such as tRNA, snRNA and miRNA) that function directly as structural, catalytic or regulatory RNAs (Eddy 2001). The genome of an organism refers to the entire complement of DNA in any of its cells. The process by which a gene exerts its effect on a cell or organism via the synthesis of mRNA and protein is termed *gene expression*. Gene expression is regulated by a complex array of molecules, and only occurs when a specific protein is required. Knowing the alteration in patterns of gene expression in various tissues, developmental stages and under different physiological conditions can offer new insights concerning regulatory mechanisms and biochemical pathways, which is important for addressing questions such as: "what are the

[1] DNA is the hereditary material in all present day cells, except for some RNA viruses, which use RNA instead of DNA to carry the hereditary information from one generation to another.

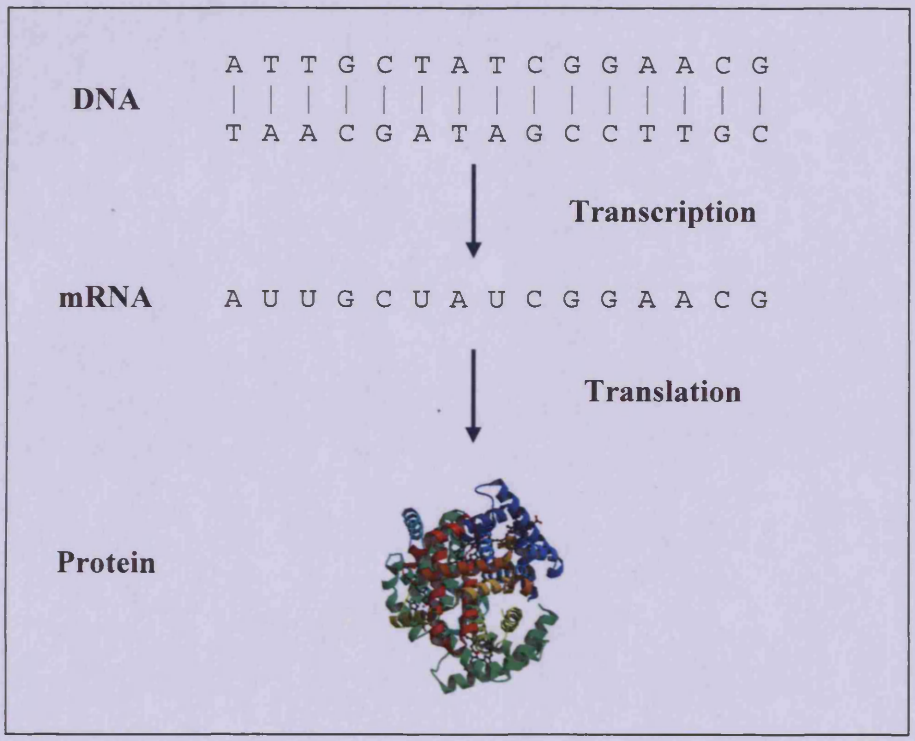

DNA
A T T G C T A T C G G A A C G
T A A C G A T A G C C T T G C
Transcription
mRNA
A U U G C U A U C G G A A C G
Translation
Protein

Figure 1.1: The "central dogma" of molecular biology: genetic information flows from DNA to RNA to proteins
Protein structure image was obtained from RCSB Protein Data Bank (PDB); http://www.rcsb.org/pdb/explore.do?structureId=1GZX.

functional roles of different genes and what cellular processes are they involved in", "how are different genes regulated and with what molecules do they interact", "how is gene expression changed by various diseases or compound treatments".

Although mRNA is not the final product of a gene, changes in mRNA levels usually result in phenotypic and morphological differences because transcription is the first step in gene regulation. Therefore knowing the abundance of mRNA in the cell is useful. The correlation between the mRNA and protein levels in the cell is not straightforward due to factors like mRNA stability, post-transcriptional splicing, post-translational modifications and degradation; nevertheless the absence of mRNA in a cell is likely to indicate a not very high level of the respective protein, therefore at least qualitative estimates about the proteome can be based on the transcriptome (the collection of mRNA in a cell) information (Brazma and Vilo 2000).

Numerous techniques have been developed for detecting mRNA levels within cells and to use these as a measure of gene expression. These techniques include Northern blotting (Alwine *et al.* 1977), differential display (Liang and Pardee 1992), reverse transcription-polymerase chain reaction (RT-PCR) (Somogyi *et al.* 1995) and serial analysis of gene expression (SAGE) (Velculescu *et al.* 1995). These methods, however, are only suitable for studying the expression of a small subset of genes at a time. Under physiological conditions, genes do not act in isolation. Instead, tens or thousands of them could be actively transcribed at any time within the cell.

With the advent of DNA microarray technologies, it has become possible to simultaneously monitor gene expression at the mRNA transcript level in cells. These genetic snapshots of cells in different conditions can provide insights about the response of various genes under different situations. The term 'microarray' was first introduced by Schena *et al.* in 1995. By 1999, several landmark papers from Brown's group and collaborators were published (DeRisi *et al.* 1996; Iyer *et al.* 1999; Lashkari *et al.* 1997; Schena *et al.* 1996), which described the use of microarrays as methods for monitoring gene expression in the field of high-throughput functional genomics, and set off a trend that is still gaining momentum (Figure 1.2). To date, microarrays have found widespread applications in many biological fields, including cancer prognosis and classification (Golub *et al.* 1999; van 't Veer *et al.* 2002), predicting gene function and identifying drug targets (Marton *et al.* 1998), placing genes in different pathways (Hughes *et al.* 2000; Iyer *et al.* 1999) and evaluating mechanisms of toxicity (Waring *et al.* 2001).

The next section gives a brief overview of the microarray technology, focusing on the Affymetrix GeneChip® system. Issues related to data analysis will be covered in Section 1.2.3.

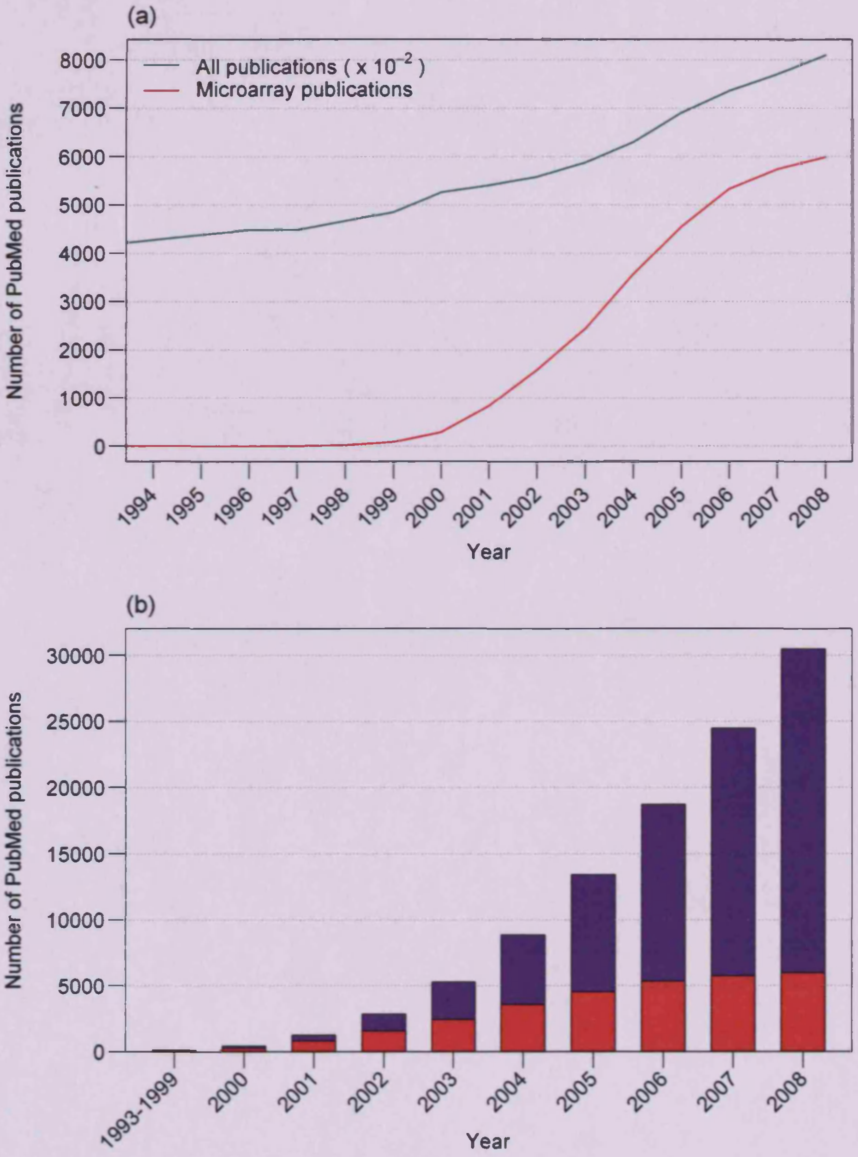


Figure 1.2: The growth in publications concerning microarrays over time
(a) The per annum total number of PubMed publications and the per annum number of publications (i.e. new additions in that year) on microarray that were published from 1994 to 2008 are shown. Note that the per annum total number of PubMed publications (green trend line) plotted has been multiplied by a factor of 0.01. For example, the actual per annum total number of citations recorded in year 1994 is 427556 but is plotted as 4275.56 here. (b) The number of publications that mention microarray increases exponentially from 119 publications at the end of 1999 to a cumulative tally of 30413 in 2008. Remarkably, there are more than 5000 additional microarray publications published each year between 2006 and 2008. In red is the per annum number of publications, in blue is the cumulative count.

1.2.1 Introduction to microarray technology

A typical microarray comprises a solid support (which can be a glass slide, a custom surface, or a membrane) on which a large collection of distinct nucleic acid sequences, known as *probes*, are attached at defined locations. Each probe on a microarray is designed to bind to a specific *target* generated from a particular biological sample under study. Microarrays are based on the principle of preferential hybridisation between complementary, single-stranded nucleic acid sequences, which follows the Watson-Crick rule such that adenine (A) binds to thymine (T) (or uracil (U), in the case of mRNA), and cytosine (C) binds to guanine (G). The idea is that when the array is interrogated with labelled mRNA, and hybridisation will take place between the mRNA targets that contains sequences complementary to the sequences of the probes deposited on the surface of the array. Since the target is labelled with a fluorescent dye or a radioactive element, the hybridisation spot can be detected and quantified. The key to microarray technology is that a probe is detected at a level that is proportional (in a predictable manner) to the abundance of its target RNA in the labelled extract.

Although many microarray systems have been developed by academic groups and commercial suppliers, the field has been dominated in the past by two major systems: the complementary DNA (cDNA) and the high-density oligonucleotide microarrays.

cDNA microarrays (spotted arrays)

This is the oldest microarray technology and was developed at Standford University (Schena *et al.* 1995; Shalon *et al.* 1996). In this platform, cDNA of characterised genes or expressed sequence tags (ESTs) are used as probes. A cDNA is a single-stranded DNA molecule synthesised in the laboratory using mRNA as a template and the enzyme reverse transcriptase, while an EST is a short sub-sequence of a transcribed cDNA known to be expressed in the tissue but not yet characterised as a gene. The probes, each representing a gene, are immobilised by a printer or high-speed robot on a solid surface such as glass slide. Spots are typically 100 ~ 300 microns in size. Using this technique, arrays consisting of 30,000 ~ 40,000 cDNAs can be fitted onto the surface of a conventional microscope slide.

cDNA arrays are also referred to as two-channel microarrays because with this method, experimental and control samples are typically labelled with two different fluorescent dyes, for instance, a red dye (Cy5) for the RNA from the experimental population and a green dye (Cy3) for that from the control population. Both extracts are hybridised on the same microarray, and a measurement is obtained from each DNA spot on the array. The intensity differences of the two fluorescent images are read out as differences in gene transcript abundance between the experimental and control samples. If the RNA from the experimental sample was in excess, it will appear red; if the RNA from the control sample was in abundance, the spot will appear green. If experimental and control samples bind equally, the spot will be yellow; the spot will appear black if neither binds.

High-density oligonucleotide microarrays

These microarrays are manufactured commercially to an extremely high density and accuracy using short oligonucleotides of length between 20 ~ 25 bases as probes (in the case of Affymetrix). In contrast to spotted cDNA arrays, high-density oligonucleotide microarrays use a set of probes to represent a gene, thus providing independent measurement of expression changes for a particular gene. High-density oligonucleotide arrays are synthesised *in situ*, either by photolithography onto silicon wafers or by inkjet printing technology. The former fabrication technique was developed by Affymetrix (www.affymetrix.com/index.affx), while the latter was developed by Rosetta Inpharmatics (www.rii.com) and licensed to Agilent Technologies (www.agilent.com). An important difference between high-density oligonucleotide microarrays and spotted cDNA arrays lies in target preparation. The high reproducibility of *in situ* synthesis of high-density oligonucleotide arrays allows accurate comparison of signals generated by samples hybridised to separate arrays, as opposed to simultaneous hybridisations of two different samples on the same array as with spotted cDNA arrays. Moreover, single-channel technology used by high-density oligonucleotide microarrays offers the advantage of simpler and more flexible experimental design.

The Affymetrix GeneChip® expression arrays are amongst the most popular single-channel platforms available on the market, with the largest panel of microarrays designed for a variety of different organisms, as reflected by the high number of datasets being deposited in public microarray repositories such as Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) (Barrett *et al.* 2009; Edgar *et al.* 2002) and ArrayExpress (www.ebi.ac.uk/microarray-as/ae/) (Brazma *et al.* 2003; Parkinson *et al.* 2009).

Other technologies

Microarray technology has been evolving rapidly in recent years due to the development of more powerful robots for arraying, new fabrication techniques, new labelling protocols, and the ever-increasing genome-sequence information for different organisms. Great efforts have been made in recent years to create higher density microarrays with more features; this means more genes covered and hence wider scope and more comprehensive results. An example is the exon arrays introduced by Affymetrix recently, which have about six times as many features as the previous generation of arrays, and provide comprehensive gene expression data at the level of individual exons. For instance, on the Human Exon 1.0 ST array, approximately 5.5 million probes (forming 1.4 million probesets) are used to separately interrogate one million known and predicted exon clusters (Affymetrix 2005). Other new technologies that have recently entered the market include the bead-array system by Illumina (www.illumina.com) and the digital micromirror arrays by NimbleGen (www.nimblegen.com). A comprehensive comparison of the various microarray platforms from the different manufacturers can be found in Wheelan *et al.* (2008) and Hardiman (2004).

1.2.2 Affymetrix GeneChip® expression array

A number of different GeneChip® expression arrays or array sets are commercially available from Affymetrix, including arrays for human, mouse, rat, *Arabidopsis*, *Drosophila*, yeast and zebrafish. The main one used in this project is the Human Genome U133A array (HG-U133A), which contains 22,215 probesets and provides coverage for 14,500 well-substantiated genes in the human genome. There are several publications discussing the fundamentals of the Affymetrix GeneChip® expression array technology (Lipshutz *et al.* 1999; Lockhart *et al.* 1996). However, for the purpose of this thesis, it will be useful to review some of the elements of this microarray platform.

Array design

Affymetrix GeneChip® expression arrays record the presence of a transcript in a solution by measuring the level of hybridisation between the transcript and a set of short (typically 25mer) oligonucleotide probes anchored to the array surface. As shown in Figure 1.3, each probeset consists of a series of perfect match (PM) probes and mismatch (MM) probes. The PM probe is designed to perfectly match the target transcript, while the MM probe is identical in sequence to its counterpart PM probe except for the central (13th) nucleotide, which is replaced with a mismatched nucleotide. Hybridisation conditions are controlled with the aim of maximising the binding between a transcript and its PM probes, while minimising the binding to its MM probes. The intention is that the PM probes record the presence of the transcript, while MM probes measure background and non-specific hybridisations. However, there is a non-linear functional relationship between the paired PM and MM probe intensities, so for weakly expressed mRNAs where the signal-to-noise ratio is smallest, subtracting MM from PM adds considerably to the noise in the data (Schadt *et al.* 2000). Nevertheless, one advantage of this approach is that the combination of short oligos and strict hybridisation conditions makes it possible to use *in silico* searches to predict which probes are likely to bind to which transcripts - information that is important because many transcripts have similar sequences (e.g. alternate splicing can lead to a set of transcripts being encoded by a single gene).

Each probeset is typically designed to match the more variable 3' non-coding region of its target transcript; however, it is not always possible to identify a set of probes that reliably and uniquely identify a particular transcript, and the design criteria are

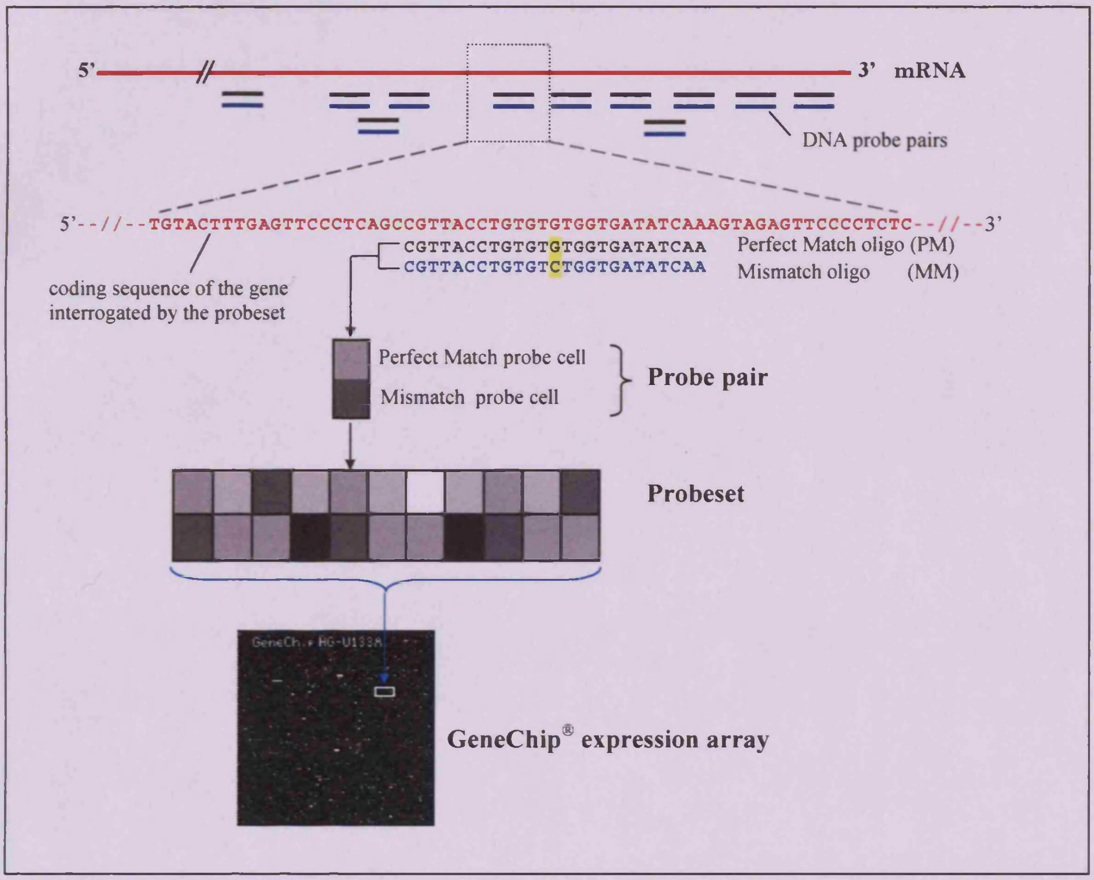

5'
3' mRNA
DNA probe pairs
5'--//--TGTACTTTGAGTTCCCTCAGCCGTTACCTGTGTGTGGTGATATCAAAGTAGAGTTCCCCTCTC--//--3'
CGTTACCTGTGTGTGGTGATATCAA Perfect Match oligo (PM)
CGTTACCTGTGTCTGGTGATATCAA Mismatch oligo (MM)
coding sequence of the gene interrogated by the probeset
Perfect Match probe cell
Mismatch probe cell
Probe pair
Probeset
GeneChip HG-U133A
GeneChip® expression array

Figure 1.3: The design of a typical Affymetrix GeneChip® expression array
Distinct transcript (mRNA) is represented on the array by 11 to 20 probe pairs mapping on 600 bases of the most 3' end of the transcript. Each of these probe pairs consists of a perfect match (PM) probe and a mismatch (MM) probe. A probe is a single-stranded oligonucleotide of 25 bases long that serves to detect the complementary target sequence. The majority of the eukaryotic 3' chips are antisense-target arrays, in which the PM probe is identical to the coding sequence of the gene it represents (an example is shown in the diagram). The MM probe has the same sequence as the PM probe except that the central base is replaced with a mismatched nucleotide (highlighted in yellow). The group of PM and MM probe pairs that together represent a gene or mRNA transcript is called a probeset. This example is based on the HG-U133A chip and is broadly applicable to the previous generation of expression arrays. The designs of other Affymetrix arrays are slightly different. For example, in the newly introduced exon arrays (e.g. Human Exon 1.0 ST array), the probes are designed to interrogate the entire length of a transcript. Also, they no longer have a paired MM probe for each PM probe, and the number of probes per probeset is reduced from 11 to 4.

relaxed accordingly, as reflected in the naming convention used for the probesets:

- Those ending with '_at' are designed to recognise transcripts uniquely.
- Those ending with '_s_at' or '_a_at' are designed to recognise multiple transcripts from the same gene family.
- Those ending with '_x_at' may cross-hybridise into completely unrelated sequences.

Other suffixes exist, and the exact meaning can be dependent on the array type.

The success of an array design is highly dependent on the quality of sequence information used. In the HG-U133A design, UniGene (Build 133) clusters were used as a starting point for the design process but were not used as the main source of sequence information. Instead, sequence information were collected from a number of primary sequence databases, including GeneBank, RefSeq, dbEST and Washington University EST trace repository (WUSTL). The draft assembly of the human genome (April 2001 Release) was used to verify sequence selection, sequence orientation and the quality of the sequence clustering (for more information, see Affymetrix 2001). Since the arrays are designed against sequence databases that are in a state of continual growth, each array therefore represents a snapshot based on the knowledge available at the time it was created. This should be kept in mind when designing and interpreting any microarray study.

Array hybridisation and image processing

Target RNA for array hybridisation is prepared by incorporating fluorescently labelled biotin in an *in vitro* transcription (IVT) reaction process. Following hybridisation, a scanning confocal microscope is used to detect fluorescence from the bound target molecules. An image data (DAT) file is created by the Affymetrix scanning software (e.g. MAS 5.0, GCOS) in which the raw image data are stored. After that, the Affymetrix software aligns a grid on the DAT files and computes the probe-level signal intensity from the pixel values, which is subsequently stored in a CEL file.

1.2.3 Gene expression data analysis

The appropriate choice of data analysis technique depends both on the data and on the goals of the experiment. This section provides an overview of some of the common themes in gene expression data analysis, including pre-processing, detection of differential expression, and extraction of biological knowledge. Despite the differences between different microarray technology platforms, many issues discussed here are broadly applicable to all microarray technology.

Pre-processing

After image processing, the raw signal intensity or "probe-level" data (for the Affymetrix system, this is stored in the CEL file) need to be modified and normalised before multiple microarray measurements can be combined into a single analysis. This procedure is commonly referred to as pre-processing. The main steps of pre-processing are background adjustment, normalisation, summarisation and quality assessment. Background adjustment is needed to account for non-specific hybridisation and noise in the optical detection system. Two commonly used background adjustment methods are the MAS 5.0 algorithm (Affymetrix 2002) and the Robust Multi-chip Average (RMA) algorithm (Irizarry *et al.* 2003). Measurements made across different arrays are not directly comparable because the hybridisations might be obscured by various sources of variations, such as physical problems with the arrays, difference in the efficiencies of reverse transcription, labelling or hybridisation reactions, and the differences in the quantity of initial RNA in the samples. Normalisation adjusts for these variations and makes measurements from different arrays comparable. Many normalisation methods have been proposed in the microarray literature, including Affymetrix's scaling method, dChip (Li and Wong 2001) and quantile normalisation (Bolstad *et al.* 2003). Summarisation is the process of combining the multiple probe intensities for each gene to produce an expression value that estimates an amount proportional to the amount of RNA transcript. This step is needed when each transcript is represented by multiple probes (such as Affymetrix arrays). The final step in pre-processing is quality assessment, which identifies divergent measurements that are beyond the acceptable level of random fluctuations. Non-informative data are usually flagged up or removed at this stage.

After pre-processing, the measurements from different arrays can be combined into a gene expression matrix with rows representing gene transcripts, columns representing different study conditions (e.g. various tissues, developmental stages and treatments), with each element in the matrix corresponding to the abundance (or relative abundance) of a particular gene in a particular condition (Figure 1.4). The set (row) of expression measurements for a gene in the microarray study is commonly referred to as the *expression profile* of that gene. The next task is to analyse the expression values in the matrix and try to extract from it some biological insights regarding the underlying microarray experiment.

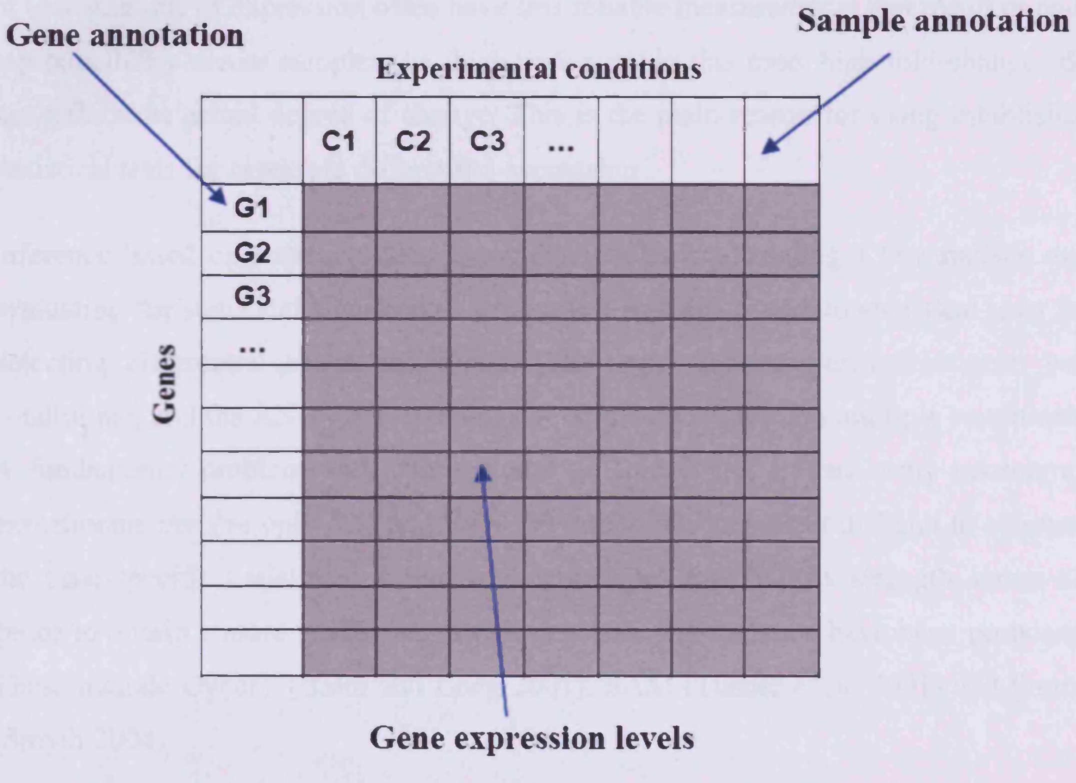


Figure 1.4: Schematic representation of a sample gene expression matrix
Each cell in the matrix corresponds to the expression levels of a particular gene measured under a particular experimental condition. The expression profile of a gene refers to the set (row) of expression measurements for that gene in the microarray study.

Detection of differential expression

Fundamental to the task of analysing gene expression data is the need to identify genes whose levels of expression change significantly according to the phenotype or experimental condition. The choice of statistical approach depends on several issues, such as experimental design (e.g. two-sample comparison, multiple samples or time series), and the number of replicates available. The simplest and most intuitive approach is to select genes using a fold-change criterion, calculated as the ratio of expression levels between two samples (for example, control versus treatment conditions). Genes with a fold-change above a fixed cutoff, typically two- or three-fold, might be considered to be differentially regulated (Chu *et al.* 1998; Schena *et al.* 1996), although fold-change is widely considered as an inadequate test statistic for inference because it does not incorporate the assessment of variance. Genes measured at low amounts of expression often have less reliable measurements that result in poor reproducibility across samples (i.e. high variance); in this case, high fold-changes do not reflect the actual degree of change. This is the main reason for using established statistical tests for assessing differential expression.

Inference based on statistical tests generally involves calculating a test statistic and evaluating the statistical significance of that test statistic. Standard statistical tests for detecting differential expression include the *t*-test (for comparison between two conditions), and the ANOVA F-statistic (for comparison between multiple conditions). A fundamental problem with conventional statistical tests is that many microarray experiments involve only few replicates per condition, making it difficult to estimate the gene-specific variations. Alternative approaches that borrow strength across all genes to obtain a more stable estimate of gene-specific variance have been proposed. These include CyberT (Baldi and Long 2001), SAM (Tusher *et al.* 2001) and limma (Smyth 2004).

Parametric approaches such as those described above generally assume that the data are sampled from normal populations with equal variances. These constraints may only be partially fulfilled in practice. Often logarithmic transformation is used in order to make the distribution of replicated measurements per gene roughly symmetric and close to normal. Alternatively, one can consider nonparametric approaches, such as a permutation test that make less stringent assumptions on the data-generating distribution. However, as with conventional parametric approaches, the power of permutation tests will also be hampered by small sample size when there are not enough replicates to obtain an accurate estimate of experimental variance.

One common differential expression problem that has received much attention recently is time series analysis. These types of experiments are designed to study gene expression changes over time and trend differences between the various experimental groups. The complexity of study design and dynamic nature of time course study pose great challenges to data analysis. Approaches based on the use of ANOVA and Bayesian models have been proposed for analysing such data (Angelini *et al.* 2007; Bar-Joseph 2004; Nueda *et al.* 2007).

The approach of testing each gene for differential expression is popular, because it is relatively straightforward and a standard repertoire of methods is available. However, there are several issues concerning such an approach, including the problem of multiple hypothesis testing. Since microarrays typically monitor the expression levels of thousands of genes in parallel, a large number of hypotheses are tested, increasing the chance of finding false positives (Benjamini and Hochberg 1995). Multiple hypothesis testing procedures are often used to assess the overall significance of the results of a family of hypothesis tests. For examples, Bonferroni and FDR corrections can be applied to control the family-wise error rates and the false discovery rate, respectively. This topic is covered in detail by Dudoit *et al.* (2003). It should be noted that there is a trade-off between specificity and sensitivity. Multiple hypothesis testing methods improve specificity (by adjusting *p*-value) at the expenses of sensitivity (that is, a reduced chance of finding true positives). One way to mitigate such problem is to reduce the number of hypotheses to be tested using some form of non-specific filtering strategy, such as by eliminating genes that do not show sufficient variation in expression across the samples, as these genes tend to provide little discriminatory power.

Biological knowledge extraction

When the microarray experiment consists of a simple comparison between two conditions (such as control versus test), the subsequent data analysis will usually be limited to the identification of differentially expressed genes using the methods introduced above. When there are multiple experimental conditions (such as more than two phenotypes, or different time points), it is useful to group the significantly expressed genes into clusters that behave similarly over the different conditions.

A popular technique for detecting groups of genes demonstrating similar expression pattern is clustering. Clustering is the process of grouping together similar entities according to some distance metric that is computed for one or more features. Most microarray clustering algorithms use the Euclidean distance, Manhattan distance, angle between vectors or the correlation distance, as the distance metric to compute the similarity between two expression profiles. Changing the underlying distance metric may alter the number of clusters and the relationship between them, because each distance metric has specific properties that can be used to emphasise certain characteristics of the data (Draghici 2003).

Clustering can reveal potentially meaningful relationships among genes. Hierarchical clustering (Eisen *et al.* 1998), k-means clustering (Tavazoie *et al.* 1999), or self-organising maps (Tamayo *et al.* 1999), have all been used to derive putative functional clusters of genes from gene expression data. In addition, clustering can be used to infer the biological functions of new genes. The assumption motivating such an approach is that simultaneously expressed genes often share a common function. If an uncharacterised gene is clustered with a group of genes known to be involved in a particular biological process, then it can be assumed that the uncharacterised gene is also involved in the same process. However, there are several core issues that cannot be addressed by using clustering alone. Co-expressed genes do not always share a common function. The reverse is also true: genes that are functionally related may demonstrate strong anti-correlation in their expression levels and cluster into different groups, blurring the relationship between them. Even when expression and function correlate well, the underlying biological mechanism is not always apparent.

Additional database or literature searches are required to explore the underlying cellular responses in the context of available knowledge.

Various approaches have been proposed to incorporate existing biological knowledge into the analysis. The vast majority of these seek to infer whether a functionally-defined set of genes are enriched within the list of differentially expressed genes. Functionally-related genes are usually defined based on structured, controlled vocabularies such as Gene Ontology (GO) classifications (Ashburner *et al.* 2000) or pathways information from KEGG (Kanehisa and Goto 2000). This mode of analysis is commonly referred to as "functional enrichment analysis". Such enrichment approaches offer a summary of the most pertinent biology in a group of differentially expressed genes, and provides mechanistic clues regarding the biological processes underlying the observed change. To date, more than 60 functional enrichment tools have been developed, reflecting the popularity of this approach (Huang *et al.* 2008). The statistical models used and the current state-of-the-art in enrichment analysis are reviewed in Section 1.3.

Another approach to superimposing biological knowledge upon microarray results comes with efforts to find associations between genes in the scientific literature. Peer-reviewed published scientific text contains a distilled version of the most relevant biological discoveries and is a potent source of functional information. This information is invaluable in guiding the investigator in interpreting genomic data. A number of text mining approaches have been developed to process this textual information, and to link groups of genes found in microarrays based on the knowledge extracted from the literature (Blaschke *et al.* 2001; Chaussabel and Sher 2002; Jenssen *et al.* 2001; Shatkay *et al.* 2000). Section 1.4 provides some background on text mining, and describes how it can be integrated with gene expression data mining framework to facilitate the biological interpretation of microarray results.

1.3 Functional enrichment analysis

Despite increasingly elegant statistical approaches to analyse microarray data, making biological sense of microarray results remains a conundrum. To address this, a range of bioinformatics approaches have been developed over the past few years to help with the biological interpretation of microarray gene lists. One of the most widely used approaches is a family of statistical methods collectively known as "Over-Representation Analysis" (ORA). This approach takes a list of differentially expressed genes and test statistically whether particular functionally-defined groups of genes or *gene sets* are over- (or under-) represented in the condition under study. The gene sets are formed prior to the statistical analysis, and can be defined in a number of ways, for examples, by grouping together genes that are part of the same cellular components, involved in the same pathway or biological process, have the same molecular function, or are located on the same chromosome. Two of the most used resources for defining gene sets are Gene Ontology (GO) (Ashburner *et al.* 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000). The significance of over-representation can be assessed using the hypergeometric distribution or its binomial approximation. Since its inception, a large number of ORA-based methods have been published; many of these have been implemented (with minor variations) as web-tools, including EASE (Hosack *et al.* 2003), DAVID (Dennis *et al.* 2003), FatiGO (Al-Shahrour *et al.* 2004), GoMiner (Zeeberg *et al.* 2003), Onto-Express (Draghici *et al.* 2003; Khatri *et al.* 2002) and GeneMerge (Castillo-Davis and Hartl 2003). In 2005, Khatri and Draghici (2005) and Curtis *et al.* (2005) collected 14 ORA-based ontological tools and performed a thorough comparison of their scope, statistical models, visualisation capabilities, corrections for multiple comparisons and reference sets. The development in this field continues to grow. Currently, there are over 40 tools implementing ORA; a comprehensive and up-to-date review can be found in Huang *et al.* (2008).

An alternative statistical procedure to test for functional enrichment is Gene Set Enrichment Analysis (GSEA) (Mootha *et al.* 2003; Subramanian *et al.* 2005). This method has received a great deal of attention from researchers in the field in recent years. The appealing feature of GSEA is its 'threshold-free' strategy, which takes all genes from a microarray experiment without the need to pre-select a list of significantly differentially expressed genes prior to the enrichment analysis. GSEA follows a basic procedure: expression information on all the genes under study is retained; then an enrichment score is calculated for a given gene set using ranks of genes, and the statistical significance of this enrichment score is inferred against a background distribution generated by permuting the labels of the original dataset. The works of Mootha *et al.* (2003) and Subramanian *et al.* (2005) have inspired the development of various GSEA methods, including globaltest (Goeman *et al.* 2004), PAGE (Kim and Volsky 2005) and GSA (Efron and Tibshirani 2007). Many of these were recently reviewed in Nam and Kim (2008).

A list of 46 functional enrichment tools using ORA and/or GSEA is shown in Table 1.1. The general features associated with each tool, such as the statistical methods employed, source of annotations, scope of analysis and web links are also listed. Both ORA and GSEA rely heavily on existing functional annotations about genes. Therefore, the data structure, quality and comprehensiveness of the annotation resources used to define gene sets are crucial to the success of these methods. Section 1.3.1 provides an overview of some of the key annotation resources used in functional enrichment analysis. Sections 1.3.2 and 1.3.3 describe the statistical and methodological aspects of ORA and GSEA, respectively.

Table 1.1: Existing function enrichment tools

Software	Class	Methods	Annotations (gene sets)	Platform	References	URL
BayGO	ORA	Bayesian, Goodman and Kruskal's gamma factor	GO	Web, R	Vencio *et al.* 2006	http://blasto.iq.usp.br/~tkoide/BayGO/
BiNGO	ORA	Hypergeometric, binomial	GO	Java plug-in	Maere *et al.* 2005	http://www.psb.ugent.be/cbd/papers/BiNGO/
CLENCH	ORA	Hypergeometric, binomial, chi-squared	GO	Perl	Shah and Fedoroff 2004	http://www.stanford.edu/~nigam/cgi-bin/dokuwiki/doku.php?id=clench#clench
DAVID	ORA	Fisher's exact test (modified as EASE score)	Over 40 annotation categories, including GO, protein-protein interactions, protein domains, disease associations, pathways, sequence features, homologies	Web	Dennis *et al.* 2003	http://david.abcc.ncifcrf.gov/summary.jsp
EASE	ORA	Fisher's exact test (modified as EASE score)	GO; KEGG; chromosomal locations	Windows standalone	Hosack *et al.* 2003	http://david.abcc.ncifcrf.gov/content.jsp?file=/ease/ease1.htm&type=1
EasyGO	ORA	Hypergeometric, binomial, chi-squared	GO	Web	Zhou and Su 2007	http://bioinformatics.cau.edu.cn/easygo/
eGOn/GeneTools	ORA	Fisher's exact test	GO	Web	Beisvag *et al.* 2006	http://www.genetools.microarray.ntnu.no/egon/index.php
FuncAssociate	ORA	Fisher's exact test	GO	Web	Berriz *et al.* 2003	http://llama.med.harvard.edu/funcassociate/
FunSpec	ORA	Hypergeometric	GO; MIPS; SMART and Pfam domains	Web	Robinson *et al.* 2002	http://funspec.med.utoronto.ca
g:Profiler	ORA	Hypergeometric (support ranked list of genes)	GO; pathways; transcription factor binding sites; Reactome	Web	Reimand *et al.* 2007	http://biit.cs.ut.ee/gprofiler/
GeneCodis	ORA	Hypergeometric, chi-squared	GO; KEGG; InterPro motifs; microRNA; transcription factors; user-defined annotation	Web	Carmona-Saez *et al.* 2007; Nogales-Cadenas *et al.* 2009	http://genecodis.dacya.ucm.es/analysis/
GeneMerge	ORA	Hypergeometric	GO; MeSH; KEGG; chromosomal locations; RNAi phenotypes	Web, Perl standalone	Castillo-Davis and Hartl 2003	http://genemerge.cbcb.umd.edu/
GFINDer	ORA	Fisher's exact test, binomial, chi-squared	GO; pathways; protein families and domains; genetic disorders and phenotypes	Web	Masseroli *et al.* 2005	http://www.bioinformatics.polimi.it/GFINDer/
GO::TermFinder	ORA	Hypergeometric	GO	Perl standalone	Boyle *et al.* 2004	http://search.cpan.org/dist/GO-TermFinder/lib/GO/TermFinder.pm
GObar	ORA	Hypergeometric	GO	Web	Lee *et al.* 2005b	http://katahdin.cshl.org:9331/GO/GO.cgi
goCluster	ORA	Hypergeometric	GO	R/Bioconductor	Wrobel *et al.* 2005	http://www.biozentrum.unibas.ch/gocluster/
GOEAST	ORA	Hypergeometric, Fisher's exact, chi-squared	GO	Html	Zheng and Wang 2008	http://omicslab.genetics.ac.cn/GOEAST

(continued over the page)

Table 1.1: Existing function enrichment tools (continued)

Software	Class	Methods	Annotations (gene sets)	Platform	References	URL
GOLEM	ORA	Hypergeometric	GO	Web, Java standalone	Sealfon *et al.* 2006	http://function.princeton.edu/GOLEM/
GoMiner	ORA	Fisher's exact test	GO	Java standalone	Zeeberg *et al.* 2003	http://discover.nci.nih.gov/gominer/
GOStat	ORA	Fisher's exact, chi-squared	GO	Web	Beissbarth and Speed 2004	http://gostat.wehi.edu.au/cgi-bin/goStat.pl
Gostats	ORA	Classical and conditional hypergeometric tests	GO	R/Bioconductor	Falcon and Gentleman 2007	http://www.bioconductor.org/packages/bioc/1.8/html/GOstats.html
GOTM	ORA	Hypergeometric	GO	Web	Zhang *et al.* 2004	http://bioinfo.vanderbilt.edu/gotm/
GOToolBox	ORA	Hypergeometric, Fisher's exact, binomial	GO	Web	Martin *et al.* 2004	http://burgundy.cmmt.ubc.ca/GOToolBox/
L2L	ORA	Binomial	GO; Reactome protein-protein interactions, cancer gene modules, L2L microarray datasets	Web, Perl application	Newman and Weiner 2005	http://depts.washington.edu/l2l/
MAPPFinder	ORA	Hypergeometric *Z*-score	GO; pathways from GenMAPP	Java standalone	Doniger *et al.* 2003	http://www.genmapp.org/
Onto-express	ORA	Hypergeometric, binomial, chi-squared, Fisher's exact test	GO	Web	Draghici *et al.* 2003; Khatri *et al.* 2002	http://vortex.cs.wayne.edu/projects.htm#Onto-Express
PageMan	ORA	Fisher's exact, chi-squared	GO; MIPS, KEGG	Web, Java standalone	Usadel *et al.* 2006	http://mapman.mpimp-golm.mpg.de/general/ora/ora.html
ProbCD	ORA	Yule's Q, Goodman-Kruskal's gamma, Cramer's T	GO	Web, R	Vencio and Shmulevich 2007	http://xerad.systemsbiology.net/ProbCD
STEM	ORA	Hypergeometric	GO	Java standalone	Ernst and Bar-Joseph 2006	http://www.cs.cmu.edu/~jernst/stem/
THEA	ORA	Hypergeometric, binomial	GO	Java standalone	Pasquier *et al.* 2004	http://thea.unice.fr/index-en.html
ErmineJ	ORA, GSEA	Permutations, Wilcoxon rank sum test	GO	Java standalone	Lee *et al.* 2005a	http://www.bioinformatics.ubc.ca/ermineJ/index.html
FatiGO FatiScan Babelomics	ORA, GSEA	Fisher's exact test, segmentation test	GO, KEGG pathways, InterPro motifs, Swiss-Prot keywords, microRNA, transcription factor and cisRED *cis*-regulatory elements	Web	Al-Shahrour *et al.* 2004; Al-Shahrour *et al.* 2006	http://www.babelomics.org/
GeneTrail	ORA, GSEA	Hypergeometric, Kolmogorov-Smirnov statistic	GO; pathways from KEGG and TRANSPATH; transcription factor from TRANSFAC	Web	Backes *et al.* 2007	http://genetrail.bioinf.uni-sb.de/enrichment_analysis.php?js=1&cc=1

(continued over the page)

Table 1.1: Existing function enrichment tools (continued)

Software	Class	Methods	Annotations (gene sets)	Platform	References	URL
JProGO	ORA, GSEA	Fisher's exact, Kolmogorov–Smirnov test, Student's t-test, unpaired Wilcoxon's test	GO	Java standalone	Scheer *et al.* 2006	http://www.jprogo.de
Catmap	GSEA	Permutations, Wilcoxon rank sum test	GO	Perl standalone	Breslin *et al.* 2004	http://bioinfo.thep.lu.se/catmap.html
FIVA	GSEA	Fisher's exact test	GO, metabolic pathways, COG classes, regulatory interactions, UniProt keywords, InterPro domains	Java standalone	Blom *et al.* 2007	http://bioinformatics.biol.rug.nl/standalone/fiva/
GAzer	GSEA	Permutations, *Z*-statistic	GO; pathways; chromosomal locations; InterPro domains; cis-regulatory elements	Web	Kim *et al.* 2007	http://integromics.kobic.re.kr/GAzer/
globaltest	GSEA	Bayesian generalized linear model, sample permutations	GO; KEGG	R/Bioconductor package	Goeman *et al.* 2004	http://www.bioconductor.org/packages/2.4/bioc/html/globaltest.html
GOdist	GSEA	Two-sample Kolmogorov–Smirnov test	GO	MATLAB	Ben-Shaul *et al.* 2005	http://basalganglia.huji.ac.il/links.htm
GO-Mapper	GSEA	Gaussian distribution, expression quotient (EQ) score	GO	Perl standalone	Smid and Dorssers 2004	http://www.gatcplatform.nl/gomapper/index.php
GSA	GSEA	maxmean statistic, sample permutations	User-defined gene sets	R package	Efron and Tibshirani 2007	http://www-stat.stanford.edu/~tibs/GSA/
GSEA	GSEA	Kolmogorov–Smirnov statistic	GO; pathways from KEGG, BioCarta and GenMAPP; transcription factors; microRNA; cancer modules; MSigDB	Java standalone, R	Subramanian *et al.* 2005	http://www.broad.mit.edu/gsea/
iGA	GSEA	Permutations, hypergeometric, probability of change (PC) values	GO	Perl standalone	Breitling *et al.* 2004	Windows executable of iGA is available as additional file of the original publication
MetaGeneProfiler	GSEA	*Z*-score, inverse standard normal cumulative distribution function	GO	Web	Gupta *et al.* 2007	http://metagp.ism.ac.jp/
SAFE	GSEA	Local and global statistics, sample randomisations	GO; KEGG; Pfam domains	R/Bioconductor package	Barry *et al.* 2005	http://www.bioconductor.org/packages/release/bioc/html/safe.html
T-profiler	GSEA	*t*-test	GO; transcription factor binding motif; chromosomal location	Web	Boorsma *et al.* 2005	http://www.t-profiler.org/

1.3.1 Resources for functional enrichment analysis

Gene Ontology

The Gene Ontology (GO) project (http://www.geneontology.org/) (Ashburner *et al.* 2000) is a collaborative effort aimed at providing controlled vocabularies to describe gene products in the major model organisms. GO terms are organised into three broad areas of cell biology:

1. *Molecular function*, which describes the biochemical reactions that a protein catalyses.
2. *Biological process*, which describes the global physiological process that a protein is involved in.
3. *Cellular component*, which describes the compartment of a cell that a protein product is situated in.

As of 13 July 2009, there are 27,784 ontologies in GO, of which 16,774 are biological process terms, 2385 are cellular component terms, and 8625 are molecular function terms.

The GO ontology is organised hierarchically as a directed acyclic graph (DAG) in which the terms are nodes and the relationships among them are edges. The key characteristic of a DAG in the context of GO is the parent-child relationship between terms, with parent terms representing more general entities than their child terms, and a term can have multiple parents. So, if a gene is assigned a particular specific GO term, it will have the property associated with that term, and also inherits the properties of all the parent terms. For example, if a gene is known to be specifically involved in “glycolysis”, it will be annotated directly to that term, and implicitly annotated to its ancestor term “carbohydrate metabolic process”. All genes assigned the term “glycolysis” are by default involved in “carbohydrate metabolic process”. However, all “carbohydrate metabolic process” genes are not necessarily “glycolysis” genes.

The association between a gene and GO terms is established either by manual curation, or computationally through predictive methods (Rhee *et al.* 2008). Genes are associated with as many terms as appropriate, and with the most specific terms available to reflect what is currently known about a gene. So, depending on the current status of knowledge, GO annotations for a gene can be made as specific or general as necessary.

Pathway information databases

The most popular source of pathway information is provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) (Kanehisa and Goto 2000; Kanehisa *et al.* 2008). KEGG is a database of manually compiled pathway maps representing the knowledge on the metabolic pathways, chemical reactions that genes are involved in, and protein-protein interaction networks. KEGG pathways are divided into six broad categories: metabolism, genetic information processing, environmental information processing, cellular processes, human disease, and drug development. As of July 2009, there are 328 reference pathways in KEGG.

Other resources of biological pathways and reactions besides KEGG include GenMAPP (http://www.genmapp.org/) (Salomonis *et al.* 2007), Biocarta (http://www.biocarta.com), and Reactome (http://www.reactome.org) (Joshi-Tope *et al.* 2005).

The Molecular Signatures Database

The Molecular Signatures Database (MSigDB) (http://www.broadinstitute.org/gsea/msigdb/index.jsp) is a collection of gene sets for use with GSEA software (Subramanian *et al.* 2005). As of this writing, MSigDB contains 5452 gene sets, which are divided into five major collections:

1. *Positional gene sets.* This collection contains 386 gene sets corresponding to genes in the human chromosomes and cytogenetic bands.
2. *Curated gene sets.* This collection contains 1892 gene sets. These are genes whose products are involved in specific metabolic and signalling pathways reported in 12 manually curated pathway databases. This set also catalogs genes co-expressed in response to genetic or chemical perturbations as reported in literature.
3. *Motif gene sets.* This collection contains 837 gene sets representing genes sharing conserved regulatory motifs in the promoter regions of human genes.

4. *Computational gene sets.* This collection contains 883 gene sets, which were defined by mining large compendium of cancer-related microarray data.
5. *GO gene sets.* This collection contains 1454 gene sets, which were derived from the Gene Ontology terms.

Other annotation resources

In addition to the resources described above, a wide range of heterogeneous annotation data can be incorporated into the enrichment analysis framework to increase the comprehensiveness of the enrichment analysis results. As shown in Table 1.1, some recently developed enrichment tools or newly released early-generation tools have extended their backend annotation databases to include biological information coming from InterPro (http://www.ebi.ac.uk/interpro/) and Pfam (http://pfam.sanger.ac.uk/) protein domains; TRANSFAC (http://www.biobase-international.com/pages/index.php?id=transfac) transcription factors binding sites; OMIM (http://www.ncbi.nlm.nih.gov/omim/) gene-disease associations; and cisRED (http://www.cisred.org/) regulatory motifs.

1.3.2 Over-representation analysis

Given a list of differentially expressed genes pre-selected by the user, this method compares the number of differentially expressed genes found in a certain functional category of interest (gene set) with the number of genes expected to be found in that category just by chance. If the observed number is markedly different from that expected just by random chance, the category is considered as significant. Several statistical approaches can be used to calculate the probability of observing the actual number of genes just by change (p-value) for a given functional category. As can be seen from Table 1.1, the hypergeometric, binomial, chi-squared and Fisher's exact test are the most widely adopted models. The chi-squared test describes how the observed proportion of hits deviates from what is expected due to chance, but it only gives an approximate p-value and is restricted to situations where the number of observations of each type (e.g. differentially expressed genes that appear in the category) is greater than five. If there are fewer than five observations, alternative approaches that calculate the exact p-values, such as the Fisher's exact test and the hypergeometric test, can be used. On the other hand, the binomial distribution is only suitable for large arrays containing tens of thousands of transcripts (e.g. whole-genome arrays). In most cases, there will not be dramatic differences between the models (Khatri and Draghici 2005). A detail discussion of these tests can be found in Draghici (2003) and Rivals *et al.* (2007).

Approaches based on ORA continue to evolve and more sophisticated algorithms have been proposed to account for the dependencies among GO terms. Alexa *et al.* (2006) developed a conditional hypergeometric test that calculates the significance of a GO term based on its neighbourhood; they showed that integrating the DAG structure of the GO terms in testing for enrichment reduce the false positive rate and enhanced inference. Grossmann *et al.* (2007) measured the over-representation of a given GO term relative to its parent terms.

Existing ORA methods have been effective in adding value to expression results, but they remain limited for a number of reasons. First, its 'threshold-based' strategy means that only the genes that are pre-selected (based on their differential expression) are considered; the enrichment results therefore depend on the stringency of the cutoff used and the quality of the gene list produced, making ORA unstable to a certain degree. A second issue is that order of genes on the significant gene list is not taken into consideration, potentially leading to loss of information. These problems are addressed in the GSEA approach, as described below.

1.3.3 Gene set enrichment analysis

The goal of GSEA is to identify functional categories or gene sets that display modest but coordinated expression changes. GSEA is performed by first ranking all genes in the dataset based on the correlation between their expression and the given phenotypes. Then the rank positions of all members in a given gene set are identified. An enrichment score that reflects the difference between the observed rankings and that expected due to chance is calculated. After determining the enrichment score for each gene set across the phenotype, GSEA iteratively permutes the sample labels and re-evaluate the enrichment across the random classes. Statistical significance (*p*-value) of the enrichment score is established with respect to a background distribution constructed by permutations of the class labels. The key idea is to determine whether the members of a given gene set are randomly distributed throughout the ranked list or primarily found at the extremes (top or bottom) of the entire ranked list.

The various GSEA methods that have been developed in recent years use the same basic procedure described above, but vary by the test statistic use to compute the enrichment score. For examples, Mootha *et al.* (2003) and Subramanian *et al.* (2005) used an enrichment score based on the Kolmogorov-Smirnov statistic as the test statistic. Tian *et al.* (2005) and Kim and Volsky (2005) proposed a similar approach but instead of using the enrichment score, they used the two-sample *t*-statistic. Different extensions of GSEA have been proposed to make its application both simpler and richer. For examples, the SAFE procedure developed by Barry *et al.* (2005) extends GSEA to cover multiclass, continuous and survival phenotypes; Jiang and Gentleman (2007) extend GSEA to allow for covariate adjustments based on the use of linear modeling and posterior probabilities. Another interesting extension is the absolute enrichment score that accounts for gene sets with bi-directional changes (Efron and Tibshirani 2007; Saxena *et al.* 2006). In many homeostatic processes, when one component of the process is up-regulated, there is a controlling down-regulation in response and *vice versa* to maintain constancy of the system (Saxena *et al.* 2006). Such patterns are poorly captured by the standard formulation of GSEA because the effect of genes altered in the opposite directions in a gene set will cancel each other to make the enrichment score insignificant. By using absolute enrichment score (e.g. absolute signal-to-noise ratio) as the ranking metric, gene sets with bi-directional changes can be detected.

GSEA is complementary to the traditional ORA approach in several aspects: (i) it minimises the arbitrary factors in the typical gene selection step that could impact the conventional ORA approach; (ii) it adjusts for the correlation structure of gene sets; and (iii) it considers all information obtained from microarray experiments by allowing genes showing minimal expression changes to contribute to the enrichment analysis.

While the "threshold-free" strategy is the main advantage of GSEA, it is also becoming its limitation in some analyses. In many cases, the upstream data processing and comprehensive gene selection statistics simply cannot be duplicated by GSEA. For example, many clinical studies involve multiple factors and variants, such as ages, sex, drug treatment/control, disease/normal, *etc.* Under such complex situations, it can be difficult to summarise the effect of the various biological aspects into one meaningful score for use with the enrichment analysis. A second issue concerns the null hypothesis at work in the GSEA permutation. Many authors have discussed the differences between gene and sample randomisation in inferring the statistical significance of gene set scores (Goeman and Bühlmann 2007; Tian *et al.* 2005). Goeman and Bühlmann (2007) showed that sample randomisation is more appropriate than gene randomisation. However, sample randomisation requires a certain amount of sample replication to attain the desired levels of significance, and this condition often is not met in many studies (for example, time series analyses).

1.3.4 Other methods

More recently, a different approach called Signalling Pathway Impact Factor analysis (SPIF) was proposed for finding significant pathways in a dataset (Draghici *et al.* 2007; Tarca *et al.* 2009). Both ORA and GSEA, when used to mine pathway information, will consider only the set of genes on any given pathway and ignore their positions in those pathways. Therefore, these techniques will provide identical results as long as the pathway diagram involves the same genes, even if the interactions between these genes were completely redefined over time. In contrary to ORA and GSEA, SPIF incorporates the pathway topology into the analysis procedure, and calculates a global p-value for each pathway based on evidence obtained from the enrichment analysis (such as the fold-change of differentially expressed genes and the statistical significance of the set of pathway genes) and the actual perturbation on a given pathway.

1.4 Text mining and its applications in microarray data analysis

Almost every known or postulated piece of functional information about genes and their role in biological processes is encapsulated in the peer-reviewed published literature. This ever expanding knowledge base constitutes a valuable resource for interpreting genome-wide experiments. However, the volume of published literature is growing at an exponential pace, making it increasingly difficult (or impossible) for biologists to stay abreast of their own field of expertise, let alone keeping up-to-date with publications in other related disciplines. Therefore, the ability to efficiently access, retrieve and manage the information in this published literature is a necessary step towards the biological interpretation of any genome-wide experiment. Automated text mining systems are indispensable tools in this regards.

The primary goal of text mining is to retrieve relevant information that is hidden in text and to present the distilled knowledge to users in a concise form without compromising the integrity of published data. This effectively shifts the burden of "information overload" from the researcher to the computer. Since the literature covers all aspects of biological science, there is almost no limit to the types of information that may be discovered through careful mining. Currently, text mining is being applied to various aspects of biological research, including the identification of proteins and gene names, construction of putative gene networks, extraction of protein-protein interactions, and analysis of microarray data.

The next section describes the main resources for biological text mining. This is followed by an overview of the disciplines involved in text processing along with the techniques and methods they use. Finally, we will look at how textual information can be incorporated into data mining framework to aid the interpretation of gene expression analysis.

1.4.1 Resources for biological text mining

There are an increasing number of public resources from which the text of articles is readily available online for analysis. The primary source of textual information for biological text mining applications is the PubMed database (http://www.ncbi..nlm. nih.gov/pubmed/) developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). Currently, PubMed contains over 19 million citations for biomedical articles indexed from 1948 to present[2]. PubMed provides free access to MEDLINE (Medical Literature Analysis and Retrieval System Online; http://www.nlm.nih.gov/databases/databases_medline.html). MEDLINE is NLM's premier bibliographic database; it contains over 16 million references to journal articles in life sciences with a concentration on biomedicine.

Most of the text mining systems available today direct their focus on the analysis of the texts of abstracts from the PubMed/MEDLINE citations, because abstracts summarise the results of the scientific work in a concise form and are readily accessible. Each PubMed citation has a unique identifier (PMID) and can be conveniently searched using Entrez, a web-based search and retrieval system provided by NCBI (Wheeler *et al.* 2008). Entrez enhances the keyword searches by translating the user query to Medical Subject Heading (MeSH) terms (Nelson *et al.* 2004). MeSH is a hierarchical set of controlled vocabulary terms assigned by expert curators who attempt to summarise the information presented in each indexed article and the genes described therein. In addition to the regular web query interface, PubMed also offers a programmatic access to its content through the Entrez Programming Utilities (eUtils) [3], which can be used alongside other popular open source projects, such as the BioPerl and BioJava integrated libraries, to build customised data pipelines for text mining applications.

Article abstracts contain short descriptions that highlight the most relevant aspects of a given article; however, they only cover a small fraction of the information contained in full-text articles (Schuemie *et al.* 2004). Therefore interests have begun to shift from PubMed/MEDLINE abstracts to full texts. Several efforts have been undertaken to build literature databases providing access to full texts. Such open access initiatives include PubMed central (PMC) (http://www.pubmedcentral.nih.gov/) and HighWire Press (http://highwire.stanford.edu/). PMC provides free online access to full-text

[2] At the time of this writing (July 2009), PubMed contains 19,033,990 entries.

[3] http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

articles deposited by voluntary publishers from the life sciences and biomedical journals. Since every article in PMC has a corresponding entry in PubMed, articles archieved in PMC are therefore also accessible via PubMed. HighWire Press is a complementary resource to PubMed for accessing peer-reviewed science articles, and provides a search interface to over 1.9 million free full-text articles[4].

Although a significant part of the current text mining efforts focuses on the analysis of biomedical and scientific literature, as we will see in Section 1.4.2 the use of domain-specific terminologies such as GO, MeSH and UMLS (Unified Medical Language System) (Bodenreider 2004)[5] are sometimes required to support tasks such as entity recognition (e.g. the identification of gene and protein names in text) and relation extraction (e.g. the identification of relationships among genes or proteins). Many of these lexical and ontological resources are highly specialised, and to provide an exhaustive list of these resources is beyond the scope of this thesis. For a detailed discussion of the roles and applications of these resources in biomedical text mining, see Bodenreider (2006).

1.4.2 Text mining in modern biology

Text mining is a modular process involving different computational and linguistic disciplines; the main building blocks are information retrieval (IR), information extraction (IE) and knowledge discovery. The first step in a text mining process is usually the identification of relevant documents (IR). Once documents are collected for computational analysis, an IE module is used to extract specific types of information or entities of interest from the texts. Then the information extracted is subjected for data mining to discover new knowledge.

[4] As of July 2009, HighWire Press provides access to 1,920,292 free full-text articles and 6,059,906 total articles.

[5] The Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/) is the National Library of Medicine's biological ontology. It contains information about many aspects of the biomedical domain, such as diseases, tissues and drugs.

Information retrieval

Information retrieval (IR) is the activity of identifying documents (these can be full-text articles or abstracts) that are relevant to a certain topic (for example, all the articles mentioned a certain disease) within a very large set of documents. Many IR tools have been developed specifically to query the biomedical literature databases, of which the best-known system is PubMed (see Section 1.4.1 for more details on the PubMed database). PubMed uses Boolean query searches based on index look-up techniques, and a document-similarity search technique based on the vector-space model (Jensen *et al.* 2006). The Boolean model allows the user to retrieve all documents that contain certain combinations of query terms (for example, "cancer" and "p53"). In the vector-space model, each document is represented as a vector of weighted terms. The weight is a function of the frequencies of the term in the document and in the whole corpus. Document-similarity is calculated by comparing these document vectors to each other; this strategy is used in PubMed to search for related articles (Wilbur and Coffee 1994).

In addition to PubMed, many advanced IR tools have been developed to support large-scale biomedical analysis, such as for explaining large-scale relationships among genes or other biological entities. Such systems include iHOP (http://www.ihop-net.org/UniPub/iHOP/) (Hoffmann and Valencia 2004) and Textpresso (http:// www.textpresso.org/) (Muller *et al.* 2004). iHOP links the interacting proteins to their corresponding PubMed records and allows navigation in the resulting network of protein interactions, whereas Textpresso uses a custom ontology to search a collection of documents for information on specific classes of biological concepts (e.g. gene, allele, phenotype) and their relations (e.g. association, regulation).

Information extraction

Information extraction (IE) techniques are used to identify relevant phrases and pre-defined types of facts in texts. Early efforts in biomedical IE were focused on two areas: (i) name entity recognition (NER), which involves the recognition of terms denoting specific classes of biological entities; and (ii) the extraction of specific relationships between such entities. Biological entities can be genes, proteins, chemicals, diseases or other pre-defined biological concepts. The identification of these entities in text is an important first step in many IE systems.

The main hurdle in biomedical NER is the lack of standardisation of names, especially in the case of genes and proteins. A variety of alternative names and abbreviations that refer to the same gene or protein are often encountered. For example, the official symbol for 'cyclin-dependent kinase inhibitor 1A' is 'CDKN1A'; but it is also referred to in the literature as 'p21', 'CIP1' or 'WAF1'. Moreover, some gene and protein symbols are ambiguous and might be confused with common English words (for example, 'lush' in *Drosophila* that mediates responses to alcohol) or other biological terms (for example, the abbreviation 'APC' might correspond to 'adenomatous polyposis coli' or 'anaphase promoting complex').

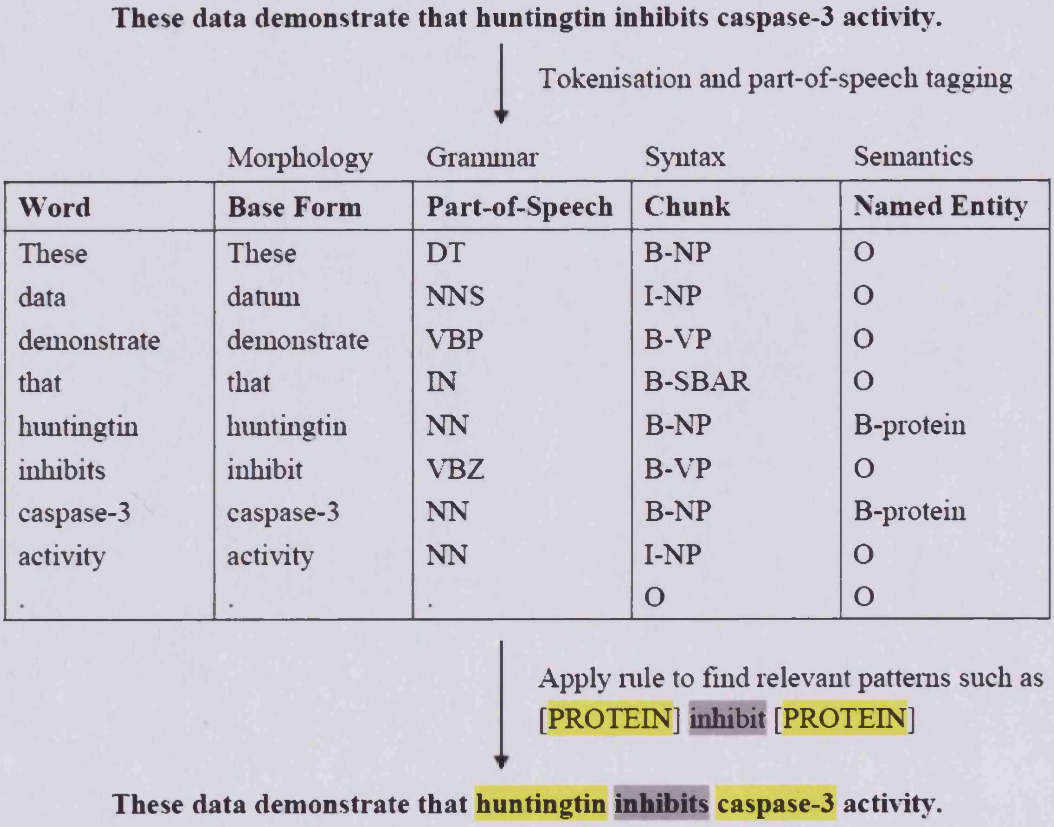
A range of different approaches to gene and protein names recognition and disambiguation have been developed. These methods fall into four general categories:

1. Rule-based approaches, which use some combination of regular expressions and patterns to match name entities in the literature (Fukuda *et al.* 1998).
2. Dictionary-based approaches, which use dictionaries to identify name entities in the literature (Krauthammer *et al.* 2000). The dictionaries are usually based on a publicly available source of standardised, structured data curated by human experts, such as the HUGO gene nomenclature (Wain *et al.* 2002) or UMLS (Unified Medical Language System) (Bodenreider 2004).
3. Machine learning-based approaches, which employ machine learning techniques, such as hidden Markov models (HMMs), support vector machines (SVMs), Bayesian learning and decision trees, to develop statistical models for gene and protein name recognition (Hatzivassiloglou *et al.* 2001).
4. Hybrid approaches, which combine two or more of the above approaches (Proux *et al.* 1998).

The next step after identification and disambiguation of biological entities is the extraction of relationships between those entities. Two fundamentally different approaches are currently being used for this task: the first approach is based on co-occurrence, while the second approach relies on natural language processing (NLP).

Co-occurrence-based approaches assume that there are some forms of biological relationships between entities that occur in the same abstracts or sentences. It is possible that two unrelated entities were mentioned together simply due to chance, therefore most systems employ a frequency-based scoring scheme to rank the extracted relationship (Hoffmann and Valencia 2004; Jenssen *et al.* 2001). The underlying rationale is that if two entities are repeatedly mentioned together, then it is likely that they are related. Co-occurrence-based approaches can be used to extract relationships of almost any type. However, this approach has difficulty in detecting directional relationships. Consider an example sentence: "These data demonstrate that huntingtin inhibits caspase-3 activity". A co-occurrence-based approach will not be able to tell whether huntingtin inhibits caspase-3 or *vice versa*. In addition, it is unable to distinguish between direct and indirect relationships between the entities in complex sentences that contain multiple entities. These problems can, however, be resolved by using natural language processing (NLP), which involves the analysis of syntax (the order in which words are put together to form phrases and sentences) and semantics (the meaning that is implied by words and sentences).

The NLP techniques used to extract information from text is illustrated in Figure 1.5. The first step is tokenisation, which breaks the text up into basic textual units called 'tokens'. This is followed by an optional stemming step that reduce the tokens to their base form (for example, 'inhibits' becomes 'inhibit'). Then the individual words (or tokens) are tagged with their part-of-speech, which are a set of word categories based on the role that words may play in the sentence in which they appear (such as noun, verb or adjective). On the basis of their part-of-speech tags, a syntax tree can be derived for each sentence to delineate noun phrases and represent their inter-relationships. Syntactically related words are subsequently grouped together; dictionaries and ontologies are then used to semantically tag the relevant biological entities (e.g. genes and proteins). Finally, a rule set is applied to identify relationships based on the syntax tree and the semantic tags.



These data demonstrate that huntingtin inhibits caspase-3 activity.

Tokenisation and part-of-speech tagging

Word	**Base Form**	**Part-of-Speech**	**Chunk**	**Named Entity**
These	These	DT	B-NP	O
data	datum	NNS	I-NP	O
demonstrate	demonstrate	VBP	B-VP	O
that	that	IN	B-SBAR	O
huntingtin	huntingtin	NN	B-NP	B-protein
inhibits	inhibit	VBZ	B-VP	O
caspase-3	caspase-3	NN	B-NP	B-protein
activity	activity	NN	I-NP	O
.	.	.	O	O

Apply rule to find relevant patterns such as
[PROTEIN] inhibit [PROTEIN]

These data demonstrate that huntingtin inhibits caspase-3 activity.

Figure 1.5: Information extraction using natural language processing techniques
The different natural language processing (NLP) layers, from tokenisation to semantics, for a given example sentence "These data demonstrate that huntingtin inhibits caspase-3 activity." are shown here. This sentence was taken from PMID: 17124493. First, the text was tokenised to produce individual words, which were subsequently stemmed to their base form. The stemmed words were tagged with their part-of-speech (i.e. grammatical tags), such as noun and verb. Dictionaries are then used to semantically tag the relevant biological entities (i.e. genes and proteins). In the last step, a rule set is applied to identify relationships on the basis of the syntax tree and the semantic labels. This example was based on the analysis performed with the GENIA tagger at http://text0.mib.man.ac.uk/software/geniatagger/. DT, determiner; IN, preposition; NN, Noun (singular); NNS, Noun (plural); VBP, Verb (present tense); VBZ, Verb (third person singular present). The B/I/O terminology refers to begin phrase (B), internal to phrase (I), or outside of phrase (O).

NLP-based systems are able to provide information on multiple types of associations, but suffer the limitation that a large number of rules are required to cover the many slightly different ways of expressing a particular relationship. These rules can either be manually crafted or learned automatically from a manually tagged corpus such as the GENIA corpus (Kim *et al.* 2003). Either way, it is a labour-intensive task.

So far, IE techniques have been used to extract various types of information from literature including protein-protein interactions (Thomas *et al.* 2000), drug-protein interactions (Rindflesch *et al.* 2000), information on gene regulation and protein phosphorylation (Saric *et al.* 2006).

Knowledge discovery

While information extraction focuses on extracting relationships between biological entities explicitly stated in the text, knowledge discovery attempts to uncover 'hidden' and previously unrecognised associations between these entities. The goal is to combine facts and information extracted from multiple publications to infer novel, indirect relationships worthy of further investigation. Most of the work in knowledge discovery follows the framework initiated by Swanson in the mid 1980s. Swanson proposed a simple model for detecting complementary structure in disjoint literatures, which state that "if A influences B, and B influences C, then A may influence C". Based on this model, Swanson found a connection between fish oil and Raynaud's disease (Swanson 1986). This literature-based hypothesis was later corroborated experimentally and clinically (Smalheiser and Swanson 1998). Since Swanson's pioneering work, this kind of knowledge discovery work has attracted the attention of other researches (Lindsay and Gordon 1999; Srinivasan 2004; Weeber *et al.* 2001).

1.4.3 Integrated mining of text and other non-text biological data

During recent years, several integrated text mining frameworks have been developed to perform synergistic mining of biological literature with other non-text data, in an attempt to facilitate the discovery of new knowledge. For examples, MacCallum *et al.* (2000) used text mining to assist in the identification of remote homolog proteins by combining sequence-similarity scores with document-similarity scores; Stapley *et al.* (2002) used a machine learning approach (support vector machines) to predict the sub-cellular location of yeast proteins based on both sequence information about the proteins and term frequencies in their associated MEDLINE abstracts; Krauthammer *et al.* (2004) integrated literature-based protein networks with genetic linkage-mapping studies to find candidate genes for Alzheimer's disease.

Most efforts to integrate the literature with biological data have so far been directed towards the annotation of data that has been obtained from high-throughput functional genomics studies such as microarrays. Given the diversity of genes involved and the amount of data generated by such studies, manually inspecting the literature for relevant information is equivalent to "attempting to drink from a fire hose". In this respect, text mining can be employed as an integrated knowledge source to help guide the mining of gene expression data, as described in the following section.

1.4.4 Enhancing gene expression data analysis with literature knowledge

During the last few years, there has been a surge of interest in leveraging the valuable information from the literature in the data mining of microarray experiments. This has lead to the development of various integrated text mining systems for the biological interpretation of gene expression data (see Table 1.2 for examples). An early such system was MedMiner (Tanabe *et al.* 1999), which was designed to perform automatic literature searches on large number of genes found to be of significance in a microarray experiment. For each gene in the query gene list, MedMiner creates a Boolean search based on user-defined combinations of keywords, and retrieves citations of matching articles. This system offers a useful aid for searching information about a few genes at a time, but it does not address the need for finding links and functional relationships among genes.

More advanced and sophisticated approaches have been developed to address this problem. Most of them have focused on the identification of explicitly stated or

Table 1.2: Text mining tools for microarray analysis

Software	Description/Methods	References	URL
Anni 2.0	For each gene in a gene list, a concept profile is created based on the context in which the gene is mentioned in MEDLINE literature. Then genes associated with similar topics in the literature are identified by hierarchical clustering of the corresponding gene concept profiles. Several classes of biomedical concepts are used, including genes, drugs and diseases.	Jelier *et al.* 2007	http://www.biosemantics.org/index.php?page=anni-2-0
Chilibot	Chilibot searches PubMed abstracts about specific relationships between proteins, genes, or keywords. Basic natural language processing techniques are used to identify sentences that describe stimulatory, inhibitory, and other relationships between pairs of genes.	Chen and Sharp 2004	http://www.chilibot.net/
CoPub	CoPub calculates keyword over-representation for a list of genes using the Fisher's exact test. The keywords used in CoPub were generated by searching the MEDLINE abstracts with biological concepts from different thesauri encompassing gene names, GO terms, pathways, diseases, drugs, *etc.*	Frijters *et al.* 2008	http://services.nbic.nl/cgi-bin/copub/CoPub.pl
CoPub Mapper	CoPub Mapper cluster genes and keywords extracted from MEDLINE abstracts based on their similarity in co-publications.	Alako *et al.* 2005	http://copub.gatcplatform.nl/
GEISHA	GEISHA identify keywords significantly associated with cluster of similarly-expressed genes by comparing the frequencies of the words present in the associated MEDLINE abstracts.	Blaschke *et al.* 2001	http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha
MILANO	MILANO performs automatic searches in PubMed and the GeneRIF for texts containing co-occurrences of search terms with a list of genes.	Rubinstein and Simon 2005	http://simon4.md.huji.ac.il/
PubGene	PubGene constructs functional association network for a group of genes based on their co-occurrence in the titles and abstracts of PubMed articles.	Jenssen *et al.* 2001	http://www.pubgene.org/
TXTGate	TXTGate performs gene-based text profiling and clustering using the information extracted from MEDLINE abstracts.	Glenisson *et al.* 2004	http://tomcat.esat.kuleuven.be/txtgate/

indirect associations between genes and other biomedical concepts, based on keyword co-occurrences in the associated abstract texts. For examples, Shatkay *et al.* (2000) used a probabilistic algorithm to find PubMed literature most relevant to each gene in a gene cluster, and then generated a list of keywords summarising the recurring theme in each set of the retrieved literature. Genes were associated with each other if their corresponding gene-by-article representations were similar. Jenssen *et al.* (2001) constructed a gene-abstract index and then used it to identify possible functional associations between genes on the basis of the co-occurrence of gene names in the titles and abstracts of PubMed articles. As opposed to Jenssen's gene-gene co-occurrence approach, Chaussabel and Sher (2002) analysed the gene-term co-occurrences in indexed abstracts and demonstrated that such approach can produce a coherent picture of the functional relationships among genes. Their method involved generating a list of co-occurring keywords for each gene in the gene list, and then clustering the genes based on the keyword co-occurrences. Glenisson *et al.* (2004) used the vector space model to cluster a list of genes into functional categories on the basis of textual information extracted from MEDLINE abstracts. To obtain different views on the associations of a gene, they used concepts from five different thesauri as features, and identified terms in abstract texts referring to these thesaurus concepts.

An alternative to co-occurrence-based methods is to identify significantly over-represented keywords within the texts. An illustrative example is the GEISHA system developed by Blaschke *et al.* (2001), which analyses the statistical properties of words present in the underlying literature corpus and determines keywords or biological terms significantly associated with groups of genes exhibiting similar expression patterns. Blaschke *et al.* grouped genes according to the similarity of their expression patterns, and then MEDLINE abstracts that mentioned at least one gene in the cluster were collected to generate an associated literature cluster for each gene cluster. The words in the various literature clusters were subjected to a series of text analyses to account for morphological variations and composite word terms. A *Z*-score was calculated for each term in each gene cluster by comparing the frequency of the term in the cluster with the frequency of the term in the other clusters.

A new perspective on the problem has been adopted by Raychaudhuri *et al.* (2002) who used the literature to evaluate the functional coherence of gene clusters. Specifically, they developed a computational method called the neighbor divergence per gene (NDPG), which calculates a numerical score that indicates how functionally coherent a group of genes is from the perspective of the published literature, such that groups of genes with shared function will receive a high score. In subsequent work Raychaudhuri *et al.* (2003) showed that NDPG can be used to determine which level of the tree to cut during hierarchical clustering to produce biologically relevant cluster boundaries.

1.5 Summary

Technological advancements during the past decades have revolutionised genomic research. The combined abundance of genes discovered by genome sequencing projects, and the literature discussing them, represents a major bottleneck for interpreting genome-wide experiments such as microarrays. The current challenge lies in converting this voluminous data and information, which is stored in both structured and unstructured annotation resources (such as GO, KEGG, PubMed), into useful biological knowledge. Various functional enrichment tools have been developed to aid the biological interpretation of microarray data. As we have seen from the survey presented in Table 1.1, almost all are limited to the mining of pre-defined, controlled vocabularies, and do not fully exploit the wealth of biological knowledge about genes in the scientific literature. Examples described in Section 1.4.4 suggest that free-text can be used as a potentially more informative knowledge base for interpreting gene expression data. This shows that there is a need to expand the existing ORA framework beyond the mining of pre-defined functional annotations to use free-text. Explorations into this approach are presented in subsequent Chapters.

1.6 Thesis layout

The remainder of this thesis is structured as follows:

Chapter 2, *Data acquisition and text processing*, describes the datasets used to evaluate the performance of the statistical methods presented in this thesis. This Chapter also details the text processing steps and how the unstructured textual data was converted into the appropriate numerical format expected by the data mining algorithms.

Chapter 3, *Classical hypergeometric distribution-based ORA*, reports initial explorations regarding whether the classical hypergeometric distribution-based ORA framework can be expanded to mine text-based information.

Chapter 4, *Exploration of factors contributing to annotation bias, and its effect on ORA*. This Chapter describes an unexpected feature of gene lists derived from a typical microarray experiment, which is that they tend to have a greater level of associated PubMed articles than would be expected by chance. This bias is a major problem for the classical hypergeometric test-based ORA approach described in Chapter 3, as it leads to many common and non-specific terms to appear significantly enriched within the gene list. The potential causes of annotation bias are investigated in this Chapter, with particular reference to gene age and the historical development of scientific research activity.

Chapter 5, *ORA based on the use of a permutation test*, describes a permutation test-based approach for overcoming annotation bias. The strengths and limitations of this approach are discussed.

Chapter 6, *ORA based upon the detection of outliers*, describes an outlier detection-based approach for identifying terms in PubMed abstracts that are significantly over-represented in a list of differentially regulated genes.

Chapter 7, *Extended hypergeometric distribution-based ORA*, presents a statistical framework that uses the extended hypergeometric distribution to model token frequency data associated with a list of differentially expressed genes, and to identify terms that are significantly over-represented.

Chapter 8, *Performance properties of OutlierDM and ExtendedHG.* In this Chapter, the performance of the outlier detection-based method (OutlierDM) and the extended hypergeometric distribution-based method (ExtendedHG) are compared to existing publicly available literature- and ontology-based approaches. Issues concerning the feasibility of extending the proposed methods to other organisms outside human are described.

Chapter 9, *PAKORA: a web application for interpreting microarray gene lists using text mining.* This Chapter describes the implementation of a graphical user interface for accessing the text-based ORA algorithms developed in this work.

Chapter 10, *Discussion and conclusions*, reviews and discusses the main issues in this work and provides indications for future research.

Chapter 2

Data acquisition and text processing

2.1 Introduction

The main objective of this project was to develop algorithms to access and utilise the very large amount of information that is available as free-text in the published biomedical literature, in order to assist in the biological interpretation of lists of differentially expressed genes generated as the output from gene expression microarray experiments. The application of statistical algorithms to biomedical literature for enrichment analysis is part of a more general procedure, which involves three main phases shown in Figure 2.1 and described below:

1. Phase I involves the retrieval of relevant PubMed articles, followed by the processing and conversion of the unstructured textual information into a more computer readable format.
2. Phase II involves the development of statistical algorithms for performing text-based over-representation analysis (ORA).
3. Phase III involves the interpretation of results produced by the statistical algorithms developed in Phase II.

To evaluate the performance of the statistical methods presented in this thesis, a diverse range of microarray datasets were used. The characteristics and sources of these datasets are detailed in Section 2.2 of this Chapter. In Section 2.3, the protocols for document retrieval and text processing are provided. Section 2.4 explains how numerical values were extracted from the text-based data associated with a gene list that could then be passed to the statistical algorithms developed in this work (Chapters 3, 5, 6 and 7).

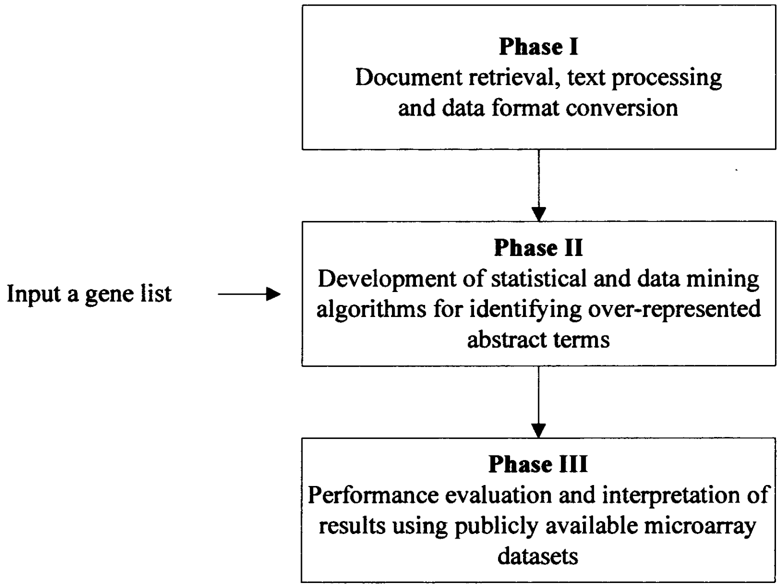

Phase I
Document retrieval, text processing and data format conversion
Input a gene list
Phase II
Development of statistical and data mining algorithms for identifying over-represented abstract terms
Phase III
Performance evaluation and interpretation of results using publicly available microarray datasets

Figure 2.1: The three phases of the proposed text-based ORA framework

2.2 Public datasets

When developing a text mining algorithm, it is important to evaluate its potential merit by comparing its performance to that of related methods by reference to one or more benchmarks. However, for exploratory procedure such as text-based ORA, there is currently a lack of "gold standard", which is dataset or benchmark for which "ground truth" is known. Therefore a hybrid approach was undertaken in which the principle results presented in this thesis are complemented by biological discussions. The datasets used to demonstrate the distinct features and utility of the text-based ORA methods presented in this thesis are described below.

2.2.1 ISG gene list

The main dataset used in this work is a list of interferon (IFN)-stimulated genes for which a substantial literature on the topic is available. This gene list is referred to as the ISG (interferon-stimulated genes) gene list throughout this thesis. The ISG gene list was used as the primary testbed because it constitutes a relatively simple and well-studied system of transcriptional regulation with well-known transcriptional targets.

The ISG gene list used in this thesis was derived from the gene expression study of Sanda *et al.* (2006). The authors examined the genes induced by type I and type II interferons in A549 lung cells at 6h and 24h following treatment. The experiment was conducted using Affymetrix HG-U133A GeneChip® arrays, with four replicates per condition. For the purpose of this study, only the effect of type I interferon was considered. The MAS5 expression data reported by the authors were downloaded from the Gene Expression Omnibus (GEO) database (accession number GSE5542) and analysed with the regularised *t*-test method (Baldi and Long 2001). Separate gene-level analyses were performed for each time period, i.e. IFN-treatment versus control at 6h, IFN-treatment versus control at 24h. Using the false discovery rate (FDR) method to correct for multiple testing, 194 and 118 probesets were found to be significantly differentially expressed (FDR p-value ≤ 0.05) at 6h and 24h, respectively. Then, a set of 106 probesets that were called significant at both time points were identified, which corresponds to a final gene list consisting of 78 unique genes.

2.2.2 Mitosis gene list

This gene list was extracted from Supplementary Table 3-1 reported in Lee *et al.* (2004). It contains 134 probesets on the Affymetrix HG-U133A array, representing 82 different genes that showed more than three-fold induction in the intrathymic T progenitor (ITTP) and double positive (DP) thymocytes subpopulations than the subpopulations of SP4 thymocytes and CB4 and AB4 T cells. Lee *et al.* concluded that these genes were predominantly involved in mitosis, cell cycle regulation and progression, DNA replication, recombination, or repair. For convenience this gene list is referred to below as the "mitosis gene list", although recognising that the range of biological functions included is somewhat broader than this.

2.2.3 Glycolysis gene list

This gene list was reported in Vanharanta *et al.* (2006). It contains 179 probesets on the Affymetrix HG-U133A array, representing 147 genes that were up-regulated in fumarate hydratase mutant relative to wild-type fibroids. According to the conclusions drawn by the authors, this list includes genes involved in carbohydrate metabolism, in particular glycolysis. Therefore, this gene list is referred to as the glycolysis gene list throughout this thesis.

2.2.4 Nishimura gene list

This gene list was as reported in Nishimura *et al.* (2003). It contains 685 probesets on the Affymetrix Arabidopsis Ath1 array, representing 679 different genes that were differentially expressed in *pmr4* mutant relative to wild type plants.

2.2.5 Literature gene lists

This is a collection of 402 gene lists that were collated from a wide range of published microarray experiments. These gene lists were used to study various aspects of the text-based ORA methods that cannot be reflected by using a single dataset alone, such as assessing their performance across different species. The 402 literature-derived gene lists are based on:

- 170 scientific papers
- 8 model organisms: human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, *Xenopus* and zebrafish.
- 10 Affymetrix arrays: HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Xenopus laevis, C. elegans and Zebrafish.

The 10 Affymetrix platforms were selected for their gene coverage and popularity, with an emphasis on those arrays with the highest gene coverage and the most substantial numbers of available public datasets for exploration. A comparison of the comprehensiveness and popularity of these and several other major Affymetrix arrays is shown in Tablė 2.1. The number of probesets represented on a particular chip type was used as a measure of comprehensiveness, and the number of GEO series records[1] available for a chip type is used as an approximate measure of popularity.

Table 2.1: A comparison of gene coverage and popularity for several major Affymetrix GeneChip® arrays

Species	Array name	# Probesets	# GEO series
Human	HG-Focus	8793	40
	HG-U95A	12626	266
	HG-U133A	**22283**	**589**
	HG-U133B	22645	99
	HG-U133Av2	22277	82
	HG-U133 Plus 2.0	**54675**	**580**
Mouse	Mu11K-A	6584	26
	Mu11K-B	6595	24
	MG-U74A	12654	21
	MG-U74Av2	12488	419
	MG-U74Bv2	12477	66
	MG-U74Cv2	11934	56
	MG-U430A	22690	336
	MG-U430B	22575	73
	MG-U430 2.0	**45101**	**647**
Rat	RN-U34	1322	7
	RG-U34A	8799	135
	RG-U34B	8791	10
	RG-U34C	8789	10
	RAE230A	15923	96
	RAE230B	15333	10
	RAT230 2.0	**31099**	**128**
Arabidopsis	Ag	8297	19
	Ath1	**22810**	**312**
Drosophila	**DrosGenome1**	**14010**	**88**
	Drosophila 2.0	**18952**	**44**
C. elegans	**Celegans**	**22625**	**23**
Xenopus laevis	**Xenopus laevis**	**15611**	**17**
Zebrafish	**Zebrafish**	**15617**	**26**

# Probesets refers to the number of Affymetrix probeset identifiers on a chip; # GEO Series refers to the number of series records submitted to the GEO database for a specific chip type (correct as of 29 September 2008). Highlighted in grey are chips for which literature gene lists were collected.

[1] A GEO series record is an original submitter-supplied record that describes a microarray experiment.

Scientific publications relevant to each of the 10 selected chip types were obtained by using the "Scientific Publication" search tool provided by Affymetrix[2]. The number of individual gene lists reported per publication varies. For each chip type, up to twenty (if possible) different publications were identified, and from which lists of differentially expressed genes reported in the form of Affymetrix probeset identifiers were extracted. The numbers of gene lists and papers collected are summarised in Table 2.2. For the Drosophila 2.0, Xenopus laevis, C. elegans and Zebrafish arrays, there are less than 20 papers for which gene lists are readily extractable.

Details of all 402 gene lists, including their size and the PubMed articles from which they were extracted, can be found in Appendix A. The mitosis, glycolysis and Nishimura gene lists described in the previous sections are also part of the literature gene lists collection and were assigned the identifiers 'hs2c', 'hs6b' and 'ath1', respectively.

Table 2.2: Number of literature gene lists collected for each chip type

Array name	# Publication	# Gene list
HG-U133A	20	52
HG-U133 Plus 2.0	20	54
MG-U430 2.0	20	40
RAT230 2.0	20	45
Ath1	20	67
DrosGenome1	20	44
Drosophila 2.0	14	29
Celegans	14	28
Xenopus laevis	8	18
Zebrafish	14	25

2.3 Text processing

Texts in abstracts are by their nature almost completely unstructured. However, the text-based ORA algorithms proposed in this thesis require structured data, i.e. in numerical format that describes the number of times a term occurs in a gene list and in

[2] http://www.affymetrix.com/publications/index.affx

the background. Therefore it is necessary to transform the unstructured textual data into an appropriate numerical format that could be fed into the data mining algorithms. A first step into this approach involved the creation of a text corpus that connects the relevant PubMed abstracts with genes included in the microarray analysis. These abstracts were then processed through a successive number of steps and re-structured to generate a list of unique tokens for mining, as detailed below.

2.3.1 PubMed articles retrieval

The first step is to map all the genes represented on an array to their corresponding Entrez Gene identifiers (EGID). For all arrays listed in Table 2.2 except Ath1, the mapping schemes provided by the appropriate Bioconductor metadata packages were used. The data versions of these metadata are listed in Table 2.3. A mapping scheme is not available for the Ath1 array at the time this research was carried out, and thus the associations between its probesets and genes were extracted from the annotation file 'ATH1-121501.na23.annot.csv' provided by Affymetrix[3] instead. Once the identities of genes represented on an array platform have been identified, PubMed articles associated with these genes were obtained from the 'gene2pubmed' file from NCBI[4] in the form of EGID to PubMed identifier (PMID) mappings. The relevant citations were then retrieved from the PubMed database using a Perl script that implements modules from the Entrez Programming Utilities (E-Utilities)[5].

[3] http://www.affymetrix.com/support/technical/byproduct.affx?product=arab; time stamp: 12 Jul 2007.
[4] ftp://ftp.ncbi.nih.gov/gene/DATA; time stamp: 25 Oct 2007.
[5] http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

Table 2.3: Bioconductor packages used to convert Affymetrix probeset ID to EGID

Array name	Package name	Version
HG-U133A	hgu133a	2.0.0
HG-U133 Plus 2.0	hgu133plus2	2.0.0
MG-U430 2.0	mouse4302	2.0.0
RAT230 2.0	rat2302	2.0.0
Ath1	N/A	N/A
DrosGenome1	drosgenome1	2.0.0
Drosophila 2.0	drosophila2	2.0.0
Celegans	celegans	2.0.0
Xenopus laevis	xenopuslaevis	2.0.0
Zebrafish	zebrafish	2.0.0

2.3.2 Construction of a text corpus formed by unique PMIDs

Upon retrieval, only the text present in the abstracts was extracted from the citation records; PMIDs that had only a title and no abstract were discarded. PMIDs that are associated with more than one EGID were also omitted from the corpus; based on manual inspection, these tend to be large-scale sequencing, nomenclature, or protein family characterisation reports. As can be seen from the examples shown in Table 2.4, such abstracts contain little explicit information about gene functions, and this lack of specificity might adversely affect the performance of the text mining algorithms. In an attempt to maximise the number of abstracts that deal specifically with the biological role of a given gene, only PMIDs that cross-reference to one EGID were retained. The number of PMIDs passing this filtering criterion is listed in Table 2.5, and the resulting text corpus will be referred to as the "Unique PMID corpus" hereafter.

Table 2.4: Examples of non-specific PMIDs and their abstract text

PMID: 14702039 Number of associated EGID: 5353	PMID: 16260502 Number of associated EGID: 8
Complete sequencing and characterization of 21,243 full-length human cDNAs. As a base for human transcriptome and functional genomics, we created the "full-length long Japan" (FLJ) collection of sequenced human cDNAs. We determined the entire sequence of 21,243 selected clones and found that 14,490 cDNAs (10,897 clusters) were unique to the FLJ collection. About half of them (5,416) seemed to be protein-coding. Of those, 1,999 clusters had not been predicted by computational methods. The distribution of GC content of nonpredicted cDNAs had a peak at approximately 58% compared with a peak at approximately 42%for predicted cDNAs. Thus, there seems to be a slight bias against GC-rich transcripts in current gene prediction procedures. The rest of the cDNAs unique to the FLJ collection (5,481) contained no obvious open reading frames (ORFs) and thus are candidate noncoding RNAs. About one-fourth of them (1,378) showed a clear pattern of splicing. The distribution of GC content of noncoding cDNAs was narrow and had a peak at approximately 42%, relatively low compared with that of protein-coding cDNAs.	**Cytoplasmic dynein nomenclature.** A variety of names has been used in the literature for the subunits of cytoplasmic dynein complexes. Thus, there is a strong need for a more definitive consensus statement on nomenclature. This is especially important for mammalian cytoplasmic dyneins, many subunits of which are encoded by multiple genes. We propose names for the mammalian cytoplasmic dynein subunit genes and proteins that reflect the phylogenetic relationships of the genes and the published studies clarifying the functions of the polypeptides. This nomenclature recognizes the two distinct cytoplasmic dynein complexes and has the flexibility to accommodate the discovery of new subunits and isoforms.

Table 2.5: Number of PubMed abstracts in corpus before and after filtering

Array name	# PMID in unfiltered corpus	# Unique PMID (% of unfiltered corpus)
HG-U133A	153520	107517 (70%)
HG-U133 Plus 2.0	159275	110811 (70%)
MG-U430 2.0	105594	64171 (61%)
RAT230 2.0	30156	23597 (78%)
Ath1	6411	4101 (64%)
DrosGenome1	19998	6303 (32%)
Drosophila 2.0	20074	6303 (31%)
Celegans	1710	1161 (68%)
Xenopus laevis	2209	1689 (76%)
Zebrafish	2825	1030 (36%)

2.3.3 Tokenisation

Abstracts in the text corpus were fragmented into single-word units (*tokens*) via a process known as tokenisation. This process involved breaking up the texts in the abstracts on whitespace or punctuation characters. Hyphens were treated as separable punctuation except in the following cases:

- Names of chemicals (e.g. *4-tert-butylphenol*), enzymes (e.g. *6-phosphofructo-2-kinase*), and gene aliases (e.g. *Bcl-2*, *IL-8*, *E-cadherin*) were not split at hyphens.
- Hyphenated words formed by connecting a noun to certain prefix such as *anti-apoptotic*, *trans-activation*, *N-glycosylation*, *O-glycosylation*, and *pre-mRNA* were treated as single tokens.

In order to reduce the size of the vocabulary list, the following tokens were eliminated:

- Extremely long amino acid or nucleotide sequences, such as *Val-Lys-Ala-Val-Cys-Val-Ile-Asn-Gly* and *ACACCACCATCAT*.
- Tokens composed exclusively of numbers.
- Non-alphanumeric characters, such as "@()<>$!?&#".

Non-scientific English words that carry low domain-specific information content, such as *the*, *of* and *is*, were retained. They serve as control condition for testing if the text-based ORA methods developed are capable of extracting information against a background of uninformative "noise" in the system.

Biological concepts are often expressed as composite words, such as *cell cycle*, *DNA polymerase*, and *CRE binding protein*. However, the identification of multiword features adds an additional layer of complexity to the processing of text that is beyond the scope of the initial studies in this work. Possibilities for the extension of the current approach to the analysis of multiword and more complex text corpora will be discussed in Chapter 10.

2.3.4 Stemming

After tokenisation, the tokens were subjected to a simple stemming procedure, which truncates suffixes from related terms and collapse them to a standard form. Porter's stemming algorithm[6] was adapted to perform the following operations:

- Reduce plurals to singular forms, e.g. convert *kinases* to *kinase*.
- Stem verb tenses to their root, e.g. collapse *phosphorylates*, *phosphorylated* to *phosphorylate*.
- Tokens that are less than three characters long were left unchanged as these usually represent gene symbols or protein names (e.g. *FAS*, *FOS*).

All tokens were converted into upper case at this stage. Other more elaborate analyses of spelling variants (e.g. *catalyze*, *catalyse*) was not explored at this stage.

2.4 From gene list to token frequencies

2.4.1 Linking Entrez Gene ID to tokens

Using the Affymetrix probeset ID–to-EGID mapping scheme described in Section 2.3.1, all the probeset IDs for a particular array were converted to a non-redundant set of EGIDs. For each of the 10 Affymetrix arrays listed in Table 2.2, a binary gene-to-term matrix was constructed as depicted in Figure 2.2.

2.4.2 Calculation of *Chip* and *List* frequencies

Based on the gene-to-term matrix, the number of genes containing a certain token of interest on a given chip type can be calculated. This is the background frequency, and is denoted as *Chip* frequency throughout this thesis.

Given a gene list for which the query identifiers are in the form of Affymetrix probeset ID, the first step is to map the probeset IDs to their corresponding EGIDs based on the mapping scheme described in Section 2.3.1. By subsetting these EGIDs

[6] Perl version, release 1; http://tartarus.org/~martin/PorterStemmer/

Gene-to-PMID mapping

	PMID 1	PMID 2	PMID 3	PMID 4
Gene 1	0	1	1	1
Gene 2	1	1	0	0
Gene 3	1	1	1	1
Gene 4	0	0	0	0
Gene 5	1	1	0	1

+

PMID-to-token mapping

	Token 1	Token 2	Token 3
PMID 1	1	1	0
PMID 2	0	0	0
PMID 3	0	0	1
PMID 4	0	1	1

↓

Gene-to-token mapping

	Token 1	Token 2	Token 3
Gene 1	0	1	1
Gene 2	1	1	0
Gene 3	1	1	1
Gene 4	0	0	0
Gene 5	1	1	1

Figure 2.2: Organising gene and token data into a numerical format suitable for use by downstream text mining algorithms

For each gene on a selected array, the relevant PubMed abstracts were retrieved and converted into a list of unique tokens as described in Section 2.3. These genes can then be represented in token space, in which multiple occurrence of a token associated with the same gene was reduced to a single binary count. The values 1 and 0 signify 'present' and 'absent', respectively.

from the gene-to-term matrix, the number of genes that are associated with the token of interest in the query gene list can be calculated. This measurement is denoted as *List* frequency throughout this thesis.

The end result is that each token in a gene list will be associated with two scores - its *Chip* and *List* frequencies - which are then used as inputs to the statistical algorithm for over-representation analysis.

Chapter 3

Classical hypergeometric distribution-based ORA

3.1 Introduction

One of the challenges in the analysis of gene expression data is to use a list of differentially expressed genes (DEG) to gain an insight into the signalling and regulatory mechanisms that generate such changes, and the biological consequences of that change in gene expression. A first step towards (at least partly) addressing this issue is often to categorise a list of DEG into known functionally related groups, perhaps based on a combination of manual literature and database searches, or using prior familiarity with a gene(s) and plausible links to the biology under study. Literature databases such as MEDLINE or PubMed, and functional annotation and pathway databases such as Gene Ontology (GO) (Ashburner *et al.* 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), are amongst the most widely used resources. The result of this manual annotation process is usually a list of biological processes or pathways that are common between the regulated genes. Naively, one might reasonably expect that biological processes that occur more frequently in the list would be more relevant to the biology of the studied system. For example, if 100 genes were found to be differentially expressed and 60 of them are known to be involved in apoptosis, then one might be tempted to conclude that apoptosis is an important biological process in the context of the experiment. However, there is a significant confounding factor; consider whether such an interpretation would be valid if 60% the genes on the array used were part of the apoptotic pathway? Clearly, the answer is 'no', because the observed number of genes involved in apoptosis is no better than by chance alone. This example illustrates that statistical tests are required to guide interpretation, so that one can distinguish between significant and random events.

The need to address this problem systematically has led to the development of a family of methods collectively known as Over-Representation Analysis (ORA), which seeks to establish statistically whether particular functionally-defined groups of genes are significantly enriched within the DEG relative to all the genes in the background population. ORA generally consists of three steps. First, the number of genes associated with each of a series of functional categories or annotation terms of interest is determined for the DEG and the background population. Then, a *p*-value for assessing over-representation is calculated for each annotation term using an appropriate statistical test. Finally, the *p*-values are corrected for multiple testing in order to keep the number of false positives at an acceptable significance level. Significant annotation terms are then those whose adjusted *p*-values pass some pre-determined cutoff threshold.

A number of statistical methods have been proposed for calculating the enrichment *p*-values, including the chi-squared test for equality of proportions, the hypergeometric test, Fisher's exact test and the binomial test. These approaches have been implemented (with minor variations) in several publicly available software tools for performing ontological analysis on gene lists; some popular ones include DAVID/EASEonline (Hosack *et al.* 2003), FatiGO (Al-Shahrour *et al.* 2004), GO::TermFinder (Boyle *et al.* 2004), GenMAPP (Dahlquist *et al.* 2002), GoMiner (Zeeberg *et al.* 2003) and OntoTools (Draghici *et al.* 2003). A comprehensive review of these tools, including the scope of analysis, underlying statistical models, sources of metadata and visualisation capabilities, can be found in Khatri and Draghici (2005).

Improvements of the methodological aspects of ORA continue to emerge. However, existing applications of ORA are largely limited to the mining of pre-defined functional annotations such as GO terms and KEGG pathways. These resources are, to a large extent, generated from manual literature reading by experts, with the aim of providing a structured, condensed and reduced description of the biological knowledge about genes in the scientific literature. Due to its labour-intensive nature, such pre-defined functional annotations are inevitably limited in scope and flexibility, and cannot fully reflect the detail of all areas of biology that might be of interest, nor are they designed to do so. The most up-to-date and perhaps richest source of biological knowledge about genes is primarily contained in the biomedical literature, which is readily accessed in the form of PubMed abstracts, and increasingly as full-text articles from selected biomedical journals. The use of free-text as a potentially more informative knowledge base for interpreting gene expression data has been demonstrated previously. For examples, Shatkay *et al.* (2000) used a similarity-based scheme to map functionally descriptive kernel documents to genes and established connections between genes. Blaschke *et al.* (2001), Chaussabel and Sher (2002) and Glenisson *et al.* (2003) showed that clustering genes by literature profiles or keyword association derived from MEDLINE records can further discern informative pictures about the nature of genes and their functional relationships.

Therefore, the objective of this study is to determine whether the successful applications of ORA can be extended beyond the mining of controlled vocabularies to a wider mining of free-text. Initial exploration into this approach was based on a simple mining of tokens extracted from PubMed abstract using the classical hypergeometric test as statistical model. The statistical framework and implementation details of this text-based ORA approach are described in the following section.

3.2 Statistical framework for mining PubMed abstract terms based on the classical hypergeometric test

The general framework for performing over-representation analysis on textual information extracted from PubMed abstracts consists of two core components:

1. A text corpus and the corresponding annotation terms that are associated with the genes under study.
2. A statistical model for determining whether there are a higher proportion of genes with certain abstract terms among the DEG relative to genes on the entire array.

The methodologies for retrieving the relevant abstracts and creating the text corpus are given in Section 2.3. The remainder of this section focuses on the second component, that is the theoretical basis underlying the classical hypergeometric distribution-based ORA approach, the construction of inputs to this model, and the *p*-value correction methods.

3.2.1 Hypergeometric test: formulations and assumptions

Consider a microarray containing *N* genes, of which *M* are associated with a certain annotation term of interest *T*. Suppose *K* genes on the array are called differentially expressed based on some preliminary gene-level analysis, and *x* of them are found to be annotated with the term *T*. The question we are concerned with is: what is the probability of this happening due to chance? This statistical problem can be illustrated as a 2 × 2 contingency table (Table 3.1).

Table 3.1: A classical 2 × 2 contingency table

	Annotated to *T*	Not annotated to *T*	Total
In DEG	x	$K-x$	K
Not in DEG	$M-x$	$N-K-M+x$	$N-K$
Total	M	$N-M$	N

If *K* genes were sampled randomly from the array without replacement under the assumption that all the genes are independent and identically distributed, then the number of genes associated with term *T* in the sample can be modelled by the hypergeometric distribution. The probability of observing exactly *x* genes associated with *T* just by chance in the DEG is given by the formula:

$$\Pr(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \qquad \text{Equation (3.1)}$$

where max(0, $M + K - N$) ≤ x ≤ min(K, M). This range assures that all counts in the contingency table are non-negative.

The p-value for testing whether the DEG presents an enrichment of genes annotated to T is given by the probability of observing x or more genes containing T, assuming that the null distribution specified in equation (3.1) is the true count distribution:

$$p = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \qquad \text{Equation (3.2)}$$

This corresponds to a one-sided test in which small p-value satisfying certain pre-defined cutoff value is indicative of over-representation (this point is discussed further in Section 3.2.5).

Alternative statistical approaches for calculating the enrichment p-value have been proposed. These include the Fisher's exact test, the binomial test and the chi-squared test. The choice of statistical model depends on the size of the background population (N) and that of the sample. In the binomial model, the probability of picking a gene annotated to T is assumed to be fixed and is equal to the proportion of genes in the background. When N is large, the probability of picking a gene annotated to T from the background barely changes after each gene is picked. However, when N is small, this probability is influenced substantially by whether the previously picked genes were annotated to T (Rhee *et al.* 2008). Therefore, the binomial test is only appropriate for large N (for example, whole-genome microarrays). On the other hand, the chi-squared test is not suitable when sample size is small. A general guideline is that at least 80% of the expected frequencies should be greater than or equal to 5 and all the expected frequencies must exceed 1 for the test to provide valid conclusions (Agresti 1984). Under these circumstances, the hypergeometric distribution or Fisher's exact test are better models. When the marginal totals in the 2 × 2 contingency table are fixed, the Fisher's exact test is in fact equivalent to the hypergeometric test (Rivals *et al.* 2007).

Out of all the statistical models mentioned above, the hypergeometric test appears to be the most robust and adaptable, and was therefore chosen for exploring the feasibility of this initial text-based ORA framework.

3.2.2 Gene universe

An important consideration when using the hypergeometric test to identify significantly enriched terms is the specification of an appropriate "gene universe", i.e. the gene reference background against which the enrichment *p*-values are calculated. Rhee *et al.* (2008) and Falcon and Gentleman (2007) both pointed out that the gene universe has a substantial effect on the test results, and an incorrect gene universe can lead to incorrect conclusions.

For the text-based ORA approach described here, only those genes that are monitored in the microarray experiment and associated with at least one abstract term were included in the gene universe. This gene universe was constructed using the following procedures:

1. *Correcting for multiple occurrences of a gene.* On some arrays, such as those from Affymetrix, the same gene can be represented by multiple probesets (see Section 1.2.2 for details). This multiplicity problem was resolved by mapping all probesets to their corresponding Entrez Gene identifiers (EGID) to ensure that each gene is represented only once on the array and will receive only one "vote" during the analysis. The metadata used for this conversion is detailed in Table 2.3. Probesets that do not map to any EGIDs were omitted.
2. *Removal of genes without annotations.* The universe was further refined by removing genes that are not associated with any abstract terms. This is because the inclusions of EGIDs that have no chance to appear in any of the annotation terms will introduce an additional source of bias into the system. Their inclusion appears to produce apparently more significant enrichment *p*-values, but "noisier" results (as illustrated in Section 3.3.2).

The phrases "gene universe", "reference background" and "background population" are used interchangeably throughout this thesis.

3.2.3 *Chip* and *List* frequencies

Each abstract term (token) associated with a given gene list is subjected to the same fixed set of two measurements. The first measurement, *Chip* frequency, is the number of genes that contains the token of interest on the chip (i.e. background). The second measurement, *List* frequency, is the number of genes that are associated with the token of interest in the query gene list. Only genes present in the gene universe were counted.

3.2.4 Jackknife adjustment

A concern with the hypergeometric test is that terms supported by a low number of genes can be over-weighted. Consider a list of DEG containing 200 members that were selected from a background population of 12500 genes. Suppose that the term 'lymphogenesis' is associated with only one gene in the background population, and that gene happens to be on the DEG list. This term would obtain a *p*-value of 0.016 and appear to be significantly enriched in the list of DEG. However, a term or biological theme supported by just a single gene is neither stable nor interesting because it contains no information about the functional relationship amongst genes in the gene list. To address this problem, a more conservative *p*-value was calculated for each term based on the idea of jackknifing a probability. This method involves removing one gene from the set of genes associated with the given term on the gene list, and then calculates the hypergeometric *p*-value based on the truncated sample. This is analogous to the calculation of EASE score as proposed in Hosack *et al.* (2003). The theoretical basis underlying jackknifing is that, by arbitrarily omitting a single observation at a time from the original sample and repeatedly recalculating the given statistic, one can ascertain the stability of the statistic and detect fluctuations in sampling error that may occur due to a single deleted observation's uniqueness. The jackknifing operation entails a conservative adjustment to the hypergeometric *p*-values by penalising the significance of terms supported by few genes whilst negligibly penalising terms supported by many genes.

Taken together, the hypergeometric p-value associated with each term was calculated by substituting the following values into Equation (3.2):

x = *List* frequency – jackknife_score

N = Number of genes in the gene universe

M = *Chip* frequency

K = Number of genes in the gene list – jackknife_score

where jackknife_score is set at 1.

3.2.5 Multiple testing correction

The p-value threshold for determining whether a particular annotation term is over-represented depends on the required significance level. The significance level is the acceptable probability of making a type I error (false positive). For example, by setting the significance level at (say) 0.05, we are prepared to accept that there is a 1 in 20 chance of calling a term significantly over-represented when in fact it is not. When many annotation terms are tested at the same time (which is usually the case in ORA), the number of false positive test results is expected to grow with the number of tests performed. If 1000 annotation terms are tested at a significance level of 0.05, then we would expect to find approximately $1000 \times 0.05 = 50$ false positives just by chance.

Classical multiple testing procedures call for the control of the probability of committing any type I error in the entire family of hypotheses under simultaneous consideration. A widely used methodology for controlling this family-wise error rate (FWER) is the Bonferroni correction (Hochberg 1988), in which the chosen significance level (α) is divided by the total number of terms being tested (R). In other words, to ensure that FWER $\leq \alpha$, the p-value from individual test (p) must satisfy $p \leq \alpha/R$. Based on this, the Bonferroni adjusted p-values can be calculated as $\min(p \times R, 1)$, and any terms with adjusted p-values $\leq \alpha$ are deemed significant.

The Bonferroni correction is a conservative approach because it assumes that the individual tests are independent, and it becomes increasingly difficult to detect true positives as the number of tests increases. A less conservative approach to multiple testing is to control the false discovery rate (FDR) instead of the FWER. The FDR is the expected proportion of false positives amongst the rejected hypotheses. There is a potential gain in power with FDR-controlling procedures (Benjamini and Hochberg 1995) because the FDR is a less stringent condition than the FWER.

For all analyses presented in this Chapter, a term is considered significantly over-represented in the gene list if its *p*-value is less than or equal to 0.05 after Bonferroni correction (i.e. at 95% confidence level).

3.3 Experiments and results

3.3.1 Performance on real datasets

The performance of the hypergeometric distribution-based ORA approach described above was tested using the ISG gene list from Sanda *et al.* (2006) and the glycolysis gene list reported in Vanharanta *et al.* (2006). Details of these two gene lists are given in Sections 2.2.1 and 2.2.3, respectively. Both studies used the HG-U133A microarray platform from Affymetrix. The text corpus used for this analysis was created according to the methodology given in Section 2.3. Briefly, PubMed articles associated with genes represented on the Affymetrix HG-U133A array were collected and filtered to give a corpus consisting of 107,517 abstracts. These abstracts were tokenised and stemmed to produce 219,857 unique single-word tokens for mining. Of these, 11,709 tokens are associated with the ISG gene list, while 17,088 tokens are associated with the glycolysis gene list. To reduce the token space, tokens with *List* frequency equal to 1 after jackknifing were removed because a token can only be useful in defining relationships among genes if it is shared by at least two of them. After this filtering, 3411 and 4927 tokens remained for testing in the ISG gene list and glycolysis gene list, respectively.

Example 1: ISG gene list

Initial use of the classical hypergeometric distribution-based method produced encouraging results when applied to the ISG gene list. 81 tokens were identified as significantly over-represented at the 0.05 significance level after the *p*-values were corrected for multiple testing using the Bonferroni method. The results are shown in Table 3.2. Biologically-plausible terms such as 'interferon', 'IFN', 'antiviral', 'IFN-alpha', 'IFN-beta', 'IFN-gamma', 'viral' and 'immune' were amongst the most significant hits being called over-represented. These terms are related in principle to the role of interferon in modulating host immune responses against viruses and infection. For examples, 'MHC', 'histocompatibility', 'HLA-A', 'HLA-B', 'HLA-G', 'antigen', 'beta2-microglobulin', 'LMP2' and 'LMP7' are related to the induction and regulation of the complement and antigen presentation pathways by interferons, and terms such as 'OAS', 'oligoadenylate', 'PKR' and 'MxA' are related to interferon-inducible antiviral effectors.

However, these biologically relevant terms were interspersed with what appeared to be relatively uninformative terms, for which it seemed less plausible that they were specifically associated with the biology of interferon-regulated gene expression. These include common English words such as 'treatment' (rank 29), 'line' (rank 60), 'intact' (rank 72), 'after' (rank 73) and non-specific biological words such as 'beta' (rank 31), 'response' (rank 37), 'monoclonal' (rank 55), 'synthesis' (rank 66). A similar mix of apparently specific and non-specific terms was generally seen for other gene lists that were analysed (data not shown), suggesting that this problem is not unique to the ISG gene list.

The amount of "uninformative" terms that were called over-represented varied substantially from gene list to gene list. In the ISG gene list, the uninformative terms were mostly found at the lower half (less significant portion) of the hit list. In some other gene lists tested, the uninformative terms were found to be given high rankings and scattered throughout the result table, as illustrated by the glycolysis gene list described below.

Table 3.2: Significantly over-represented abstract terms in the ISG gene list as identified using the classical hypergeometric test

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
INTERFERON	414	46	3.30E-46	1.13E-42	1
IFN	245	35	4.00E-37	1.36E-33	2
IFN-BETA	71	18	1.81E-22	6.16E-19	3
ANTIVIRAL	176	23	2.00E-22	6.81E-19	4
IFN-ALPHA	114	19	2.97E-20	1.01E-16	5
INDUCIBLE	1068	37	9.19E-18	3.14E-14	6
INTERFERON-ALPHA	59	14	8.18E-17	2.79E-13	7
INFECTION	1177	36	1.83E-15	6.25E-12	8
VIRAL	892	32	2.55E-15	8.69E-12	9
IMMUNE	1275	35	1.57E-13	5.36E-10	10
INNATE	363	21	2.25E-13	7.69E-10	11
TREAT	1817	40	8.76E-13	2.99E-09	12
IFN-GAMMA	443	22	9.32E-13	3.18E-09	13
DSRNA	60	11	7.52E-12	2.57E-08	14
IMMUNITY	387	20	7.85E-12	2.68E-08	15
OLIGOADENYLATE	18	8	1.70E-11	5.81E-08	16
VIRUS	1408	34	1.73E-11	5.90E-08	17
ISRE	31	9	2.46E-11	8.41E-08	18
LYMPHOBLASTOID	239	16	5.84E-11	1.99E-07	19
INDUCTION	2048	39	2.23E-10	7.60E-07	20
ISG	14	7	2.58E-10	8.80E-07	21
HLA-A	30	8	1.02E-09	3.48E-06	22
MHC	353	17	1.53E-09	5.22E-06	23
HOST	800	24	1.58E-09	5.39E-06	24
STOMATITIS	52	9	2.10E-09	7.15E-06	25
HLA-CLASS	11	6	6.24E-09	2.13E-05	26
EVASION	65	9	1.31E-08	4.47E-05	27
HLA-B	25	7	1.43E-08	4.89E-05	28
TREATMENT	3120	45	2.16E-08	7.35E-05	29
ENCEPHALOMYOCARDITIS	16	6	5.74E-08	0.0002	30
BETA	2127	36	5.82E-08	0.0002	31
INFECT	825	22	8.62E-08	0.0002	32
CYTOKINE	1266	27	1.14E-07	0.0003	33
HISTOCOMPATIBILITY	303	14	1.25E-07	0.0004	34
HEPATITIS	366	15	1.59E-07	0.0005	35
ANTIGEN	1687	31	1.88E-07	0.0006	36
RESPONSE	3630	47	2.44E-07	0.0008	37
MELANOMA	581	18	2.89E-07	0.0009	38
BETA2-MICROGLOBULIN	42	7	3.87E-07	0.0013	39
OAS	10	5	4.34E-07	0.0014	40
HLA-G	10	5	4.34E-07	0.0014	41
REPLICATION	830	21	4.70E-07	0.0016	42
EPSTEIN-BARR	233	12	5.08E-07	0.0017	43
GAMMA-INTERFERON	44	7	5.15E-07	0.0017	44

(continued over the page)

Table 3.2: Significantly over-represented abstract terms in the ISG gene list as identified using the classical hypergeometric test (continued)

Term	*Chip* frequency	*List* frequency	p-value	Bonferroni p-value	Rank
MXA	11	5	6.79E-07	0.0023	45
EBV	194	11	7.93E-07	0.0027	46
INDUCE	4669	53	8.93E-07	0.0030	47
TAPASIN	12	5	1.01E-06	0.0034	48
HLA	253	12	1.15E-06	0.0039	49
INTERFERON-GAMMA	313	13	1.35E-06	0.0045	50
LMP7	13	5	1.45E-06	0.0049	51
LMP2	13	5	1.45E-06	0.0049	52
INDIGENOUS	29	6	1.46E-06	0.0049	53
PKR	30	6	1.74E-06	0.0059	54
MONOCLONAL	1365	26	2.02E-06	0.0068	55
UPREGULATE	1087	23	2.09E-06	0.0071	56
PROMYELOCYTIC	216	11	2.11E-06	0.0072	57
LYSIS	169	10	2.26E-06	0.0077	58
INDUCIBILITY	131	9	3.13E-06	0.0107	59
LINE	4667	52	3.16E-06	0.0108	60
OR-C	5	4	3.18E-06	0.0108	61
IMMUNODEFICIENCY	472	15	3.37E-06	0.0115	62
TAP	61	7	3.70E-06	0.0126	63
AUTOIMMUNE	557	16	4.66E-06	0.0159	64
PROTEASOME	490	15	5.21E-06	0.0178	65
SYNTHESIS	2200	33	6.06E-06	0.0207	66
P69	6	4	6.33E-06	0.0216	67
MICROGLOBULIN	39	6	6.69E-06	0.0228	68
DEFENSE	370	13	7.62E-06	0.0260	69
VSV	19	5	7.64E-06	0.0261	70
LEUKEMIA	1182	23	8.37E-06	0.0286	71
INTACT	1382	25	9.14E-06	0.0312	72
AFTER	3913	46	9.47E-06	0.0323	73
CTL	154	9	1.04E-05	0.0355	74
LOAD	383	13	1.08E-05	0.0369	75
ISG15	7	4	1.10E-05	0.0376	76
PEPTIDE-MHC	7	4	1.10E-05	0.0376	77
AND-C	44	6	1.23E-05	0.0419	78
INFLUENZA	75	7	1.24E-05	0.0424	79
REACTIVITY	534	15	1.39E-05	0.0475	80
C1R	22	5	1.42E-05	0.0484	81

Over-represented terms were defined as having p-value $\leq$ 0.05 after Bonferroni correction. Terms are ordered by increasing p-values. This analysis was performed using genes existed on the HG-U133A chip and associated with at least one term in the corresponding text corpus as gene universe (N = 9638).

Example 2: Glycolysis gene list

When the glycolysis gene list was analysed with the classical hypergeometric distribution-based ORA approach, 48 terms were called significantly enriched (Table 3.3). Despite the successful identification of biologically relevant terms such as 'glycolytic', 'dehydrogenase', 'reductase', 'peroxidation', 'isoenzyme' and 'NAD', approximately 50% of the hits identified appears to be noise, such as 'resolution' (rank 1), 'level' (rank 7), 'had' (rank 11), 'one' (rank 16), 'correspond' (rank 18), 'high' (rank 20) and 'library' (rank 27). The enrichment of these uninformative terms would severely hamper the usefulness of the proposed approach. The explanation for this artefactual enrichment of uninformative terms, and methods to avoid it, are the subject of subsequent Chapters.

Table 3.3: Significantly over-represented abstract terms in the glycolysis gene list as identified using the classical hypergeometric test

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
RESOLUTION	1006	39	1.21E-09	5.95E-06	1
GLYCOLYTIC	71	12	1.92E-09	9.44E-06	2
DEHYDROGENASE	465	25	7.91E-09	3.90E-05	3
NORMAL	4514	92	1.92E-08	9.47E-05	4
CRYSTAL	1292	42	3.74E-08	0.0001	5
REDUCTASE	328	20	5.71E-08	0.0002	6
LEVEL	5876	107	6.25E-08	0.0003	7
PEROXIDATION	102	12	9.41E-08	0.0004	8
NAD	187	15	1.66E-07	0.0008	9
LIVER	2545	62	1.75E-07	0.0008	10
HAD	3782	80	1.93E-07	0.0009	11
ISOENZYME	195	15	2.78E-07	0.0013	12
WESTERN	2347	58	4.03E-07	0.0019	13
ESCHERICHIA	1100	36	5.04E-07	0.0024	14
OXIDATIVE	672	27	5.83E-07	0.0028	15
ONE	5620	102	6.10E-07	0.0030	16
OXYGEN	632	26	6.65E-07	0.0032	17
CORRESPOND	1126	36	8.81E-07	0.0043	18
OXIDATION	431	21	9.00E-07	0.0044	19
HIGH	4674	90	1.01E-06	0.0049	20
NICOTINAMIDE	80	10	1.01E-06	0.0050	21
COLI	1297	39	1.13E-06	0.0055	22
CULTURE	3076	68	1.15E-06	0.0056	23

(continued over the page)

Table 3.3: Significantly over-represented abstract terms in the glycolysis gene list as identified using the classical hypergeometric test (continued)

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
METABOLITE	439	21	1.20E-06	0.0059	24
MUSCLE	2184	54	1.47E-06	0.0072	25
ABDOMINAL	162	13	1.49E-06	0.0073	26
LIBRARY	3443	73	1.50E-06	0.0074	27
PARAMETER	1008	33	1.91E-06	0.0093	28
RATE	2334	56	2.11E-06	0.0104	29
CANCER	2991	66	2.21E-06	0.0109	30
PHOSPHATE	680	26	2.54E-06	0.0125	31
CATALYSIS	345	18	2.89E-06	0.0142	32
NADP	69	9	3.29E-06	0.0162	33
KCAT	121	11	4.36E-06	0.0215	34
THREE	4884	91	4.37E-06	0.0215	35
GROUP	3121	67	4.79E-06	0.0236	36
PERCENT	401	19	5.35E-06	0.0264	37
EXPERIMENTAL	1218	36	5.44E-06	0.0268	38
OBSERVE	4529	86	6.49E-06	0.0320	39
REACTIVITY	534	22	6.55E-06	0.0323	40
CATALYZE	1072	33	7.11E-06	0.0350	41
SUBSTRATE	2428	56	7.60E-06	0.0374	42
TISSUE	6027	104	8.00E-06	0.0394	43
COMPARE	4021	79	8.10E-06	0.0399	44
ISOLATE	5186	94	8.34E-06	0.0411	45
PROFILE	1408	39	8.37E-06	0.0413	46
POINT	2122	51	8.67E-06	0.0427	47
NADPH	194	13	9.71E-06	0.0478	48

Over-represented terms were defined as having p-value ≤ 0.05 after Bonferroni correction. Terms are ordered by increasing p-values. This analysis was performed using genes existed on the HG-U133A chip and associated with at least one term in the corresponding text corpus as gene universe ($N = 9638$).

3.3.2 Impact of gene universe on ORA results

Setting up an appropriate gene universe is important in hypergeometric test-based ORA method as it can have a marked effect on the outcome. The analyses presented in Section 3.3.1 were performed with a strictly-defined gene universe $G_{annotated}$., which consist of genes that are not only existed on the chip but also associated with at least one term in the corresponding text corpus.

To examine the impact of gene universe on the classical hypergeometric distribution-based ORA approach, the ISG gene list was re-analysed with a larger set of genes as background, in which all genes existing on the HG-U133A array were included in the gene universe regardless of whether they are associated with any annotation term in the corresponding text corpus. Let this more broadly-defined gene universe be G_{total}.

As can be seen from Table 3.4, an analysis performed with G_{total} leads to a greater amount of non-specific terms being called significantly enriched when compared to the results obtained with $G_{annotated}$. This finding is in accordance with the views of Huang *et al.* (2009), who pointed out that larger backgrounds tends to produce more significant *p*-values, as compared with a narrowed-down set of genes as a population background. This example highlights the importance of background reference against which the hypergeometric *p*-values were calculated, an issue that has constantly been overlooked by existing ORA tools.

Table 3.4: Results of re-analysing the ISG gene list with G_{total} as gene universe

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
INTERFERON	414	46	1.67E-48	5.71E-45	1
IFN	245	35	2.90E-39	9.89E-36	2
ANTIVIRAL	176	23	5.39E-24	1.84E-20	3
IFN-BETA	71	18	9.52E-24	3.25E-20	4
IFN-ALPHA	114	19	1.38E-21	4.72E-18	5
INDUCIBLE	1068	37	7.02E-20	2.40E-16	6
INTERFERON-ALPHA	59	14	8.13E-18	2.77E-14	7
INFECTION	1177	36	1.60E-17	5.44E-14	8
VIRAL	892	32	3.09E-17	1.05E-13	9
IMMUNE	1275	35	1.58E-15	5.38E-12	10
TREAT	1817	40	6.32E-15	2.16E-11	11
INNATE	363	21	9.16E-15	3.13E-11	12
IFN-GAMMA	443	22	3.43E-14	1.17E-10	13
VIRUS	1408	34	2.07E-13	7.05E-10	14
IMMUNITY	387	20	3.78E-13	1.29E-09	15
DSRNA	60	11	1.25E-12	4.25E-09	16
INDUCTION	2048	39	1.92E-12	6.56E-09	17
OLIGOADENYLATE	18	8	4.64E-12	1.58E-08	18
LYMPHOBLASTOID	239	16	4.80E-12	1.64E-08	19
ISRE	31	9	5.67E-12	1.93E-08	20
HOST	800	24	5.44E-11	1.85E-07	21
ISG	14	7	8.39E-11	2.86E-07	22
MHC	353	17	1.16E-10	3.97E-07	23
TREATMENT	3120	45	1.42E-10	4.83E-07	24
HLA-A	30	8	2.81E-10	9.58E-07	25
STOMATITIS	52	9	4.91E-10	1.67E-06	26
BETA	2127	36	7.56E-10	2.58E-06	27
RESPONSE	3630	47	1.53E-09	5.21E-06	28
HLA-CLASS	11	6	2.43E-09	8.30E-06	29
EVASION	65	9	3.10E-09	1.06E-05	30
CYTOKINE	1266	27	3.37E-09	1.15E-05	31
ANTIGEN	1687	31	3.93E-09	1.34E-05	32
INFECT	825	22	4.13E-09	1.41E-05	33
INDUCE	4669	53	4.27E-09	1.46E-05	34
HLA-B	25	7	4.71E-09	1.61E-05	35
HISTOCOMPATIBILITY	303	14	1.51E-08	5.17E-05	36
LINE	4667	52	1.62E-08	5.53E-05	37
HEPATITIS	366	15	1.71E-08	5.82E-05	38
MELANOMA	581	18	2.22E-08	7.59E-05	39
ENCEPHALOMYOCARDITIS	16	6	2.25E-08	7.67E-05	40
REPLICATION	830	21	2.66E-08	9.08E-05	41
AFTER	3913	46	7.39E-08	0.00025	42
MONOCLONAL	1365	26	7.45E-08	0.00025	43
EPSTEIN-BARR	233	12	8.20E-08	0.00028	44

(continued over the page)

Table 3.4: Results of re-analysing the ISG gene list with G_{total} as gene universe (continued)

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
UPREGULATE	1087	23	1.03E-07	0.00035	45
SYNTHESIS	2200	33	1.28E-07	0.00043	46
BETA2-MICROGLOBULIN	42	7	1.29E-07	0.00044	47
EBV	194	11	1.47E-07	0.00050	48
GAMMA-INTERFERON	44	7	1.72E-07	0.00058	49
HLA	253	12	1.89E-07	0.00064	50
INTERFERON-GAMMA	313	13	1.95E-07	0.00066	51
OAS	10	5	2.04E-07	0.00069	52
HLA-G	10	5	2.04E-07	0.00069	53
TYPE	4725	50	3.15E-07	0.00107	54
MXA	11	5	3.19E-07	0.00109	55
ALPHA	2422	34	3.50E-07	0.00119	56
DEFINE	2823	37	3.81E-07	0.00130	57
IMMUNODEFICIENCY	472	15	3.95E-07	0.00135	58
PROMYELOCYTIC	216	11	4.00E-07	0.00136	59
INTACT	1382	25	4.03E-07	0.00137	60
LEUKEMIA	1182	23	4.44E-07	0.00151	61
INDEPENDENT	2840	37	4.45E-07	0.00152	62
EACH	3117	39	4.60E-07	0.00157	63
TAPASIN	12	5	4.76E-07	0.00162	64
LYSIS	169	10	4.92E-07	0.00168	65
AUTOIMMUNE	557	16	4.94E-07	0.00169	66
INDIGENOUS	29	6	5.77E-07	0.00197	67
PROTEASOME	490	15	6.20E-07	0.00212	68
LMP7	13	5	6.84E-07	0.00233	69
LMP2	13	5	6.84E-07	0.00233	70
PKR	30	6	6.89E-07	0.00235	71
INDUCIBILITY	131	9	7.86E-07	0.00268	72
CORRESPONDING	2800	36	1.04E-06	0.00353	73
MOLECULE	3217	39	1.07E-06	0.00365	74
DEFENSE	370	13	1.16E-06	0.00396	75
DIFFERENTIAL	1923	29	1.17E-06	0.00398	76
ACTION	1806	28	1.18E-06	0.00403	77
TAP	61	7	1.25E-06	0.00428	78
STIMULATE	2564	34	1.35E-06	0.00460	79
CONFER	1265	23	1.41E-06	0.00480	80
LOAD	383	13	1.67E-06	0.00569	81
REACTIVITY	534	15	1.72E-06	0.00588	82
OR-C	5	4	1.79E-06	0.00612	83
MEDIATE	4505	47	2.06E-06	0.00702	84
RECOMBINANT	2880	36	2.06E-06	0.00704	85
CTL	154	9	2.67E-06	0.00910	86
MICROGLOBULIN	39	6	2.67E-06	0.00912	87
STRAND	1108	21	2.76E-06	0.00942	88

(continued over the page)

Table 3.4: Results of re-analysing the ISG gene list with G_{total} as gene universe (continued)

Term	*Chip* frequency	*List* frequency	*p*-value	Bonferroni *p*-value	Rank
RECOGNIZE	2007	29	2.77E-06	0.0094	89
ALSO	6842	60	3.03E-06	0.0103	90
DERIVE	3496	40	3.12E-06	0.0106	91
P69	6	4	3.57E-06	0.0122	92
VSV	19	5	3.61E-06	0.0123	93
DOUBLE	1235	22	3.76E-06	0.0128	94
INFLUENZA	75	7	4.27E-06	0.0146	95
AND-C	44	6	4.93E-06	0.0168	96
ADDITION	4483	46	5.24E-06	0.0179	97
NK	288	11	5.39E-06	0.0184	98
TNF	357	12	5.54E-06	0.0189	99
ISG15	7	4	6.22E-06	0.0212	100
PEPTIDE-MHC	7	4	6.22E-06	0.0212	101
NEW	4044	43	6.40E-06	0.0218	102
TRANSCRIPTION	4045	43	6.44E-06	0.0220	103
C1R	22	5	6.73E-06	0.0230	104
NATURAL	979	19	7.61E-06	0.0260	105
EXTRACT	1876	27	8.41E-06	0.0287	106
NB4	49	6	8.47E-06	0.0289	107
MOREOVER	2653	33	9.13E-06	0.0311	108
CD8	378	12	9.51E-06	0.0325	109
HLA-C	24	5	9.70E-06	0.0331	110
EVADE	87	7	1.01E-05	0.0346	111
NECROSIS	809	17	1.09E-05	0.0373	112
ANALOGOUS	462	13	1.14E-05	0.0388	113
DAUDI	52	6	1.14E-05	0.0389	114
2-MICROGLOBULIN	25	5	1.15E-05	0.0392	115
INFLAMMATORY	1219	21	1.17E-05	0.0399	116
UBIQUITIN	552	14	1.35E-05	0.0460	117
FORM	4963	48	1.42E-05	0.0484	118
PATHOGEN	322	11	1.43E-05	0.0489	119

Over-represented terms were defined as having *p*-value $\leq$ 0.05 after Bonferroni correction. Terms are ordered by increasing *p*-values. This analysis was performed using all genes represented on the HG-U133A array as gene universe (N = 13,441).

3.4 Discussion

This Chapter reports initial explorations regarding whether the classical hypergeometric distribution-based ORA framework can be expanded to mine text-based information, initially in the form of tokens extracted from PubMed abstracts. Analyses performed on selected public datasets show that plausible results that convey useful biological insights can be obtained using the proposed approach, provided that the gene universe is specified correctly. However, the usefulness of this approach is compromised by a marked over-representation of many additional and apparently non-specific and uninformative terms, which interspersed with the biologically relevant terms. These uninformative terms typically have relatively high frequencies in the background and in the studied gene list. It would thus appear that the probabilities of picking genes associated with these terms are higher than expected. This points to the background frequencies of these terms being under-estimated in the current statistical model.

As will be described in the following Chapters, the explanation for this effect appears related to an unequal representation of textual information across different genes on the array. An effect of this is that if a particular microarray experiment were focused on a particularly well-studied area of biology this can lead to a greater number of PubMed abstracts associated with the resultant gene list than might otherwise occur. This would in turn introduce a bias that affects the application of the classical hypergeometric test such that even a relatively modest increase in frequency of a common word would produce a significant *p*-value. This is referred to as the annotation bias problem, and its cause is formally investigated in Chapter 4.

Chapter 4

Exploration of factors contributing to annotation bias, and its effect on ORA

4.1 Introduction

We have seen in the previous Chapter that the performance of the standard hypergeometric distribution-based ORA approach is not entirely satisfactory when it was applied to analyse terms extracted from PubMed abstracts. In addition to the biologically-plausible terms, many common and apparently somewhat uninformative terms were also identified as significantly over-represented. It was hypothesised that this undesirable effect is caused, at least in part, by an imbalance in the representation of textual information (as in the number of PubMed articles) across genes. Specifically, some areas of biology have traditionally been the subject of a greater level of research activity; this is reflected in relative coverage of different fields and topics in the biological literature. Gene lists generated from microarray experiments focusing on well-studied areas of biology are therefore more likely to be associated with a higher number of PubMed articles than might otherwise be expected. This in turn would introduce a bias that might affect the performance of the hypergeometric test.

The regulation of gene expression by the interferon-induced signalling pathway is an example of a well-researched biological system. Thus, the use of the ISG gene list for assessing the performance of the classical hypergeometric distribution-based ORA approach reflects what is a relatively well-studied area, with a substantial published literature. This can be illustrated by the number of PMIDs associated with the genes in this list. The 78 genes in the ISG gene list are annotated by 1382 PMIDs according to the EGID-to-PMID mapping scheme and the unique PMID corpus compiled as detailed in Sections 2.3.1 and 2.3.2. However, if one were to create a 78-gene list by random sampling from the same set of genes as the background, then (on average) only 625 PMIDs are expected to be associated with such a random gene list. In this example, the "real world" ISG gene list has more than twice the number of PMIDs associated with it than would be expected by chance. This phenomenon will be explored formally in the following sections with the use of a larger dataset.

4.1.1 Comparing the amount of PubMed citations in biologically-derived versus random gene lists

52 gene lists (the gene identifiers are in the form of Affymetrix probeset IDs) that were based on the use of the human HG-U133A array were collected from the published literature. The details of each list are summarised in Table A.1 in Appendix A. For each of these gene lists, the probeset IDs were first reduced to a set of unique Entrez Gene IDs (EGIDs), then 1000 size-matched random gene lists were created by sampling the same number of unique EGIDs (without replacement) randomly from the same set of background genes on the reference array; the mean number of PMIDs associated with these random gene lists was compared with the amount of PMIDs in the corresponding literature gene list.

The results are summarised in Figure 4.1 and reveal two notable features. First, gene lists derived by experimental means (i.e. the result of mining a real biological dataset) tend to have more PMIDs associated with them than equivalently-sized random gene lists (Figure 4.1(a)). For some gene lists this effect is quite marked, with 22 (out of 52) gene lists have at least 1.5-fold more PMIDs than expected by chance (Figure 4.1(b)). The same effect is seen in Figure 4.1(d), which shows a related metric, the annotation density (the number of PMIDs divided by the number of *annotated* genes in the gene list).

Second, one might expect the annotation bias to operate in two directions, insofar as a gene list from a popular area of research would be expected not only to have more well-annotated genes but also fewer very poorly-annotated genes. As shown in Figure

4.1(c) there is evidence that this effect is present. Based on the text corpus used in this analysis, one would expect to find 26-30% unannotated genes (i.e. genes that do not have any PMID) in a random gene list of any size. However, as shown in Figure 4.1(c), the experimentally-derived gene lists tend to contain a lower proportion of unannotated genes than would be expected by chance. Similar trends towards an over-representation of well-annotated genes, and an under-representation of poorly-annotated genes, was also seen for gene lists derived from experiments using the Affymetrix human HG-U133 Plus 2.0 array (Figure 4.2).

To determine if the same applies to other model organisms, the analysis was repeated on gene lists derived from microarray experiments using the mouse, rat, *Arabidopsis*, *Drosophila*, zebrafish, *Xenopus* and *C. elegans* arrays (details of these gene lists can be found in Appendix A). Indeed, gene lists based on popular model organisms such as mouse and rat produced similar trends as that seen in the human data (Figures 4.3 and 4.4). The effect is also seen, albeit with a less dramatic fold difference between the observed and expected PMID counts, for literature gene lists derived from less well-annotated species such as *Arabidopsis* (Figure 4.5), *Drosophila* (Figures 4.6 and 4.7), zebrafish (Figure 4.8) and *Xenopus* (Figure 4.9). The only exception is *C. elegans*, which produced a somewhat inconclusive picture (Figure 4.10), with a substantial number of smaller gene lists showing fewer associated PMIDs than would be expected by chance.

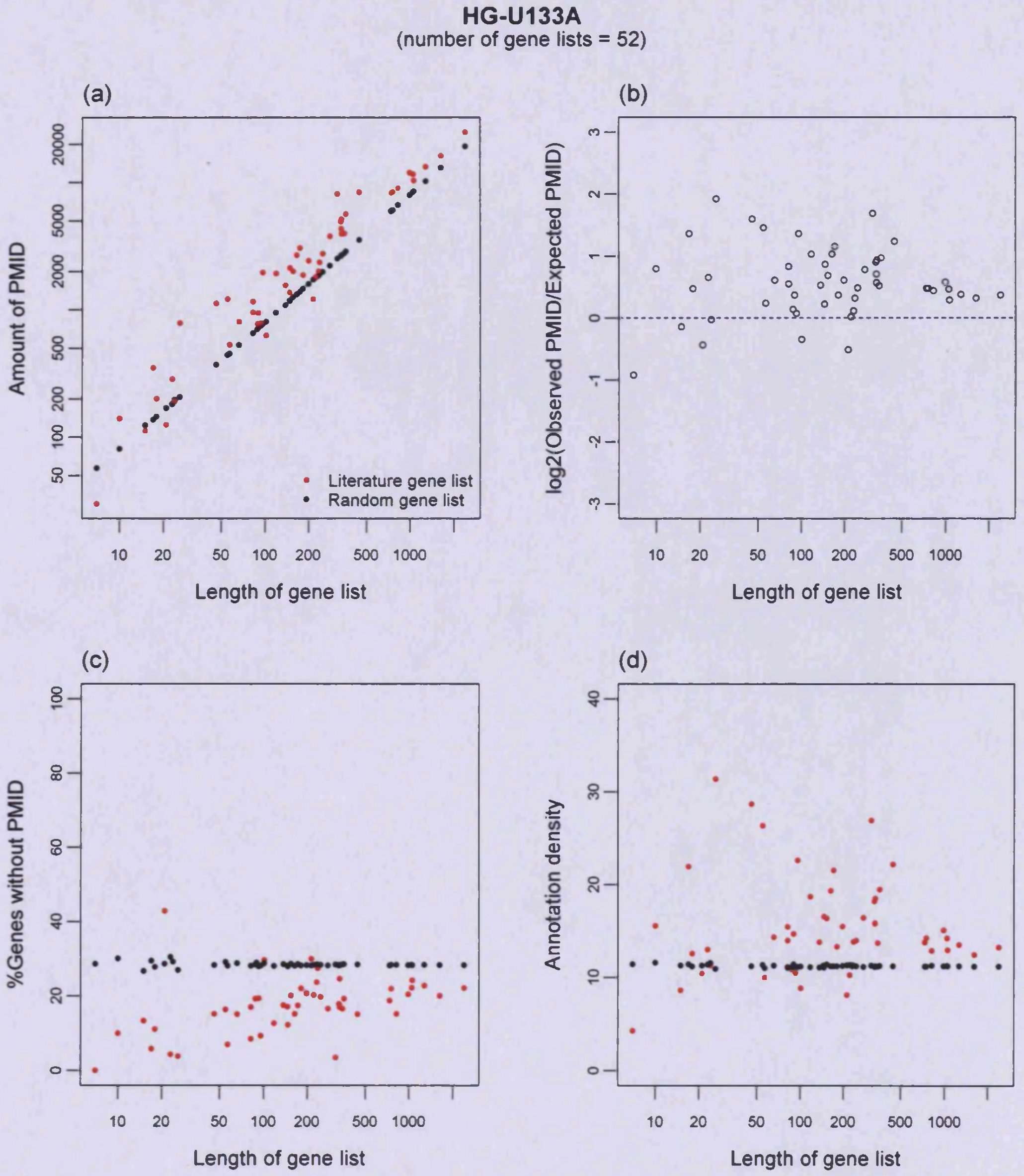


Figure 4.1: Annotation bias analysis for gene lists from HG-U133A
52 gene lists that were based on the human HG-U133A chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

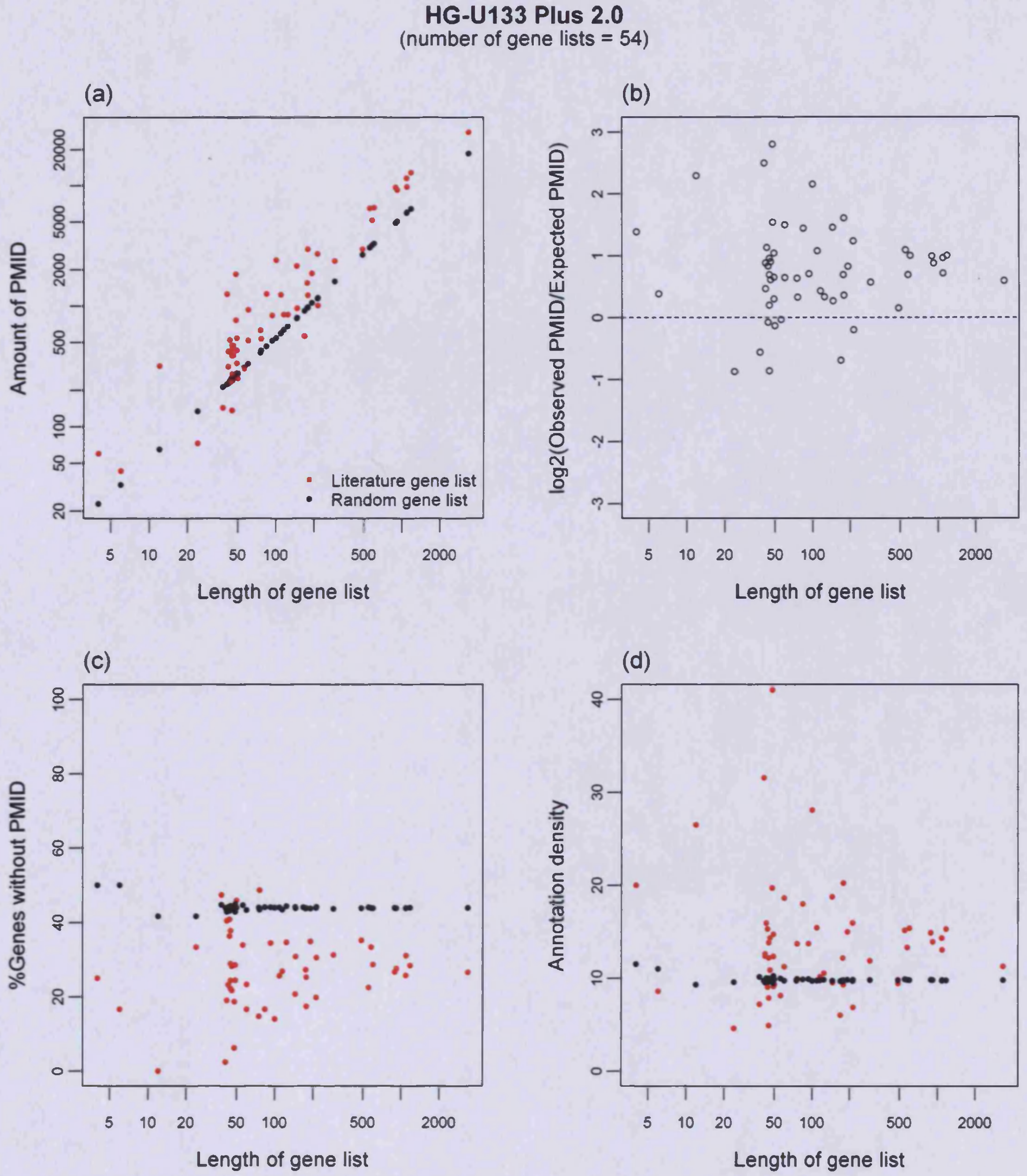


Figure 4.2: Annotation bias analysis for gene lists from HG-U133 Plus 2.0
54 gene lists that were based on the human HG-U133 Plus 2.0 chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

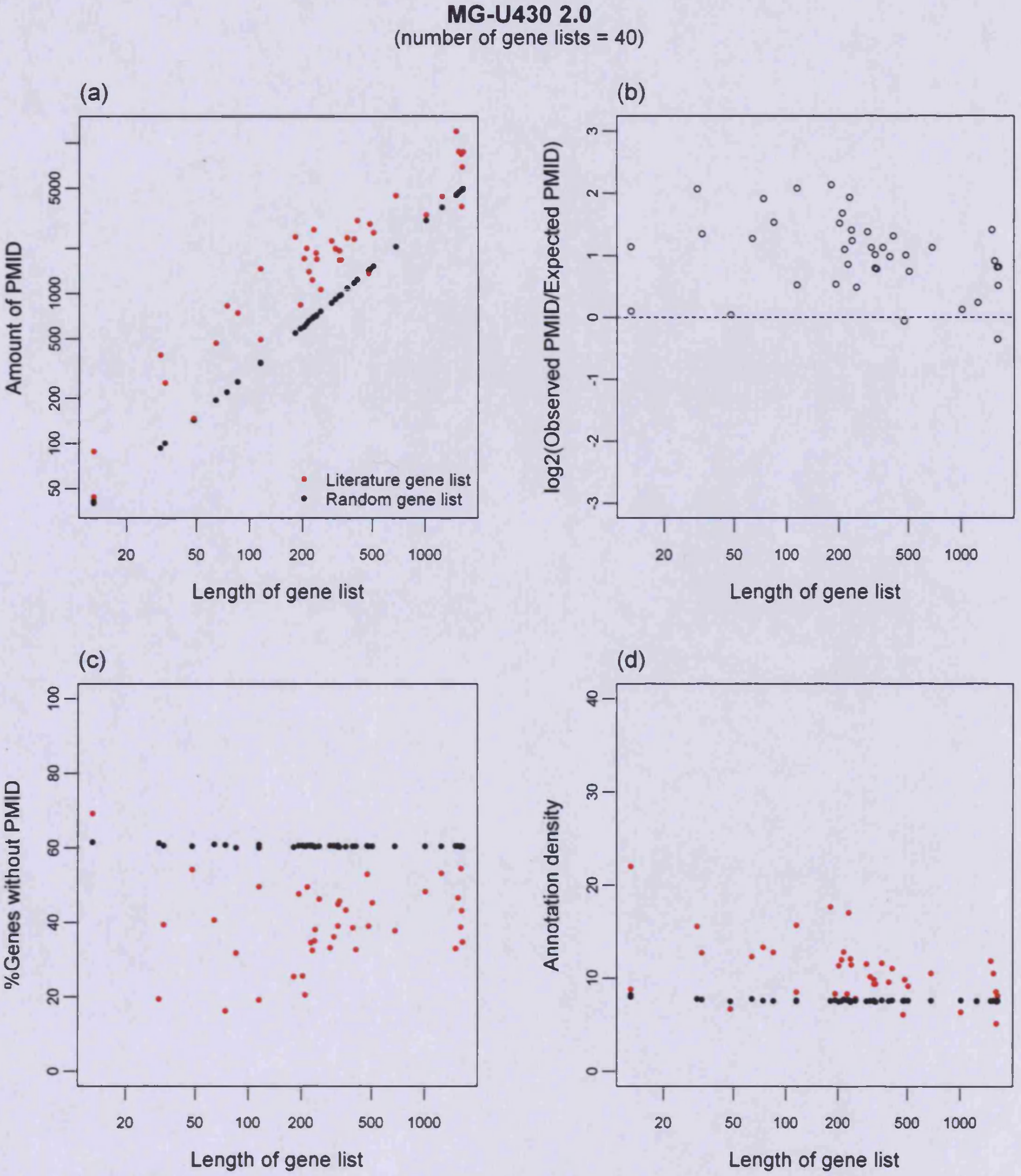


Figure 4.3: Annotation bias analysis for gene lists from MG-U430 2.0

40 gene lists that were based on the mouse MG-U430 2.0 chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

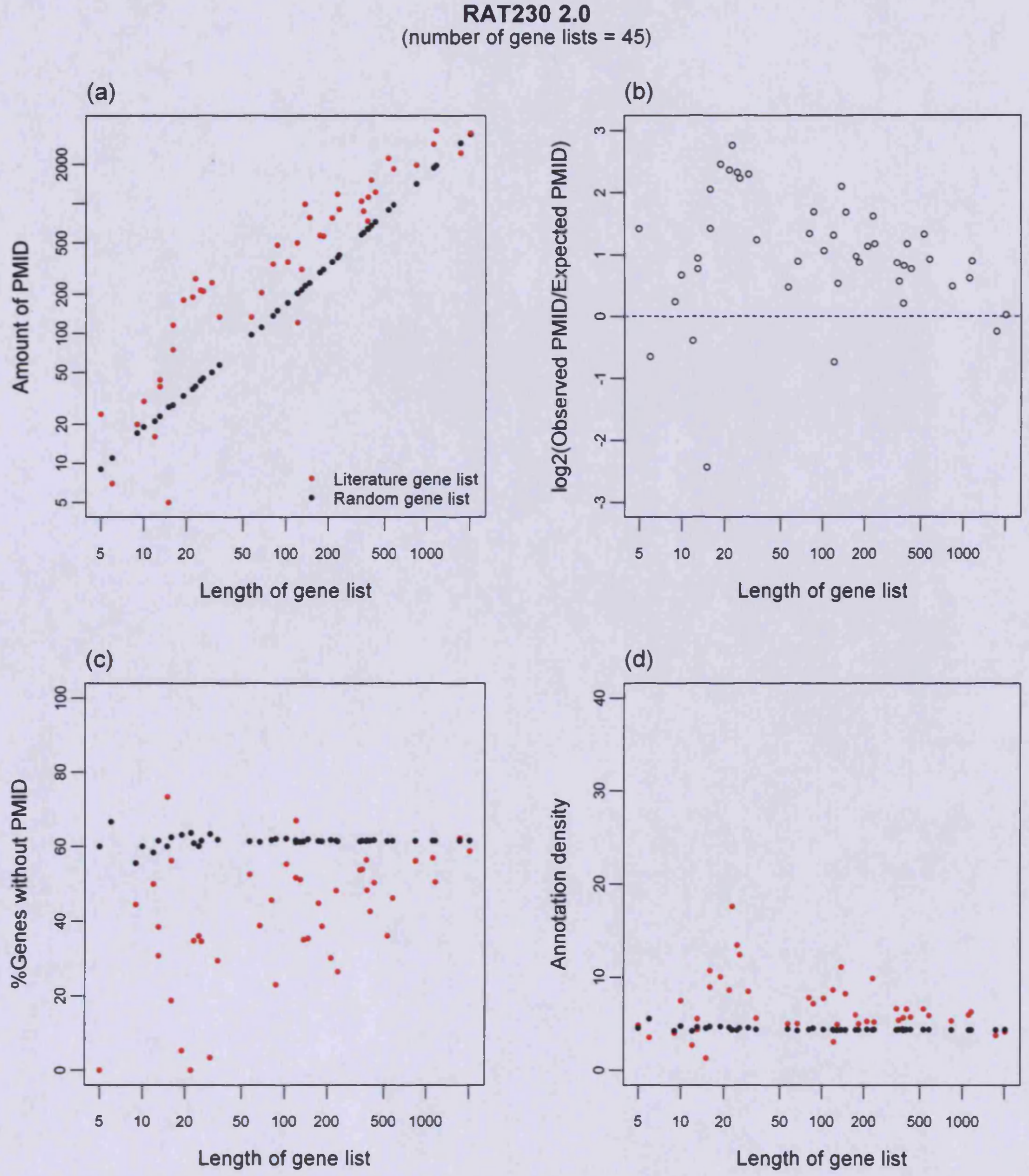


Figure 4.4: Annotation bias analysis for gene lists from RAT230 2.0

45 gene lists that were based on the RAT230 2.0 chip were collated from published the literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

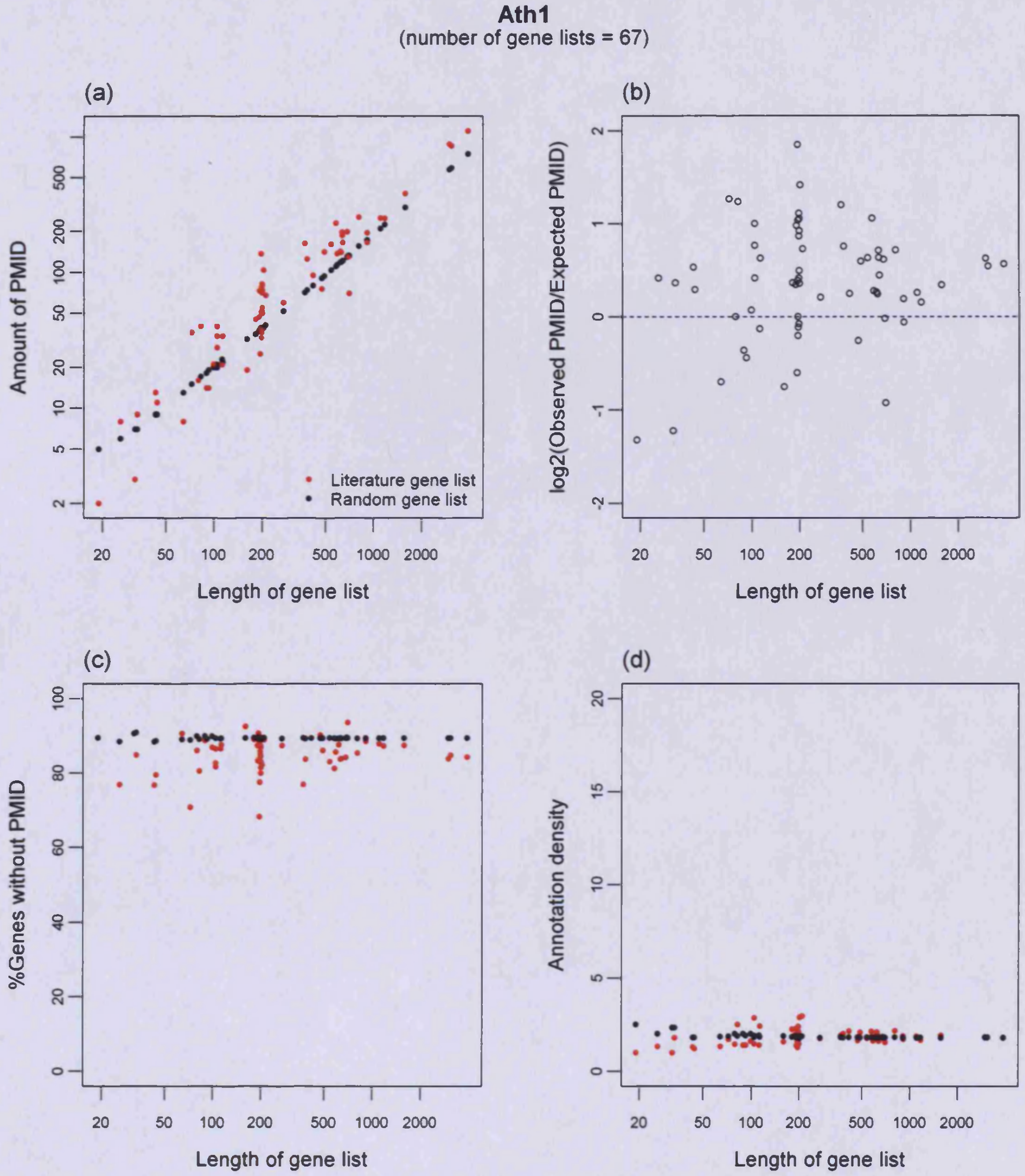


Figure 4.5: Annotation bias analysis for gene lists from Ath1

67 gene lists that were based on the *Arabidopsis* Ath1 chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

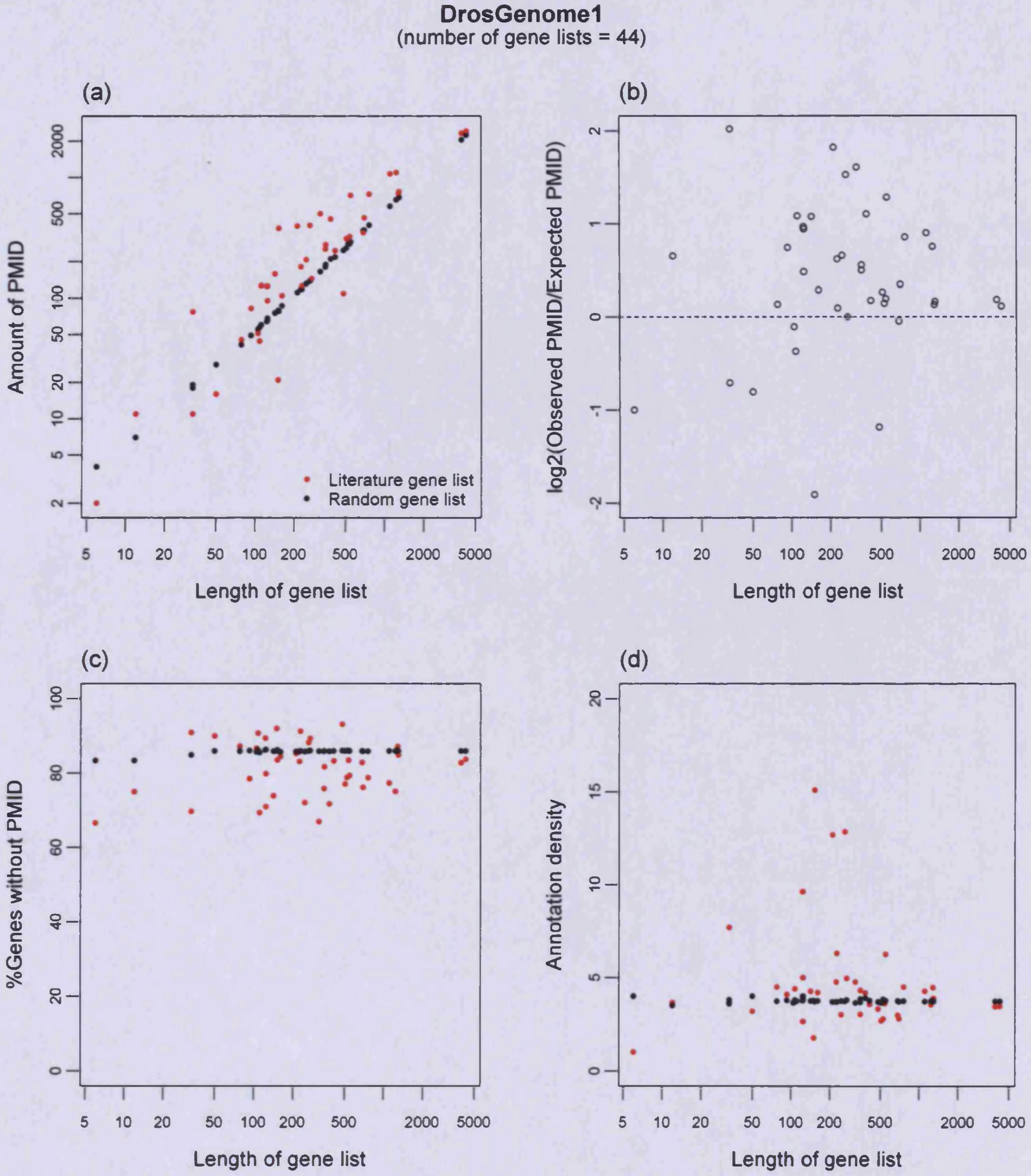


Figure 4.6: Annotation bias analysis for gene lists from DrosGenome1

44 gene lists that were based on the *Drosophila* DrosGenome1 chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

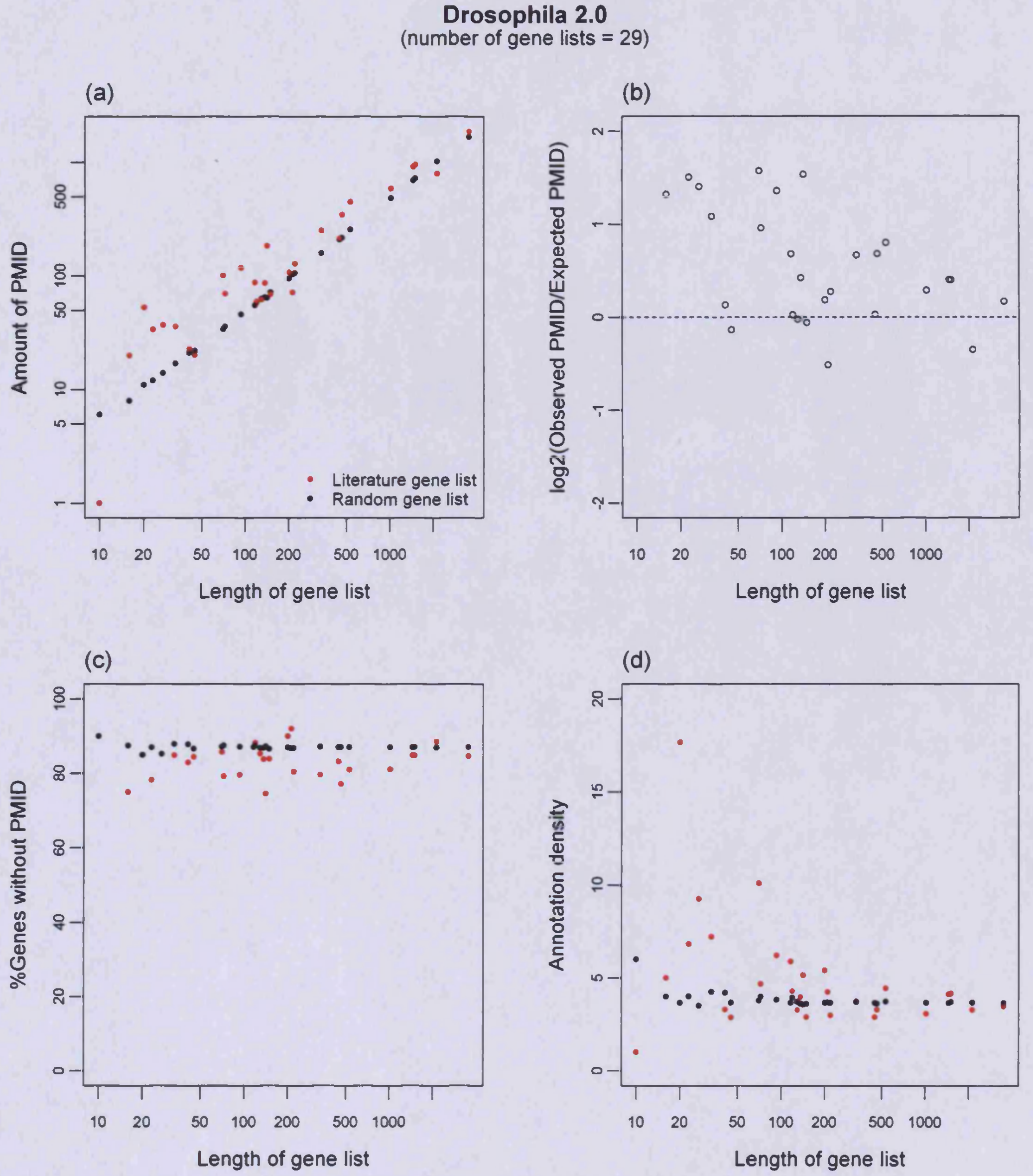


Figure 4.7: Annotation bias analysis for gene lists from Drosophila2

29 gene lists that were based on the Drosophila2 chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

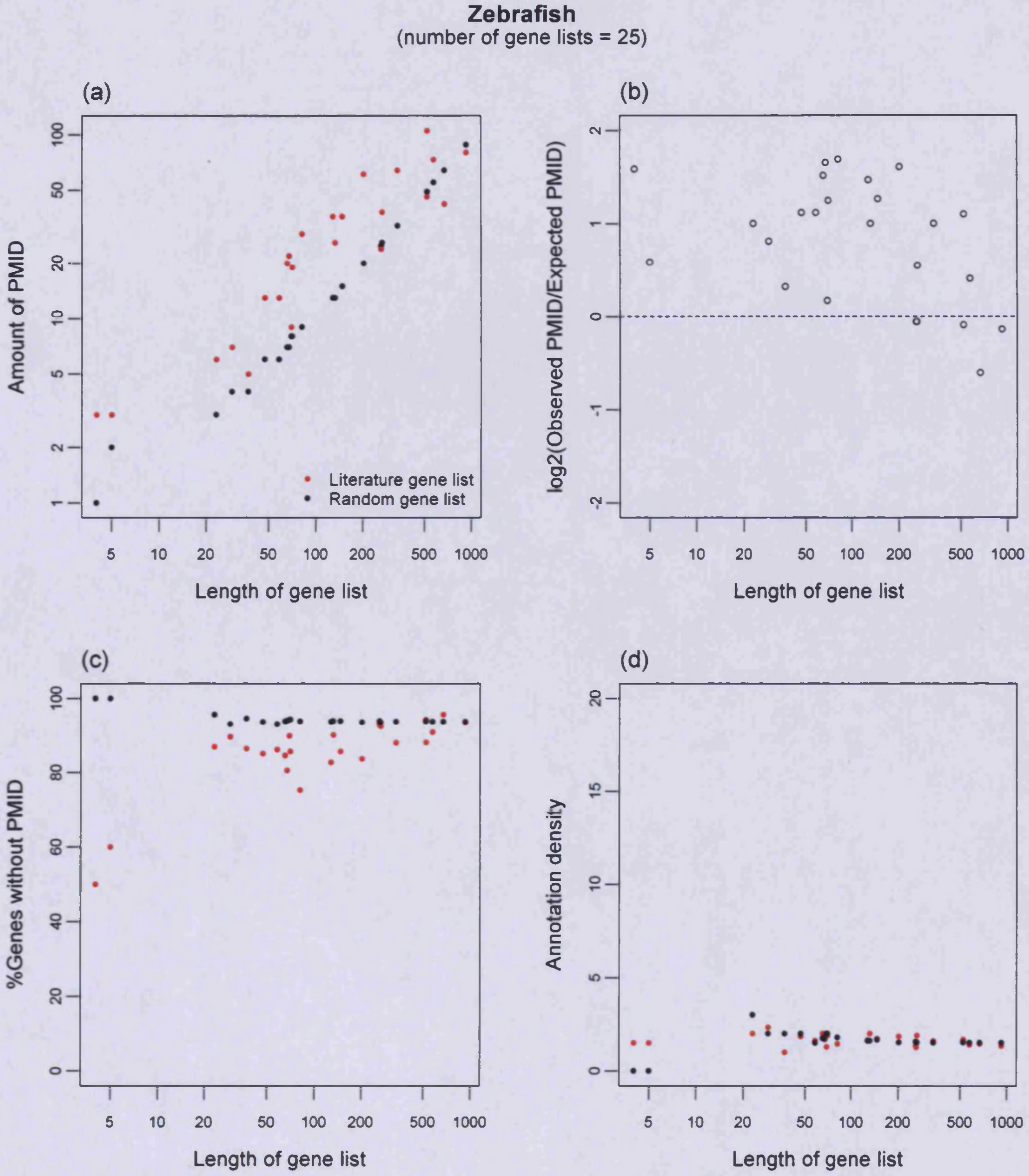


Figure 4.8: Annotation bias analysis for gene lists from Zebrafish

25 gene lists that were based on the Zebrafish chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

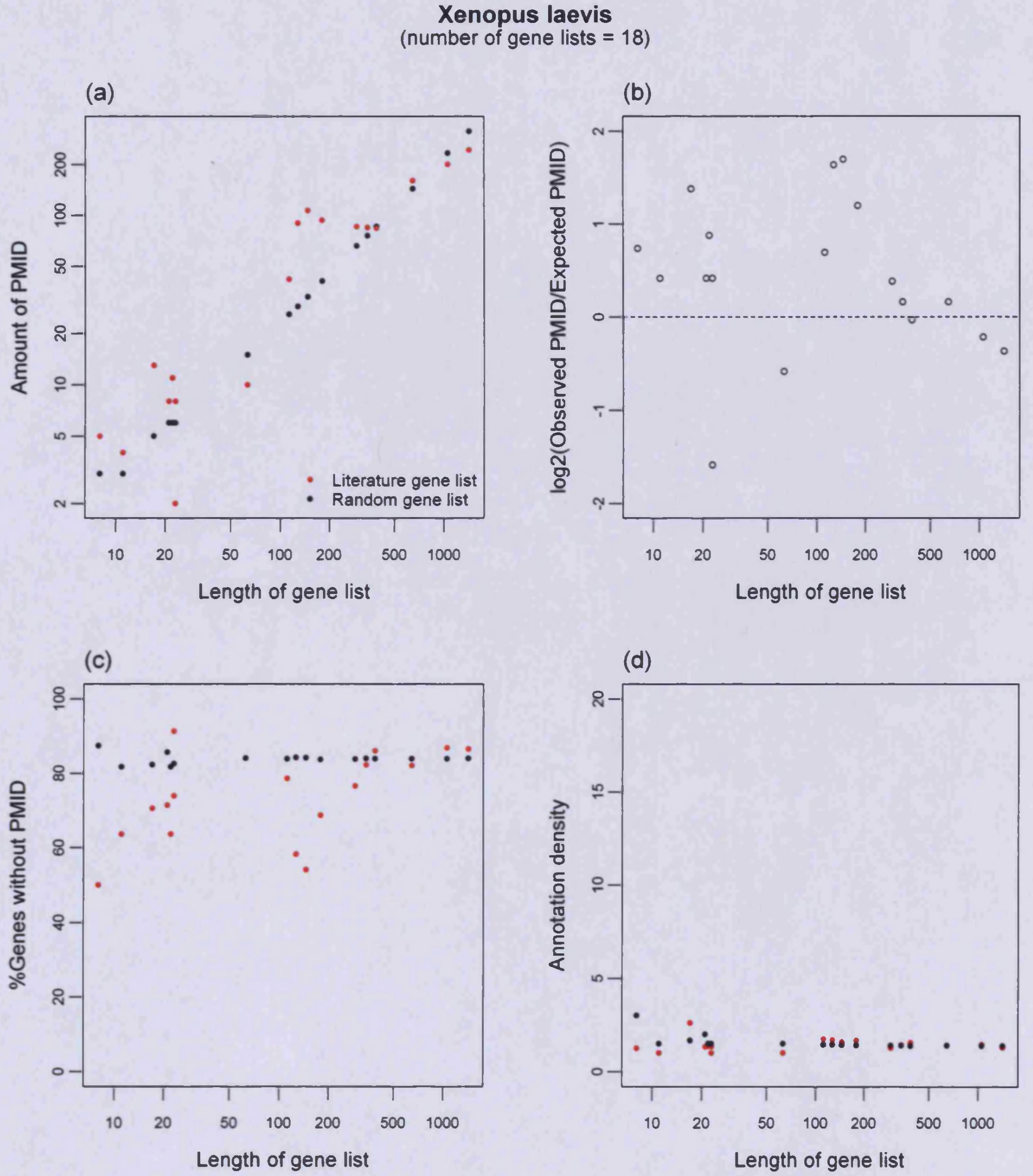


Figure 4.9: Annotation bias analysis for gene lists from Xenopus laevis

18 gene lists that were based on the Xenopus laevis chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

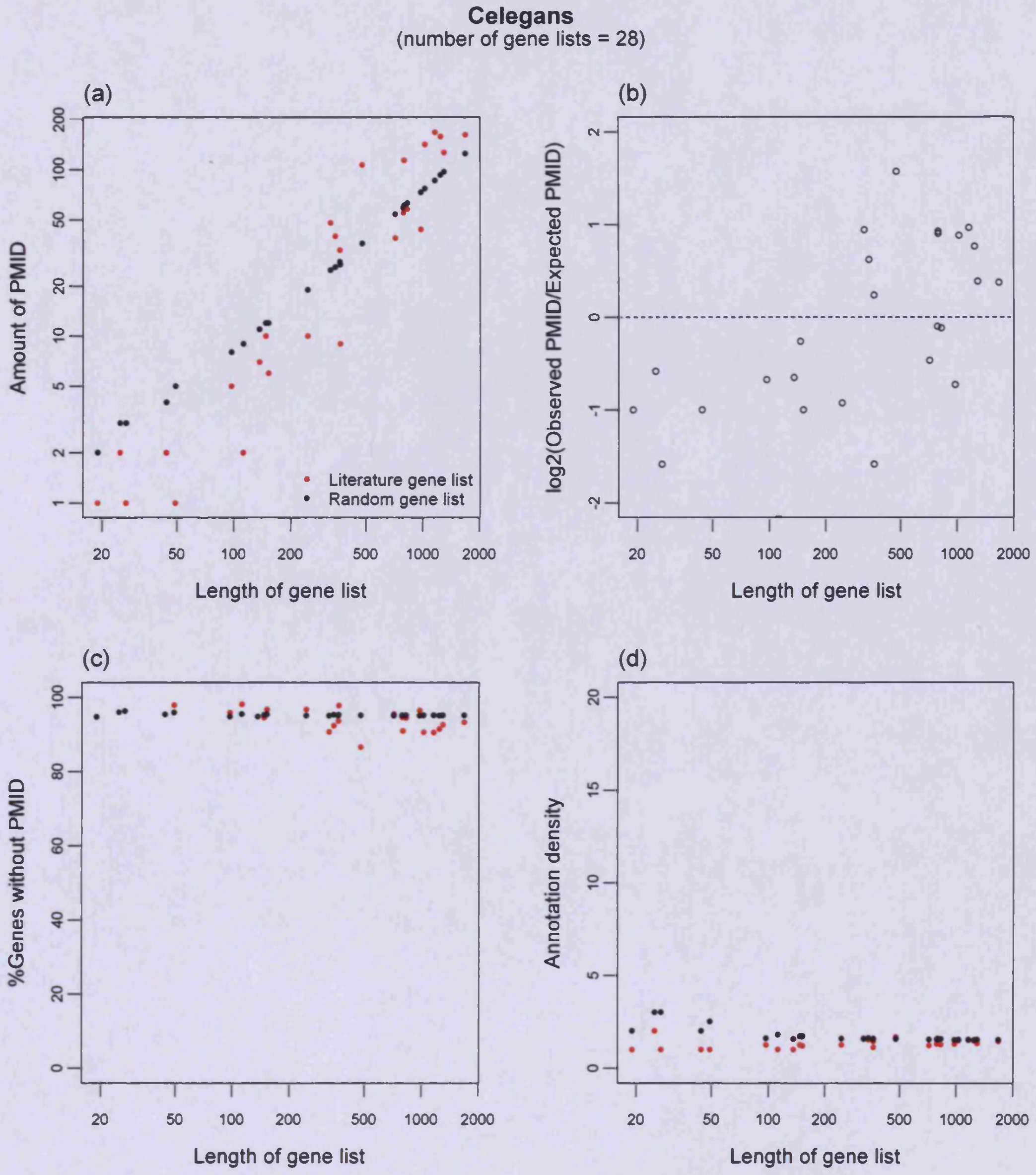


Figure 4.10: Annotation bias analysis for gene lists from Celegans

28 gene lists that were based on the Celegans chip were collated from the published literature, and for each of these 1000 equivalently-sized random gene lists were created. (a) A comparison of the number of PMIDs associated with the literature and random gene lists. (b) Fold-change in the amount of PMIDs between the literature (observed) and random (expected) gene lists. (c) A comparison of the proportions of genes without PMID citation in the literature and random gene lists. (d) A comparison of the annotation densities in the literature and random gene lists. Annotation density was calculated as the number of PMIDs divided by the number of annotated genes in the gene list.

4.1.2 Comparing the amount of GO annotations in biologically-derived versus random gene lists

Due to its comprehensiveness and convenient data structure for high-throughput data mining, Gene Ontology (GO) is one of the most popular annotation resources used in existing ORA tools. To determine whether there is also an excess of GO annotations inherent with the literature gene lists, GO terms associated with genes represented on the selected human, mouse, rat, *Drosophila*, zebrafish, and *C. elegans* arrays were identified using the mapping scheme given in the 'gene2go' file[1]. Due to the lack of GO-to-EGID mappings for *Arabidopsis* and *Xenopus laevis*, these two species were omitted from this analysis. All aspect of GO (biological process, molecular function and cellular component) was used, while GO terms without evidence codes were removed. Using the same approach as that described in Section 4.1.1, the number of GO terms annotated to genes in the literature gene lists were determined and compared with that expected in a set of equivalently-sized random gene lists.

Results based on the well-annotated and less well-annotated species were shown in Figures 4.11 and 4.12, respectively. With the exceptions of *Drosophila* and *C. elegans*, literature gene lists derived from the remaining species were generally associated with more GO terms than random gene lists. However, the magnitudes of the differences between the observed and expected GO counts are typically less dramatic than that seen for PubMed citations. To illustrate this point, consider the 52 literature HG-U133A literature gene lists. The fold-change between the observed and expected GO counts is between 0.8 and 1.3, with only three gene lists having 1.3-times more GO terms than expected by chance. In contrast, for the same set of gene lists, as many as 22 gene lists (42%) have at least 1.5-times more PMIDs than expected by chance.

[1] ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz; time stamp: 24 Jan 2008.

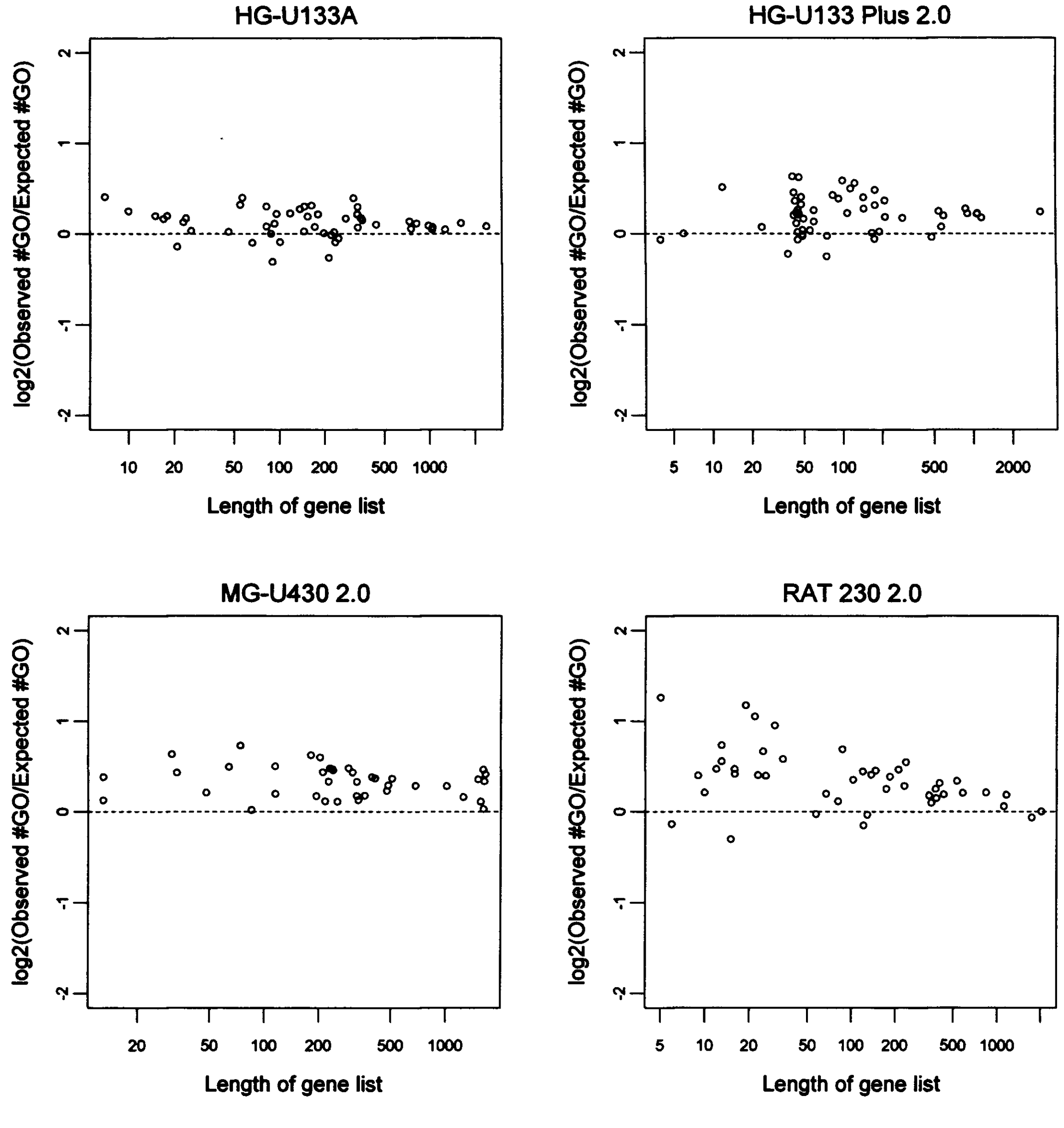


Figure 4.11: GO-based annotation bias analysis for gene lists derived from well-annotated species

The amount of GO terms associated with genes in the literature and random gene lists derived from human (HG-U133A, HG-U133 Plus 2.0), mouse (MG-U430 2.0) and rat (RAT 230 2.0) were compared. 'Observed #GO' represents the number of GO terms in the literature gene lists; 'Expected #GO' represents the mean number of GO terms calculated from 1000 size-matched random gene lists.

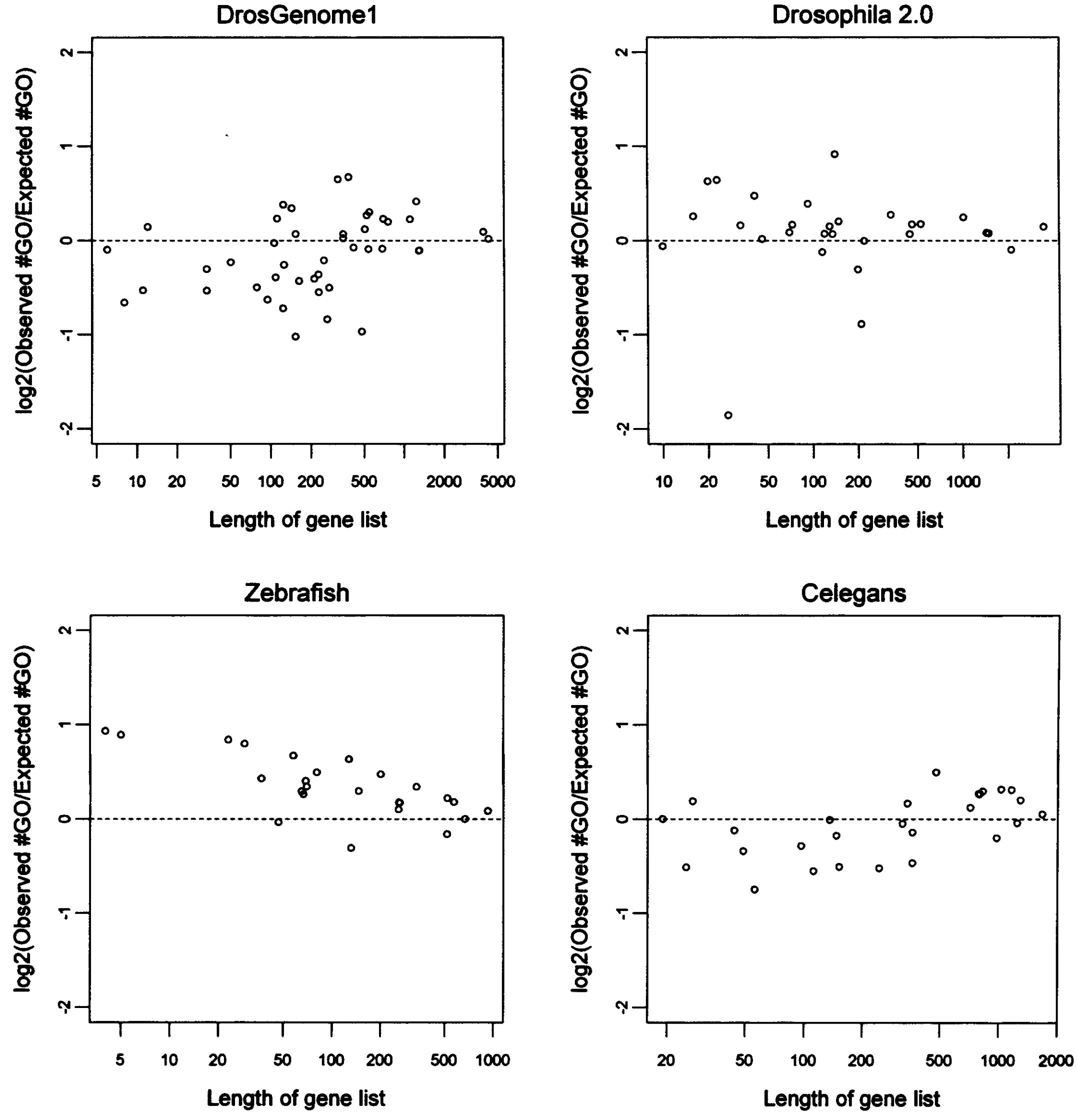


Figure 4.12: GO-based annotation bias analysis for gene lists derived from less well-annotated species

The amount of GO terms associated with genes in the literature and random gene lists derived from *Drosophila* (DrosGenome1, Drosophila 2.0), zebrafish (Zebrafish) and *C. elegans* (Celegans) were compared. 'Observed #GO' represents the number of GO terms in the literature gene lists; 'Expected #GO' represents the mean number of GO terms calculated from 1000 size-matched random gene lists.

4.1.3 Implications for existing ORA approach

The above findings suggest that there is an excess of annotation associated with highly-annotated gene lists. The predicted consequence of this on the use of classical hypergeometric test for identifying over-represented PubMed tokens is that certain tokens (in particular common terms) would be shifted towards appearing over-represented simply because the background frequencies of these tokens are artefactually under-estimated (this phenomenon is illustrated with examples in Section 5.1.1). Therefore even a modest increase in token frequency of these common terms would appear to yield a significant hypergeometric *p*-value. This would then result in a mixture of biologically-plausible and non-specific terms being called significantly over-represented by the classical hypergeometric test-based ORA approach, as is seen.

Since GO terms and PubMed tokens are fundamentally different in structure and nature, it is difficult to assess the extent to which annotation bias would affect the performance of GO-based ORA. In particular, GO involves a systematic representation of knowledge and links between processes and genes; whether a particular GO term is associated with a gene based on a handful of published papers, or several thousand, the result is still a single link. As such, one might expect some of the potential for annotation bias to have been removed in GO. However, if a particular biological process is better studied, it would seem intuitive that there may therefore be more genes involved in it to be known (all other things being equal), and thus the GO term associated with that process would be more likely to appear significant than others. Both Blaschke *et al.* (2001) and Khatri and Draghici (2005) have pointed out, such annotation bias present in the ontological databases should be taken into account during enrichment analysis. Unfortunately, this problem remains unresolved, and to date has largely been overlooked by existing ontological tools that implement ORA.

In the next sections, the relationships between annotation bias, gene age and trends of biological research will be investigated. Solutions for overcoming annotation bias and novel approaches for identifying significantly enriched PubMed tokens within a gene list will be presented in subsequent Chapters.

4.2 Gene age

A well-studied gene may in part reflect one that has been known for many years, thus allowing time for a substantial corpus of literature regarding it to be accumulated. To investigate the possible effect of this aspect of the history of recent scientific research, the "ages" for 28,424 human genes were determined and the numbers of PMIDs associated with each were compared. Here, "gene age" is defined as an approximate measure of how long a gene has been "known" and researched upon relative to other genes, and this usage should not be confused with the concept of the molecular timescale of evolution in the genome.

In this analysis, the age of a gene was inferred on the basis of two criteria: i) when the gene was first cited in the published literature, and ii) when the gene was first integrated into the public databases. An overview of how the gene ages were derived is shown in Figure 4.13, while the methodological details are described below.

4.2.1 Gene age indicators

Using a combination of PubMed, OMIM (Online Mendelian Inheritance in Man) (Amberger *et al.* 2009) and HGNC (HUGO Gene Nomenclature Committee) (Bruford *et al.* 2008), four types of possible age indicators were derived for each of the 28,424 human genes.

1. *PubMed_earliest*: date of the earliest PubMed article that described the gene
2. *OMIM_earliest*: date of the earliest article cited in an OMIM record for which the gene is described
3. *OMIM_creation*: date on which the gene first appeared in an OMIM record
4. *HGNC_approved*: date on which the gene symbol was first approved by HGNC

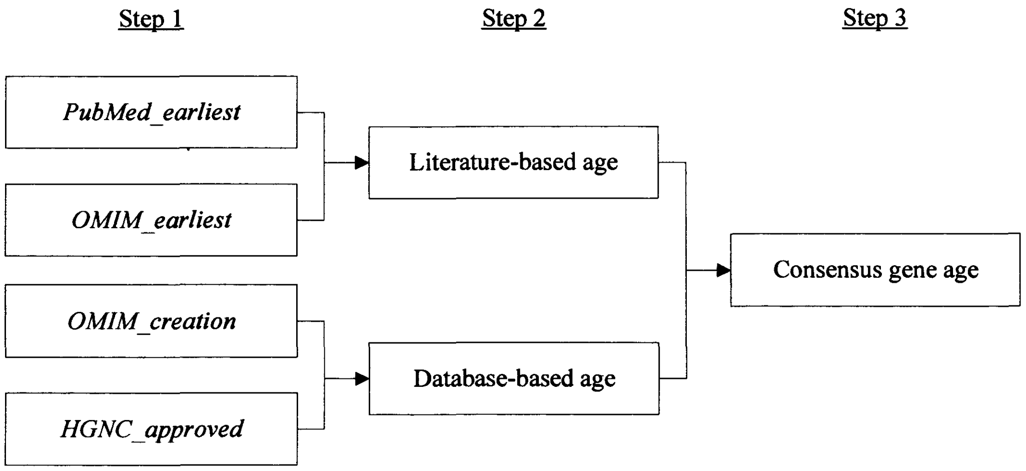


Figure 4.13: Overview of the steps in the computation of gene age
Step 1: Four age indicators, denoted as *PubMed_earliest*, *OMIM_earliest*, *OMIM_creation* and *HGNC_approved*, were derived using a combination of resources including PubMed, OMIM and HGNC. Step 2: *PubMed_earliest* and *OMIM_earliest* were combined to give the age estimate called 'Literature-based age'; while *OMIM_creation* and *HGNC_approved* were combined to give a second age estimate called 'Database-based age'. Step 3: The literature- and database-based ages were combined to give the final gene age measure known as the 'Consensus gene age'.

To obtain *PubMed_earliest*, PubMed citations that are associated with the set of human genes were identified based on the mappings provided in the 'gene2pubmed' file[2]. For each gene, the earliest article describing it was determined and the year of publication of this article was recorded. The values of *PubMed_earliest* lie within the range [1950, 2007].

OMIM_earliest and *OMIM_creation* were inferred from OMIM. OMIM is a database of human genes and genetic disorders. Its content is based exclusively on the published biomedical literature. OMIM entries that are relevant to the set of human genes were downloaded in XML format from the NCBI database based on the mappings provided in the 'mim2gene' file[3]. Each OMIM gene entry contains sections

[2] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz; time stamp: 25 Oct 2007.
[3] ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene; file downloaded on 15 Nov 2007.

of text describing the current knowledge about the gene (e.g. structure, function, location and linkage information) and literature references. *OMIM_earliest* and *OMIM_creation* were derived from the sections "References" and "Creation Date", respectively. The "References" section contains a list of all the articles cited in the OMIM entry. From this, the publication date of the earliest article was determined and this corresponds to *OMIM_earliest*. The "Creation Date" field specifies the date on which the OMIM entry was created and this corresponds to *OMIM_creation*. The values of *OMIM_earliest* and *OMIM_creation* lie within [1798, 2007] and [1986, 2007], respectively.

HGNC_approved was derived from the HGNC database. The "All Data" version of the database output was downloaded in text format from the HGNC website[4]. HGNC assigned four types of dates to each gene: 1) *Date Approved*, which represents the date when the gene symbol and name were approved by HGNC, 2) *Date Modified*, which represents the date when the database entry was modified by HGNC, 3) *Date Symbol Changed*, which represents the date on which the gene symbol was last changed by HGNC from a previously approved gene symbol, and 4) *Date Name Changed*, which represents the date on which the gene name was last changed by HGNC from a previously approved gene name. *HGNC_approved* was based solely on *Date Approved* because the other dates are less informative as to when a gene was first integrated into the HGNC database. The values of *HGNC_approved* lie in the range [1986, 2007].

4.2.2 Consensus gene age

Based on *PubMed_earliest*, *OMIM_earliest*, *OMIM_creation* and *HGNC_approved* two age estimates were derived for each of the 28,424 human genes:

1. Literature-based age. This provides an approximate measure of the date on which a gene was first described in the scientific literature. It was calculated by averaging the values from *PubMed_earliest* and *OMIM_earliest*. Its values range from 1892 to 2007.

[4] http://www.genenames.org/data/gdlw_index.html; file downloaded on 15 Nov 2007.

2. Database-based age. This represents the date on which a gene was first integrated into the public database. It was calculated as the mean of *OMIM_creation* and *HGNC_approved.* Its values range from 1986 to 2007.

The distributions of the literature- and database-based ages are shown in panels (a) and (b) of Figures 4.14, respectively. Literature-based age registered a pronounced peak at year 2002. This could in part be due to the prominence of Human Genome Project around that period, which triggered an active phase of research into the sequencing and characterisation of human genes.

As shown in Figure 4.14(c), there is a good correlation between the literature- and database-based ages. Therefore, in order to provide a single measure of relative (rather than absolute) gene age, a 'consensus gene age' was calculated for each gene by taking average of the literature- and database-based ages. The values of consensus gene age calculated using this method range between 1939 and 2007 and its distribution is shown in Figure 4.14(d).

In the context of this analysis, a gene with a consensus age of 1990 implies that it was discovered in approximately 1990, and is considered "older" (that is, has been studied for longer) than a gene with a consensus age of 1998. Throughout the rest of this Chapter, the concept "older genes" will be used to refer to genes that have been known for longer, whilst "younger genes" referring to genes that have only been described more recently.

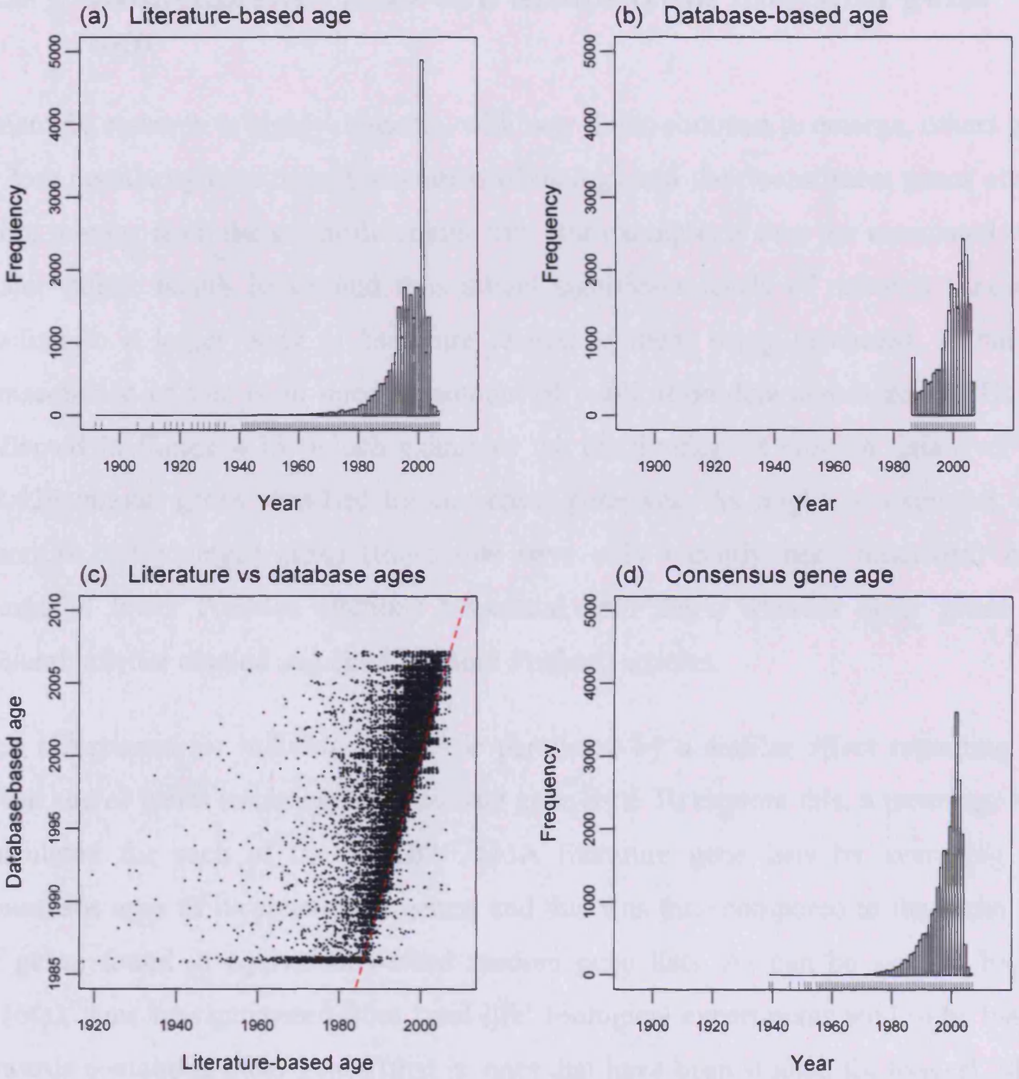


Figure 4.14: Distributions and characteristics of the literature-based age, database-based age and consensus gene age

(a) Histogram of literature-based age. (b) Histogram of database-based age. (c) Scatter plot of literature- and database-based ages. The red dashed line represent the $y = x$ line. Note that a small amount of 'noise' has been added to the data to break ties. This operation is only for plotting purpose and does not affect the calculation of consensus gene age. (d) Histogram of consensus gene age.

4.3 Relationship between annotation bias and gene age

Scientific research is highly dynamic, with new fields continue to emerge, others gain or lose popularity over time. Some areas of biology and their constituent genes attract more interest from the scientific community (for example, if they are associated with major public health issues and thus attract significant levels of research funding), leading to a larger body of literature related to them being generated. A natural consequence of this is an unequal amount of publication data across genes. This is reflected in Figure 4.15, which examines the distribution of citation data over the 28,424 human genes stratified by consensus gene age. As might be expected, one trend is that younger genes (those that have only recently been described) have markedly fewer PubMed citations associated with them, whereas older genes are generally better studied and cited by more PubMed articles.

The effect seen for individual genes is paralleled by a similar effect regarding the mean age of genes in biologically-derived gene lists. To explore this, a mean age was calculated for each of the 52 HG-U133A literature gene lists by averaging the consensus ages of its constituent genes, and this was then compared to the mean age of genes found in equivalently-sized random gene lists. As can be seen in Figure 4.16(a), gene lists generated from "real-life" biological experiments tend to be biased towards containing older genes (that is, ones that have been studied for longer). Most literature-derived gene lists have an overall consensus age that is older than the mean age of a random gene list (i.e. 1996 in this case). There is a strong trend whereby those gene lists showing an excess of PMID annotation are also those whose constituent genes have been known for longer (Figure 4.16(b)). Similar trends were also observed for gene lists derived from the human HG-U133 Plus 2.0 array when a similar analysis was carried out (Figure 4.17).

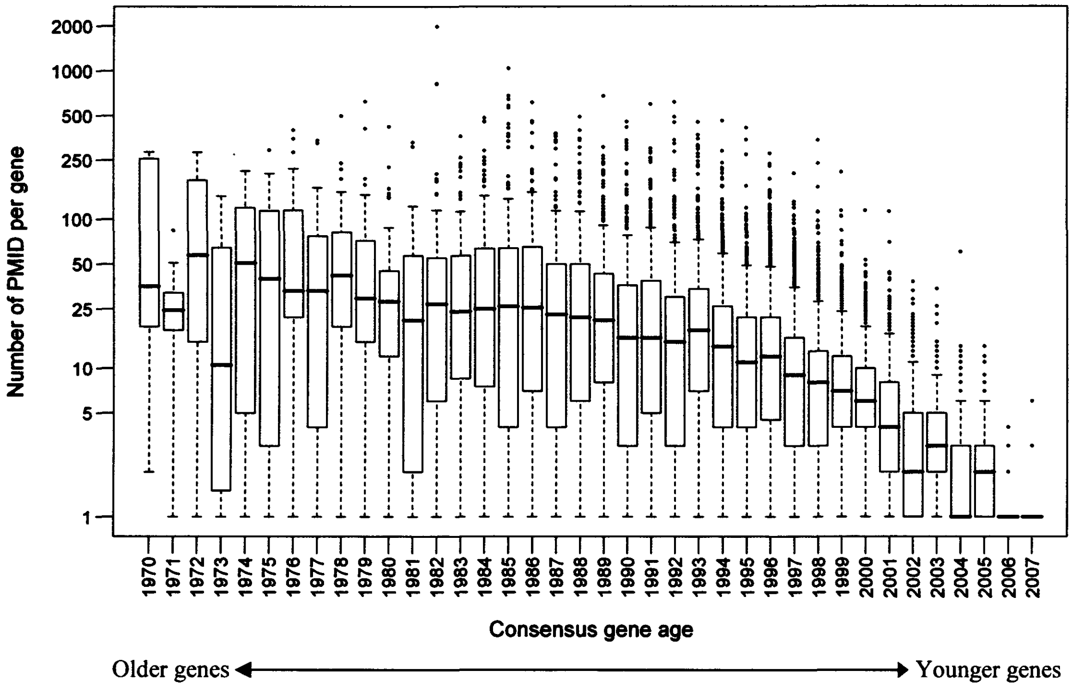


Figure 4.15: Amount of citation data by consensus gene age

The 28,424 human genes were split into groups according to their consensus gene age and the amount of PMIDs per gene across these age groups were compared. Age groups corresponding to consensus gene age below 1970 were excluded because there are not enough observations in them (less than five in most cases) for producing representative plot. The *y*-axis is on logarithmic scale.

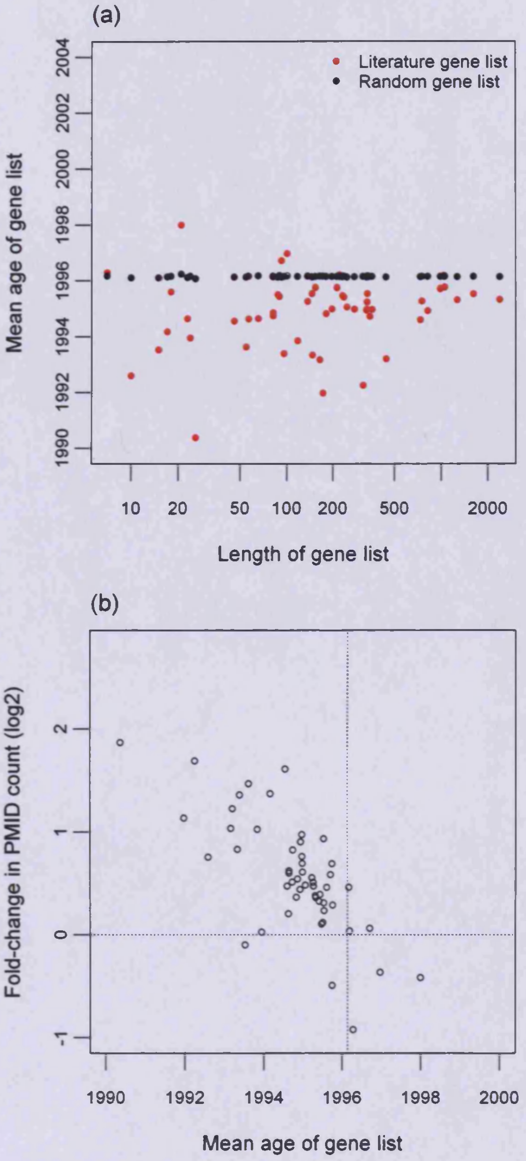


Figure 4.16: Gene age and annotation bias analysis for gene lists derived from HG-U133A

A mean age was calculated for each of the literature gene lists by averaging the consensus ages of its constituent genes. (a) The mean age of the literature gene lists were compared to that of the random gene lists. (b) Fold-change in PMID was calculated by dividing the number of PMIDs associated with a literature gene list by the average PMID count in 1,000 equivalently-sized random gene lists. The vertical dashed line represents the mean age of a random gene list, which is 1996 in this case; the horizontal dashed line represents the level at which there is no difference between the numbers of PMIDs associated with the literature and random gene lists.

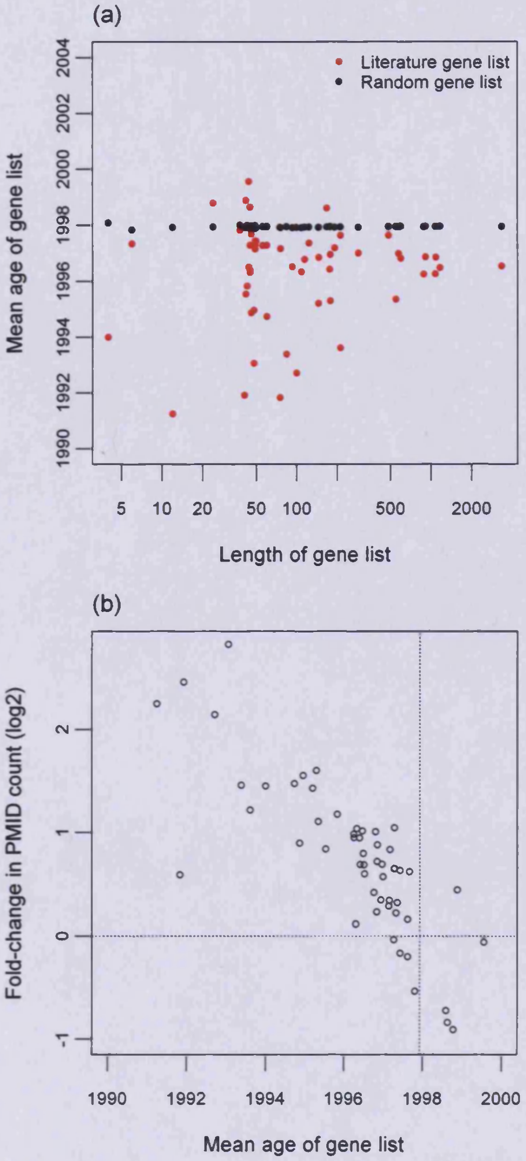


Figure 4.17: Gene age and annotation bias analysis for gene lists derived from HG-U133 Plus 2.0

A mean age was calculated for each of the literature gene lists by averaging the consensus ages of its constituent genes. (a) The mean age of the literature gene lists were compared to that of the random gene lists. (b) Fold-change in PMID was calculated by dividing the number of PMIDs associated with a literature gene list by the average PMID count in 1,000 equivalently-sized random gene lists. The vertical dashed line represents the mean age of a random gene list, which is 1998 in this case; the horizontal dashed line represents the level at which there is no difference between the numbers of PMIDs associated with the literature and random gene lists.

4.4 The temporal dynamics of scientific research and annotation bias

To gain an insight into the trends in biological research over the years, a basic analysis was performed based on an examination of the distribution of gene age and the level of research activity across different biological topics. In this analysis, biological process terms in the GO annotations were used as the basis for defining biological topics. The results are presented in Figure 4.18.

In Figure 4.18(a), the rows of the heatmap correspond to the biological processes and the columns correspond to the consensus gene ages. Each cell in the heatmap represents the fraction of genes that falls into the corresponding consensus age group for a specific biological process. For example, the GO term "immune response (GO:0006955)" contains 43 genes with an age of 1997, and there are a total of 2,847 genes in that age group across all data; the fractional gene count in this case is thus 43/2,847 = 0.015. The colour intensity is proportional to the value of the cell, with a more intense colour (deeper orange) indicates a higher fractional gene count. For each biological process, a mean age was calculated by averaging the consensus gene age of its constituent genes, and this was then used to arrange the rows such that the "younger biological processes" (i.e. processes that are biased towards more recently discovered genes) are positioned at the top of the heatmap while the "older biological processes" (i.e. processes that are biased towards genes that have been known for longer) are at the bottom. The dynamics of these biological processes were estimated by determining the number of papers associated with genes involved in these biological processes for each publication year between 1985 and 2007. Due to the skewed nature of the citation count distributions, the absolute counts were subsequently divided by the total number of papers published in each year so that the citation counts can be compared directly across time. The dynamics of the 20 oldest and 20 youngest biological processes are shown in panels (b) and (c) of Figure 4.18, and the identities of these biological processes were shown underneath the plots.

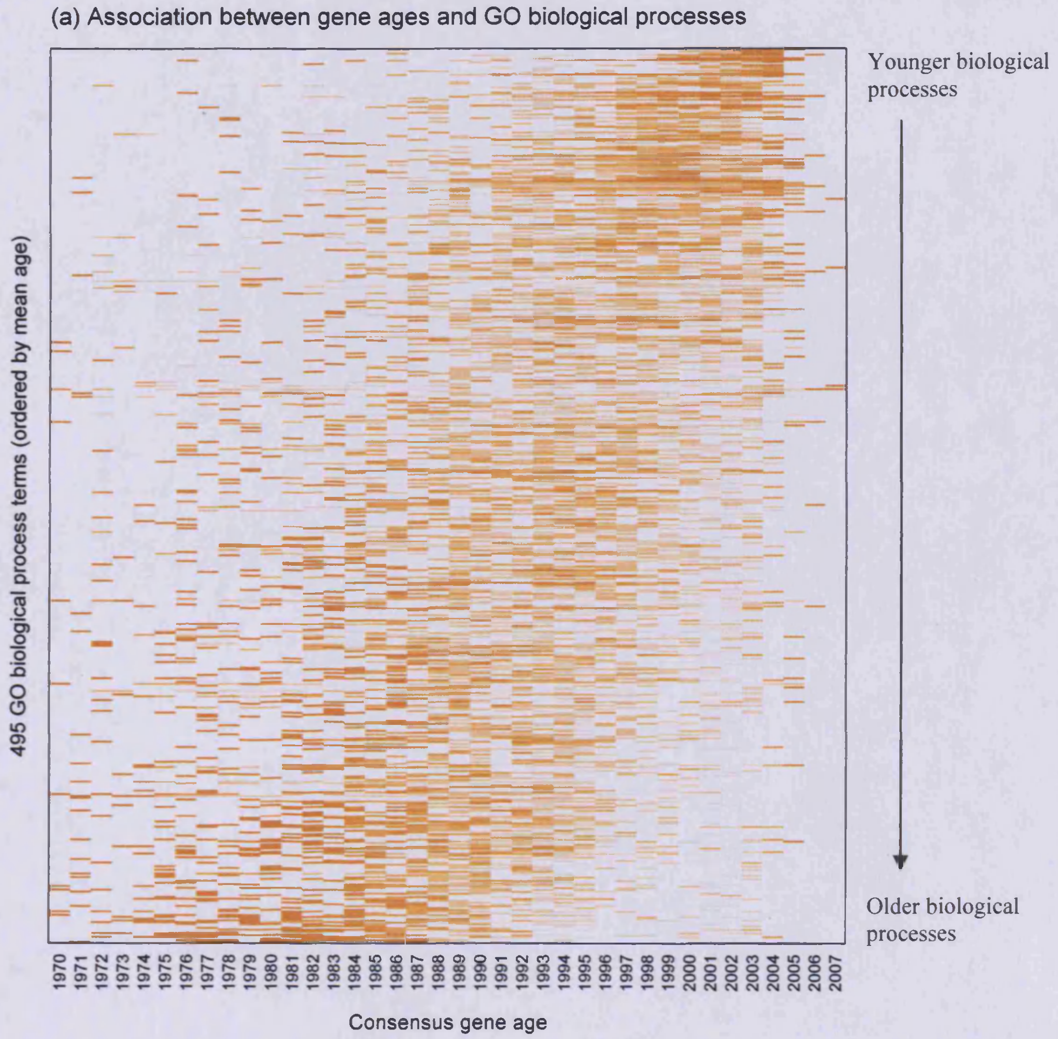
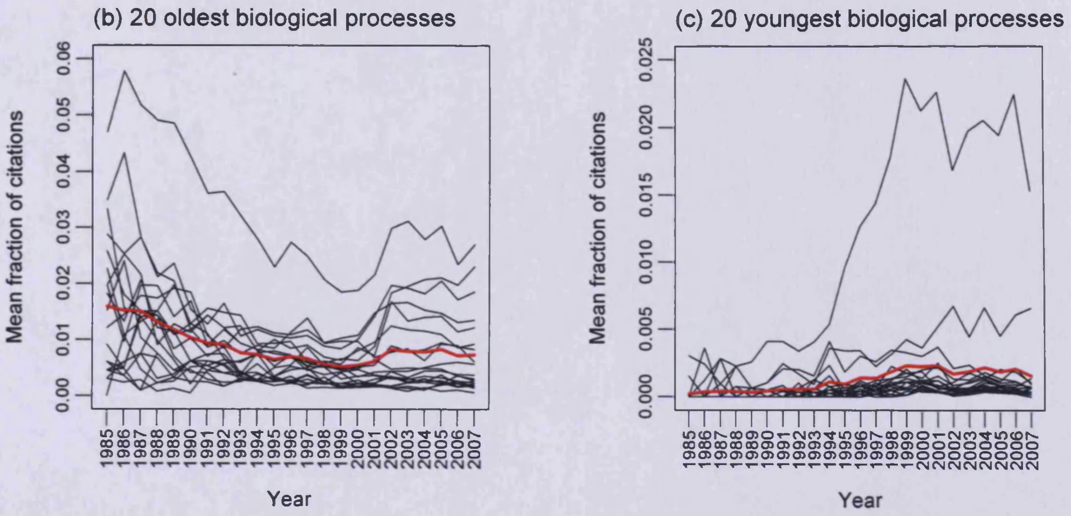


Figure 4.18: Trends in biological research and gene age

(a) Heatmap showing the age distributions of genes annotated to 495 GO biological process (bp) terms. The colour intensity in the heatmap was proportional to the value of the cell for which darker colour (deeper orange) indicates higher fractional gene count. (b) Temporal dynamics of 20 oldest biological processes. The identities of these processes are listed in the table below the plot. (c) Temporal dynamics of 20 youngest biological processes. The identities of these processes are listed in the table below the plot. For (b) and (c), an overall trend was calculated as the mean fraction of citations per year and shown as a red trend line in each plot.

(continued over the page)

Figure 4.18 (continued)


(b) 20 oldest biological processes

(c) 20 youngest biological processes

20 oldest GO (bp) terms	Mean Age
Heme biosynthetic process	1983.10
Complement activation, classical pathway	1984.07
Complement activation, alternative pathway	1985.09
Acute-phase response	1986.52
Antigen processing and presentation of peptide antigen via MHC class I	1986.62
Oxygen transport	1986.78
Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1986.83
Glycolysis	1987.12
Neutrophil chemotaxis	1987.36
Cytolysis	1987.43
Purine nucleotide biosynthetic process	1987.54
Blood coagulation	1987.61
Blood pressure regulation	1987.92
C21-steroid hormone biosynthetic process	1988.00
Iron ion homeostasis	1988.05
Pregnancy	1988.10
Positive regulation of peptidyl-tyrosine phosphorylation	1988.20
Platelet activation	1988.32
Gluconeogenesis	1988.42
Sex determination	1988.80

20 youngest GO (bp) terms	Mean Age
snRNA processing	2002.92
Glycerol metabolic process	2001.61
Cell wall catabolic process	2001.14
Regulation of Rab GTPase activity	2000.94
Aromatic compound metabolic process	2000.58
tRNA processing	2000.42
Sensory perception of smell	2000.29
Gamete generation	2000.20
Sensory perception of taste	2000.16
Autophagy	2000.11
Protein targeting to membrane	1999.91
Acyl-CoA metabolic process	1999.90
Lipid biosynthetic process	1999.52
Carbohydrate biosynthetic process	1999.42
Ubiquitin cycle	1999.41
GPI anchor biosynthetic process	1999.10
Regulation of ARF protein signal transduction	1999.06
Sulfur metabolic process	1999.00
mRNA catabolic process, nonsense-mediated decay	1998.96
rRNA processing	1998.93

It is apparent in Figure 4.18(a) that some biological processes tend to be associated with older genes. These include themes such as complement activation/antigen processing, heme biosynthesis and oxygen transport. Interestingly, analysis at the level of research activities associated with the 20 oldest biological processes shows that they undergone a gradual decrease in proportional popularity over time (see Figure 4.18(b) and table below the plot). On the other hand, the 20 "youngest: biological processes such as those involving snRNA/tRNA processing, regulation of signal transduction and perception of smell, show a continuing rise in popularity between 1994 and 2007 (see Figure 4.18(c) and table below the plot).

These findings are consistent with the idea that biological science is in a state of continuous flux, where some areas of biology are more thoroughly researched, with genes participating in these processes being generally discovered earlier in time and better studied.

4.5 Discussion

This Chapter describes a hitherto under-appreciated feature of gene lists derived from a typical microarray experiment, which is that they tend to have a greater level of associated PubMed articles than would be expected by chance alone. In this Chapter, potential causes of this bias were investigated, with particular reference to gene age and the historical development of scientific research activity. It was found that gene lists generated from real-life biological experiments tends to favour groups of genes and more established areas of biology that have been studied for longer, and for which a greater amount of published literature are available.

Whilst giving an insight into trends within biological research and the progress of scientific endeavour, the consequence of annotation bias for text-based ORA is a negative effect on the performance of standard hypergeometric distribution-based approach (see Chapter 3 for examples). This is because the background frequencies of some tokens, in particular the common and uninformative terms, can become under-estimated, resulting in more impressive hypergeometric p-values.

The next Chapters describe several solutions to address this, by taking into consideration the imbalance in the published literature. The first is based on the use of a permutation test, which makes no assumption about the underlying data distribution. The features and performance of this approach are presented in the next Chapter.

Chapter 5

ORA based on the use of a permutation test

5.1 Introduction

An underlying concept in text-based ORA is that if a list of differentially expressed genes contain subsets involved in one or more common functional roles, then it should be possible to detect and identify this based on similarities in the text of the abstracts corresponding to these genes. The abstracts associated with a gene list would therefore have words and concept terms in common, which can be broadly categorised into three groups: (i) biological terms that carry meaning related to specific biological functions, such as 'mitosis', 'apoptosis' and 'methylation'; (ii) biological terms that are relatively non-specific with regard to biological function, such as 'gene', 'protein' and 'clone'; (iii) common English words that occur frequently in all abstracts but contain no useful biological information, such as 'the', 'of' and 'analysis'.

The first group of tokens is potentially the most semantically useful for establishing any functional relationships among genes. One goal of text-based ORA is therefore to determine if any biologically-specific terms as described in (i) above is significantly over-represented in a list of differentially expressed genes. Initial attempts to address this (Chapter 3) were based on the use of a parametric approach that assumes the distribution of token frequency (i.e. the number of genes associated with a specific token) in a gene list follow the hypergeometric distribution. While this approach has been successfully applied to many ontology-based enrichment analyses, it was found to be inappropriate for mining tokens extracted from PubMed abstracts due to an under-appreciated annotation bias problem. This results in biologically-derived gene lists having more PubMed citations associated with them than equivalently-sized gene lists created by random sampling from the genes included on the microarray that was used (Chapter 4).

5.1.1 Effect of annotation bias on token frequency distribution

To devise an appropriate method for mining tokens, it is important to understand the potential effect of annotation bias on token frequency distribution. As discussed in Section 4.1.3, the significance of certain tokens (especially the common terms) in highly-annotated gene list tends to be artefactually inflated under the classical hypergeometric distribution model. This is because the frequency distributions under the null hypothesis (i.e. the reference distributions used to derive the *p*-values) for these tokens are distorted when the gene list is associated with more PubMed articles (PMIDs) than expected by chance. This phenomenon is illustrated in the following analysis, using tokens in the ISG gene list as examples.

There are 78 genes (unique Entrez Gene IDs) in the ISG gene list, of which 68 are linked to at least one PMIDs in the text corpus used in this analysis (see Sections 2.3.1 and 2.3.2 for details in the construction of text corpus). These 68 genes have excellent coverage in scientific literature: they are cited by a total of 1382 PMIDs in the text corpus and the mean number of PMIDs per gene is 20.3. However, if one were to create a large number of random gene lists by sampling 78 genes without replacement from the HG-U133A chip, then on average only 625 PMIDs are expected to be associated with these random gene lists, and the mean number of PMIDs per gene will drop to 11.3. Therefore, there is a 2-fold difference between the observed and expected numbers of PMIDs. This excess annotation needs to be taken into consideration when deriving the reference distribution for inferring *p*-value. To demonstrate this idea, consider the simple analysis described below, in which the theoretical null distribution simulated under the hypergeometric model was compared with the empirical null distribution generated under two different settings.

In the first setting, the empirical null distribution for a specific token was generated based on 1000 random gene lists, each of which was created by selecting 78 genes randomly without replacement from the HG-U133A chip. When this empirical null distribution was compared to the theoretical null distribution generated with the `rhyper` function, it was found that there is a good agreement in terms of location and shape between the two distributions. This is shown in Figure 5.1 for three tokens 'interferon', 'replication' and 'after', each corresponding to a case from biological, non-specific biological and common term, respectively. The empirical and theoretical null distributions appear to come from the same population in all three types of terms.

In the second setting, the empirical null distribution for a specific token was derived from 1000 random gene lists created by a biased sampling approach, in an attempt to emulate the situation where there are more PMIDs than expected by chance. Random gene lists were created by sampling without replacement from the HG-U133A array but only gene list that contains exactly 68 annotated genes and 1382 PMIDs was retained during the randomisation process. As such, the random gene lists created were biased towards those for which the constituent genes are better annotated. Figure 5.2 compares the empirical null distribution derived from these 'highly-annotated random gene lists' with the theoretical null distribution generated with the `rhyper` function. It can be seen that the theoretical and empirical distributions have similar shapes; however, the excess PMIDs caused a rightward shift in the empirical null distribution; tokens with high occurrence in the background (such as the common terms) were generally affected to a greater extent.

The above observations suggest that if annotation bias is not present, then the frequency distributions of tokens in a given gene list will follow the hypergeometric distribution. However, if the gene list is generated from a microarray experiment focusing on well-studied areas of biology and are associated with a higher number of PMIDs than might otherwise be expected, then the null distributions could be under-estimated under the simple hypergeometric distribution model. A consequence of this is an apparent over-representation of many common and non-specific biological terms, as the hypergeometric p-values were inferred from distorted null distributions.

Figure 5.1: A comparison of the empirical null distribution and theoretical null distribution simulated under the classical hypergeometric distribution model
This figure shows the empirical and theoretical null distributions for three tokens in the ISG gene list. They correspond to three types of terms that one might find in a typical abstract: (a) 'interferon' represents a specific biological term; (b) 'replication' represents a non-specific biological term; (c) 'after' represents a frequently occurring English word. The theoretical null distribution was simulated with the R function `rhyper(nn, m, n, k)`. In this function, the parameter `nn` is the number of random observations to create, `m` is the number of genes that contain the token of interest in the reference population, `n` is the number of genes that are not associated with the token of interest in the reference population, and `k` is the number of annotated genes in the gene list. As an example, for the token 'interferon' shown in panel (a), the null hypergeometric distribution were created by using `rhyper(nn=1000, m=414, n=13441-414, k=78)`, where 13441 is the number of unique genes present on the HG-U133A chip. The empirical null distribution was generated from 1000 random gene lists, each of which was created by sampling 78 genes without replacement from the HG-U133A chip. Then, the numbers of genes that are associated with the selected tokens in each random gene list were determined. The theoretical (hypergeometric) null distribution is shown as a grey histogram and the empirical null distribution as a blue histogram.

(continued over the page)

Figure 5.1 (continued)

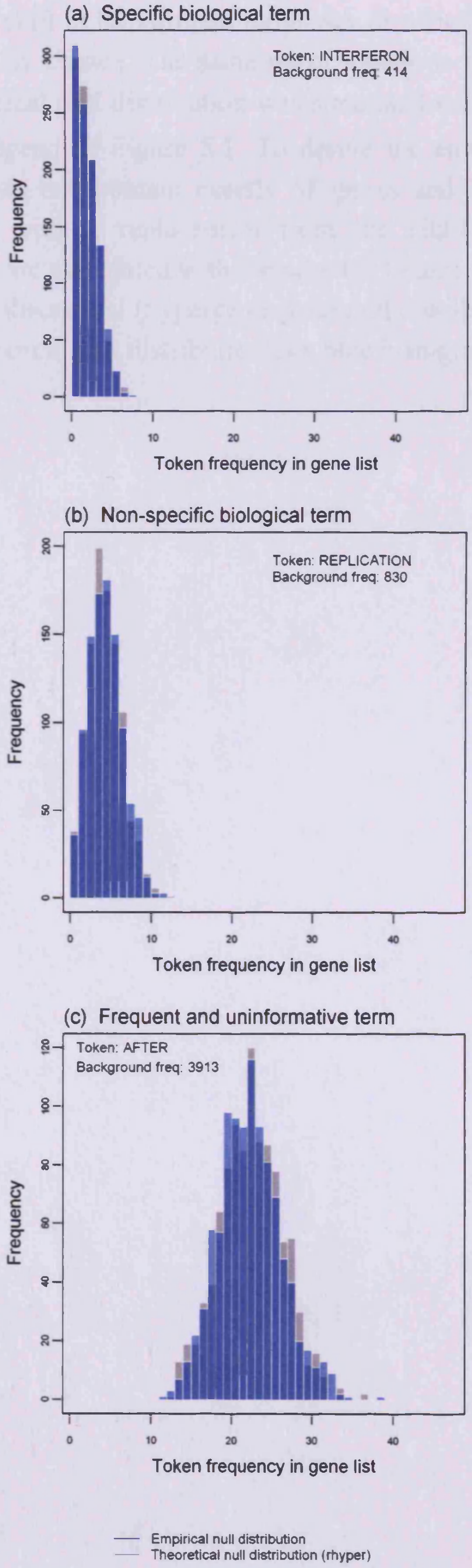
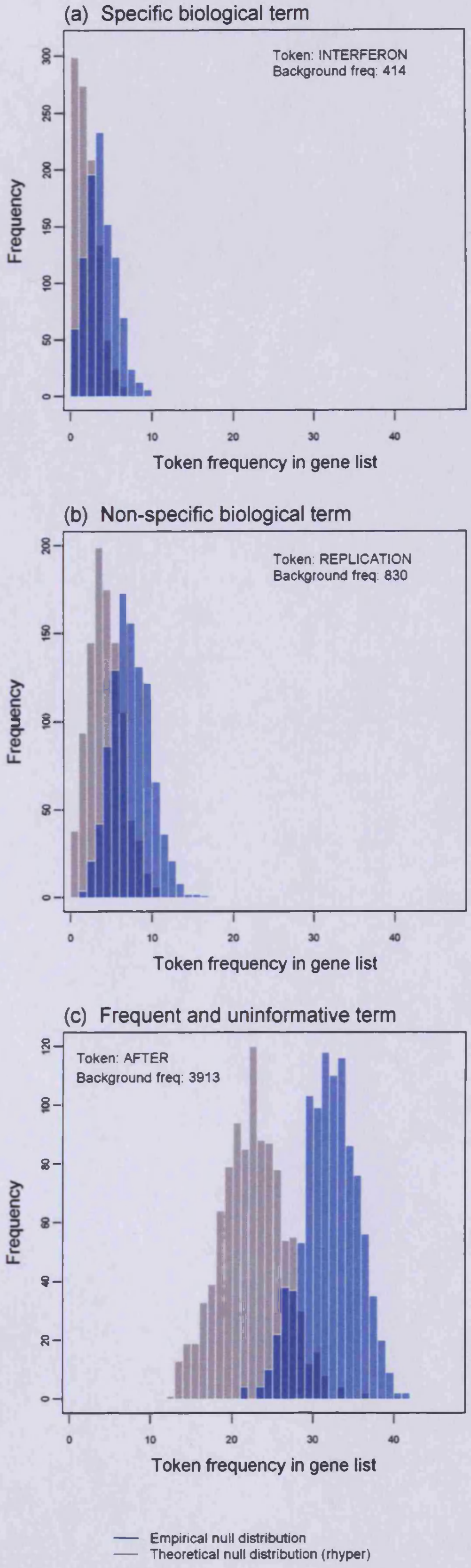


Figure 5.2: A comparison of the empirical null distribution generated with a biased sampling approach and theoretical null distributions simulated under the classical hypergeometric distribution model
This figure shows the shift in background frequency distribution when there are more PMIDs than expected by chance. The same set of tokens as shown in Figure 5.1 was used here. The theoretical null distribution was simulated with the `rhyper` function as described in the legend of Figure 5.1. To derive the empirical null distribution, 1000 random gene lists that contain exactly 68 genes and 1382 PMIDs each were created by sampling without replacement from the HG-U133A chip. Then, the numbers of genes that are associated with the selected tokens in each random gene list were determined. The theoretical (hypergeometric) null distribution is shown as a grey histogram and the empirical null distribution as a blue histogram.

(continued over the page)

Figure 5.2 (continued)



(a) Specific biological term

(b) Non-specific biological term

(c) Frequent and uninformative term

5.1.2 Parametric versus nonparametric approaches

In hypotheses testing, a statistical problem is classified as parametric if the underlying distribution is known, and it is nonparametric if the distribution from which the observations are sampled is unclear (Gibbons and Chakraborti 2003). The hypergeometric test is an example of a parametric method because it is based on a specific set of assumptions and parameters regarding the nature of the underlying population distributions. A shortcoming of this approach, as highlighted in the analysis presented in previous section, is that the patterns of token occurrence in a gene list cannot be fully captured by using the simple hypergeometric probability distribution model. This is because the classical hypergeometric distribution does not have enough distribution parameters to account for the excess citations inherent with a highly annotated gene list.

Having encountered problems with the classical parametric approach, a nonparametric approach that does not make formal assumptions about the underlying population distribution was explored. Briefly, this method generates a reference distribution for the token of interest by means of data permutation, based on which a p-value is then calculated and used to evaluate if the token is over-represented. The theoretical basis of this permutation test-based approach and how it is integrated into the ORA framework for mining tokens are the subjects of this Chapter. The performance of this approach is illustrated on two public microarray datasets in Section 5.4. Its practicalities and limitations will be discussed in Section 5.5.

5.2 Permutation-based testing

A popular technique for establishing statistical significance when the data do not meet the assumed parametric distribution is to resample the data. Resampling methods are computer-intensive procedures for probability estimation in which the significance of a test statistic is evaluated based on the empirical distributions generated from the observed data instead of the unknown parametric distribution (Fortin *et al.* 2002; Roff 2006).

5.2.1 Basic concepts

Permutation tests are a variant of resampling-based significance tests that draw random samples from the original data without replacement. An important assumption behind permutation tests is that the observations are independent, such that the rearrangements of the data are equally likely and exchangeable under the null hypothesis (Good 2005). This assumption does not apply perfectly in the context of text-based ORA analysis, as will be discussed later in Section 5.5.

The principle underlying a permutation test is to determine if a certain type of pattern or tendency that appears in data is simply a chance effect of randomness. Much of the pioneering work about permutation tests were made in the 1930s by Fisher (1935) and Pitman (1937). Recent work and a detailed exposition of permutation tests can be found in Good (2005) and Edgington and Onghena (2007).

A permutation test consists of first calculating a test statistic, S_{obs}, from the observed data. The test statistic is chosen to measure the extent to which the data show the pattern in question. Under the null model of no effect, the data are randomly rearranged and the test statistic recalculated, S_{rand}. This process is repeated a large number of times, via resampling without replacement, to produce a set of S_{rand} for constructing the reference distribution of the test statistic. By comparing the observed test statistic S_{obs} to the reference distribution, the probability of obtaining a value of S_{obs} as extreme or more extreme under the null hypothesis can be estimated. This probability is the *p*-value of the permutation test, which is simply the proportion of S_{rand} that are greater than or equal to S_{obs} amongst all the randomisation runs performed. The idea behind this is that if the pattern seen in the data is indeed due to chance alone (i.e. the null hypothesis is true), then S_{obs} should appear as a typical value drawn from the reference distribution. If S_{obs} takes an extreme value in the reference distribution, then the permutation test will produce a small *p*-value.

5.2.2 Exact and approximate randomisations

There are two ways of permuting the data to calculate *p*-values by means of randomisation: an exact method and an approximate method. The first calculates exact

p-values by exhaustive computation of all possible combinations of the data, and is therefore also known as the "exact randomisation test". Since the number of permutations increases as the factorial of the sample size, an exact test is not computationally practical for large datasets. For example, a sample size of 8 results in 40,320 possible permutations; but with only 2 extra observations the number of permutations increases to 3,628,800.

When the number of observations in a dataset precludes an exhaustive randomisation, an "approximate randomisation test" that estimates the p-values based on a subset of all possible permutations is often used. This approach is an asymptotic approximation of the exact test. The precision of the p-value thus calculated is determined by the number of randomisations performed. For instance, the smallest p-value that can be resolved based on 1000 randomisations is given by 1/1000 = 0.001.

5.2.3 How many randomisations are required?

Many researchers advocate using 10,000 or more randomisations to ensure that the estimated p-value is stable (Manly 2006). However, accepting this recommendation depends on how precise the need to estimate p-values is viewed. With ORA, there is an additional factor that needs to be taken into consideration when choosing the number of randomisations to use, which is that many tokens are tested simultaneously in a typical over-representation analysis, so the p-values derived from the permutation test have to be adjusted for multiple testing. This therefore suggests the use of a greater number of permutations than might otherwise be the case.

To illustrate this, consider a list of differentially expressed genes containing T tokens. To keep the overall probability of making any type I error at $\alpha = 0.05$ with the Bonferroni correction procedure, the empirical p-value derived from individual permutation test (p) must satisfy the condition $p \leq \alpha/T$. If the gene list is associated with, say, 1000 tokens, then the precision of p-value that must be resolved by individual test is $0.05/1000 = 5 \times 10^{-5}$. As mentioned previously, the smallest possible p-value that can be directly measured by permutation test is given by $1/N$, where N is the number of randomisations carried out. Therefore at least $1/(5 \times 10^{-5}) = 20{,}000$

randomisations would be needed in this case. Using less than 20,000 randomisations would result in none of the tokens could be called significant at the 0.05 significance level after Bonferroni correction.

For reasons of computational efficiency one would wish to keep *N* as low as possible. When the Bonferroni procedure is used to correct for multiple testing, the minimum number of randomisations N_{min} required can be estimated as $N_{min} = T/\alpha$. According to this formula, we can see that for a chosen significance level, the number of randomisations required is directly proportional to the amount of tokens subjected for testing. To help minimise the number of randomisations required it is possible to remove tokens that are associated with only one gene in the input gene list prior to performing the permutation test, because by definition they are of no utility in defining links between genes since this requires shared tokens. In the permutation-based ORA approach described below, *N* is provisionally set at 100,000. This setting is sufficient to detect a Bonferroni adjusted *p*-value as small as 0.05 when the proposed approach is applied to reasonably-sized gene lists containing less than 5000 tokens.

5.3 A permutation test-based ORA framework

The fundamental idea underlying the permutation-based ORA method described below is to create random gene list that matches the experimentally-derived gene list not only in the number of genes but also the amount of associated PMIDs. This is achieved by replacing the abstracts for each gene in the experimentally-derived gene list with other abstracts selected randomly without replacement from the text corpus. As such, the number of genes and abstracts (hence PMID) for each random gene list are kept the same as that of the experimentally-derived gene list. This permutation procedure is repeated 100,000 times to generate a null reference distribution for each token, and based on which the significance of each token is evaluated.

The permutation-based ORA method consists of the following steps:

1. *Token space reduction.* To minimise the multiple hypotheses testing issue, tokens with *List* frequency equal to one are removed.
2. *Specify null hypothesis and calculate the observed test statistic.* The null hypothesis is that the number of genes associated with a particular token of interest in the input gene list is not significantly different from that seen in an equivalently-sized gene list for which the constituent genes were picked at random from the background gene population. The test statistic used to test this hypothesis is the *List* frequency of each token, denoted as S_{obs}.
3. *Create random gene lists and recalculate the test statistic for the randomised data.* To account for the effect of annotation bias, the null reference distribution of S_{obs} is derived from random gene lists that match the input gene list in terms of (i) number of annotated genes, (ii) number of PMIDs, and (iii) partition pattern of associated PMIDs. Here, partition pattern is defined as the distribution of PMIDs per gene in the gene list. For example, if the first gene on the list is cited by 20 PMIDs, the second gene by 110 PMIDs, the third gene by 15 PMIDs, and so on. Then the partition pattern is expressed as {20, 100, 15,...}. Maintaining the partition pattern of PMIDs ensures that the annotation density of the random gene list is identical to that of the input gene list. Random gene list that meet the three criteria stated above is obtained by substituting the abstracts for each gene in the input gene list with other abstracts picked randomly without replacement from the text corpus. For each token under testing, its *List* frequency in the random gene list is determined.
4. *P-value calculation.* Step 3 is repeated for n = 100,000 times. A jackknifed *p*-value is calculated for each token as the proportion of times its frequency in the random gene list (R_i) is equal to or greater than that seen in the input gene list (S_{obs}) after a single observation is arbitrarily omitted from S_{obs}:

$$p = \frac{1}{n}\sum_{i=1}^{n}(R_i \geq S_{obs} - 1) \qquad \text{Equation (5.1)}$$

The jackknifing operation ascertains the stability of the empirical *p*-value by minimising the influence of fluctuations in sampling error that may occur due to a single deleted observation's uniqueness. The rationale underlying jackknife adjustment and the reasons for using it can be found in Section 3.2.4.

5. *Multiple testing correction and criterion of over-representation.* A token is considered as significantly enriched if its *p*-value is less than 0.05 after Bonferroni multiple hypothesis correction.

A flowchart summarising the computational steps described above is presented in Figure 5.3. The scripts for this method were developed and tested under R-2.6 and Perl v5.8.7.

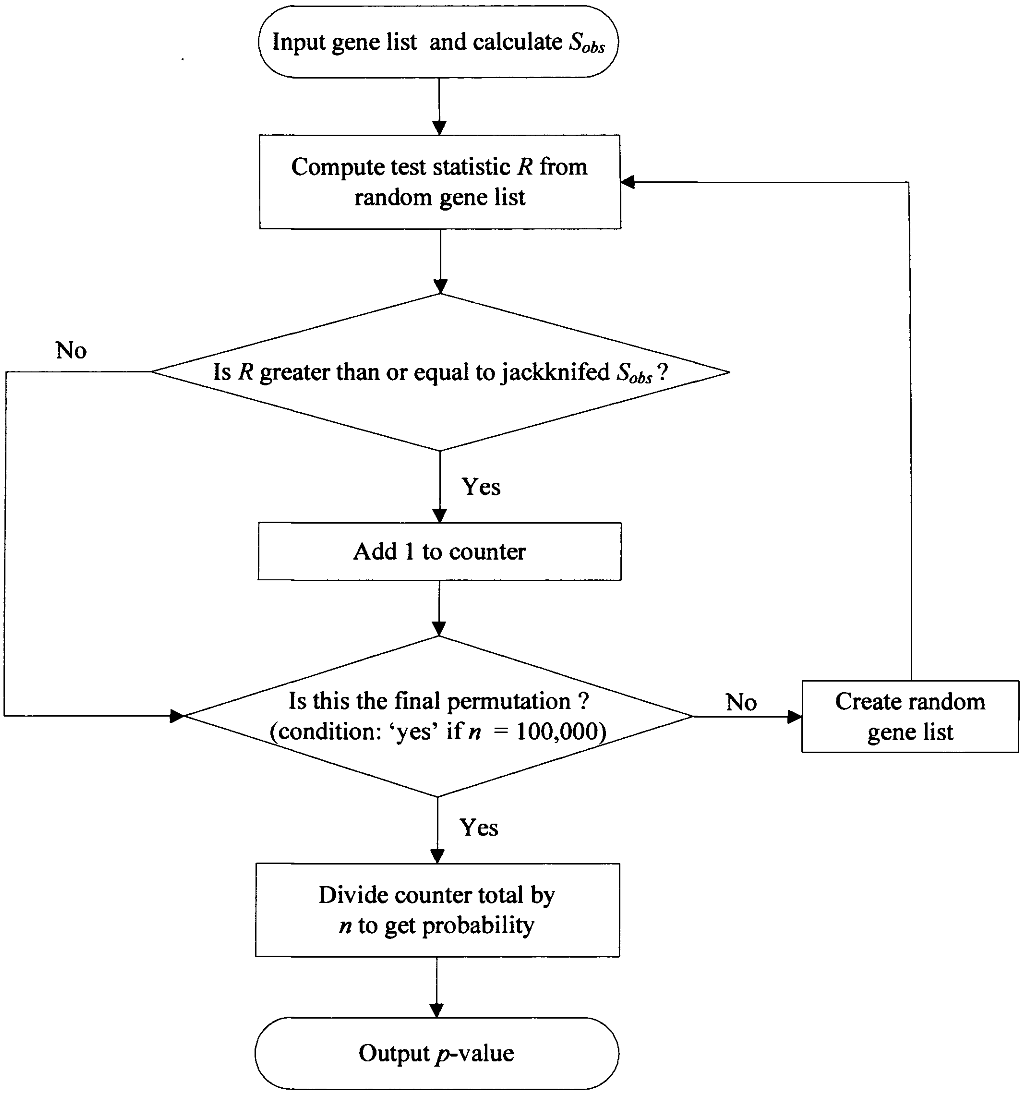

Input gene list and calculate S_{obs}
Compute test statistic R from random gene list
Is R greater than or equal to jackknifed S_{obs} ?
No
Yes
Add 1 to counter
Is this the final permutation ? (condition: 'yes' if n = 100,000)
No
Create random gene list
Yes
Divide counter total by n to get probability
Output p-value

Figure 5.3: Computational steps of the permutation test-based ORA approach
The first step is to determine the observed test statistic S_{obs}, which is the *List* frequency associated with each token in the input (experimentally-derived) gene list. Then random gene lists that match the input gene list in both gene and PMID counts are created, based on which the test statistic is recalculated. The test statistic derived from the randomised data is denoted as R. This permutation process is repeated 100,000 times. For each token under test, a counter is initialised and the tally is increased by 1 when the value of R obtained is greater than equal to the value of the jackknifed S_{obs}. Finally, a p-value is calculated by dividing the counter's sum by the total number of permutations performed.

5.4 Experiments and results

5.4.1 Performance on real datasets

The performance of the permutation test-based approach was tested on the ISG gene list from Sanda *et al.* (2006) and the mitosis gene list reported in Lee *et al.* (2004). Details of these two gene lists can be found in Sections 2.2.1 and 2.2.2, respectively.

Example 1: ISG gene list

The ISG gene list is associated with 11,709 unique tokens according to the text corpus created for the HG-U133A array. To reduce the multiple hypotheses testing problem the token space was reduced by removing tokens that are associated with only one gene in the gene list. After filtering, 4841 tokens remained for testing.

The significance of each of the 4841 tokens was determined by calculating an empirical *p*-value based on the creation of 100,000 random gene lists, each of which was matched for the number of genes and the amount of associated PMIDs. Since the smallest *p*-value that can be directly measured based on 100,000 permutations is 10^{-5}, the best possible Bonferroni adjusted *p*-value that could be detected is $10^{-5} \times 4841 = 0.04841$. The *p*-value of token for which none of the random gene lists (out of the total 100,000) has *List* frequency greater than that seen in the real gene list was provisionally set to $<10^{-5}$, and the corresponding Bonferroni adjusted *p*-value was set to <0.04841.

The results of this analysis are shown in Table 5.1. 23 tokens were identified as significantly over-representation (Bonferroni *p*-value ≤ 0.05). A comparison of these results with those obtained by using the classical hypergeometric test-based approach (cf. Table 3.2) reveal a good agreement between the two. For example, nine out of the top ten hits found by the hypergeometric test were also called significant by the permutation test. In fact, the set of significant tokens listed in Table 5.1 is a subset of the hits produced by the hypergeometric test-based method. However, the permutation test-based approach produced an improvement over the classical hypergeometric test-based approach, insofar as it successfully retained those biologically-plausible terms such as ‘interferon’, ‘IFN’, ‘antiviral’, whilst no longer called those less-specific terms as significant such as ‘line’, ‘intact’, ‘after’, ‘response’ and ‘synthesis’.

Table 5.1: Significantly over-represented abstract terms in the ISG gene list as identified using the permutation test

Term	*Chip* frequency	*List* frequency	Empirical p-value	Bonferroni p-value	Rank from ClassicalHG
INTERFERON	414	46	$< 10^{-5}$	< 0.04841	1
IFN	245	35	$< 10^{-5}$	< 0.04841	2
IFN-BETA	71	18	$< 10^{-5}$	< 0.04841	3
ANTIVIRAL	176	23	$< 10^{-5}$	< 0.04841	4
IFN-ALPHA	114	19	$< 10^{-5}$	< 0.04841	5
INDUCIBLE	1068	37	$< 10^{-5}$	< 0.04841	6
INTERFERON-ALPHA	59	14	$< 10^{-5}$	< 0.04841	7
INFECTION	1177	36	$< 10^{-5}$	< 0.04841	8
VIRAL	892	32	$< 10^{-5}$	< 0.04841	9
TREAT	1817	40	$< 10^{-5}$	< 0.04841	12
DSRNA	60	11	$< 10^{-5}$	< 0.04841	14
IMMUNITY	387	20	$< 10^{-5}$	< 0.04841	15
OLIGOADENYLATE	18	8	$< 10^{-5}$	< 0.04841	16
ISRE	31	9	$< 10^{-5}$	< 0.04841	18
LYMPHOBLASTOID	239	16	$< 10^{-5}$	< 0.04841	19
ISG	14	7	$< 10^{-5}$	< 0.04841	21
STOMATITIS	52	9	$< 10^{-5}$	< 0.04841	25
HLA-CLASS	11	6	$< 10^{-5}$	< 0.04841	26
ENCEPHALOMYOCARDITIS	16	6	$< 10^{-5}$	< 0.04841	30
OAS	10	5	$< 10^{-5}$	< 0.04841	40
EVASION	65	9	10^{-5}	0.04841	27
GAMMA-INTERFERON	44	7	10^{-5}	0.04841	44
INDIGENOUS	29	6	10^{-5}	0.04841	53

Over-represented terms were defined as having p-value ≤ 0.05 after Bonferroni correction. 100,000 randomisations were performed for the permutation test. 4841 tokens were being tested during the permutation test and the best possible Bonferroni p-value attainable is $10^{-5} \times 4841 = 0.04841$. Any term with an empirical p-value less than 10^{-5} is provisionally assigned a value of $< 10^{-5}$, and the corresponding Bonferroni p-value is set to be < 0.04841. For the purpose of comparison the rankings of the significant tokens as determined by the classical hypergeometric test-based approach are also shown and listed in the column ‘Rank from ClassicalHG’.

Example 2: Mitosis gene list

The mitosis gene list was based on a study by Lee *et al.* (2004), in which the gene expression profiles of human CD4$^+$ T cells at the early and late stages of differentiation were investigated using the Affymetrix HG-U133A array. Lee and coworkers annotated the gene list manually and suggested that these genes were mainly involved mitosis, cell cycle regulation or progression, DNA replication, recombination, or repair.

The 82 genes in the mitosis gene list are cited by 1160 unique PMIDs and associated with 10,725 unique tokens. This set of tokens were trimmed by omitting candidates linked to only one gene in the gene list, leaving 4424 tokens for mining. When these tokens were subjected to permutation testing, 13 were identified as significantly over-represented (Bonferroni p-value ≤ 0.05), as shown below.

Table 5.2: Significantly over-represented abstract terms in the mitosis gene list as identified using the permutation test

Term	*Chip* frequency	*List* frequency	Empirical p-value	Bonferroni p-value	Rank from ClassicalHG
MITOTIC	485	28	$< 10^{-5}$	< 0.04424	1
SPINDLE	298	23	$< 10^{-5}$	< 0.04424	2
MITOSIS	443	26	$< 10^{-5}$	< 0.04424	3
ANAPHASE	126	17	$< 10^{-5}$	< 0.04424	4
CHECKPOINT	267	19	$< 10^{-5}$	< 0.04424	5
KINETOCHORE	64	11	$< 10^{-5}$	< 0.04424	6
CENTROMERE	181	13	$< 10^{-5}$	< 0.04424	7
DIVISION	426	18	$< 10^{-5}$	< 0.04424	8
HELA	1393	31	$< 10^{-5}$	< 0.04424	9
PROLIFERATING	523	19	$< 10^{-5}$	< 0.04424	10
PROMETAPHASE	60	8	$< 10^{-5}$	< 0.04424	16
INTERPHASE	253	13	10^{-5}	0.04424	17
CONGRESSION	21	6	$< 10^{-5}$	< 0.04424	18

100,000 randomisations were performed for the permutation test. 4424 tokens were being tested during the permutation test and the best possible Bonferroni p-value attainable is $10^{-5} \times 4424 = 0.04424$. Any term with an empirical p-value less than 10^{-5} is provisionally assigned a value of $< 10^{-5}$, and the corresponding Bonferroni p-value is set to be < 0.0424. For the purpose of comparison the rankings of the significant tokens as determined by the classical hypergeometric test-based approach are also shown and listed in the column ‘Rank from ClassicalHG’.

As shown in Table 5.2, the significant terms appear relevant to the biology under studied. For examples, 'mitosis', 'mitotic', 'spindle', 'kinetochore' and 'anaphase' are related to the process of cell cycle, which are also comparable to the biological themes reported by the authors.

5.5 Discussion

Analyses in this Chapter based on selected datasets suggests that the proposed permutation test-based ORA approach is capable of producing biologically-plausible results, while at the same times avoids calling as significant common terms that are artefactually over-represented due to annotation bias. This approach has the advantages of flexibility and relative ease of implementation as it supports significance testing without making formal distributional assumptions.

The main limitation of the proposed approach is that it is extremely computationally intensive, requiring six hours on a standard desktop computer to analyse each of the two datasets described above. The computation time increases dramatically with the size of gene list and the number of tokens to be tested. There are two additional issues that merit attention and offer the possibility of further improvements.

The first issue is related to the precision of the p-values produced by the proposed method. Since the empirical p-value is computed as the fraction of simulated counts more extreme or equal to that observed (Equation 5.1), poor estimations may arise in cases where the actual p-values are small, or when the token being tested has very low background occurrence frequency. For example, the p-value associated with token for which none out of 100,000 of the random gene lists created has *List* frequency greater than that seen in the real gene list would be computed as 0/100,000 = 0. In the current implementation, the empirical p-values associated with such tokens are provisionally set to 1/100,000 (i.e. $<10^{-5}$). A problem with this remedy is that due to the ties in the assigned p-values, there is no way to rank these tokens according to their relative importance in the gene list (see Table 5.1 for examples). The precision of the empirical p-values estimated could be improved by increasing the number of permutations performed; but this would entail an impractical number of permutations in order to control the family-wise error rates, especially if the input gene list is associated with a large number of unique tokens, making the permutation test-based ORA procedure computationally-intractable and extremely time-consuming.

The second issue is related to the assumption behind the proposed method. The independence assumption is used in bioinformatics and text mining because it allows the construction of an easy-to-fit probabilistic framework, even if it does not always apply perfectly (Raychaudhuri 2006). The widely implemented statistical models for testing the over-representation of GO terms within lists of differentially expressed genes, such as the hypergeometric distribution and Fisher's exact test, are built on the assumption that genes are independent. As with these conventional ORA approaches, the independence assumption is essential to the construction of the null distribution in permutation-based testing. The permutation test-based ORA approach described in this Chapter assumes that the PMIDs (and hence tokens) associated with the genes in a given gene list are independent. However, this assumption may not be perfect, because lists of differentially expressed genes derived by experimental means are likely to have correlated expression levels and hence similar biology, with the result that the tokens in the associated PMIDs may also be correlated.

To account for any inter-relatedness among genes that might exist, an alternative permutation approach, which treats the subjects (e.g. class labels, phenotypes) as opposed to the genes as the sampling units, could be considered. This method would require the actual gene expression measurements. The idea would be to permute the class labels among individuals, re-run the gene-level analysis and rank the genes in order of significance, followed by generating random gene lists by running down the ranked list until the total number of PMIDs is identical to that of the actual (real) gene list. The reference distribution for each token would then be derived as usual by repeating the above permutation procedure a large number of times, based on which a p-value could be calculated.

While a subject sampling-based permutation model might allow for the correlation structure between genes, it suffers several limitations: 1) the random gene list created with this approach will match the real gene list only in the number of associated PMIDs but not the number of genes, thus introducing another source of variation into the analysis; 2) the total number of permutations that can be performed is limited by the number of subjects featured in the microarray experiment, so the proposed method may not be applicable to small dataset. For example, a microarray experiment consisting of eight samples (say four treatments versus four controls) would entail 8! = 40,320 different permutations, which is insufficient if 5000 tokens are to be tested simultaneously at a family-wise error rate of 0.05.

In conclusion, the results presented in this Chapter suggest that it is possible to account for the effect of annotation bias by means of a permutation test-based ORA framework. However, the practicality of this approach is limited by it being computationally intensive and time-consuming, making it unsuitable for routine analysis. The independence assumption remains a subtle and controversial issue; how to formally adjust for the correlation structure in the data during permutation testing is beyond the scope of this work and is an issue for future study.

In Chapters 6 and 7 two computationally tractable approaches, which are based on the detection of outliers and the extended hypergeometric distribution, will be presented and discussed.

Chapter 6

ORA based upon the detection of outliers

6.1 Introduction

In this Chapter, an outlier detection-based approach for identifying terms in PubMed abstracts that are significantly over-represented in a list of differentially regulated genes is described. This method is motivated by the observation that, on a scatter plot of *Chip* frequency (the number of genes that are associated with a token of interest on the entire chip) versus *List* frequency (the number of genes that are associated with a token in the query gene list), there are a set of biologically-plausible terms that deviate substantially from the main data cluster and appear as outliers. This feature is illustrated in Figure 6.1(a), where the *List* and *Chip* frequencies associated with tokens in the ISG gene were plotted in two dimensions. On this plot the majority of the data points, each of which represents a token, form a funnel-shaped distribution characterised by a high variance in the values measured at the low end of the scale (to the left) and a low variance in the values measured at the high end of the scale (to the right). Common English words like 'the', 'of' or 'and' are always located at top right portion of the plot (tip of the funnel) because they are present in almost all abstracts, and therefore have high *Chip* and *List* frequencies. On the other extreme (to the left of the plot) are rarely observed tokens that are shared by only few genes in the gene list, but may be associated with a few or many instances on the entire chip. A set of outlying observations can be seen to deviate substantially from the main cloud of data (the funnel-shaped cluster). Upon closer inspection, it can be seen that these tokens are pertinent to the biology of the ISG gene list, including 'interferon', 'ISRE' and 'antiviral'. ISRE corresponds to the interferon-stimulated response element; it is a

DNA sequence in the promoters of ISGs (interferon-stimulated genes). The antiviral effect of interferon is mainly mediated via the JAK-STAT signalling pathway, which leads to the activation of transcription factor complexes such as interferon-stimulated gene factor 3 (ISGF3). ISGF3 then binds to the ISRE sequence in ISGs promoters, resulting in the transcriptional activation of many antiviral proteins such as MxA, 2'-5' oligoadenylate synthetase. These proteins are able to interfere with viral replication and therefore protect cells from infections by viruses (Platanias 2005).

When the same type of scatter plot was produced for tokens associated with an artificial gene list that matched the ISG gene list in terms of gene and PMID counts, a similar funnel-shaped distribution was obtained. However, this time all tokens were found to be tightly clustered, and as expected no obvious outliers are present (Figure 6.1(b)). One explanation for these observations is that the majority of the tokens in a gene list will not be over-represented because they are either common words or non-specific biological terms that are shared by most abstracts, or rare words that are not biologically interesting. These will form the main cloud of data points on the plot (the funnel-shaped cluster). On the other hand, biologically-plausible terms in a list of differentially expressed genes will be associated with more genes in the gene list and assume different token frequency distributions, thus appear to be separated from other observations in the background.

In order to identify these biologically-interesting terms, an outlier detection procedure that tests for discordancy based on the calculation of *Z*-scores and *p*-values was developed. The rest of the Chapter is constructed as follows. Three outlier labelling techniques that are related to the proposed approach will be introduced in Section 6.2. The outlier detection-based ORA framework developed for mining textual information will be detailed in Section 6.3. In Section 6.4, the overall performance of the proposed approach will be evaluated through simulation studies and application to three biological datasets. The normality assumption underlying the proposed approach will also be tested.

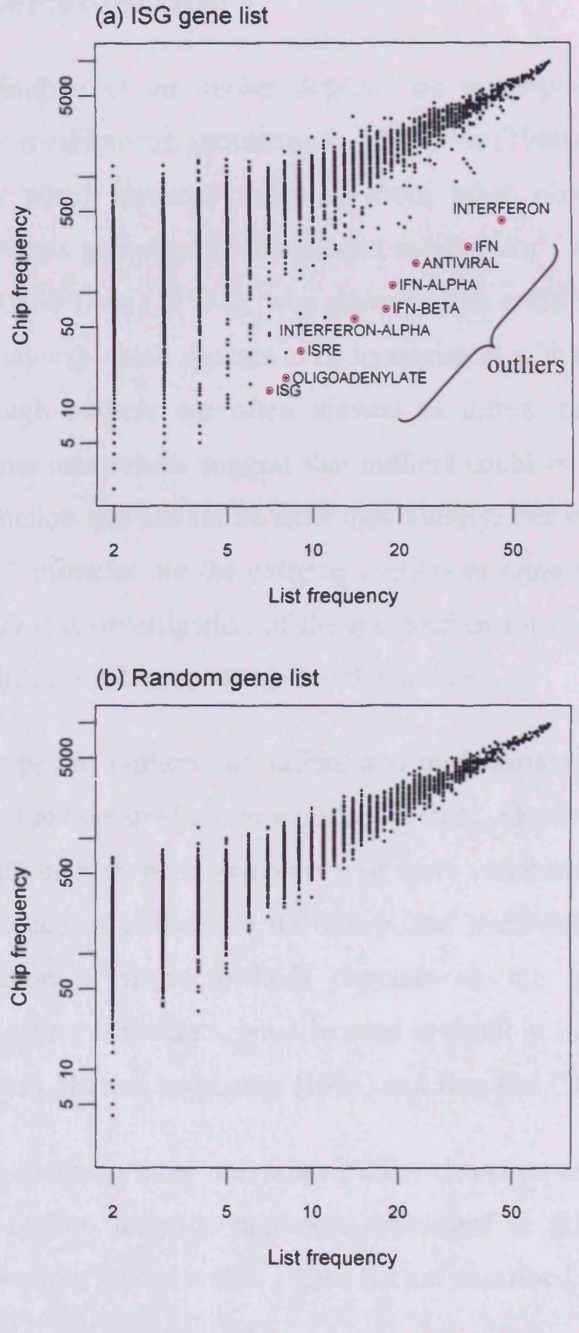


Figure 6.1: Scatter plots of *Chip* versus *List* frequencies showing (a) the presence of outliers in a real gene list, and (b) the absence of outliers in a random gene list
Each observation on the scatter plot represents a token. Only tokens that are associated with at least two genes in the gene list were plotted. For comparison purposes, the random gene list was created in such a way that it matches the ISG gene list in the amounts of genes and associated PMIDs. Both axes are on logarithmic scale ($\log_{10}$).

6.2 Outlier detection

The precise definition of an outlier depends on assumptions regarding the data structure and the mechanisms generating it. Hawkins (1980) described an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". A similar definition was given by Barnet and Lewis (1994), who described an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". Although outliers are often viewed as errors, or cases that do not fit expectations, some researchers suggest that outliers could be special cases that carry important information and should be examined closely. For example, Kruskal (1988) suggested that "...miracles are the extreme outliers of nonscientific life. It is widely argued of outliers that investigation of the mechanism for outlying may be far more important than the original study that led to the outlier".

There are two types of outliers: univariate and multivariate. Univariate outliers are cases that possess extreme values on a single variable. Multivariate outliers are cases with unusual combinations of scores on two or more variables. Various approaches to test for the presence of outliers in univariate and multivariate datasets have been proposed. Selection of these methods depends on the type of target outliers, distribution and nature of the data, are discussed in detail in Hawkins (1980), Iglewicz and Hoaglin (1993), Barnett and Lewis (1994) and Ben-Gal (2005).

In the following sections, three univariate outlier detection methods that are directly relevant to the outlier detection approach developed in this work for identifying biologically-interesting tokens within a gene list are described.

6.2.1 *Z*-scores

One approach to detect for outliers is to assume that they have a different distribution from the remaining observations generated from a target distribution (Davies and Gather 1993). With the *Z*-score method, the target distribution is taken to be a normal distribution. The outlier identification problem is then translated into the problem of identifying those observations that deviated substantially from the normal distribution.

The normal distribution has several important characteristics. The distribution is symmetrical around the mean, with approximately 68%, 95% and 99.7% of the data fall within 1, 2 and 3 standard deviations of the mean, respectively (Figure 6.2). Based on these, the likelihood of seeing extreme values in a normally distributed data can be estimated.

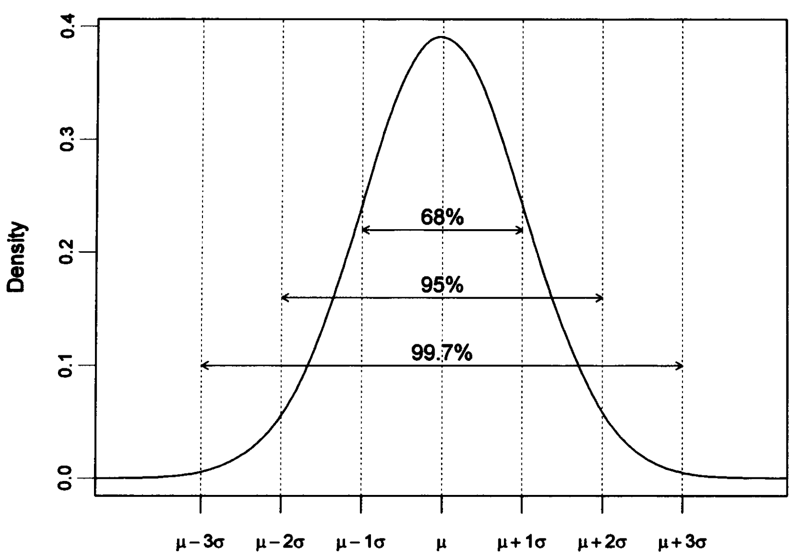


Figure 6.2: Probability density plot for a standard normal distribution
μ represents mean and σ represents standard deviation.

The *Z*-score method is based on the property of a normal distribution. Suppose the univariate variable *Y* follows a normal distribution, then the *Z*-score is distributed with mean zero and standard deviation of one, denoted as *N* (0, 1). For observations $Y = \{y_1, y_2, \ldots, y_n\}$, the *Z*-scores are given by:

$$Z_i = \frac{(y_i - \bar{y})}{sd}, \text{ where } sd = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \qquad \text{Equation (6.1)}$$

where $\bar{y}$ represents the sample mean, and *sd* is the standard deviation.

A general guide is to flag as outliers those observations with *Z*-scores exceeding 3 in absolute value, that is observations that lie 3 standard deviations away from the mean (Iglewicz and Hoaglin 1993). Alternatively, a *p*-value can be inferred from the normal distribution as a means for evaluating the significance of the *Z*-score. For example, in a two-tailed test, the absolute *Z*-score associated with a 95% confidence level are 1.96 and the corresponding *p*-value is 0.05. So if the *Z*-score lies somewhere between -1.96 and +1.96, then the *p*-value will be greater than 0.05. If the *Z*-score falls outside this range, then the *p*-value will be less than 0.05. The critical *Z*-scores for two-tailed and one-tailed test are different. In one-tailed test, the critical *Z*-score corresponding to a *p*-value of 0.05 is 1.645. Regardless of whether it is a two-tailed or a one-tailed test, the key idea is that a large absolute *Z*-score value (i.e. one located in the tails of the normal distribution) and small *p*-value are indicative of an unusual observation that deviates from the normal distribution. Therefore, an observation may be considered as outlier if the associated *p*-value is less than or equal to 0.05.

6.2.2 *M*-scores

The sample mean and standard deviation used in the *Z*-score method can be affected by a single or a few unusual observations. The standard deviation, for example, can be inflated by neighbouring extreme values, resulting in the so-called ‘masking’ effect. Masking occurs when discordant observations cancel the effect of more extreme observation and prevent the outlier detection procedure from declaring any of the observations as outliers (Acuna and Rodriguez 2004; Hadi 1992). This problem can be solved by using resistant (robust) estimators such as the sample median and MAD (median absolute deviation) in place of the mean and standard deviation (Iglewicz and Hoaglin 1993).

Median and MAD are said to have a high *breakdown point*. The breakdown point of an estimator is defined as the largest percentage of the data that can be replaced by arbitrary values without causing the estimator to become infinite (Hampel 1971). For example, the median will tolerate up to 50% gross error before it can be made arbitrarily large; we say that its breakdown point is 50%. In contrast, the breakdown point for the mean is 0%, i.e. even one observation moved to infinity would make the mean infinite. The breakdown points for sample standard deviation and MAD are approximately 0% and 50%, respectively.

The "modified *Z*-score", or *M*-score, is an alternative outlier labelling criterion to the *Z*-score. Given *n* normally distributed observations $y_i \sim N(\mu, \sigma^2)$, *M*-score is calculated by replacing the sample mean and standard deviation by the sample median and MAD, respectively:

$$M_i = \frac{0.6745(y_i - \tilde{y})}{MAD} \qquad \text{Equation (6.2)}$$

where $\tilde{y}$ represents the sample median, and $MAD = \text{median}_i\{|\, y_i - \tilde{y}\,|\}$. The constant 0.6745 is required to ensure $E(\text{MAD}) = 0.6745\sigma$ for large *n*.

Iglewicz and Hoaglin (1993) suggested an observation is flagged as an outlier when the absolute *M*-score is greater than 3.5. This criterion is based on the observation from a simulation study they performed to determine the value of *M*-score that caused 2.5%, 5% and 10% of random normal observations to be labelled as outliers in a large number of datasets with sample sizes of 10, 20 and 40. Their simulation results showed that *M*-scores can serve as a guide for labelling outliers in normally distributed data.

6.2.3 Tukey's fences and boxplot

Tukey (1977) suggested a simple graphical method for identifying outliers that is based on the boxplot and involves the construction of "inner fences" and "outer fences". This method is appealing due to its simplicity, and because it is less sensitive to extreme values (breakdown point is approximately 25%) since it uses quartiles to measure the spread of a distribution.

Boxplot provides an example summary of the location, spread and skewness of a univariate variable. A typical boxplot is illustrated in Figure 6.3, in which the central line in the box shows the position of the median. The lower and upper boundaries of the box represent the location of the lower quartile (Q1) and upper quartile (Q3), respectively. The interval between Q1 and Q3 is called the inter-quartile range (IQR).

Since Q1 and Q3 are the 25th and 75th percentiles, thus IQR contains the central 50% of the data.

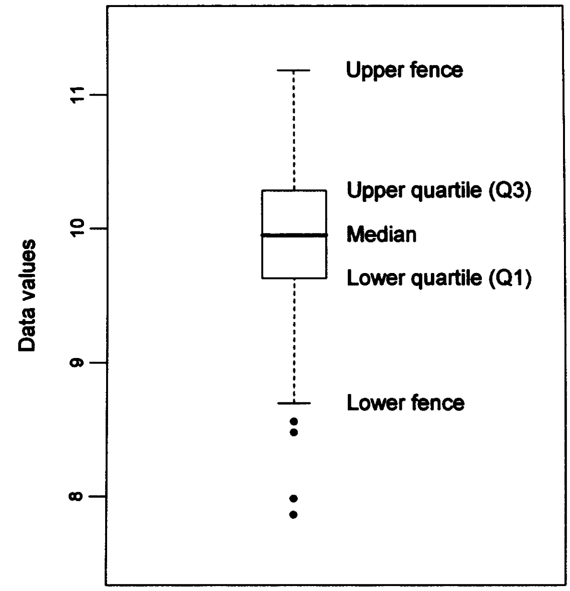


Figure 6.3: A typical boxplot showing the definitions of its components

Tukey (1977) proposed the following rules for labelling outliers:

1. Inner fences are defined as Q1 – 1.5 IQR, Q3 + 1.5 IQR.
2. Outer fences are defined as Q1 – 3 IQR, Q3 + 3 IQR.
3. Observations falling between the inner and outer fences are "outside" outliers, while those falls beyond the outer fences are "far out" outliers.

There is no statistical basis for Tukey's choice of 1.5 and 3 regarding the IQR to construct the inner and outer fences. However, if the data is normally distributed, then 95% of the observations would fall within the interval of the inner fences, and 99% would be within the range of the outer fences.

A limitation with the boxplot method is that when the data are skewed, the information about the tails as given by the boxplot can become unreliable. Hoaglin *et al.*, (1986) studied the probability properties of the boxplot rule and found that for a sample size in the interval (5, 20), approximately 25% of the samples simulated from a normal distribution will contain at least one outlier based on the cutoff points defined by the inner fences. Therefore, the boxplot rule provides only an exploratory tool for labeling outliers; observations flagged as "outside outliers" would need to be followed up and confirmed by more powerful methods such as the *Z*-scores or *M*-scores methods, the generalised extreme studentised deviate (ESD), or Dixon-type tests proposed by Iglewicz and Hoaglin (1993).

6.2.4 Data transformation considerations

The outlier detection methods described above implicitly or explicitly assume that the data or the outlier scores are normally distributed. When the distribution of a variable is skewed, or in other regards departs from the normal distribution, it is sometimes possible to transform the values of that variable to create a new variable that is more closely normal in shape. There are a variety of possible data transformations, such as adding constants, multiplying, squaring or taking the square root of the values, or converting to logarithmic scales. In some cases, an appropriate transformation can be chosen based on theoretical considerations or knowledge of the process generating the data. For example, taking the logarithm of observations from a lognormal or positively skewed distribution tends to make the data appear nearly normal in shape. Unfortunately, the choice of the "best" transformation is not always obvious. This was recognised by Box and Cox in 1964 when they proposed a family of parametric power transformations techniques for formally estimating a suitable transformation so that the transformed variable approximates a normal distribution as closely as possible (Box and Cox 1964).

Box-Cox transformation

Given a random variable y from some distribution with only positive data values, the Box-Cox power transformation is defined as:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y^{\lambda}-1)}{\lambda} & ,\lambda \neq 0 \\ \log(y) & ,\lambda = 0 \end{cases} \qquad \text{Equation (6.3)}$$

where lambda λ denotes the power of the transformation, and $y^{(\lambda)}$ is the transformed observations. Consider the column vector $\mathbf{y}^{(\lambda)} = \{y_1^{(\lambda)},\ldots, y_n^{(\lambda)}\}$ with n transformed observations that satisfies the linear model:

$$E\{\mathbf{y}^{(\lambda)}\}= \mathbf{a}\boldsymbol{\theta} \qquad \text{Equation (6.4)}$$

where **a** is a known matrix and **θ** a vector of unknown parameters associated with the transformed observations. The central assumption was that for some unknown λ, the transformed observations $y_i^{(\lambda)}$ (where $i = 1,\ldots, n$) can be treated as independently normally distributed with constant variance σ^2 and with expectations specified by Equation (6.4).

Two methods were proposed by Box and Cox (1964) for estimating the transformation parameter λ. The first approach is based on maximising the likelihood function, which leads directly to point estimates of the transformation parameters. The second approach is based on the Bayes's theorem. Since the maximum likelihood (MLE) approach is conceptually simpler than the Bayesian method and the profile likelihood function is easier to compute, therefore only the MLE method is considered here. The general idea behind MLE inference of λ is outlined below; for further details about the Bayesian approach, see the original paper by Box and Cox (1964) or Pericchi (1981).

Box and Cox (1964) showed that the profile likelihood function for λ is

$$L_{\max}(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i) \qquad \text{Equation (6.5)}$$

where $\hat{\sigma}^2(\lambda)$ is the residual sum of squares in the analysis of variance of $y^{(\lambda)}$. Box and Cox suggested that one way to select the power of transformation is to use λ that maximises the log-likelihood $L_{\max}(\lambda)$ in Equation (6.5).

In general, easily interpretable λ values such as 0 (log), 0.5, (square-root) or -1 (inverse) would be preferred. Practical considerations from the context will also provide further guidance in the choice of λ. For examples, values of $\lambda < 1$ tend to deflate large values of Y and are useful for transforming positively skewed observations. In contrary, values of $\lambda > 1$ tend to inflate Y and are useful for transforming negatively skewed observations.

Since the work of Box and Cox (1964) a number of modifications have been proposed. For examples, Manly (1976) suggested an alternative version that is valid for negative data values; John and Draper (1980) presented the so-called "modulus transformation" that can be used with distributions that are already somewhat symmetric. A review of the works relating to these alternative approaches can be found in Sakia (1992).

6.3 OutlierDM: an outlier detection framework for identifying over-represented PubMed abstract terms based on *Z*-scores

6.3.1 Basic ideas

As shown in Section 6.1, biologically-interesting terms that convey useful information about the functional relationships amongst genes in a gene list tend to lie away from the main data cluster formed by the remaining terms when the *Chip* frequencies of these terms are plotted against the corresponding *List* frequencies. This is because of the inherent assumption that the majority of tokens in a gene list would be largely irrelevant to the biology under study (e.g. common English words), thus forming a background cloud of points, with enriched and potentially interesting terms appearing as outliers.

A feature that distinguishes the biologically-interesting terms from the others is that they often have lower *Chip* frequency than expected by chance, and appear as outliers. This phenomenon is illustrated in Figure 6.4. Motivated by this observation, an outlier detection-based procedure, which seeks within a group of tokens corresponding to the same *List* frequencies any tokens that have lower than expected *Chip* frequencies, was developed.

Given a gene list, all tokens associated with it are first divided into groups according to their *List* frequency. Each such group is equivalent to a vertical slice in the scatter plot, as illustrated in Figure 6.4(b) and (c). A local mean and standard deviation of the

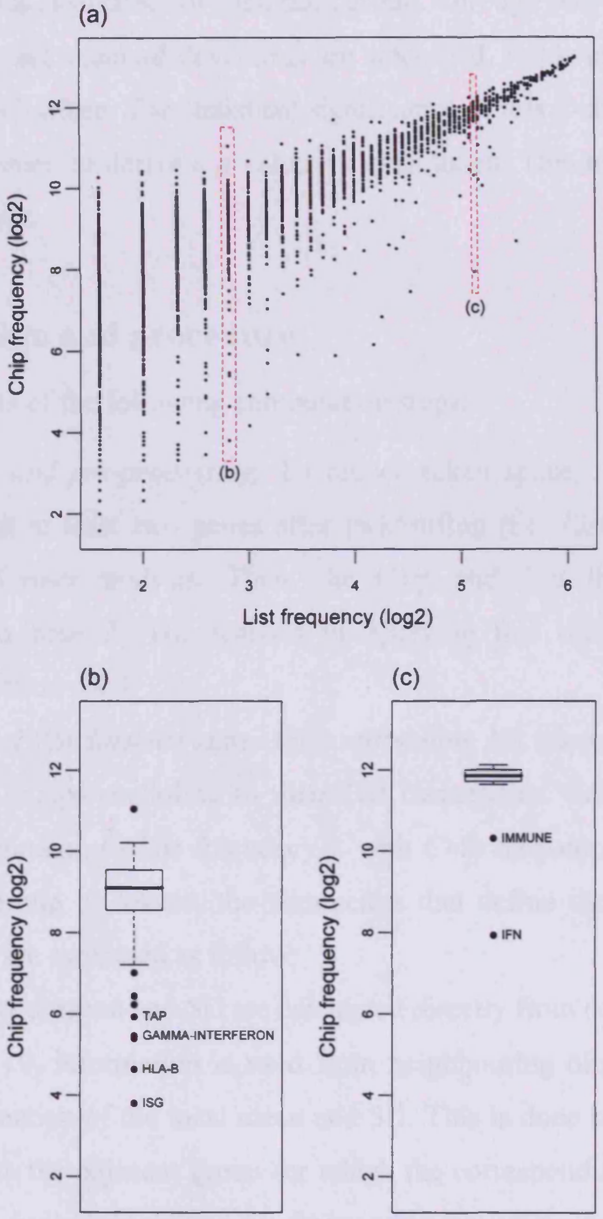


Figure 6.4: Outlying tokens tend to have lower *Chip* frequencies than expected
(a) The log2-transformed *Chip* frequencies for tokens in the ISG gene list are plotted against their corresponding log2-transformed *List* frequency. Highlighted in red boxes are two groups of tokens corresponding to *List* frequency 7 and 35. The distributions of their *Chip* frequencies are illustrated as boxplots in panels (b) and (c), respectively. It can be seen that biologically-plausible terms that are relevant to the ISG gene list such as 'ISG', 'gamma-interferon' and 'IFN' tend to have lower *Chip* frequencies than other tokens in the same group and appear as potential outliers.

Chip frequencies are calculated for each such group. Through polynomial curve fitting, these local means and standard deviations are smoothed, followed by the calculation of a *Z*-score for each token. The statistical significance of this *Z*-score is inferred from the normal distribution to derive a *p*-value for each token. This method is denoted as OutlierDM hereafter.

6.3.2 Algorithm and procedure

OutlierDM consists of the following computation steps:

1. *Data filtering and pre-processing.* To reduce token space, only tokens that are associated with at least two genes after jackknifing (i.e. *List* frequency - 1) are retained for further analysis. Then, the *Chip* and *List* frequencies are log-transformed to base 2. The reasons of applying this transformation will be provided in Section 6.4.1.
2. *Local mean and standard deviation (SD) estimation.* All tokens in the gene list are stratified into groups according to their *List* frequencies. Consider a group of n tokens corresponding to *List* frequency L with *Chip* frequencies given by $(x_1,\ldots, x_n)$. For this group of tokens, the parameters that define the outlier region, i.e. mean and SD, are estimated as follow:
    - If $n \geq 10$, local mean and SD are calculated directly from $(x_1,\ldots, x_n)$.
    - If $1< n < 10$, information is used from neighbouring observations to give a better estimation of the local mean and SD. This is done by capturing $(10 - n)$ tokens from the adjacent group for which the corresponding *List* frequency is less than L and combine the *Chip* frequencies from them with $(x_1,\ldots, x_n)$. Then, a local mean and SD are calculated based on the combined data.
3. *Local mean and SD smoothing.* The means and SDs estimated in Step 2 are smoothed as a function of *List* frequency by fitting polynomial curves to the frequency data. Locally smoothed mean and SD are computed for each group based on the fitted values derived from the best-fitting curves as illustrated in Figures 6.5 (a) and (b). The purpose of smoothing is to stabilise the variance in the data so as to obtain representative mean and SD values for *Z*-score calculation. The effect of smoothing is shown in Figure 6.5 (c) and (d).

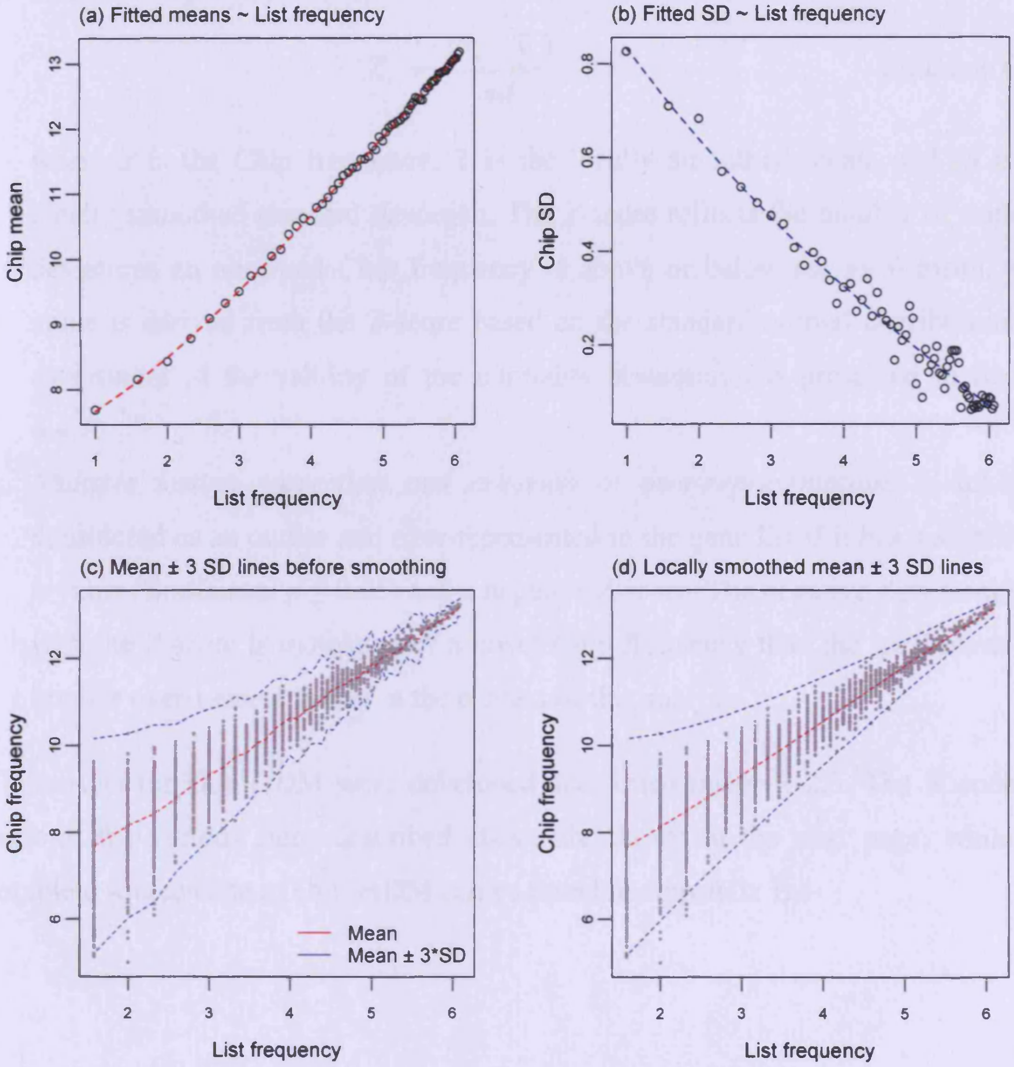


Figure 6.5: Diagnostic plots for outlier detection procedure

The data shown here is based on a random gene list that contains 78 genes and 1382 PMIDs, which was created by sampling without replacement from the HG-U133A array. The *Chip* and *List* frequencies of the tokens from this gene list were log-transformed to base 2. Then, these tokens were stratified into groups according to their *List* frequencies. For each group, the local means and standard deviations (SD) were calculated. (a) The local means are smoothed as a function of *List* frequencies using polynomial fitting. The red dashed line represents the curve of best fit. (b) The local SDs are smoothed as a function of *List* frequencies using polynomial fitting. The blue dashed line represents the curve of best fit. (c) Local mean ± 3 SD lines before smoothing. (d) Local mean ± 3 SD lines after smoothing. Both axes are plotted on log2 scale.

4. *Z-score and p-value calculation.* A *Z*-score for token *i* is calculated as:

$$Z_i = \frac{(x_i - \bar{x})}{sd} \qquad \text{Equation (6.6)}$$

   where *x* is the *Chip* frequency, $\bar{x}$ is the locally smoothed mean, and *sd* is the locally smoothed standard deviation. The *Z*-score reflects the number of standard deviations an observed *Chip* frequency is above or below the local mean. A *p*-value is derived from the *Z*-score based on the standard normal distribution. An assessment of the validity of the normality assumption is presented in Section 6.4.1.

5. *Multiple testing correction and criterion of over-representation.* A token is considered as an outlier and over-represented in the gene list if it has a significant *p*-value (Bonferroni $p \leq 0.05$) and a negative *Z*-score. The negative sign associated with the *Z*-score is indicative of a lower *Chip* frequency than the local mean and implies over-representation in the context of this analysis.

The scripts for OutlierDM were developed and tested under R-2.6. The R codes to perform the various steps described above are shown in the next page, while the complete source code of OutlierDM can be found in Appendix B.

Sample R codes of OutlierDM

Input: dat, a data matrix containing tokens and their raw *List* and *Chip* frequencies

```
# Step 1: Data filtering and pre-processing
dat$List <- dat$List - 1
dat <- dat[dat$List > 1, ]
dat$List <- log2(dat$List)
dat$Chip <- log2(dat$Chip)
dat <- dat[order(dat$List, decreasing=T), ]

# Step 2: Local mean and standard deviation (SD) estimation
tokenGroups <- sort(unique(dat$List))
uniqMean <- vector(length=length(tokenGroups))
uniqSD <- vector(length=length(tokenGroups))
smoothedMean <- vector(length=nrow(dat))
smoothedSD <- vector(length=nrow(dat))

for(j in 1:length(tokenGroups)){
    i <- dat$List == tokenGroups[j]
    n <-  sum(i)
   if(n >= 10){
      x <- dat[i, "Chip"]
      uniqMean[j] <- mean(x)
      uniqSD[j] <- sd(x)
   }else{
      weHaveGot <- n
      k <- dat$List < tokenGroups[j]
      resDat <- dat[k, ]
      x <- c(dat[i, "Chip"], resDat[1:(10-weHaveGot), "Chip"])
      uniqMean[j] <- mean(x)
      uniqSD[j] <- sd(x)
   }
}

# Step 3: Local mean and SD smoothing
chipMean.model <- lm(uniqMean ~ tokenGroups + I(tokenGroups^2) +
                  I(tokenGroups^3))
chip.mean <- chipMean.model$coef %*% rbind(1, tokenGroups,
             tokenGroups^2, tokenGroups^3)
chipSD.model <- lm(uniqSD ~ tokenGroups + I(tokenGroups^2) +
                I(tokenGroups^3))
chip.sd <- chipSD.model$coef %*% rbind(1, tokenGroups, tokenGroups^2,
           tokenGroups^3)

for(p in 1:length(tokenGroups)) {
    sel <- dat$List == tokenGroups[p]
    smoothedMean[sel] <- chip.mean[p]
    smoothedSD[sel] <- chip.sd[p]
}

# Step 4: Z-score and p-value calculation
Z.score <- (dat$Chip - smoothedMean)/smoothedSD
Raw.pval <- pnorm(Z.score)
Bonferroni.pval <- p.adjust(Raw.pval, method="bonferroni")
Outliers <- paste(dat[which(Bonferroni.pval < 0.05), "Token"])
```

Output: *Z*-scores, *p*-values and significant tokens

6.4 Experiments and results

6.4.1 Assessments of the normality assumption

A critical assumption behind OutlierDM is that the *Z*-scores derived from the token frequency data can be modelled by a normal distribution with mean zero and standard deviation of one. For this assumption to hold, the *Chip* frequencies of tokens associated with a specific *List* frequency are expected to be approximately normal (or have a nearly symmetric density) because the *Z*-scores are derived from the *Chip* frequencies. However, empirical observations based on simulated datasets show that the raw *Chip* frequencies are not normally distributed. This is illustrated in Figure 6.6(a), in which the histograms of *Chip* frequencies for tokens with *List* frequency equal to 3, 5, 7, 10, 13, 15, 16, 19, 24, 29, 34, 39, 46, 55 and 60 in a random gene list containing 78 genes and 1382 PMIDs, are shown. Simulated gene lists were used for this analysis because the objective here is to examine the distribution of raw *Chip* frequencies under the null hypothesis (i.e. when the data is completely random and no token is significant).

Distribution of *Chip* frequencies under the null hypothesis

The distribution patterns of the *Chip* frequencies shown in Figure 6.6(a) can be broadly divided into two groups. The first group includes those cases for which the distributions are slightly positively skewed. This group of observations typically has low *List* frequency with values between 3 and 24. Histograms corresponding to *List* frequency greater than 24 are classified as the second group. Due to the limited number of observations in each case, objective judgment could not be made regarding the shape of the distributions in the second group; visually they bear no obvious resemblance to any well-known distribution.

To determine if it is possible to transform the raw *Chip* frequency data into a new variable that is more nearly normal in shape, an optimal power index λ was determined for each of the cases shown in Figure 6.6(a) by means of the Box-Cox transformation technique introduced in Section 6.2.4. Using the `boxcox.fit` function implemented in the R package geoR[1], the log-likelihood vector for a range of λ values were computed. The value of λ corresponding to the maximum log-likelihood (or equivalently minimum negative log-likelihood) is taken as the optimal λ. This is shown in Figure 6.6(b), where the maximised log-likelihood is plotted against λ for a trial series of λ values. The optimal λ can be read off from the plots.

Figure 6.6(c) shows the histograms of *Chip* frequencies after Box-Cox transformation using the optimal λ. It can be seen that there are improvements in symmetry and normality for those cases with *List* frequency less than 15, and which are predominantly positively skewed data before Box-Cox transformation. The optimal λ values determined for these cases are typically less than one. This makes theoretical sense as values of $\lambda < 1$ tend to reduce the relative spacing of scores on the right side of the distribution more than the scores on the left side. As for those cases with *List* frequency greater then 15, the distributions of the *Chip* frequencies are still not symmetric even after transformation.

The above analysis shows that, to transform the *Chip* frequencies across the range of *List* frequency in a gene list to normality, different λ values are required. For the current dataset, the values of the optimal λ range from -4.1 to 2.7 (Figure 6.7(a)). Similar trends were observed when Box-Cox transformation was applied to three other simulated gene lists of size 100, 300 and 500. These are shown in panels (b), (c) and (d) of Figure 6.7, respectively. Clearly, these are not easily interpretable λ values. Applying such complex transformations to the original *Chip* frequencies data would change the relative distances between data points in the same dataset to different extent, complicating the interpretation of the results.

As can be seen from Figure 6.7, a substantial proportion of the optimal λ determined lie between -1 and 1. It was decided to use the logarithm transformation ($\lambda = 0$) of base 2 as a compromise choice because the log transformation provides values that are easily interpretable and, as will be seen later, the *Z*-scores calculated on the basis of the log2-transformed data is close to being normally distribution.

[1] geoR: a package for geostatistical data analysis using the R software; http://leg.ufpr.br/geoR/

Figure 6.6: Box-Cox transformation of *Chip* frequencies
This is a three-part figure that shows the distributions of *Chip* frequencies associated with tokens found in a typical random gene list before and after Box-Cox transformation. The random gene list used in this analysis contains 78 genes and 1382 PMIDs, and was created by sampling without replacement from the HG-U133A array (NB: this is the same gene list as the one used to produce Figure 6.5). The examples shown here are based on tokens for which the associated *List* frequency equal to 3, 5, 7, 10, 13, 15, 16, 19, 24, 29, 34, 39, 46, 55 and 60. Panel (a) shows histograms of the raw *Chip* frequencies. Panel (b) shows the Box-Cox plots for the chosen examples. In each plot, the value of lambda, λ, that correspond to the maximum log-likelihood value was read off as the optimal λ. Panel (c) shows the histograms of *Chip* frequencies after Box-Cox transformation with the optimal λ. Abbreviations used in the plot: *List* freq = *List* frequency; n = number of observations in the example.

(continued over the page)

Figure 6.6 (continued)

(a) Histograms of raw Chip frequencies

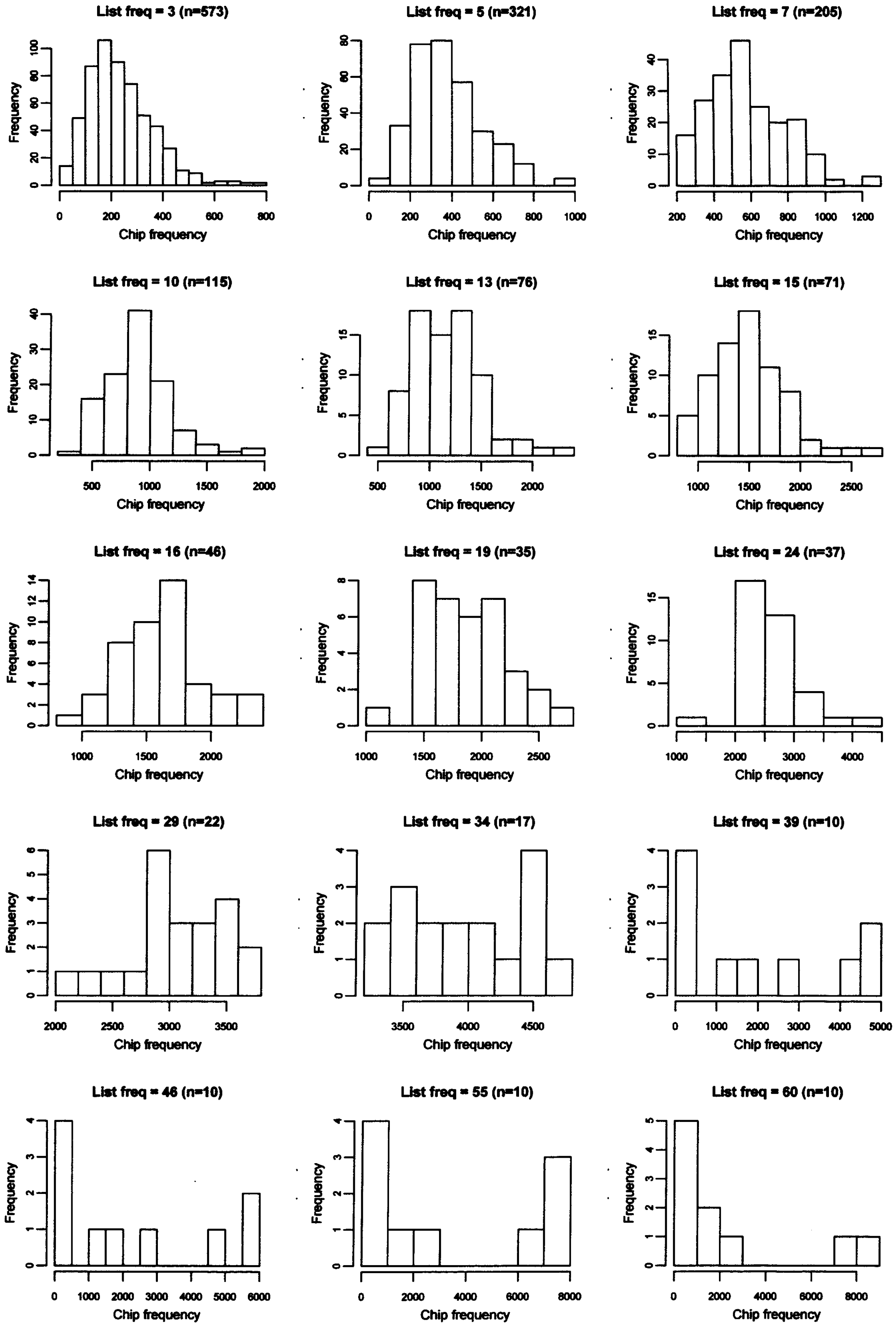


Figure 6.6 (continued)

(b) Box-Cox plots

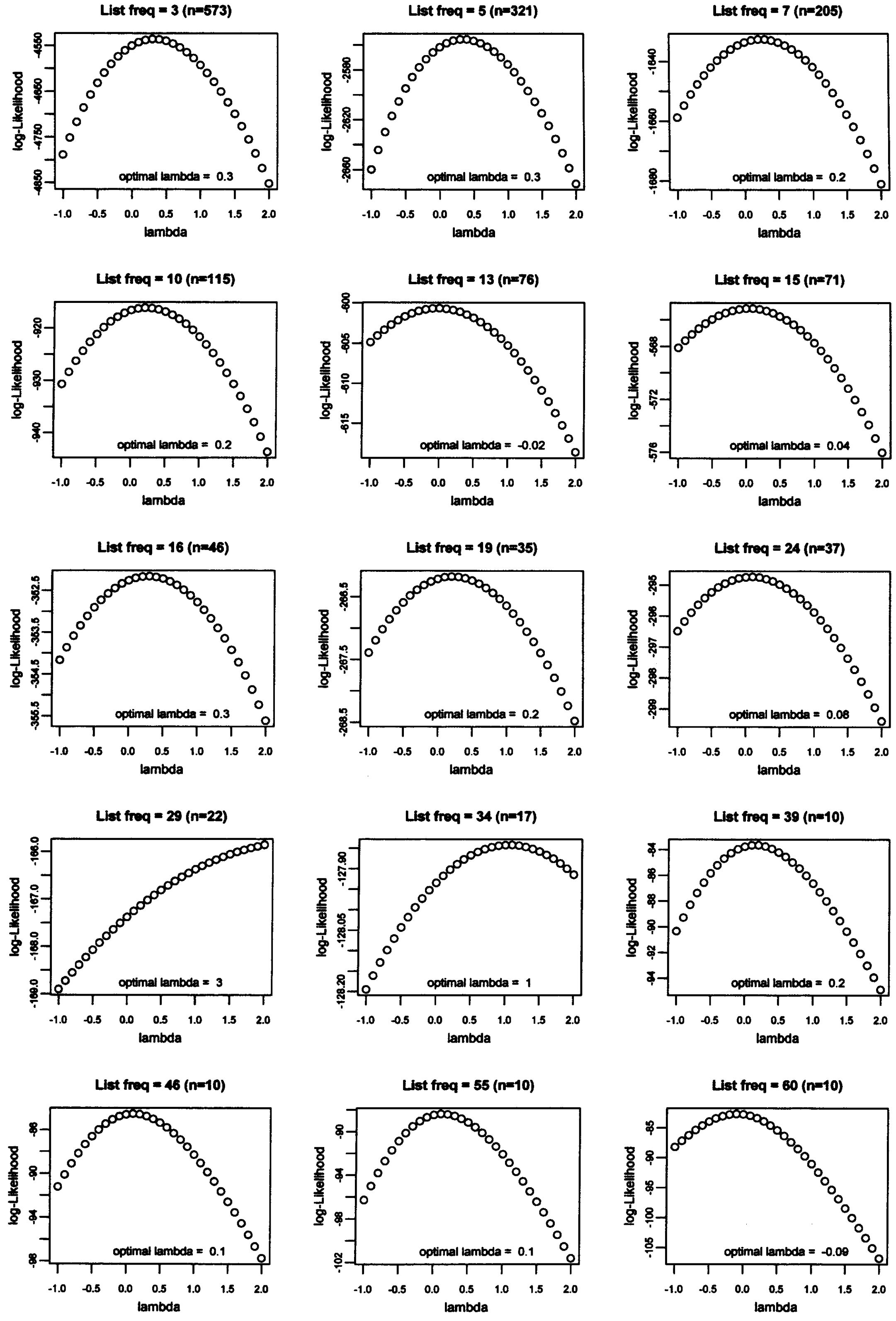
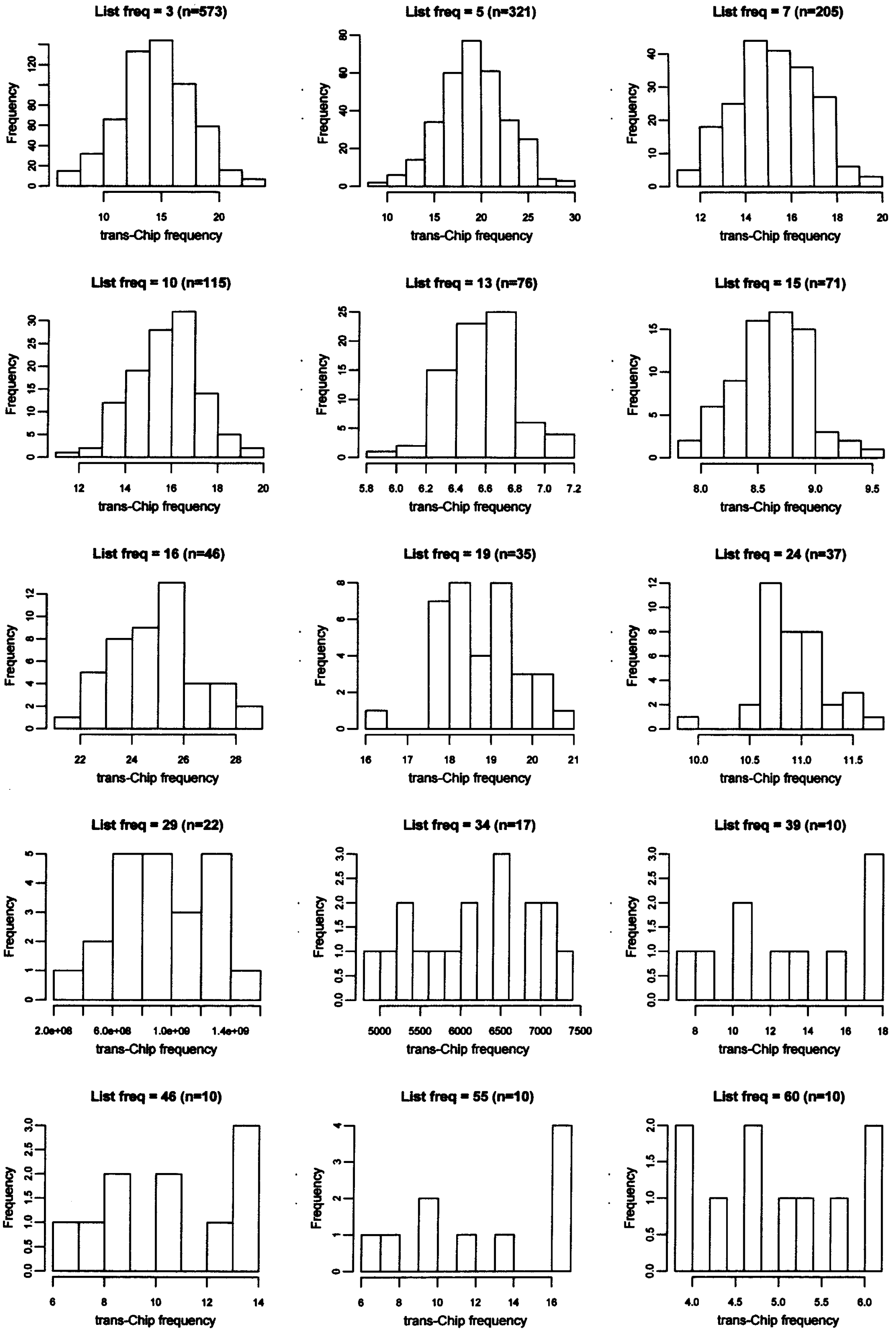


Figure 6.6 (continued)

(c) Histograms of Chip frequencies after Box-Cox transformation



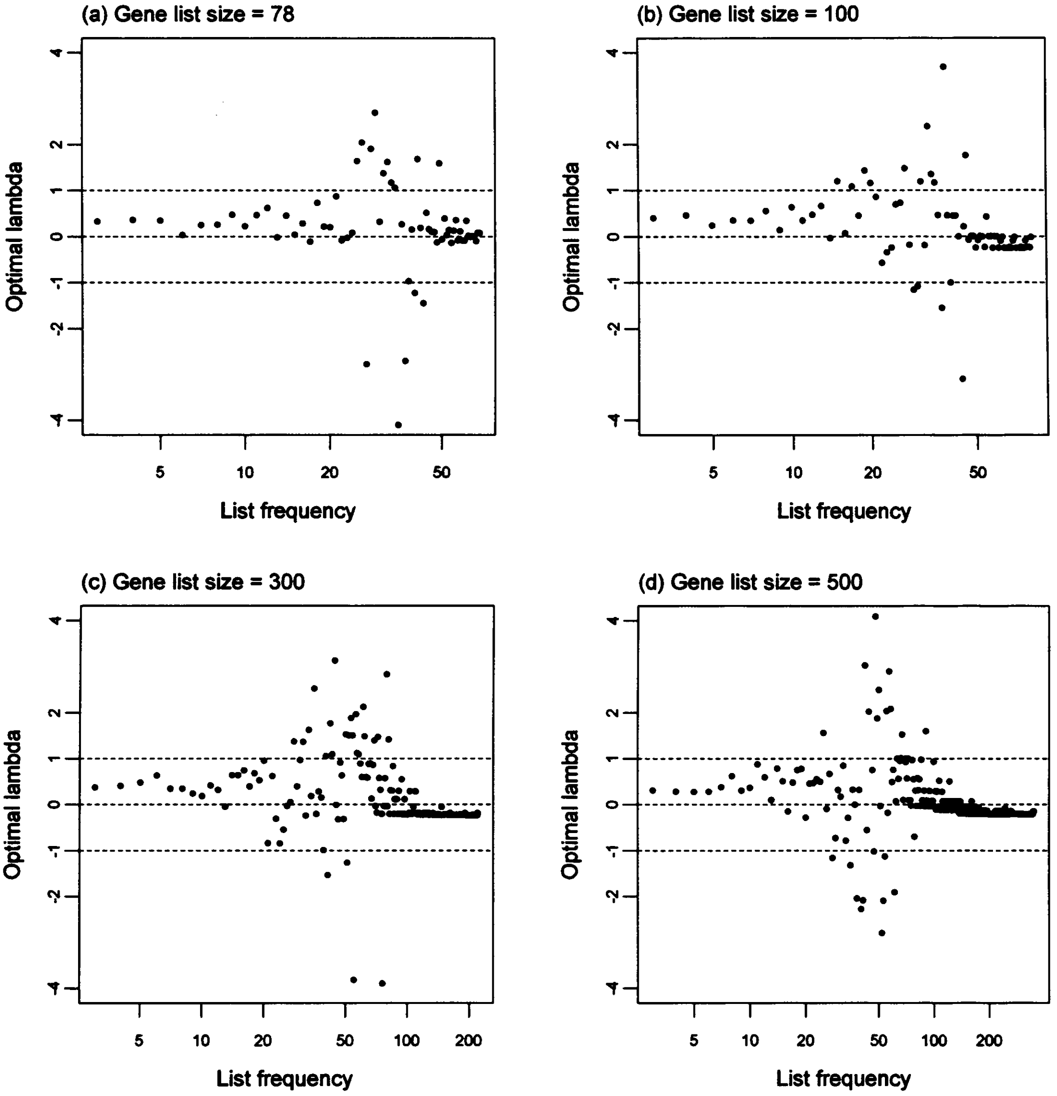


Figure 6.7: Plots of optimal lambda versus *List* frequency

The values of optimal λ (lambda) required to transform the *Chip* frequencies to normality were plotted against the corresponding *List* frequencies. Random gene lists of different sizes were created by sampling without replacement from the HG-U133A array. Details of the gene lists used to produce these plots are as follow: Panel (a) is based on a simulated gene list with 78 genes and 1382 PMIDs. This is the same gene list used in the analysis described in Figure 6.6. Panel (b) is based on a simulated gene list with 100 genes and 1181 PMIDs. Panel (c) is based on a simulated gene list with 300 genes and 2528 PMIDs. Panel (d) is based on a simulated gene list with 500 genes and 3973 PMIDs. The x-axis is on logarithmic scale.

Distribution of *Z*-scores under the null hypothesis

Six random gene lists of lengths (as in number of unique genes) 50, 100, 300, 500, 1000 and 2000, were created by sampling without replacement from the reference HG-U133A array. The *List* and *Chip* frequencies associated with tokens in these random gene lists were determined and a log2 transformation applied to them. A *Z*-score was calculated for each token using the outlier detection method outlined in Section 6.3.2.

To assess the validity of the normality assumption under the null hypothesis, histograms and normal probability plots for the *Z*-scores derived from these random gene lists were examined. The results are presented in Figure 6.8. In the histogram, the sample's *Z*-scores (black density curve) are plotted alongside a normal distribution simulated with mean of 0 and standard deviation (SD) of 1 (red density curve). On top of the histogram, the mean and SD of the sample's *Z*-scores are shown. It can be seen that the sample's *Z*-scores have a nearly normal distribution, with the fit becoming better as the size of gene list grows from 50 to 2000.

The distributions of *Z*-scores from shorter gene lists (e.g. between 50 to 300 genes) are slightly negatively skewed. This skew is more apparent from the normal probability plot, which compares each data point in the sample with its expected value in a standard normal distribution. In order to examine if this presents a serious problem to the outlier detection procedure, a number of simulations were carried out to estimate the false positive rates associated with gene lists of specific lengths.

Figure 6.8: Distribution of *Z*-scores under the null hypothesis
To evaluate the validity of the normality assumption underlying the proposed outlier detection-based ORA approach, six random gene lists of different lengths were created. 50, 100, 300, 500, 1000 and 2000 genes were drawn at random (without replacement) from the HG-U133A array. The *List* and *Chip* frequencies associated with the tokens found in these gene lists were log2-transformed and their *Z*-scores calculated as described in Section 6.3.2. Then histograms and normal probability plots for *Z*-scores derived from the six random gene lists were produced. In the histogram, the *Z*-scores calculated using the outlier detection procedure (black density curve) was plotted alongside the normal distribution simulated with the R command `rnorm(n=10000, mean=0, sd=1)`(red density curve). The normal probability plot (QQ-plot) was created by comparing each token in the random gene list with its expected value in a theoretical normal distribution. The blue dashed line represents the $y = x$ line.

(continued over the page)

Figure 6.8 (continued)

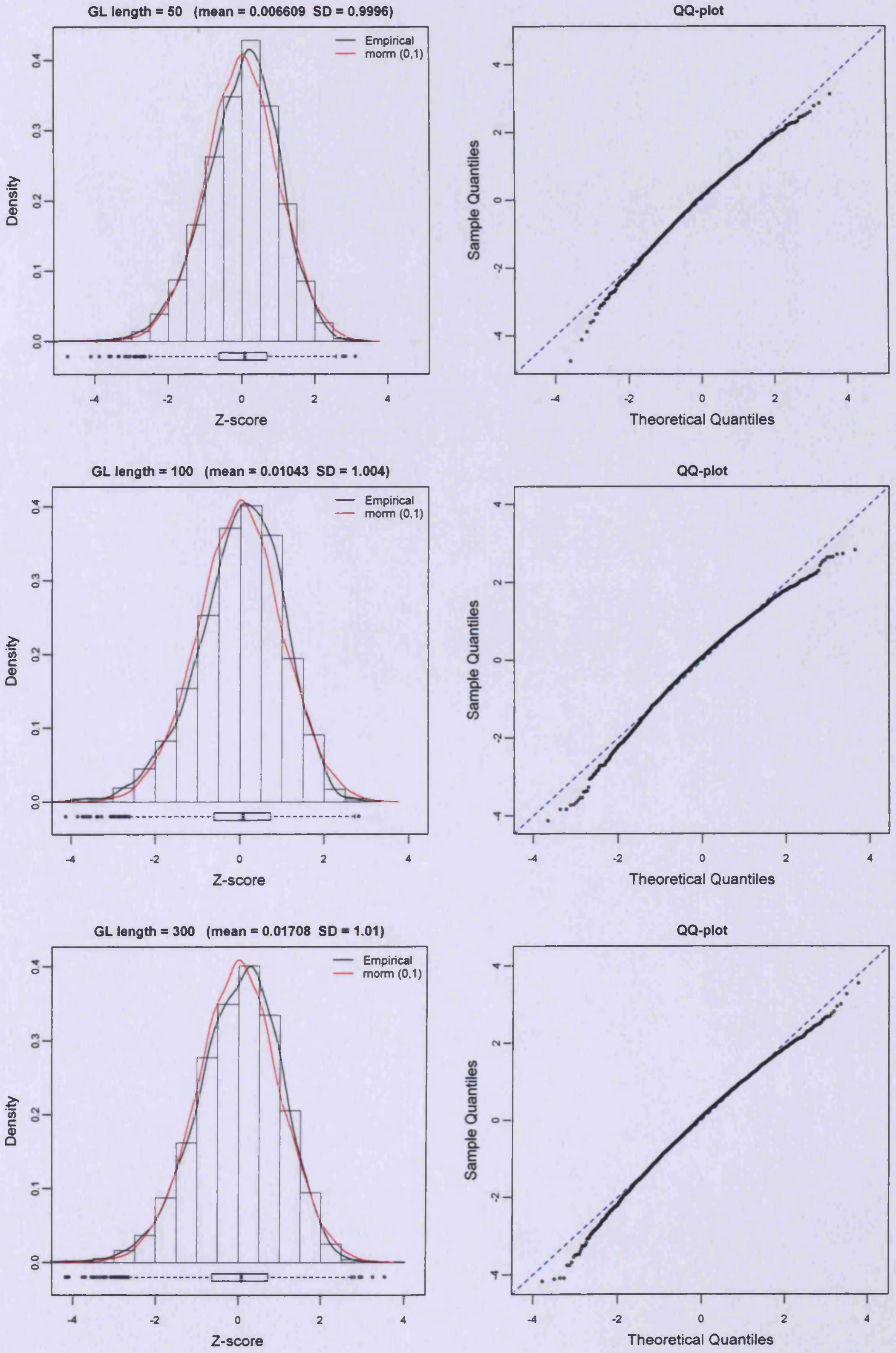
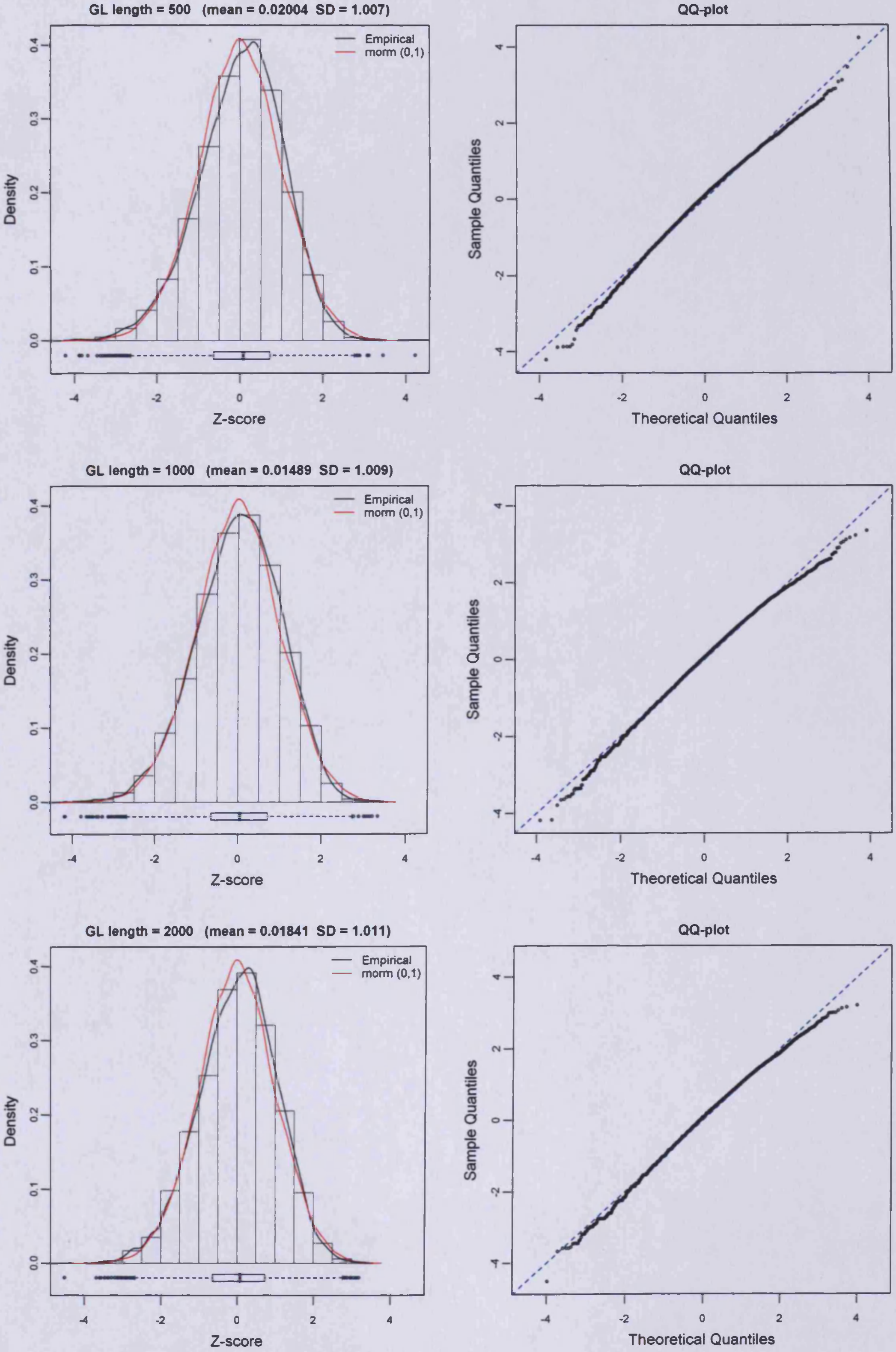


Figure 6.8 (continued)



6.4.2 False positive rates under the null hypothesis

From the specificity perspective, an ideal ORA method should not find any significant terms in a random gene list. However, even when the data is completely random (i.e. the null hypothesis is true), any statistical test will reject the null hypothesis for a number of cases directly controlled by the chosen significance level (e.g. by setting $\alpha = 0.05$). Therefore, it is important to verify that the proportion of false positives generate by the proposed method does not exceed this expected proportion.

To estimate the false positive rates associated with the proposed outlier detection-based ORA approach, 1000 random gene lists were created by randomly sampling N_{gene} unique genes from the HG-U133A array, where N_{gene} = 50, 100, 300, 500, 1000 and 2000. These random gene lists were then analysed with OutlierDM. For each N_{gene}, a mean false positive rate was determined as follows:

1. The false positive rate (FP) for each gene list was calculated as the proportion of falsely rejected cases among all the tests performed:

$$FP_i = \frac{\text{Number of false positives in gene list } i}{\text{Total number of tokens tested in gene list } i} \qquad \text{Equation (6.7)}$$

   where $i = 1, 2, \ldots n$. Here, $n = 1000$, which is the total number of random gene lists created for a specific N_{gene}. In the context of this analysis, false positive is defined as tokens that were called significantly over-represented in the random gene list.

2. The mean FP rate was calculated by averaging the FP values obtained from Step 1:

$$\text{Mean FP rate} = \frac{1}{n}\sum_{i=1}^{n} FP_i \qquad \text{Equation (6.8)}$$

The results of this analysis are summarised in Table 6.1.

Table 6.1: False positive rates of the *Z*-score-based outlier detection method

N_{gene}	Mean number of token tested per gene list	Mean number of false positives per gene list	Mean FP rate
50	1554	1.84	0.00123
100	2540	1.97	0.00079
300	5025	1.67	0.00034
500	6670	1.14	0.00017
1000	9935	0.47	4.81E-05
2000	14728	0.17	1.18E-05

The mean FP rate for OutlierDM ranges from 1.18 $\times 10^{-5}$ to 0.001 at $\alpha = 0.05$, with shorter gene lists ($N_{gene} < 300$) being more susceptible to false positives. On average, it is possible to find up to two false positives when the query gene list has less than 300 genes. As previously shown in Figure 6.8, the *Z*-score distribution tends to show a slight negative skewness for short gene lists (containing 50 to 300 genes). This slight deviation from normality could lead to a small number of observations to be falsely identified as outliers at the left tail region, and hence the higher FP rates for shorter gene lists. However, the *Z*-scores tend to assume a more nearly normal distribution as the size of gene list increases, and this leads to a fall in the FP rates.

6.4.3 Performance on real datasets

The performance of OutlierDM will now be illustrated using three datasets: the ISG gene list from Sanda *et al.* (2006), the glycolysis gene list reported in Vanharanta *et al.* (2006), and the mitosis gene list reported in Lee *et al.* (2004). These gene lists have already been introduced in previous Chapters when the classical hypergeometric distribution-based and the permutation-based ORA approaches were presented. Using the same gene lists here to demonstrate the performance of the outlier detection-based approach allow for cross comparisons of the various text-based ORA methods presented in this thesis.

Example 1: ISG gene list

In total, 23 tokens were identified as significantly over-represented (Bonferroni p-value ≤ 0.05). These terms are listed in Table 6.2 and their locations on the scatter plot are shown in Figure 6.9. All significant terms appear to be relevant to the biology of an interferon-regulated response.

Table 6.2: Over-represented abstract terms in the ISG gene list as identified using the *Z*-score-based outlier detection approach

Term	*Chip* frequency	*List* frequency	*Z*-score	*p*-value	Bonferroni *p*-value	Rank
INTERFERON	414	46	-12.6197	8.22E-37	2.81E-33	1
IFN	245	35	-9.5951	4.19E-22	1.43E-18	2
IFN-BETA	71	18	-7.5764	1.78E-14	6.06E-11	3
ANTIVIRAL	176	23	-6.7487	7.46E-12	2.54E-08	4
IFN-ALPHA	114	19	-6.7135	9.50E-12	3.24E-08	5
INTERFERON-ALPHA	59	14	-6.6209	1.78E-11	6.09E-08	6
OLIGOADENYLATE	18	8	-6.0755	6.18E-10	2.11E-06	7
ISG	14	7	-5.7749	3.85E-09	1.31E-05	8
ISRE	31	9	-5.7076	5.73E-09	1.95E-05	9
DSRNA	60	11	-5.3972	3.39E-08	0.0001	10
HLA-CLASS	11	6	-5.3042	5.66E-08	0.0002	11
HLA-A	30	8	-5.1818	1.10E-07	0.0004	12
HLA-B	25	7	-4.8328	6.73E-07	0.0023	13
INDUCIBLE	1068	37	-4.7870	8.47E-07	0.0029	14
ENCEPHALOMYOCARDITIS	16	6	-4.7508	1.01E-06	0.0035	15
STOMATITIS	52	9	-4.7475	1.03E-06	0.0035	16
OAS	10	5	-4.4576	4.14E-06	0.0141	17
HLA-G	10	5	-4.4576	4.14E-06	0.0141	18
MXA	11	5	-4.3337	7.33E-06	0.0250	19
EVASION	65	9	-4.3333	7.34E-06	0.0250	20
INNATE	363	21	-4.2752	9.55E-06	0.0326	21
TAPASIN	12	5	-4.2207	1.22E-05	0.0415	22
VIRAL	892	32	-4.1831	1.44E-05	0.0490	23

Over-represented terms were defined as having p-value ≤ 0.05 after Bonferroni correction. The results were ordered by increasing p-values. The gene universe used is that based on the HG-U133A chip and contains 9638 genes. In total, 3412 tokens were tested. Note that the negative sign associated with the Z-score is indicative of a lower *Chip* frequency than the local mean and implies over-representation in the context of this analysis.

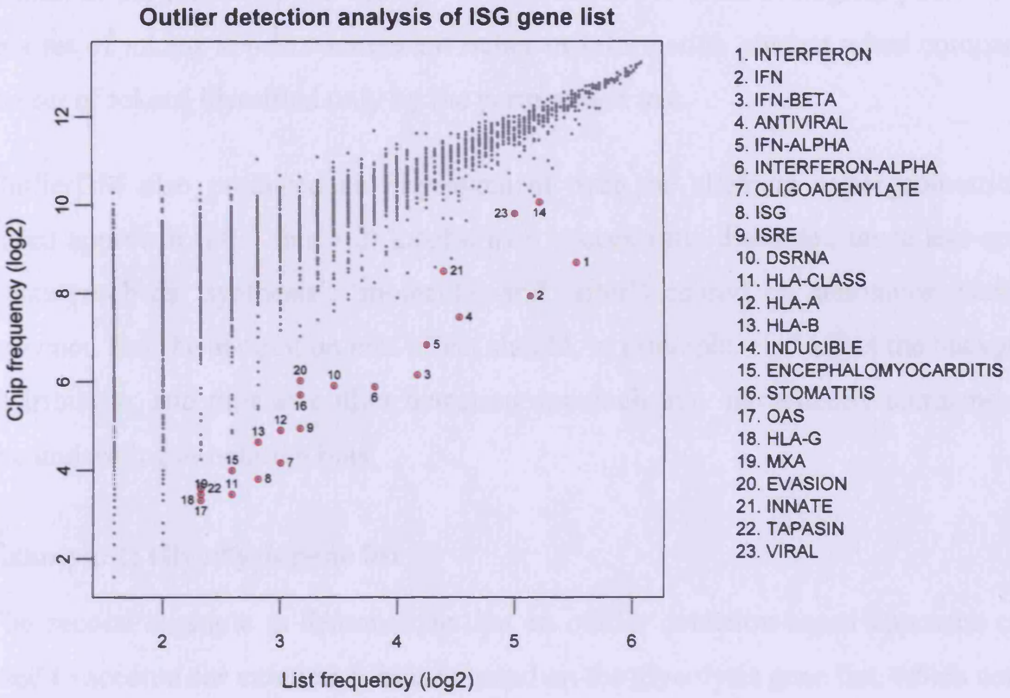


Figure 6.9: A scatter plot of *Chip* versus *List* frequencies for tokens in the ISG gene list

Terms that were identified as significantly over-represented (Bonferroni *p*-value ≤ 0.05) in the ISG gene list are circled in red and the adjacent numbers corresponding to their rankings. The *Z*-scores and *p*-values associated with the significant terms can be found in Table 6.2.

This set of result is also very similar to that produced by the permutation-based method (cf. Table 5.1). Six tokens reported as over-represented by the permutation-based approach but not found by OutlierDM are 'infection', 'treat', 'immunity', 'lymphoblastoic', 'gamma-interferon' and 'indigenous'. These tokens were assigned ranks of 24, 40, 34, 28, 31 and 33, respectively, by the outlier detection-based approach and fall short of the *p*-value cutoff used.

On the other hand, there are also six tokens that were identified as significantly enriched by OutlierDM but not by the permutation-based method. These are 'HLA-A' (rank 12), 'HLA-B' (rank 13), 'HLA-G' (rank 18), 'MxA' (rank 19), 'innate' (rank 21) and 'tapasin' (rank 22). Except for the term 'innate', these tokens are specifically related to the functions and biology of interferon. From the biological point of view, this set of tokens appears somewhat richer in information content when compared to the set of tokens identified only by the permutation test.

OutlierDM also produced an improvement over the classical hypergeometric test-based approach (cf. Table 3.2) insofar as it successfully discarded those less-specific terms (such as ‘synthesis’, ‘molecule’ and ‘after’) caused by annotation bias. It is reasoned that the annotation bias effect should, in principle, also affect the background distribution, and thus an outlier detection approach may intrinsically compensate for the underlying annotation bias.

Example 2: Glycolysis gene list

The second example to demonstrate that an outlier detection-based approach can be used to account for annotation bias is based on the glycolysis gene list, which contains genes mainly involved in carbohydrate metabolism. When the glycolysis gene list was first analysed with the classical hypergeometric distribution-based method, a mixture of biologically-specific and uninformative terms were obtained (Table 3.3). When this gene list was re-analysed using OutlierDM, only four tokens (i.e. ‘glycolytic’, ‘aldo-keto’, ‘peroxidation’ and ‘nicotinamide’) were called over-represented; but the uninformative terms were successfully avoided. The results are shown in Table 6.3 and their positions on the scatter plot are displayed in Figure 6.10.

Table 6.3: Over-represented abstract terms in the glycolysis gene list as identified using the *Z*-score-based outlier detection approach

Term	*Chip* frequency	*List* frequency	*Z*-score	*p*-value	Bonferroni *p*-value	Rank
GLYCOLYTIC	71	12	-5.6983	6.05E-09	2.98E-05	1
ALDO-KETO	12	5	-4.7037	1.28E-06	0.0063	2
PEROXIDATION	102	12	-4.6662	1.53E-06	0.0076	3
NICOTINAMIDE	80	10	-4.2967	8.67E-06	0.0427	4

Over-represented terms were defined as having p-value ≤ 0.05 after Bonferroni correction. The results were ordered by increasing p-values. The gene universe used is that based on the HG-U133A chip and contains 9638 genes. In total, 4927 tokens were tested.

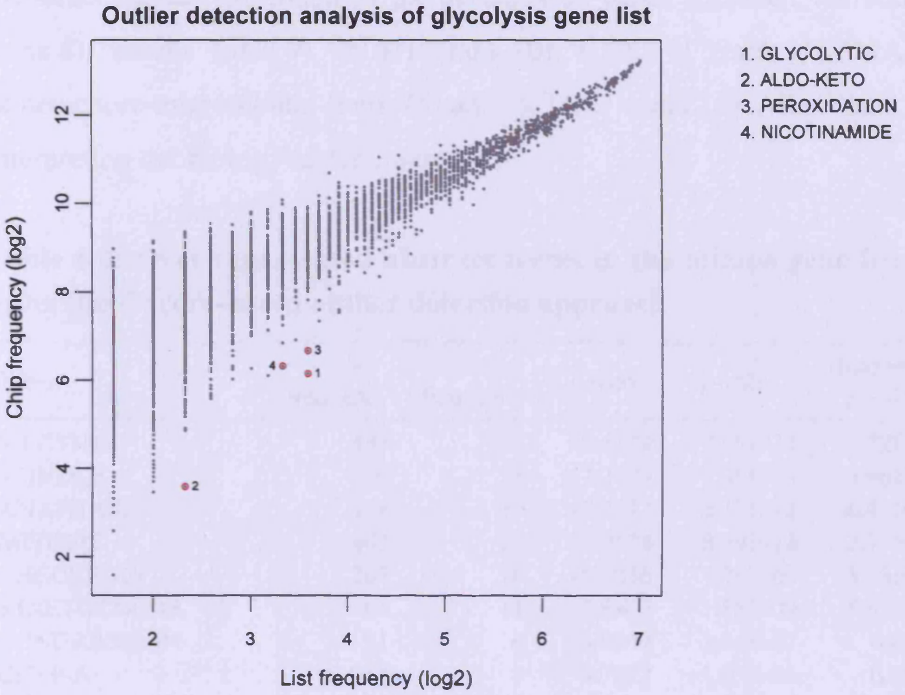


Figure 6.10: A scatter plot of *Chip* versus *List* frequencies for tokens in the glycolysis gene list

Terms that were identified as significantly over-represented (Bonferroni p-value $\leq$ 0.05) in the glycolysis gene list are circled in red and the adjacent numbers corresponding to their rankings. The Z-scores and p-values associated with the significant terms can be found in Table 6.3.

Example 3: Mitosis gene list

The third example is based on the mitosis gene list. It contains 82 genes that were differentially up-regulated during T-cells differentiation. When this gene list was analysed with OutlierDM, 16 tokens were called significantly over-represented (Bonferroni p-value $\leq$ 0.05). These terms are listed in Table 6.4 and their positions on the scatter plot are shown in Figure 6.11. The significant terms are linked to cell-cycle regulation, cell-cycle progression and mitosis, which is in good agreement with the manual annotation reported by Lee *et al.* (2004). This result is also comparable to that obtained by the permutation-based approach (cf. Table 5.2). However, the outlier detection-based approach appear to be more sensitive in that it is able to detect seven extra hits that are not found by the permutation-based approach, including 'CENP-A' (rank 8), 'aurora' (rank 9), 'BUB1' (rank 10), 'CENP-H' (rank 11), 'MAD2' (rank 12), 'kinetochore-microtubule' (rank 15) and 'MTOC' (rank 16). All of them are useful for interpreting the biology under studied.

Table 6.4: Over-represented abstract terms in the mitosis gene list as identified using the *Z*-score-based outlier detection approach

Term	*Chip* frequency	*List* frequency	*Z*-score	*p*-value	Bonferroni *p*-value	Rank
MITOTIC	485	28	-7.5344	2.45E-14	7.72E-11	1
SPINDLE	298	23	-7.2629	1.89E-13	5.96E-10	2
ANAPHASE	126	17	-7.2477	2.12E-13	6.67E-10	3
MITOSIS	443	26	-7.0524	8.79E-13	2.77E-09	4
CHECKPOINT	267	19	-5.9056	1.76E-09	5.53E-06	5
KINETOCHORE	64	11	-5.8943	1.88E-09	5.92E-06	6
CONGRESSION	21	6	-4.9647	3.44E-07	0.0011	7
CENP-A	15	5	-4.7392	1.07E-06	0.0034	8
AURORA	26	6	-4.6190	1.93E-06	0.0061	9
BUB1	9	4	-4.6112	2.00E-06	0.0063	10
CENP-H	4	3	-4.5986	2.13E-06	0.0067	11
MAD2	18	5	-4.4676	3.96E-06	0.0124	12
CENTROMERE	181	13	-4.4515	4.26E-06	0.0134	13
PROMETAPHASE	60	8	-4.3942	5.56E-06	0.0175	14
KINETOCHORE-MICROTUBULE	12	4	-4.2226	1.21E-05	0.0380	15
MTOC	12	4	-4.2226	1.21E-05	0.0380	16

Over-represented terms were defined as having *p*-value ≤ 0.05 after Bonferroni correction. The results were ordered by increasing *p*-values. The gene universe used is that based on the HG-U133A chip and contains 9638 genes. In total, 3148 tokens were tested.

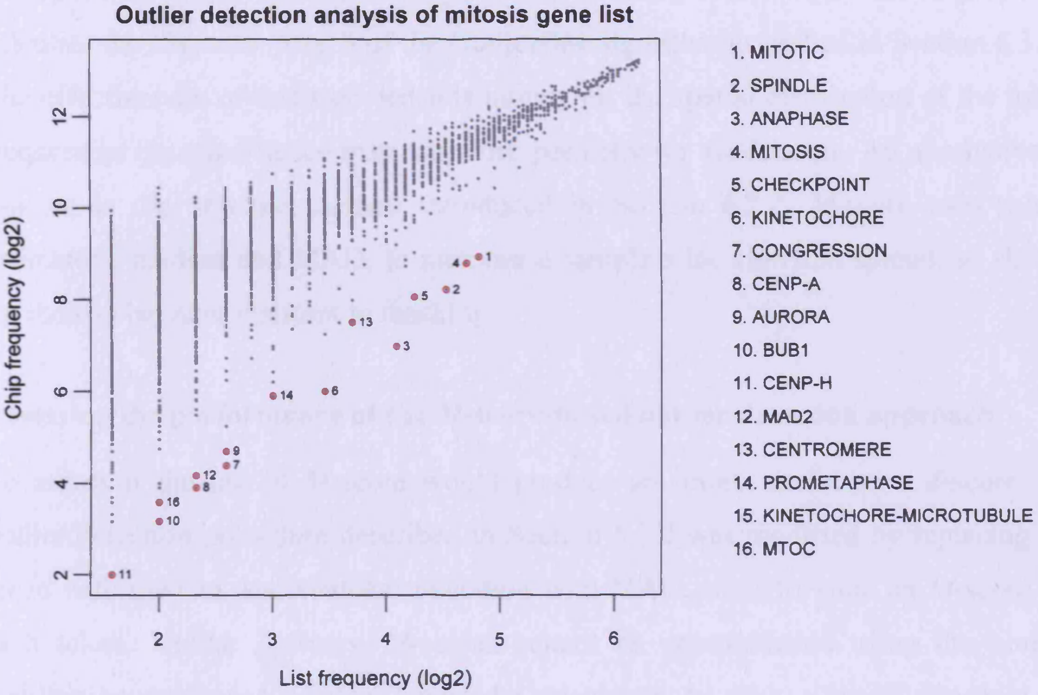


Figure 6.11: A scatter plot of *Chip* versus *List* frequencies for tokens in the mitosis gene list

Terms that were identified as significantly over-represented (Bonferroni p-value $\leq$ 0.05) in the mitosis gene list are circled in red and the adjacent numbers corresponding to their rankings. The Z-scores and p-values associated with the significant terms can be found in Table 6.4.

6.4.4 *Z*-scores versus *M*-scores in text-based ORA

A concern with the proposed outlier detection-based ORA approach is that it can be susceptible to a phenomenon termed "masking", in which presence of one (or more) outliers conceals the appearance of another outlier. Masking occurs when a small cluster of outliers attracts the mean and inflates the standard deviation in its direction, yielding small values for Z-scores (Hadi 1992). From the point of view of robust data analysis, the problem of masking is due to the fact that the mean and standard deviation used in Z-score have low breakdown point. To prevent these two estimators from being affected by a single or a few extreme values, the means and standard deviations are adjusted via local polynomial smoothing before they were employed to calculate the *Z*-scores (Step 3 of the OutlierDM algorithm described in Section 6.3.2). The effectiveness of this step depends largely on the spatial distribution of the token frequencies data and hence may not work perfectly for all datasets. An alternative to *Z*-score is the *M*-score method introduced in Section 6.2.2. *M*-score uses robust estimators, median and MAD, to measure a sample's location and spread, so should (in theory) be more resistant to masking.

Assessing the performance of the *M*-score-based outlier detection approach

To assess if the use of *M*-score would produce an improvement over *Z*-score, the outlier detection procedure described in Section 6.3.2 was modified by replacing the mean with median, and standard deviation with MAD, so as to yield an *M*-score for each token. Unlike *Z*-scores, *M*-scores cannot be approximated using the normal distribution and hence *p*-values cannot be calculated. As such, a cutoff threshold for *M*-score needs to be specified for labelling outliers. Iglewicz and Hoaglin (1993) recommended that observations with |*M*-score| > 3.5 are labelled as outliers. Using this as a guide, the false positive rates associated with the *M*-score-based outlier detection approach were examined at several cutoff points in order to find a suitable threshold that generates the least number of false positive under the null hypothesis. In this analysis, the set of random gene lists created in Section 6.4.2 were re-analysed with the *M*-score-based outlier detection approach. The proportions of tokens with *M*-scores more extreme than the following thresholds: -3.5, -3.75, -4.0, -4.25 and -4.5, were considered as significantly over-represented (i.e. false positives). The average number of false positives and the mean false positive rates were calculated based on Equations (6.7) and (6.8) specified in Section 6.4.2. The results of this analysis are shown in Table 6.5 and Table 6.6.

As expected, the mean number of false positives and the false positive rates associated with gene lists of various sizes decrease when the stringency of the *M*-score cutoff criteria increases. The simulation results showed that the cutoff threshold recommended by Iglewicz and Hoaglin (1993) is too liberal for mining token data, as up to 15 false positives could be obtained per analysis when $M < -3.5$ is used, which is too high in practice for text-based ORA. The threshold $M < -4.5$ is more appropriate for mining token data because the numbers of false positives and the corresponding FP rates are relatively low and at a level comparable to that of the Z-score-based approach (cf. Table 6.1).

Table 6.5: Mean number of false positives per gene list for several *M*-score cutoff values

N_{gene}	$M < -3.5$	$M < -3.75$	$M < -4.0$	$M < -4.25$	$M < -4.5$
50	7.63	5.16	3.43	2.29	1.54
100	10.57	6.93	4.56	2.95	1.95
300	14.42	8.83	5.32	3.16	1.84
500	14.93	8.58	4.74	2.56	1.23
1000	14.18	6.39	2.72	1.20	0.56
2000	8.35	3.46	1.54	0.61	0.27

Table 6.6: Mean false positive rates for several *M*-score cutoff values

N_{gene}	$M < -3.5$	$M < -3.75$	$M < -4.0$	$M < -4.25$	$M < -4.5$
50	0.00504	0.00343	0.00228	0.00153	0.00104
100	0.00419	0.00275	0.00181	0.00117	0.00078
300	0.00289	0.00177	0.00107	0.00064	0.00037
500	0.00224	0.00128	0.00071	0.00038	0.00019
1000	1.43E-03	6.48E-04	2.75E-04	1.21E-04	5.72E-05
2000	5.70E-04	2.36E-04	1.05E-04	4.13E-05	1.84E-05

For ease of discussion, the following notations will be used in the rest of this Chapter: *Out(Z-score, Bonferroni-P≤0.05)* refers to the original approach (i.e. OutlierDM) that is based on the use of the Z-scores and Bonferroni p-value ≤ 0.05 as cutoff. *Out(M-score<-4.5)* refers to the modified version that uses M-score < -4.5 as outlier identification criteria.

When *Out(M-score<-4.5)* was applied to the ISG gene list, 56 tokens were called significantly enriched (Table 6.7). Among these significant terms, 33 were reported as non-significant by *Out(Z-score, Bonferroni-P<0.05)*. These tokens were marked with 'No' in the column entitled 'Bonf' in Table 6.7. As can be seen from Figure 6.12, those tokens that were uniquely identified as significant by the *Out(M-score<-4.5)* method (points circled in blue) are located nearer to the main data cluster while those tokens that were called significant by both *Out(Z-score, Bonferroni-P<0.05)* and *Out(M-score<-4.5)* appear as the most extreme outliers (points circled in red). It can be reasoned that the most extreme outliers artificially inflate the standard deviations, resulting in small *Z*-scores for those 33 tokens such that their *p*-values just fall short of the cutoff after Bonferroni correction.

By using a less conservative multiple testing correction procedure such as the false discovery rate (FDR) along with the *Z*-score-based outlier detection procedure, it is possible to obtain a set of significant tokens that is comparable to that produced by *Out(M-score<-4.5)*. As shown in Table 6.7, the majority of the tokens that were reported as non-significant by *Out(Z-score, Bonferroni-P<0.05)* are called significant when the cutoff FDR-$P \leq 0.05$ were used instead. Only 8 tokens, e.g. 'virus', 'induction', 'host', 'treatment', 'hepatitis', 'Epstein-Barr' and 'EBV', remain non-significant. A strong resemblance between the results reported by *Out(M-score<-4.5)* and *Out(Z-score, FDR-P < 0.05)* were again observed for other real datasets like the mitosis gene list (Table 6.8).

Table 6.7: A comparison of the significant terms in the ISG gene list as reported by the *Z*-scores and *M*-scores-based outlier detection methods

Term	Chip frequency	List frequency	*M-scores*		*Z-scores*		
			M	Rank	*Z*	Bonf	FDR
INTERFERON	414	46	-24.95	1	-12.62	Yes	Yes
IFN	245	35	-18.43	2	-9.595	Yes	Yes
IFN-BETA	71	18	-12.52	3	-7.576	Yes	Yes
ANTIVIRAL	176	23	-11.92	4	-6.749	Yes	Yes
IFN-ALPHA	114	19	-11.28	5	-6.713	Yes	Yes
INTERFERON-ALPHA	59	14	-10.27	6	-6.621	Yes	Yes
INDUCIBLE	1068	37	-9.501	7	-4.787	Yes	Yes
OLIGOADENYLATE	18	8	-8.393	8	-6.075	Yes	Yes
INFECTION	1177	36	-8.285	9	-4.166	No	Yes
VIRAL	892	32	-8.127	10	-4.183	Yes	Yes
ISRE	31	9	-8.049	11	-5.708	Yes	Yes
DSRNA	60	11	-7.936	12	-5.397	Yes	Yes
ISG	14	7	-7.852	13	-5.775	Yes	Yes
INNATE	363	21	-7.484	14	-4.275	Yes	Yes
IMMUNE	1275	35	-7.248	15	-3.636	No	Yes
HLA-A	30	8	-7.186	16	-5.182	Yes	Yes
HLA-CLASS	11	6	-7.137	17	-5.304	Yes	Yes
IFN-GAMMA	443	22	-7.079	18	-3.982	No	Yes
TREAT	1817	40	-7.038	19	-3.439	No	Yes
STOMATITIS	52	9	-6.726	20	-4.748	Yes	Yes
IMMUNITY	387	20	-6.712	21	-3.869	No	Yes
HLA-B	25	7	-6.602	22	-4.833	Yes	Yes
LYMPHOBLASTOID	239	16	-6.594	23	-4.047	No	Yes
ENCEPHALOMYO-CARDITIS	16	6	-6.413	24	-4.751	Yes	Yes
EVASION	65	9	-6.155	25	-4.333	Yes	Yes
VIRUS	1408	34	-6.135	26	-3.062	No	No
HLA-G	10	5	-6.003	27	-4.458	Yes	Yes
OAS	10	5	-6.003	28	-4.458	Yes	Yes
MXA	11	5	-5.843	29	-4.334	Yes	Yes
TAPASIN	12	5	-5.696	30	-4.221	Yes	Yes
INDUCTION	2048	39	-5.641	31	-2.721	No	No
MHC	353	17	-5.634	32	-3.376	No	Yes
OR-C	5	4	-5.590	33	-4.099	No	Yes
LMP7	13	5	-5.562	34	-4.117	No	Yes
LMP2	13	5	-5.562	35	-4.117	No	Yes
BETA2-MICROGLOBULIN	42	7	-5.484	36	-3.990	No	Yes
GAMMA-INTERFERON	44	7	-5.384	37	-3.914	No	Yes
P69	6	4	-5.334	38	-3.901	No	Yes
INDIGENOUS	29	6	-5.266	39	-3.872	No	Yes
PKR	30	6	-5.200	40	-3.822	No	Yes
HOST	800	24	-5.180	41	-2.794	No	No
ISG15	7	4	-5.117	42	-3.734	No	Yes

(continued)

Table 6.7: A comparison of the significant terms in the ISG gene list as reported by the Z-scores and M-scores-based outlier detection methods (continued)

Term	*Chip* frequency	*List* frequency	*M-scores*		*Z-scores*		
			M	Rank	*Z*	*Bonf*	*FDR*
PEPTIDE-MHC	7	4	-5.117	43	-3.734	No	Yes
VSV	19	5	-4.923	44	-3.623	No	Yes
HISTOCOMPATIBILITY	303	14	-4.714	45	-2.961	No	No
TREATMENT	3120	45	-4.710	46	-2.186	No	No
MICROGLOBULIN	39	6	-4.694	47	-3.435	No	Yes
TAP	61	7	-4.680	48	-3.383	No	Yes
C1R	22	5	-4.676	49	-3.433	No	Yes
SP100	10	4	-4.615	50	-3.347	No	Yes
ANTI-HLA	10	4	-4.615	51	-3.347	No	Yes
ISGF3	10	4	-4.615	52	-3.347	No	Yes
HEPATITIS	366	15	-4.545	53	-2.794	No	No
HLA-C	24	5	-4.530	54	-3.320	No	Yes
EPSTEIN-BARR	233	12	-4.525	55	-2.956	No	No
EBV	194	11	-4.516	56	-3.013	No	No

This table listed all terms with *M*-score $\leq$ -4.5. The *Z*-score-based outlier detection method were used in conjunction with two multiple testing correction methods. Tokens with Bonferroni *p*-value $\leq$ 0.05 are marked with 'Yes' under the column *Bonf* and 'No' otherwise. Tokens with FDR *p*-value $\leq$ 0.05 are marked with 'Yes' under the column *FDR* and 'No' otherwise.

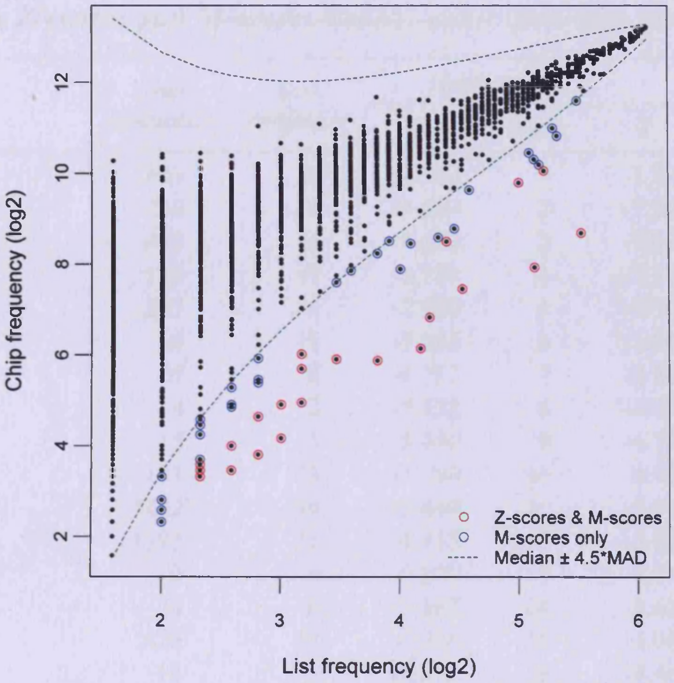


Figure 6.12: Scatter plot showing the locations of outliers identified by using *Z*-scores and *M*-scores-based methods

The ISG gene list was analysed with the two methods: *Out(Z-score, Bonferroni-P<0.05)* and *Out(M-score<-4.5)*. Terms that were identified as significantly enriched by both methods are circled in red, whereas those that were reported as significant by only the *Out(M-score<-4.5)* method are circled in blue. All significant terms are located outside the region marked by the median ± 4.5 MAD lines. The *Z*-scores and *M*-scores associated with the significant terms can be found in Table 6.7.

Table 6.8: A comparison of the significant terms in the mitosis gene list as reported by the *Z*-scores and *M*-scores-based outlier detection methods

Term	*Chip* frequency	*List* frequency	*M-scores*		*Z-scores*		
			M	Rank	*Z*	*Bonf*	*FDR*
MITOTIC	485	28	-10.300	1	-7.534	Yes	Yes
SPINDLE	298	23	-9.649	2	-7.263	Yes	Yes
MITOSIS	443	26	-9.550	3	-7.052	Yes	Yes
ANAPHASE	126	17	-9.183	4	-7.248	Yes	Yes
CHECKPOINT	267	19	-7.639	5	-5.906	Yes	Yes
KINETOCHORE	64	11	-7.083	6	-5.894	Yes	Yes
CONGRESSION	21	6	-5.757	7	-4.965	Yes	Yes
CENP-H	4	3	-5.522	8	-4.599	Yes	Yes
CENP-A	15	5	-5.496	9	-4.739	Yes	Yes
CENTROMERE	181	13	-5.490	10	-4.452	Yes	Yes
CYCLE	1882	36	-5.469	11	-3.871	No	Yes
HELA	1393	31	-5.433	12	-3.879	No	Yes
BUB1	9	4	-5.390	13	-4.611	Yes	Yes
AURORA	26	6	-5.367	14	-4.619	Yes	Yes
DIVISION	426	18	-5.286	15	-4.086	No	Yes
MAD2	18	5	-5.191	16	-4.468	Yes	Yes
PROMETAPHASE	60	8	-5.176	17	-4.394	Yes	Yes
MICROTUBULE	523	19	-4.975	18	-3.806	No	Yes
PROLIFERATING	523	19	-4.975	19	-3.806	No	Yes
MTOC	12	4	-4.950	20	-4.223	Yes	Yes
KINETOCHORE-MICROTUBULE	12	4	-4.950	21	-4.223	Yes	Yes
CENTROSOME	180	12	-4.947	22	-4.039	No	Yes
G1	468	18	-4.930	23	-3.803	No	Yes
CYTOKINESIS	152	11	-4.841	24	-3.989	No	Yes
ANEUPLOIDY	74	8	-4.730	25	-4.004	No	Yes
CHROMATID	74	8	-4.730	26	-4.004	No	Yes
M-PHASE	38	6	-4.674	27	-4.005	No	Yes
CDC20	15	4	-4.609	28	-3.921	No	Yes
INTERPHASE	253	13	-4.514	29	-3.638	No	Yes
MISSEGREGATION	16	4	-4.510	30	-3.834	No	Yes

This table listed all terms with *M*-score ≤ -4.5. The *Z*-score-based outlier detection method were used in conjunction with two multiple testing correction methods. Tokens with Bonferroni *p*-value ≤ 0.05 are marked with 'Yes' under the column *Bonf* and 'No' otherwise. Tokens with FDR *p*-value ≤ 0.05 are marked with 'Yes' under the column *FDR* and 'No' otherwise.

Reasons for using *Z*-scores instead of *M*-scores for text-based ORA

Based on the above findings, it can be concluded that the problem of masking can be addressed by using either the *M*-scores method or an outlier detection procedure that uses *Z*-scores coupled with less stringent multiple correction procedure such as the FDR. For text-based ORA, the *Z*-scores-based outlier detection framework may be preferred because of the following advantages over the *M*-scores-based method.

First, the probability of making a type I error (calling a token over-represented by mistake) can be controlled precisely at the desired significance level in the *Z*-score-based approach, whereas the threshold used in the *M*-score approach has to be chosen arbitrarily or based on empirical observations from simulation study, which is less mathematically tractable. The lack of control over the type I error could be particularly problematic when the underlying distribution has heavy tails because the *M*-score method may yield more false positives than expected.

Second, *p*-values can only be derived from *Z*-scores, and not from *M*-scores. Reporting the results of an over-representation analysis in terms of (adjusted) *p*-values, as opposed to a binary decision of whether a token is significant or not, provides a summary of the strength of evidence against the null hypothesis. Besides, *p*-values allow for direct comparisons of detection power between methods proposed in this thesis, e.g. permutation test and the hypergeometric distribution-based approaches.

Third, the *Z*-score-based outlier detection framework can be easily extended to include the common multiple testing correction procedures other than the Bonferroni and FDR, such as the Benjamini-Hochberg, Holm, Hommel and other methods. This feature is especially useful for exploratory analysis such as ORA because it enables researchers to explore the results generated from different correction methods at the desirable significance levels. The *M*-score-based method is lacking in this flexibility.

6.5 Discussion

Although outlier detection techniques have been suggested for numerous microarray-related data-mining tasks, such as to detect abnormal gene expression pattern in cancer samples (Tibshirani and Hastie 2007; Tomlins *et al.* 2005) and psychiatric disease (Ernst *et al.* 2008), the potential application of outlier detection technique in text-based enrichment analysis has not been investigated before. This idea was explored in this Chapter, including the development of a *Z*-score-based outlier detection framework, termed OutlierDM, for identifying PubMed abstract terms that are significantly over-represented in a list of differentially expressed genes. This method exploits the observation that biologically-plausible terms are often associated with a lower *Chip* frequency than expected by chance for a specific *List* frequency, and thus appear as outliers on a scatter plot of *Chip* versus *List* frequencies. These outliers can then be identified through the calculation of *Z*-scores, which are defined using the location and scale estimates of the token frequency data. Simulation studies and applications to selected microarray datasets showed that OutlierDM is appealing in terms of both detection power and false positive rates. The terms reported as significant were found in most cases to convey useful information and shed light on the biology under study. In addition, the proposed method is computationally efficient, requiring approximately 20-30 seconds to analyse a reasonably-sized gene list (e.g. less than 500 genes) on a desktop PC.

OutlierDM has several advantages over the classical hypergeometric distribution and permutation-based methods described previously:

- Unlike the classical hypergeometric distribution-based approach (Chapter 3), OutlierDM can compensate for the underlying annotation bias when applied to well-annotated gene list.
- Theoretical significance-level results can be obtained through the outlier detection-based approach; whereas only empirical *p*-value can be obtained with the permutation-based method (Chapter 5). The latter is typically computationally intractable and lower bounded by the ratio of 1 to the total number of permutations performed.

A central assumption underlying OutlierDM is that the Z-scores are normally distributed. Simulation results showed that the distributions of Z-scores from shorter gene lists are slightly negatively skewed, imposing a risk of finding up to two false positives when the query gene list has less than 300 genes (Section 6.4.2). However, it can be argued that since the text-based ORA analysis usually produces on average 8~10 tokens (this approximation is based on the HG-U133A array and may vary for other systems), one or two false positives can be easily spotted and the researchers may be ready to accept a small number of false positives as long as biologically-plausible findings can be obtained.

Chapter 7

Extended hypergeometric distribution-based ORA

7.1 Introduction

A difficulty associated with text analysis in functional genomics is that the availability and quality of literature is biased towards well-studied areas. As a consequence, the amount of articles present per gene is highly skewed. Failing to account for such bias can hamper the performance of text mining algorithms, as seen with the classical hypergeometric distribution-based ORA method described in Chapter 3.

In this Chapter, an ORA framework that uses the extended hypergeometric distribution to model token frequency data associated with a list of differentially expressed genes and to search for terms that are significantly over-represented is presented. This method is an extension of the classical hypergeometric distribution-based approach. The principle idea is to use the extra parameters in the extended hypergeometric distribution model in an attempt to account for the bias within the published literature.

This Chapter is structured as follows. Section 7.2 describes the theory of extended hypergeometric distribution in general. Section 7.3 presents the algorithm developed for assessing over-representation in token data. In Section 7.4, the performance of the proposed approach is evaluated based on an assessment of false positive rate and applications to real datasets. Finally, the advantages and limitations of the proposed method are discussed in Section 7.5.

7.2 Extended hypergeometric distribution

The extended hypergeometric distribution, also known as the Fisher non-central hypergeometric distribution, is a generalisation of the classical hypergeometric distribution that is designed for situation where the sampling procedure is biased (Fog 2008; Harkness 1965; Johnson *et al.* 2005). To illustrate key concepts, consider what is known as the urn experiment. Assume that n balls are drawn without replacement from a population (the urn) containing N balls, of which m_1 are red and m_2 are white. The balls have different weights, where the weight for each red and white ball is w_1 and w_2, respectively. When sampling is unbiased ($w_1 = w_2$), the balls have equal probability of being taken ($p_1 = p_2$) and the results will follow the classical hypergeometric distribution. However, if sampling is biased such that the probability of taking a ball of one colour is proportional to its weight but independent of the other balls, then the number of balls of a particular colour drawn will follow the binomial distribution:

$$x_i \sim \text{binomial } (m_i, p_i), \text{ where } i = 1, 2$$

On the condition that the sum of the independent binomial variables is fixed (i.e. $\sum x_i = n$), the number of red balls in our sample x will follow the extended hypergeometric distribution. The probability of seeing x red balls simply by chance is:

$$\Pr[m_1 = x \mid m_1 + m_2 = n] = \frac{\binom{m_1}{x}\binom{m_2}{n-x}\theta^x}{\sum_x \binom{m_1}{x}\binom{m_2}{n-x}\theta^x} \qquad \text{Equation (7.1)}$$

where $\max(0, n - m_2) \le x \le \min(m_1, n)$; the same limits apply to the summation in Equation (7.1). θ is the odds ratio, which is a measure of bias and is equivalent to the probability ratio of red over white balls:

$$\theta = \frac{p_1(1-p_2)}{p_2(1-p_1)} \qquad \text{Equation (7.2)}$$

Let $x_1, x_2, \ldots, x_s$ be a sample of S independent observations from distribution specified by Equation (7.1). Harkness (1965) proposed obtaining the lower and upper bounds for the maximum-likelihood estimation for θ based on the natural estimator:

$$\tilde{\theta} = \frac{\bar{x}(m_2 - n + \bar{x})}{(m_1 - \bar{x})(n - \bar{x})} \qquad \text{Equation (7.3)}$$

The most common application of the extended hypergeometric distribution is as a model for the alternative hypothesis in the analysis of contingency tables. By assuming $\theta = 1$ or equivalently $p_1 = p_2$, Equation (7.1) reduces to the classical hypergeometric mass function specified in Equation (3.1) in Chapter 3. This is the basis of the Fisher's exact test of significance in 2 × 2 contingency tables.

7.3 ExtendedHG: a statistical framework for identifying over-represented PubMed abstract terms using the extended hypergeometric distribution

7.3.1 Basic ideas

When the amount of annotation associated with a gene list is higher than expected by chance, the sampling will be biased in favour of certain types of tokens. As a consequence, some common words or non-specific terms shared by most abstracts, such as 'cell', 'analysis' and 'molecule', will have higher probabilities of being selected along with genes in the gene list. As such, the token significance inferred from the classical hypergeometric distribution is likely to be misleading and erroneous, because this approach does not account for the excess annotation. In this Chapter it is reasoned that it may be possible to account for the effect due to annotation bias by using a biased sampling approach such as the extended hypergeometric distribution. To explore this idea an ORA system that uses the extended hypergeometric distribution to establish statistically which biological concepts or tokens from the PubMed abstracts are significantly over-represented within a gene list was developed. This method is denoted as ExtendedHG hereafter.

7.3.2 Algorithm and procedure

ExtendedHG consists of the following computation steps:

1. *Odds ratio estimation.* Given a gene list of size n and a corpus of text that is relevant to the genes under study, the first step is to determine the *List* and *Chip* frequencies of tokens found in the input gene list. Let the number of genes associated with a particular token T in the background be m_1, and the number of genes associated with tokens other than T in the background be m_2. The odds ratio of T is estimated by substituting n, m_1 and m_2 into Equation (7.3) to give:

$$\text{odds ratio} = \frac{\bar{y}(m_2 - n + \bar{y})}{(m_1 - \bar{y})(n - \bar{y})} \qquad \text{Equation (7.4)}$$

   where $\bar{y}$ is the mean number of genes expected to be associated with T under the null hypothesis. There is no simple and explicit expression for $\bar{y}$ because its value depends on the degree of annotation bias inherent with the input gene list, which will vary from gene list to gene list. In the current implementation, $\bar{y}$ is determined empirically by fitting a degree 7 polynomial regression line through the token frequency data:

$$y \sim x^1 + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 \qquad \text{Equation (7.5)}$$

   In this model, the explanatory variable x represents *Chip* frequency (equivalent to m_1); the response variable y represents *List* frequency. The best-fitting curve is obtained by the method of least squares implemented in the R function `lm`. For each token, the fitted value of *List* frequency expected for a given *Chip* frequency is determined from the best-fitting curve. This fitted value is a good approximation for $\bar{y}$ and is used to calculate the odds ratio. The reason of using a degree 7 polynomial fit is given in Section 7.3.3.

2. *Jackknife adjustment.* To ascertain that terms supported by few genes are not over-weighted, the *List* frequency is jackknifed (i.e. *List* frequency - 1). This operation entails a conservative adjustment to the p-value by penalising the significance of terms supported by few genes. The rationale behind jackknife adjustment has been discussed in Section 3.2.4. Only tokens that are associated with at least two genes after jackknifing are retained for further analysis.

3. *P-value calculation.* To calculate a p-value for each token based on the extended hypergeometric distribution, the `pFNCHypergeo` function implemented in the BiasedUrn R package[1] is used:

```
pFNCHypergeo(x, m1, m2, n, odds, precision, lower.tail)
```

where

`x`	=	Jackknifed *List* frequency – 1
`m1`	=	*Chip* frequency
`m2`	=	Total number of genes in the gene universe – *Chip* frequency
`n`	=	Number of genes in the gene list – 1
`odds`	=	Odds ratio determined from Step 1

The value for the argument `precision` is set at 10^{-40}. To obtain a one-tailed test for over-representation, the argument `lower.tail` is set to `FALSE`.

4. *Multiple testing correction and criterion of over-representation.* A token is considered significantly over-represented in the gene list if the corresponding p-value is less than 0.05 after Bonferroni adjustment.

The script for ExtendedHG was developed and tested under R-2.6, and the source code can be found in Appendix B.

7.3.3 Odds ratio estimation

In order to obtain an estimate for the odds ratio associated with a particular token T in a given gene list, four pieces of information are required: (1) the mean number of genes expected to be associated with T just by chance, $\mu_{extended}$; (2) the number of genes associated with T in the background; (3) the number of genes associated with tokens other than T in the background; (4) the size of the input gene list. All parameters, except for $\mu_{extended}$, can be determined directly from the data itself. For the classical hypergeometric distribution, the mean number of genes expected to be associated with T just by chance, $\mu_{classical}$, can be obtained simply as:

[1] http://cran.r-project.org/web/packages/BiasedUrn/index.html

$$\mu_{classical} = n \times \frac{M}{N} \qquad \text{Equation (7.6)}$$

where n = number of genes in the input gene list, M = *Chip* frequency, and N = total number of genes in the background. However, there is no simple explicit expression for the mean in the extended hypergeometric distribution.

In ExtendedHG, $\mu_{extended}$ is predicted from a polynomial regression fit to the token frequency data. This idea was motivated by the findings that the hypergeometric mean $\mu_{classical}$ for a specific token in a random gene list can be estimated by fitting a regression line through the *Chip* and *List* frequencies. This is illustrated in Figure 7.1, in which a random gene list containing 253 genes was created by sampling without replacement from the HG-U133A array, and then a seventh order polynomial regression line was fitted to the token frequencies in this random gene list according to the regression model specified in Equation (7.5). As shown in Figure 7.1(a), the goodness of fit as measured by R^2 is 0.989, indicative of an apparently good fit to the data (the definition of R^2 will be given later). The fitted values were compared to the corresponding $\mu_{classical}$ values as calculated using Equation (7.6). It is clear from Figure 7.1(b) that the fitted values provide a good approximation to $\mu_{classical}$. Similar results were also found for random gene lists of various lengths. Although illustrated here using simulated gene list, it is reasonable to assume that $\mu_{extended}$ can be estimated in the same way by fitting polynomials to real gene lists. The argument made here is that the polynomial regression model should be able to capture the general tendency and intrinsically account for the shift in token distribution caused by annotation bias (if any), and provides a reasonable estimation for $\mu_{extended}$.

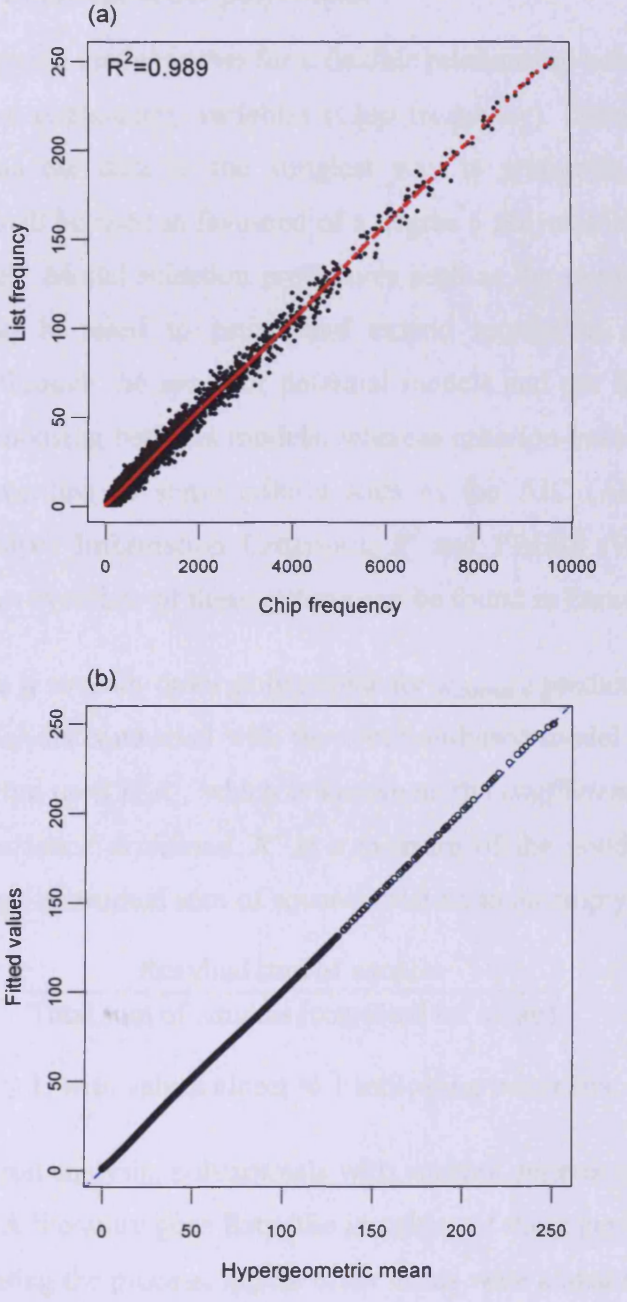


Figure 7.1: A comparison of the hypergeometric mean and fitted values predicted from a seventh order polynomials fit to the token frequency data
(a) A scatter plot of *Chip* and *List* frequencies from a random gene list containing 253 annotated genes. The red data points represent the fitted values predicted from a degree 7 polynomial regression model. (b) The hypergeometric means $\mu_{classical}$ are plotted against the corresponding fitted values derived from the polynomial regression shown in (a). The blue dashed line is the *y*=*x* line.

Reasons for using a seventh order polynomial

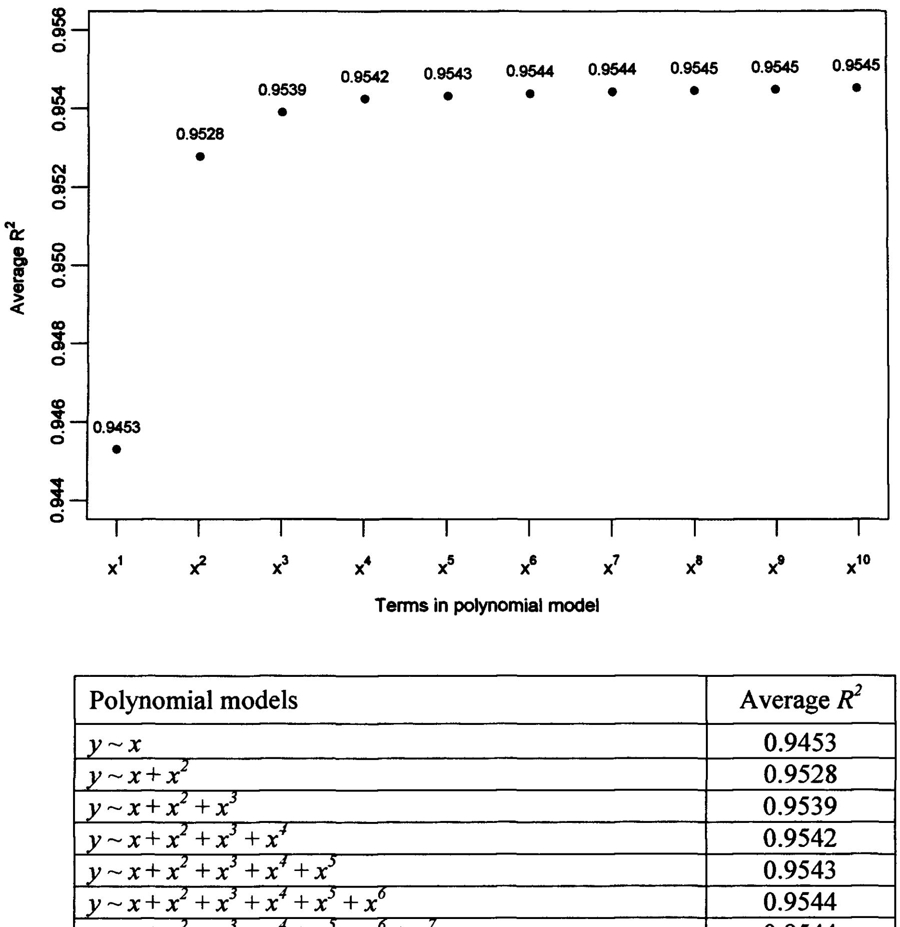
A polynomial regression model allows for a flexible relationship between the response (*List* frequency) and explanatory variables (*Chip* frequency). Usually, a polynomial model that explains the data in the simplest way is preferred, e.g. a degree 3 polynomial model will be used in favoured of a degree 6 polynomial model if both fit the data equally well. Model selection procedures such as the stepwise and criterion-based methods can be used to prune and extend regression models. Stepwise procedures search through the space of potential models and use hypothesis testing-based method for choosing between models, whereas criterion-based methods choose the best model according to some criteria such as the AIC (Akaike Information Criterion), BIC (Bayes Information Criterion), R^2 and PRESS (Predicted Residual Sum of Squares). An overview of these criteria can be found in Faraway (2002).

The decision to use a seventh order polynomial for $\mu_{extended}$ prediction was based on the results of an analysis conducted with the criterion-based model selection method. The selection criterion used is R^2, which is known as the *coefficient of determination* or *percentage of variance explained.* R^2 is a measure of the goodness of fit and is defined as the change in residual sum of squares relative to an empty model:

$$R^2 = 1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares (corrected for mean)}} \qquad \text{Equation (7.7)}$$

The range is $0 \leq R^2 \leq 1$, with values closer to 1 indicating better fits.

In this model selection analysis, polynomials with varying degrees were fitted to each of the 52 HG-U133A literature gene lists (the identities of these gene lists are detailed in Appendix A). During the process, higher order terms were added sequentially to the polynomial model such that the regression model was expanded from x to x^{10} in a stepwise manner. The R^2 values corresponding to the models tested were recorded at each step. Then, an average R^2 was calculated to yield an overall goodness of fit for each polynomial model. The results are shown in Figure 7.2. It can be seen that polynomial models containing the x^6, x^7, x^8, x^9 and x^{10} terms have very similar average R^2 values, which range from 0.9544 to 0.9545. It was observed that bigger models with higher terms are susceptible to over-fitting at high *Chip* frequency regions where data is sparse. Therefore the degree 7 polynomial model was chosen because it should be less susceptible to over-fitting compared to bigger models.



Polynomial models	Average R^2
$y \sim x$	0.9453
$y \sim x + x^2$	0.9528
$y \sim x + x^2 + x^3$	0.9539
$y \sim x + x^2 + x^3 + x^4$	0.9542
$y \sim x + x^2 + x^3 + x^4 + x^5$	0.9543
$y \sim x + x^2 + x^3 + x^4 + x^5 + x^6$	0.9544
$y \sim x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7$	0.9544
$y \sim x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8$	0.9545
$y \sim x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9$	0.9545
$y \sim x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^9 + x^{10}$	0.9545

Figure 7.2: Polynomial models selection based on R^2

Polynomials were fitted to each of the 52 HG-U133A literature gene lists in a stepwise manner. Higher order terms were added sequentially to the regression model and the corresponding R^2 values recorded. An average R^2 was calculated for each polynomial model by taking the means of the individual R^2.

Adjustment of the odds ratio

The odds ratio estimated using the procedures described above will always be in the interval (0, ∞). For example, the odds ratios for tokens in the ISG gene list lie in the range (0.036, 82.545) and have a median value of 1.633. As illustrated in Figure 7.3(a), the majority of the tokens in the ISG gene list have similar odds ratios; but a small number of tokens that have low background frequencies (e.g. *Chip* frequency < 10) appear to be associated with relatively high odds ratios. Recall that the odds ratio is calculated as a function of mean ($\bar{y}$), which is predicted from the best-fitting curve. By way of intuition, tokens that are associated very few genes in a gene list and in the background are considered unstable; therefore $\bar{y}$ values predicted for these tokens are less accurate. When a token has very low background frequency and the predicted mean $\bar{y}$ is close to the value of its *Chip* frequency, the denominator in Equation (7.4) will become small and the numerator will become large, yielding a high odds ratio. In theory, the higher the odds ratio, the harder it is for a token to be called significant. A consequence of this is that tokens with low *List* and *Chip* frequency may never be detected as significant (even though they are truly enriched in the gene list), simply because their odds ratios are inflated. To prevent this problem, an *ad hoc* tuning procedure was devised. First, the median of the odds ratio for all tokens were calculated, say *med*. Then, tokens for which the odds ratios are 3 MAD (median absolute deviation) away from *med* were identified, and the odds ratios for these tokens were replace by *med*. The interval of the odds ratio for tokens in the ISG gene list becomes (0.036, 5.374) after the above adjustment has been applied. The distribution of the adjusted odds ratio is shown in Figure 7.3(b). It can be seen that the extreme odds ratios were adjusted to a level equal to the median (i.e. 1.623), whereas for most observations their odds ratios remain unchanged.

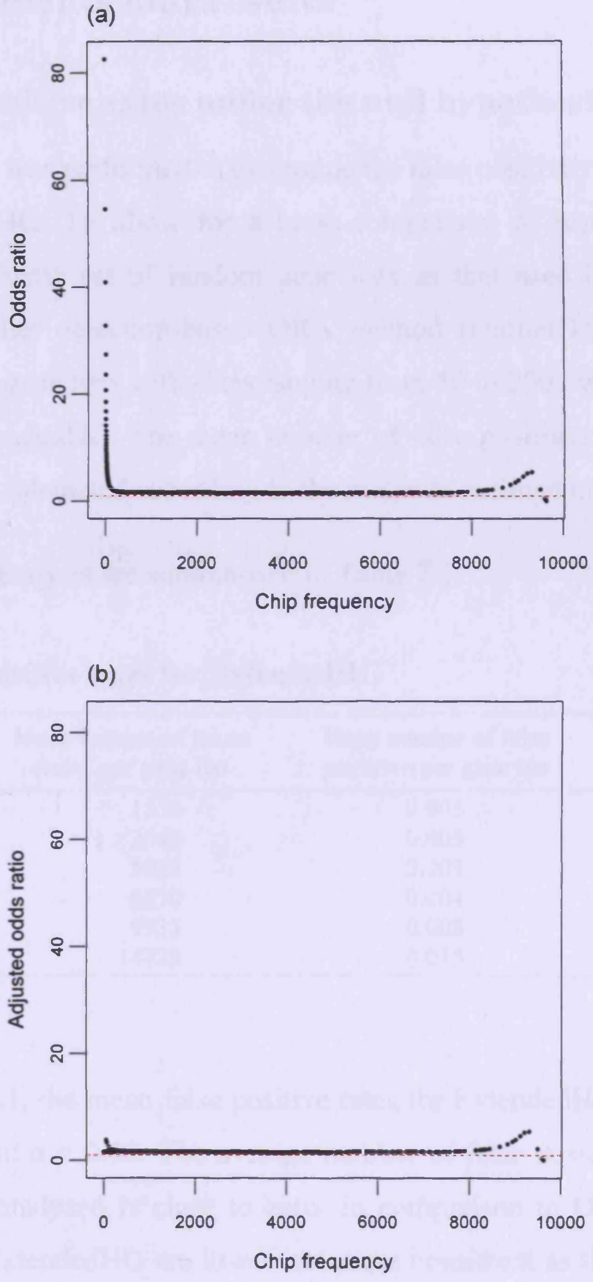


Figure 7.3: Odds ratio before and after adjustment

The (a) raw odds ratio and (b) adjusted odds ratio for tokens in the ISG gene list were plotted against their corresponding *Chip* frequencies. The red dashed line is $y = 1$.

7.4 Experiments and results

7.4.1 False positive rates under the null hypothesis

A simulation study was performed to determine the false positive rate for the proposed method, ExtendedHG. To allow for a cross-comparison of results, the study was performed on the same set of random gene lists as that used in Section 6.4.2 for evaluating the outlier detection-based ORA method (OutlierDM). Briefly, a large number of random gene lists with sizes ranging from 50 to 2000 were constructed and analysed with ExtendedHG. The mean number of false positives and the mean false positive rates were calculated according to the methods outlined in Section 6.4.2.

The results of this analysis are summarised in Table 7.1.

Table 7.1: False positive rates for ExtendedHG

N_{gene}	Mean number of token tested per gene list	Mean number of false positives per gene list	Mean FP rate
50	1554	0.005	3.90E-06
100	2540	0.005	2.12E-06
300	5025	0.003	5.98E-07
500	6670	0.004	6.33E-07
1000	9935	0.008	8.23E-07
2000	14728	0.015	1.01E-06

As seen in Table 7.1, the mean false positive rates for ExtendedHG ranges from 1.01 $\times 10^{-6}$ to 3.9 $\times 10^{-6}$ at $\alpha = 0.05$. The average number of false positives that one might find per gene list analysed is close to zero. In comparison to OutlierDM, the false positive rates for ExtendedHG are lower and more consistent as they are not affected by the size of gene list. Taken together, the results presented in Table 7.1 suggest that the proportion of false positives produced by ExtendedHG is kept within the level controlled by the chosen significance threshold α, and this is true regardless of the number of genes analysed.

7.4.2 Performance on real datasets

As with OutlierDM, the performance of ExtendedHG was assessed by testing on the three datasets: ISG gene list from Sanda *et al.* (2006), the mitosis gene list from Lee *et al.* (2004), and the glycolysis gene list from Vanharanta *et al.* (2006). A threshold of 0.05 for the Bonferroni adjusted *p*-values was used to identify significantly enriched terms.

Example 1: ISG gene list

38 tokens were identified as significantly over-represented in the ISG gene list by ExtendedHG. These tokens and their *p*-values are presented in Table 7.2. All hits, except for 'innate' (rank 12), 'treat' (rank 20) and 'host' (rank 30), are biologically-specific terms and convey useful biological insights to the functions shared by the list of genes being analysed. Non-specific terms that were called significant by the classical hypergeometric test-based approach such as 'synthesis', 'molecule', 'after', were successfully avoided by using ExtendedHG.

A comparison of the results obtained by the permutation test, OutlierDM and ExtendedHG when applied to the ISG gene list is presented in Table 7.3. There appears to be a reasonable agreement in general between the different methods, in particular for the top ranking terms. 17 biologically-meaningful terms are common between the three methods, including 'interferon', 'IFN', 'antiviral', 'IFN-beta', 'IFN-alpha', 'inducible', 'viral', 'oligoadenylate', 'interferon-alpha', 'ISRE', 'ISG', 'HLA-class', 'stomatitis', 'evasion', 'encephalomyocarditis', 'dsRNA' and 'OAS'.

The overlaps between OutlierDM and ExtendedHG were examined more closely. In many cases, OutlierDM and ExtendedHG capture similar information and appear generally comparable in terms of statistical power for this dataset (Figure 7.4(a)), although ExtendedHG did pick up some apparently relevant terms that OutlierDM missed, such as 'IFN-gamma', 'LMP2', 'LMP7' and 'beta2-microglobulin' (Table 7.3). The term ranks given to the top 100 tokens in this gene list by the two methods were compared and shown in Figure 7.4(b). Despite minor differences in rank order, there appears to be a good concordance between OutlierDM and ExtendedHG.

Table 7.2: Significantly over-represented abstract terms in the ISG gene list as identified using the extended hypergeometric test-based ORA method

Term	*Chip* frequency	*List* frequency	Odds ratio	p-value	Bonferroni p-value	Rank
INTERFERON	414	46	1.6646	4.64E-37	1.58E-33	1
IFN	245	35	1.8486	1.16E-28	3.95E-25	2
ANTIVIRAL	176	23	2.0648	4.84E-16	1.65E-12	3
IFN-BETA	71	18	3.3446	4.93E-14	1.68E-10	4
IFN-ALPHA	114	19	2.5199	1.59E-13	5.41E-10	5
INDUCIBLE	1068	37	1.6191	3.99E-12	1.36E-08	6
VIRAL	892	32	1.6129	1.88E-10	6.40E-07	7
INFECTION	1177	36	1.6228	3.39E-10	1.16E-06	8
OLIGOADENYLATE	18	8	1.6234	4.71E-10	1.61E-06	9
INTERFERON-ALPHA	59	14	3.7999	9.63E-10	3.28E-06	10
ISRE	31	9	1.6234	1.05E-09	3.58E-06	11
INNATE	363	21	1.6959	1.65E-09	5.63E-06	12
ISG	14	7	1.6234	4.47E-09	1.52E-05	13
IFN-GAMMA	443	22	1.6517	5.24E-09	1.79E-05	14
IMMUNE	1275	35	1.6255	1.28E-08	4.35E-05	15
IMMUNITY	387	20	1.6796	2.65E-08	9.05E-05	16
HLA-A	30	8	1.6234	2.70E-08	9.19E-05	17
HLA-CLASS	11	6	1.6234	6.74E-08	0.0002	18
STOMATITIS	52	9	1.6234	8.24E-08	0.0003	19
TREAT	1817	40	1.6256	9.67E-08	0.0003	20
LYMPHOBLASTOID	239	16	1.8617	1.71E-07	0.0006	21
HLA-B	25	7	1.6234	2.38E-07	0.0008	22
EVASION	65	9	1.6234	4.89E-07	0.0017	23
VIRUS	1408	34	1.6280	5.36E-07	0.0018	24
ENCEPHALOMYOCARDITIS	16	6	1.6234	6.09E-07	0.0021	25
DSRNA	60	11	3.7547	1.47E-06	0.0050	26
MHC	353	17	1.7037	1.54E-06	0.0052	27
OAS	10	5	1.6234	2.90E-06	0.0099	28
HLA-G	10	5	1.6234	2.90E-06	0.0099	29
HOST	800	24	1.6108	3.77E-06	0.0129	30
MXA	11	5	1.6234	4.52E-06	0.0154	31
INDUCTION	2048	39	1.6187	5.96E-06	0.0203	32
BETA2-MICROGLOBULIN	42	7	1.6234	6.04E-06	0.0206	33
TAPASIN	12	5	1.6234	6.72E-06	0.0229	34
GAMMA-INTERFERON	44	7	1.6234	7.98E-06	0.0272	35
LMP7	13	5	1.6234	9.62E-06	0.0328	36
LMP2	13	5	1.6234	9.62E-06	0.0328	37
OR-C	5	4	1.6234	1.33E-05	0.0455	38

Over-represented terms were defined as having p-value ≤ 0.05 after Bonferroni correction. The results were ordered by increasing p-values. The gene universe used is that based on the HG-U133A chip and contains 9638 genes. 3412 tokens were tested.

Table 7.3: A comparison of the results from permutation test, OutlierDM and ExtendedHG when applied to the ISG gene list

Term	*Chip* frequency	*List* frequency	Bonferroni *p*-value		
			Permutation	OutlierDM	ExtendedHG
INTERFERON	414	46	**< 0.0484**	**2.81E-33**	**1.58E-33**
IFN	245	35	**< 0.0484**	**1.43E-18**	**3.95E-25**
ANTIVIRAL	176	23	**< 0.0484**	**2.54E-08**	**1.65E-12**
IFN-BETA	71	18	**< 0.0484**	**6.06E-11**	**1.68E-10**
IFN-ALPHA	114	19	**< 0.0484**	**3.24E-08**	**5.41E-10**
INDUCIBLE	1068	37	**< 0.0484**	**0.0029**	**1.36E-08**
VIRAL	892	32	**< 0.0484**	**0.0490**	**6.40E-07**
INFECTION	1177	36	**< 0.0484**	0.0529	**1.16E-06**
OLIGOADENYLATE	18	8	**< 0.0484**	**2.11E-06**	**1.61E-06**
INTERFERON-ALPHA	59	14	**< 0.0484**	**6.09E-08**	**3.28E-06**
ISRE	31	9	**< 0.0484**	**1.95E-05**	**3.58E-06**
INNATE	363	21	0.0968	**0.0326**	**5.63E-06**
ISG	14	7	**< 0.0484**	**1.31E-05**	**1.52E-05**
IFN-GAMMA	443	22	0.0968	0.1160	**1.79E-05**
IMMUNE	1275	35	0.5325	0.4720	**4.35E-05**
IMMUNITY	387	20	**< 0.0484**	0.1860	**9.05E-05**
HLA-A	30	8	1	**0.0004**	**9.19E-05**
HLA-CLASS	11	6	**< 0.0484**	**0.0002**	**0.0002**
STOMATITIS	52	9	**< 0.0484**	**0.0035**	**0.0003**
TREAT	1817	40	**< 0.0484**	0.9950	**0.0003**
LYMPHOBLASTOID	239	16	**< 0.0484**	0.0885	**0.0006**
HLA-B	25	7	1	**0.0023**	**0.0008**
EVASION	65	9	**0.0484**	**0.0250**	**0.0017**
VIRUS	1408	34	0.8714	1	**0.0018**
ENCEPHALOMYOCARDITIS	16	6	**< 0.0484**	**0.0035**	**0.0021**
DSRNA	60	11	**< 0.0484**	**0.0001**	**0.0050**
MHC	353	17	1	1	**0.0052**
OAS	10	5	**< 0.0484**	**0.0141**	**0.0099**
HLA-G	10	5	1	**0.0141**	**0.0099**
HOST	800	24	1	1	**0.0129**
MXA	11	5	1	**0.0250**	**0.0154**
INDUCTION	2048	39	0.4357	1	**0.0203**
BETA2-MICROGLOBULIN	42	7	1	0.1130	**0.0206**
TAPASIN	12	5	1	**0.0415**	**0.0229**
GAMMA-INTERFERON	44	7	**0.0484**	0.1550	**0.0272**
LMP7	13	5	0.3873	0.0656	**0.0328**
LMP2	13	5	0.3389	0.0656	**0.0328**
OR-C	5	4	0.0968	0.0709	**0.0455**
INDIGENOUS	29	6	**0.0484**	0.1840	0.0502

Abstract terms that were identified as over-represented by the corresponding methods are highlighted in yellow and in bold. The cutoff used is Bonferroni *p*-value ≤ 0.05.

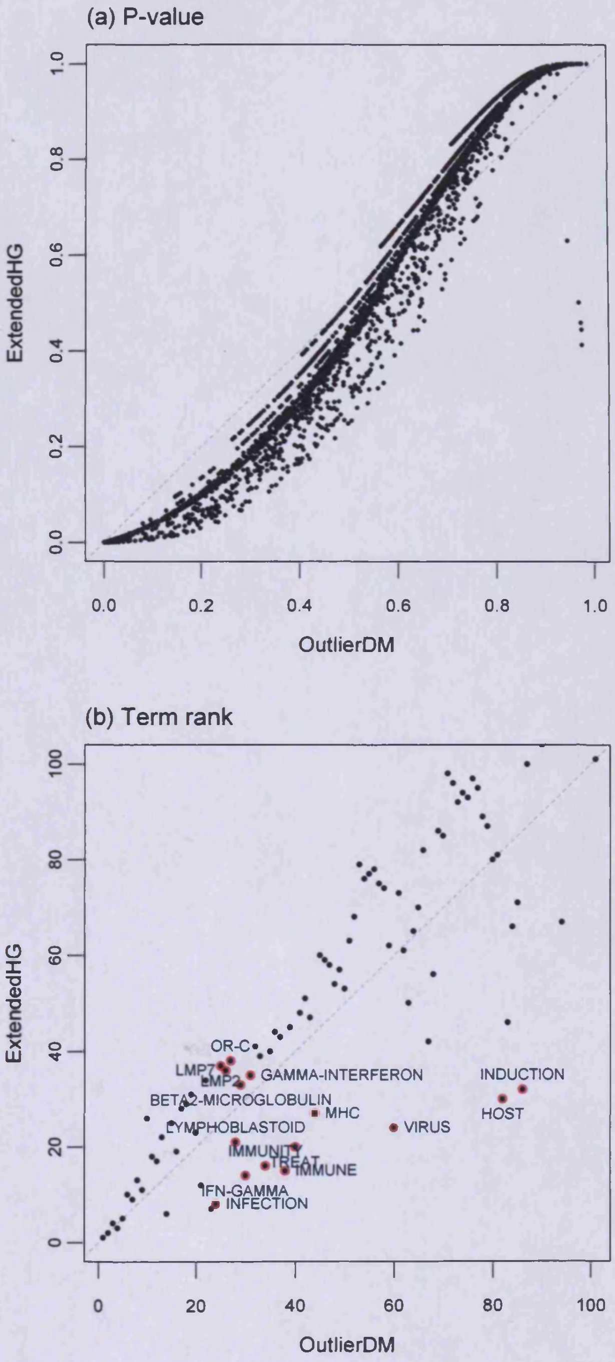


Figure 7.4: Concordance between OutlierDM and ExtendedHG

(a) A comparison of the unadjusted p-values reported by OutlierDM and ExtendedHG for all tokens in the ISG gene list. (b) The rankings for the top 100 tokens in the ISG gene list as reported by OutlierDM are plotted against that of ExtendedHG. Tokens that were called significant only by ExtendedHG are labelled and circled in red. The grey dashed line is the $y = x$ line.

Example 2: Glycolysis gene list

The classical hypergeometric distribution-based approach reported 48 terms as significantly enriched in this gene list, among which 50% were considered as "noise" due to annotation bias (Table 3.3). This problem appears to be largely avoided when using ExtendedHG; only one term, 'glycolytic' (Bonferroni *p*-value = 0.0299), was identified as significant at the chosen *p*-value threshold of 0.05. OutlierDM appears to perform better in this dataset insofar as it detects three additional biologically-plausible terms ('aldo-keto', 'peroxidation' and 'nicotiamide') as significantly over-represented.

Example 3: Mitosis gene list

ExtendedHG identified 19 tokens as having a Bonferroni *p*-value ≤ 0.05 in the mitosis gene list (Table 7.4), all of which - except for the term 'hela' (rank 15) - appear relevant to the biology of cell cycle progression and regulation. There are substantial overlaps between the significant terms produced by ExtendedHG, OutlierDM and the permutation test-based approach (Table 7.5). Terms that were assigned significant *p*-values by all three methods include 'mitosis', 'mitotic', 'spindle', 'anaphase', 'checkpoint', 'kinetochore', 'congression', 'prometaphase' and 'centromere'. It was observed that hits unique to OutlierDM are mainly associated with low *Chip* and *List* frequencies whereas hits unique to ExtendedHG tend to have higher *Chip* frequency (Figure 7.5). The latter group of tokens (e.g. 'cycle', 'division', 'hela') also appear to carry less specific biological information compared to the former group of tokens (e.g. 'BUB1' 'MAD2', 'kinetochore-microtubule').

Table 7.4: Significantly over-represented abstract terms in the mitosis gene list as identified using the extended hypergeometric test-based ORA method

Term	*Chip* frequency	*List* frequency	Odds ratio	*p*-value	Bonferroni *p*-value	Rank
MITOTIC	485	28	1.3582	6.07E-14	1.91E-10	1
SPINDLE	298	23	1.4733	4.31E-13	1.36E-09	2
MITOSIS	443	26	1.3726	6.58E-13	2.07E-09	3
ANAPHASE	126	17	2.0132	2.82E-11	8.87E-08	4
CHECKPOINT	267	19	1.5141	6.07E-10	1.91E-06	5
KINETOCHORE	64	11	3.0085	1.02E-06	0.0032	6
DIVISION	426	18	1.3797	1.13E-06	0.0036	7
CONGRESSION	21	6	1.3192	1.64E-06	0.0052	8
PROMETAPHASE	60	8	1.3192	2.01E-06	0.0063	9
CENTROMERE	181	13	1.7171	2.79E-06	0.0088	10
PROLIFERATING	523	19	1.3482	3.07E-06	0.0097	11
MICROTUBULE	523	19	1.3482	3.07E-06	0.0097	12
G1	468	18	1.3636	3.50E-06	0.0110	13
AURORA	26	6	1.3192	5.10E-06	0.0161	14
HELA	1393	31	1.3189	5.49E-06	0.0173	15
CHROMATID	74	8	1.3192	8.33E-06	0.0262	16
ANEUPLOIDY	74	8	1.3192	8.33E-06	0.0262	17
CYCLE	1882	36	1.3159	1.14E-05	0.0359	18
CENP-A	15	5	1.3192	1.20E-05	0.0379	19

Over-represented terms were defined as having p-value $\leq$ 0.05 after Bonferroni correction. The results were ordered by increasing p-values. The gene universe used is that based on the HG-U133A chip and contains 9638 genes. 3148 tokens were tested.

Table 7.5: A comparison of the results from different methods when applied to the mitosis gene list

Term	*Chip* frequency	*List* frequency	Bonferroni *p*-value		
			Permutation	OutlierDM	ExtendedHG
MITOTIC	485	28	**< 0.0442**	**7.72E-11**	**1.91E-10**
SPINDLE	298	23	**< 0.0442**	**5.96E-10**	**1.36E-09**
MITOSIS	443	26	**< 0.0442**	**2.77E-09**	**2.07E-09**
ANAPHASE	126	17	**< 0.0442**	**6.67E-10**	**8.87E-08**
CHECKPOINT	267	19	**< 0.0442**	**5.53E-06**	**1.91E-06**
KINETOCHORE	64	11	**< 0.0442**	**5.92E-06**	**0.0032**
DIVISION	426	18	**< 0.0442**	0.0691	**0.0036**
CONGRESSION	21	6	**< 0.0442**	**0.0011**	**0.0052**
PROMETAPHASE	60	8	**< 0.0442**	**0.0175**	**0.0063**
CENTROMERE	181	13	**< 0.0442**	**0.0134**	**0.0088**
PROLIFERATING	523	19	**< 0.0442**	0.2220	**0.0097**
MICROTUBULE	523	19	1	0.2220	**0.0097**
G1	468	18	0.3982	0.2250	**0.0110**
AURORA	26	6	1	**0.0061**	**0.0161**
HELA	1393	31	**< 0.0442**	0.1650	**0.0173**
CHROMATID	74	8	0.9290	0.0979	**0.0262**
ANEUPLOIDY	74	8	0.4866	0.0979	**0.0262**
CYCLE	1882	36	1	0.1710	**0.0359**
CENP-A	15	5	0.4424	**0.0034**	**0.0379**
MAD2	18	5	1	**0.0124**	0.0829
INTERPHASE	253	13	**0.0442**	0.4320	0.0893
BUB1	9	4	1	**0.0063**	0.2500
MTOC	12	4	0.7521	**0.0380**	0.6420
KINETOCHORE-MICROTUBULE	12	4	1	**0.0380**	0.6420
CENP-H	4	3	1	**0.0067**	1

Abstract terms that were identified as over-represented by the corresponding methods are highlighted in yellow and in bold. The cutoff used is Bonferroni *p*-value ≤ 0.05. The raw *p*-values and rankings of these tokens can be found in Tables 5.2, 6.3 and 7.4.

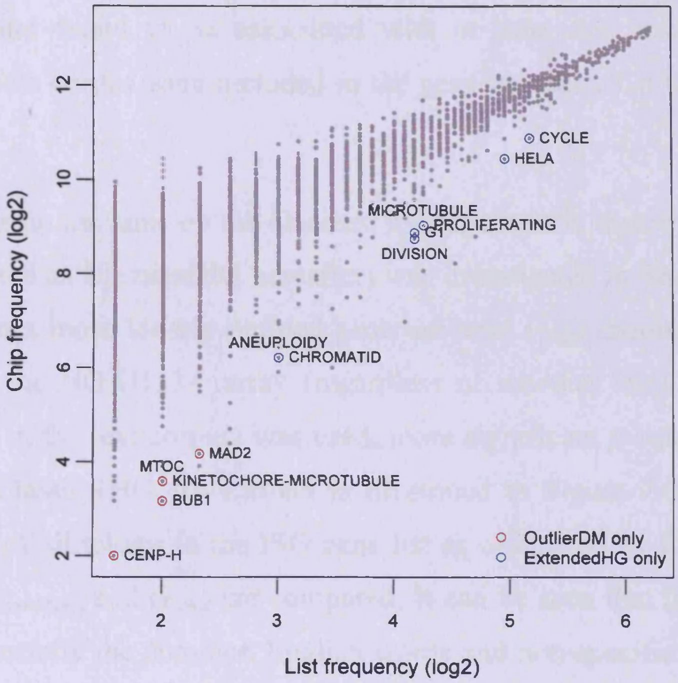


Figure 7.5: A comparison of the significant tokens in the mitosis gene list as obtained with OutlierDM and ExtendedHG

Tokens that were called significant by OutlierDM but not by ExtendedHG are circled in red. Tokens that were called significant by ExtendedHG but not by OutlierDM are circled in blue. Significant tokens are defined as those having Bonferroni *p*-value ≤ 0.05.

7.4.3 Impact of gene universe

The fundamental idea behind hypergeometric test-based enrichment analysis is that if a biological process is regulated or changed in a given study, the genes involved in the process will have a higher chance of being selected by gene-level analysis and therefore of being included into the gene list. To determine the degree of enrichment of this group of genes (and hence the biological process), a certain background or "gene universe" must be defined to perform the comparison, which means that the gene universe has an effect on the final conclusions of the analysis. The gene universe can be set up in many ways, e.g. use the total genes in the genome as a global background reference, or use a narrowed-down set of genes that only exist on a microarray. For all analyses presented above, only those genes that exist on the HG-U133A array and found to be associated with at least one abstract term in the corresponding text corpus were included in the gene universe. Let this gene universe be $G_{annotated}$.

The impact of gene universe on the classical hypergeometric distribution-based ORA approach (denoted as ClassicalHG hereafter) was investigated in Section 3.3.2. It was found that when a more loosely-defined gene universe G_{total} encompassing all genes represented on the HG-U133A array (regardless of whether they are cited by any PubMed article in the text corpus) was used, more significant *p*-values were given to the tokens by ClassicalHG. This effect is illustrated in Figure 7.6(a), in which the actual *p*-values of all tokens in the ISG gene list as calculated by ClassicalHG based on the use of $G_{annotated}$ and G_{total} are compared. It can be seen that the significance of most terms, especially the common English words and non-specific terms, tend to be artificially inflated when G_{total} was used. As a consequence, ClassicalHG becomes less conservative such that more tokens were called significant at the selected *p*-value threshold when G_{total} was used instead of $G_{annotated}$.

To examine the impact of the gene universe on the performance of ExtendedHG, the ISG gene list was re-analysed with G_{total}, and the results were compared to that obtained based on $G_{annotated}$. Interestingly, the same set of terms was identified as significantly enriched by ExtendedHG irrespective of whether $G_{annotated}$ or G_{total} was used. As can be seen from Figure 7.6(b), the *p*-value of all terms calculated based on $G_{annotated}$ and G_{total} were affected in a relatively similar manner, as opposed to the discrepancies seen for ClassicalHG. Similar trends were also observed when the rankings of the terms in the ISG gene list, analysed with different reference backgrounds, were compared (Figure 7.7). Such stable *p*-values and term ranks suggest that the extended hypergeometric distribution is a better statistical model than the classical hypergeometric distribution for assessing text-based over-representation, because it can intrinsically account for changes made to the gene universe and produces consistent output.

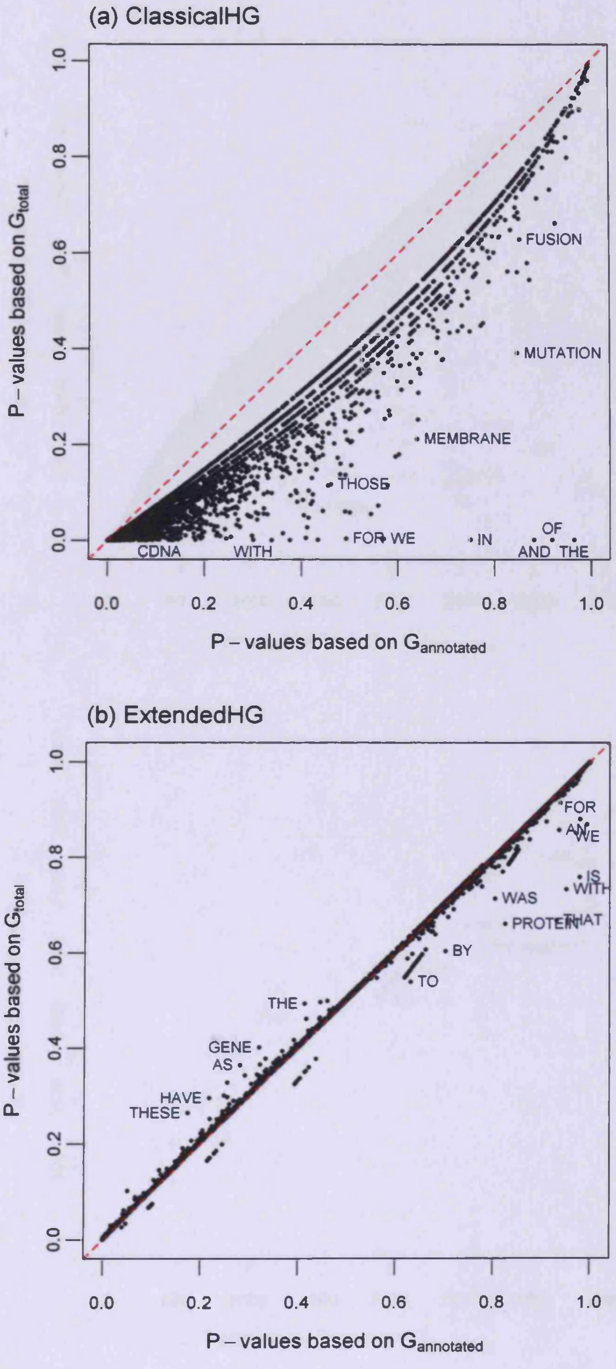


Figure 7.6: A comparison of the *p*-values generated based on different gene reference backgrounds

The raw *p*-values of all tokens in the ISG gene list as given by (a) ClassicalHG and (b) ExtendedHG based on the two different gene universes $G_{annotated}$ and G_{total} are compared. The red dashed line is the $y = x$ line.

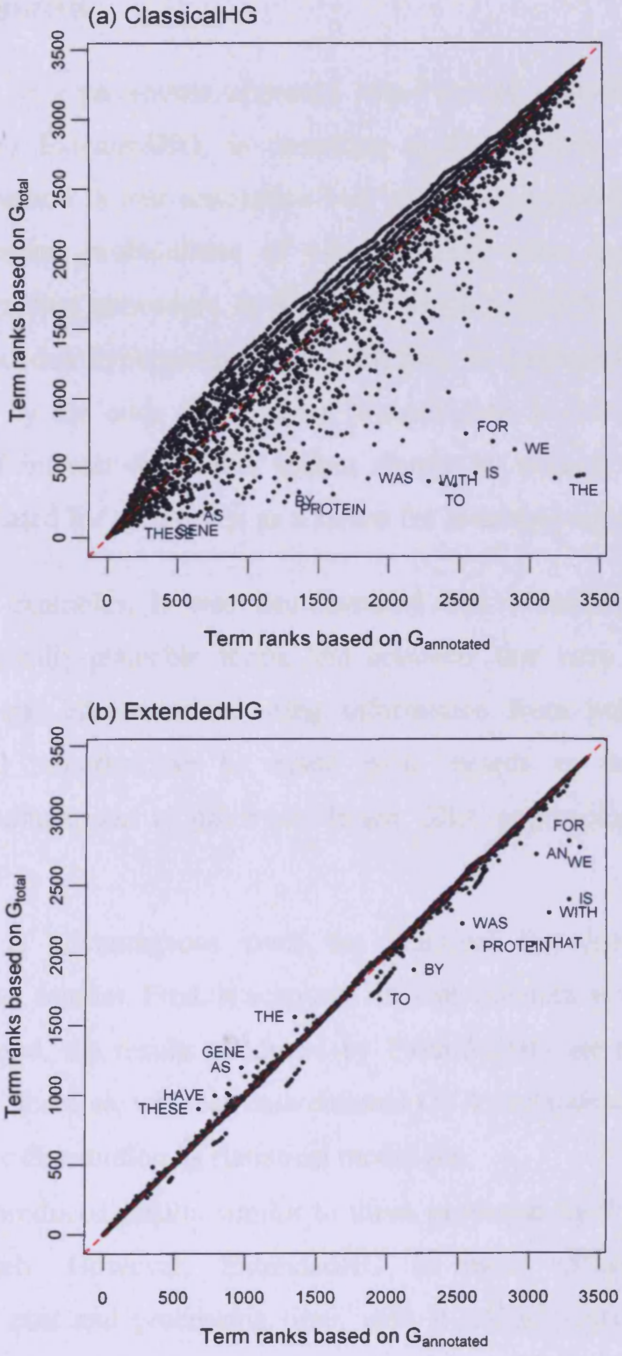


Figure 7.7: A comparison of the term ranks generated based on different gene reference backgrounds

The rank orders of all tokens in the ISG gene list as given by (a) ClassicalHG and (b) ExtendedHG based on the two different gene universes $G_{annotated}$ and G_{total} are compared. The red dashed line is the $y = x$ line.

7.5 Discussion

The development of a parametric approach based on the extended hypergeometric distribution, called ExtendedHG, is described in this Chapter. The key concept underlying this method is that annotation bias will cause common and non-specific terms to have higher probabilities of being selected than expected by chance. Therefore the sampling procedure is biased, with the token frequency distribution following the extended hypergeometric distribution. In ExtendedHG, the degree of bias is measured by the odds ratio, which is equivalent to the probability ratio of seeing a token of interest over other tokens simply by chance, and a *p*-value can therefore be calculated for each token as a means for assessing significance.

Through various examples, it was demonstrated that ExtendedHG is capable of identifying biologically-plausible terms and concepts that may then facilitate the process of gene list interpretation, using information from published biomedical literature. Several remarks can be made with regards to the performance of ExtendedHG in comparison to other text-based ORA approaches described in the previous Chapters:

- ExtendedHG is advantageous over the classical hypergeometric test-based approach in two aspects. First, it accounts for bias inherent with highly-annotated gene list. Second, the results produced by ExtendedHG are not affected by the choice of gene universe, whereas conventional ORA methods that use the classical hypergeometric distribution as statistical model are.
- ExtendedHG produced results similar to those produced by the permutation test-based approach. However, ExtendedHG is more efficient in terms of computational cost and processing time, with it taking approximately 20 to 30 seconds to analyse a 500-gene list, as opposed to a minimum of 6 hours that is typically required by the permutation test-based method on the same machine.
- The performances of ExtendedHG and OutlierDM are comparable when applied to the ISG and mitosis gene lists. However, in some cases, the two methods appear to capture slightly different components of the enrichment signal, such that terms uniquely detected by OutlierDM tend to be highly specific biological terms with

relatively low background frequencies, whereas terms uniquely detected by ExtendedHG are predominantly words with higher background frequency. This issue will be examined more closely in Chapter 8.

As with many conventional ORA approaches, ExtendedHG is built on the assumption that genes are independent. This assumption has been criticised because strong correlations between genes are frequently encountered in microarray data, especially between functionally related genes. Delongchamp *et al.* (2006) suggested that ignoring the correlations between genes can overstate the significance of the true *p*-value and proposed modified meta-analysis methods for combining *p*-values to adjust for correlation. Goeman and Bühlmann (2007) demonstrated how the use of models that implicitly or explicitly assume independence across genes can produce less conservative results. However, the extent to which this problem is a significant one is unclear. For example, Gold *et al.* (2007) have argued that ORA as conventionally applied is robust to the assumption of independence. They showed that the conventional Fisher's exact test (which does not formally adjust for correlation) and a multivariate normal approximation approach (which accounts for correlation) produced similar biological conclusions. Although not the main focus of the present study the possible effects of among-gene dependence are important, and present an interesting topic for future research.

Chapter 8

Performance properties of OutlierDM and ExtendedHG

8.1 Introduction

Evaluation is a process to determine whether a given method or system effectively achieves its stated objective, and the extent to which it succeeds in performing a task and achieving the anticipated results (Zweigenbaum *et al.* 2007). The analysis of gene lists produced from high-throughput technology like microarrays is more of an exploratory computational procedure rather than a pure statistical solution. Evaluating the performance of such approaches, which include the text-based ORA approaches proposed in previous Chapters, is challenging for several reasons. First, there is no immediately available "ground truth" which could be used (that is, there is no "gold standard"). Second, it is difficult to define reproducible evaluation metrics in biology because new knowledge is constantly being added, and whole subfields may be re-structured. Tarca *et al.* (2009) suggested that best practice in the absence of a gold standard is to (i) analyse the results produced by the proposed method in the light of the existing biological knowledge regarding the condition studied, and (ii) compare the performance of the proposed method with related approaches in the context of the same existing biological knowledge regarding the condition studied.

A focused evaluation strategy was undertaken in this Chapter to assess the capabilities of the text-based ORA approaches developed in this work and described in earlier Chapters. Specifically, the performance of OutlierDM and ExtendedHG (but not the permutation-based approach, which presents significant logistical problems for a large-scale evaluation because of its computationally intensive nature) were compared to existing literature- and ontology-based approaches. The ISG gene list and the literature gene lists from the HG-U133A array were used as test gene lists. The biological relevance and plausibility of the over-represented tokens and GO terms produced by the selected methods were then assessed against the perceived biology of the original publications. This comparative analysis is presented in Section 8.2. Section 8.3 will then focus on extending the proposed methods to other organisms outside human. In Section 8.4, the behaviours and shortcomings inherent with the OutlierDM and ExtendedHG are discussed.

8.2 Comparison with related software

8.2.1 Comparison with literature-based method

A number of text mining methods have been developed for linking groups of genes displaying interesting expression patters in microarray experiments with textual information contained in biomedical literature. Some popular ones have been discussed in Chapter 1 and a list of these tools is given in Table 1.2. Among them, the one closest in spirit to OutlierDM and ExtendedHG is the CoPub system developed by Frijters *et al.* (2008). The principle underlying CoPub is similar to conventional ontology-based ORA tools. Specifically, CoPub calculates keyword over-representation for a list of genes using the Fisher's exact test, in which the association of a given keyword with genes in the query gene list is statistically tested against a background reference. The keywords used in CoPub were generated by searching the MEDLINE abstracts with biological concepts from eleven thesauri encompassing gene names (only for human, mouse and rat), Gene Ontology (GO) terms, liver pathologies, pathways, diseases, drugs and tissues.

To examine how other popular literature-based approaches that are not based on the over-representation analysis would fare against OutlierDM and ExtendedHG, a system named TXTGate (Glenisson *et al.* 2004) is included in the comparative analysis. TXTGate is built on the ideas of textual profiling and clustering proposed in Blaschke *et al.* (2001) and Chaussabel and Sher (2002). It uses the vector space model to cluster a group of genes into functional categories based on textual information from MEDLINE and display the best-scoring terms associated with the group of genes. The document vectors used in TXTGate were restricted to abstract words or phrases that were pre-defined in functional vocabularies such as Gene Ontology (GO), Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM). All textual information and domain vocabularies were stemmed and indexed with a normalised inverse document frequency (IDF) weighting scheme.

OutlierDM and ExtendedHG versus CoPub and TXTGate

The ISG gene list was analysed with CoPub (http://services.nbic.nl/cgi-bin/copub/CoPub.pl) and TXTGate (http://tomcat.esat.kuleuven.be/txtgate/home.jsp); the results were compared to the outcome from OutlierDM and ExtendedHG. The ISG gene list was chosen as the benchmark because it constitutes a well-studied system of transcriptional regulation. As detailed in Section 2.2.1, the ISG gene list contains 78 genes induced by type I and type II interferons in A549 lung cells at 6h and 24h following treatment. These genes are principally involved in the regulation of the complement pathway, antiviral response, JAK-STAT signaling pathway, apoptosis and cytokine interactions. Therefore, if a text mining method is successful, the terms that it identifies as significant should correspond (or be related) to the aforementioned biological processes.

The results of this analysis are summarised in Table 8.1, in which the significant terms returned by each individual method are ordered by their significance. CoPub reported 40 keywords as significantly enriched in the ISG gene list at p-value ≤ 0.01 after correction for multiple testing. The top-ranking keywords from CoPub, such as ‘antigen presentation’, ‘antiviral response’, ‘peptide transport’, ‘virus-host interaction’ and ‘innate immune response’, are related to interferon-mediated immune responses. It can be seen that mutually corresponding biological concepts reflecting the immunomodulatory effects of interferons were also identified by OutlierDM and ExtendedHG; these include ‘immune’, ‘antiviral’, ‘OAS’, ‘HLA-A’ and ‘MxA’. While CoPub did pick up a few more additional biologically-plausible hits than

OutlierDM and ExtendedHG, such as 'apoptosis', 'proteasome', 'natural killer cell activity', the fact that these biological processes were due to interferon-stimulated gene expression changes is not particularly obvious. The only keyword that implicates the involvement of interferon in CoPub is 'ifn gamma signaling pathway' (ranked at 15). In contrast, the term 'interferon' is the most significant hit found by the rest of the methods shown in Table 8.1.

TXTGate output the top 10 terms with the highest IDF scores by default. As shown in Table 8.1, the term 'interferon' was reported as the best-scoring term in all three domain vocabularies used by TXTGate. The other two biologically-plausible terms returned by TXTGate are 'ifn' and 'interferon induc' (i.e. 'interferon induction'). However, except for the three cases just mentioned, the remaining hits are not very useful for the interpretation of the ISG gene list because they are predominantly non-specific biological terms like 'length', 'gene' and 'protein'. This is a surprising outcome because the adoption of specific domain vocabularies coupled with an IDF weighting scheme should (by definition) reduce the impact of common words and non-specific biological terms that occur frequently in abstracts. OutlierDM and ExtendedHG did not suffer from this problem because the apparent over-representation of these uninformative terms due to annotation bias was effectively dealt with, as described in Chapters 6 and 7.

It can be concluded from the above findings that the algorithms proposed for mining textual information as implemented in OutlierDM and ExtendedHG are competitive with current literature-mining methods.

Table 8.1: Comparison of terms found by OutlierDM, ExtendedHG with those found by TXTGate and CoPub in the ISG gene list

For OutlierDM and ExtendedHG, over-represented terms were identified using the threshold Bonferroni p-value ≤ 0.05. The most significant tokens (i.e. smallest p-values) are listed at the top. The analysis with CoPub was performed in "species-specific" mode with the human HG-UU133A chip as background; "categories analysed" was set to biological process and pathway; the "minimal number of genes associated with keyword" = at least 5; "literature threshold" = 3 or more co-publications; "R-scaled threshold" = at least 35. A p-value threshold of less than 0.01 after Benjamini-Hochberg multiple testing correction was used to assign over-represented keywords. The CoPub database used here is based on MEDLINE abstracts as of February 2008. The most significant keywords (i.e. smallest p-values) are listed at the top. The analysis with TXTGate was performed with MEDLINE abstracts annotated to Entrez Gene entries as of April 2006. Only the top ten terms with the highest inverse document frequency (IDF) scores were shown here; vocabularies with the highest term weight are listed at the top.

(continued over the page)

Table 8.1 (continued)

Outlier	ExtendedHG	CoPub	TXTGate
INTERFERON	INTERFERON	antigen presentation	GO vocabulary
IFN	IFN	antiviral response	interferon
IFN-BETA	ANTIVIRAL	peptide transport	accuraci
ANTIVIRAL	IFN-BETA	virus-host interaction	length
IFN-ALPHA	IFN-ALPHA	antigen processing and presentation	gene
INTERFERON-ALPHA	INDUCIBLE	innate immune response	sequenc
OLIGOADENYLATE	VIRAL	immune response	contain
ISG	INFECTION	viral replication	complet
ISRE	OLIGOADENYLATE	antigen processing	distribut
DSRNA	INTERFERON-ALPHA	immune system	ifn
HLA-CLASS	ISRE	proteasome	protein
HLA-A	INNATE	response to virus	
HLA-B	ISG	disease resistance	OMIM vocabulary
INDUCIBLE	IFN-GAMMA	natural killer cell activity	interferon
ENCEPHALOMYOCARDITIS	IMMUNE	ifn gamma signaling pathway	health
STOMATITIS	IMMUNITY	cell-mediated immune response	length
OAS	HLA-A	t-cell selection	gene
HLA-G	HLA-CLASS	transcription and rna-dependent	protein
MXA	STOMATITIS	cytolysis	induc
EVASION	TREAT	response to pathogen	function
INNATE	LYMPHOBLASTOID	cell maturation	cell
TAPASIN	HLA-B	cell recognition	gener
VIRAL	EVASION	lymphocyte activation	restrict
	VIRUS	b-cell activation	
	ENCEPHALOMYOCARDITIS	cell activation	MeSH vocabulary
	DSRNA	transcription	interferon
	MHC	t-cell proliferation, t-cell homeostatic proliferation	human
	OAS	t-cell differentiation	interferon induc
	HLA-G	t-cell activation	public
	HOST	humoral immune response	orf
	MXA	mrna transcription	institut
	INDUCTION	protein modification	nation
	BETA2-MICROGLOBULIN	lymphocyte differentiation, lymphocyte proliferation	clone
	TAPASIN	conjugation	gene
	GAMMA-INTERFERON	gene conversion	collect
	LMP7	apoptosis	
	LMP2	rna splicing	
	OR-C	cell development	
		cytokine biosynthesis, production, secretion	
		monocyte activation, differentiation	

8.2.2 Comparison with ontology-based method

OutlierDM and ExtendedHG versus DAVID

Using the ISG gene list as the benchmarking dataset, the performance of OutlierDM and ExtendedHG was evaluated against an ontology-based functional enrichment tool, DAVID (The Database for Annotation, Visualization and Integrated Discovery; http://david.abcc.ncifcrf.gov/home.jsp) (Huang *et al.* 2007; Huang *et al.* 2009). The analysis was performed with the Functional Annotation Chart module from DAVID, in which the significance of GO terms associated with genes in the ISG gene list was determined using the modified Fisher's exact test (or EASE score). A cutoff of 0.05 was used such that GO terms with p-values ≤ 0.05 after the Bonferroni correction were considered as over-represented.

As shown in Table 8.2, a good agreement was observed between the biology associated with the enriched GO terms reported by DAVID and the PubMed abstract terms produced by OutlierDM and ExtendedHG (cf. Table 8.1), with concepts related to immune response highly-ranked by all three approaches. As an illustration of the limitations of pre-defined ontologies such as GO it was noted that none of the significant GO terms gives an indication of the involvement of interferon, thus demonstrating how mining of PubMed abstracts can potentially reveal additional biological insight that is not possible by mining controlled vocabularies alone.

In what follows, the evaluation was expanded beyond the ISG gene list to a larger set of data comprising of the 52 literature gene lists derived from experiments using the Affymetrix human HG-U133A array. The aim is to present a focused performance review of how the results produced by OutlierDM and ExtendedHG and standard ORA approach using GO terms compare, both in depth of information and also in biological plausibility.

Table 8.2: Significant GO terms in the ISG gene list as reported by the functional enrichment tool DAVID

Term	Population Hits	List Count	*p*-value	Bonferroni *p*-value	Ranking
Response to biotic stimulus	853	49	1.98E-36	6.40E-33	1
Immune response	737	44	2.54E-32	8.30E-29	2
Defense response	816	45	8.84E-32	2.90E-28	3
Response to stimulus	1765	52	3.49E-25	1.10E-21	4
Organismal physiological process	1660	46	6.17E-20	2.00E-16	5
Response to virus	70	14	6.67E-16	2.20E-12	6
Response to pest, pathogen or parasite	503	25	7.37E-15	2.40E-11	7
Response to other organism	514	25	1.19E-14	3.90E-11	8
Response to stress	956	27	1.85E-10	6.00E-07	9
MHC protein complex	18	6	1.05E-07	6.30E-05	10
MHC class I protein complex	18	6	1.05E-07	6.30E-05	11
Antigen presentation, endogenous antigen	27	7	2.07E-08	6.70E-05	12
Antigen processing, endogenous antigen via MHC class I	28	7	2.62E-08	8.50E-05	13
MHC class I receptor activity	36	7	8.23E-08	2.00E-04	14
Antigen processing	36	7	1.30E-07	4.20E-04	15
Antigen presentation	42	7	3.38E-07	1.10E-03	16
Immunological synapse	31	6	1.95E-06	1.20E-03	17

The ontological tool DAVID 2.0 was used to identify over-represented GO terms in the ISG gene list. The analysis was performed using all levels of GO terms and HG-U133A chip as background (database version as of 19 Dec 2007). Over-represented GO terms were defined as having Bonferroni *p*-value ≤ 0.05 based on Fisher's exact test (threshold settings: Count = 2, EASE = 0.1).

OutlierDM and ExtendedHG versus GOstats

In this analysis, the GO terms identified as over-represented in the 52 HG-U133A literature gene lists were qualitatively assessed against the significantly over-represented PubMed tokens found by OutlierDM and ExtendedHG.

Experiment settings

For the text-based analysis, the 52 gene lists were analysed with OutlierDM and ExtendedHG as described in Sections 6.3.2 and 7.3.2, respectively. A threshold of Bonferroni-corrected *p*-value ≤ 0.05 was used to identify significantly over-represented tokens.

For the GO-based analysis, the Bioconductor package GOstats (Falcon and Gentleman 2007) was used. Specifically, the function `hyperGTest` implemented in GOstats (version 2.4.0) was used to identify GO terms (restricted to the biological process category only) that are significantly enriched in these gene lists, based on the classical hypergeometric test. This version of GOstats used the mappings provided in the annotation data packages hgu133a.db (version 2.0.2) and GO.db (version 2.0.0) to convert between Affymetrix probeset IDs, Entrez Gene IDs and GO terms. Only genes that exist on the HG-U133A array and found to be associated with at least one GO term were included in the gene universe. A cutoff of 0.05 (i.e. Bonferroni *p*-value ≤ 0.05) was used to identify significantly enriched GO terms.

Results

Three pieces of information were collected for each of the 52 gene list: (i) the number of GO terms reported as significantly enriched by GOstats; (ii) the number of abstract tokens reported as significantly over-represented by OutlierDM; and (iii) the number of abstract tokens reported as significantly over-represented by ExtendedHG. Out of the 52 gene lists, 4 gene lists do not have any hits in all three methods assessed; 2 gene lists have at least one significant GO term but not tokens; 16 gene lists have at least one significant token but not GO terms. Those gene lists with at least one significant GO term and token are shown in Figure 8.1, where the number of over-represented GO terms was plotted against the number of tokens found in them. To give a consensus measure for the token-based analysis, only results from the text-based method (either Outlier or ExtendedHG) that produced the most number of significant tokens are shown here.

As shown Figure 8.1, there is a good correlation between the number of significant GO terms and the number of significant tokens identified across these gene lists; those gene lists with more apparent enriched biology, as defined by GOstats, also show a higher number of over-represented tokens according to the proposed text-based approach. The biological relevance and plausibility of the over-represented tokens and GO terms were then assessed against the perceived biology of the original publications. Table 8.3 shows the outcomes of GOstats, the proposed text-based approach (OutlierDM or ExtendedHG) and biology of the original publication in parallel. In most cases, the over-represented GO terms and tokens appear to be both interesting and plausible, and also in good agreement with the biological themes extracted manually.

It was found that the level of detail offered by the two resources (GO versus abstract token) is rather different. For example, in the two gene lists 'hs5a' and 'hs5b', while GOstats reported biological processes such as 'immune response', 'lymphocyte activation' as significant; text-based analysis with ExtendedHG revealed the key players involved in these processes including 'CD3', 'CD4', 'CD8', 'TCR' and 'IL-2'. Consider another example based on the gene list 'hs11b' in which no GO terms was found to be significant enriched; OutlierDM were able to pick up the term 'Metalloproteinase-2' as significant. Indeed, matrix metalloproteinases were perceived to play an important role in mediating lung metastasis in this study.

These examples demonstrate that the text-based approaches proposed in this thesis are not a replacement to the classical GO-based ORA but a complement and extension of it, and when both were used in combination could (sometimes) give greater insights than using GO alone.

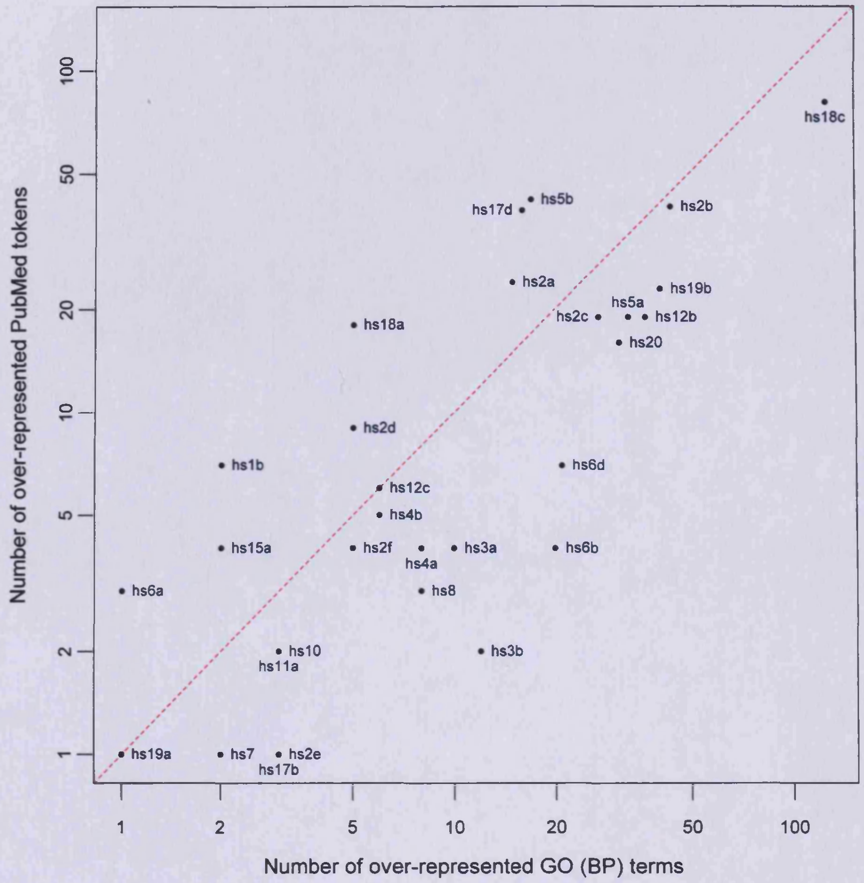


Figure 8.1: A comparison of the number of significant GO (BP) terms versus the number of significant abstract tokens in 30 HG-U133A literature gene lists

Each point in the plot represents a gene list. For each gene list, the number of over-represented GO terms reported by GOstats was plotted against the number of over-represented tokens reported by the proposed text-based ORA approaches (the results shown here are based on the text-based method - either Outlier or ExtendedHG - that produced the most number of significant tokens). Details of these gene lists, the significant GO terms and tokens found in them can be found in Table 8.3. BP = biological process terms in GO. The axes are on log scale.

Table 8.3: Significant GO terms versus PubMed tokens

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
hs1b	PMID: 16531451 Title: Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. Extracted from Supplementary Table II.	Probesets changed in the polysomal RNA sample. • Amine metabolism • Amino acid activation and translation • RNA transport and metabolism • Cell proliferation • Apoptosis • tRNA aminoacylation pathway	• Regulation of progression through cell cycle • Regulation of cell cycle	BREAST OVEREXPRESSION METASTATIC CANCER FIBROBLAST OVEREXPRESS IMMORTALIZATION	E
hs2a	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 1.	Transcripts with greater than 3-fold enrichment in every T cell subpopulation compared to TSC. • T cell differentiation and function • Immune response • TCR signalling • Intrathymic differentiation • ERK1/ERK2 activity • Intrathymic T cell selection	• Immune system process • T cell activation • Lymphocyte activation • Leukocyte activation • Positive regulation of antigen receptor-mediated signaling pathway • Hemopoietic or lymphoid organ development • Immune system development • Cell activation • Hemopoiesis • Regulation of T cell activation • Immune response • T cell differentiation • Lymphocyte differentiation • Regulation of lymphocyte activation • Regulation of cell activation	T-CELL LYMPHOID TCR THYMOCYTE CD3 LYMPHOCYTE LINEAGE HEMATOPOIETIC NK IL-2 IMMUNE JURKAT LYMPHOMA T-LYMPHOCYTE KILLER B-CELL CD45RA NAIVE CD4 CD2 CD8 LCK ENGAGEMENT ACTIVATION	E
hs2b	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 2.	Transcripts whose expression changed by more than 3-fold during T cell differentiation.	• Cell cycle • Cell cycle process • Mitotic cell cycle • M phase of mitotic cell cycle • Cell cycle phase • Mitosis • Regulation of cell cycle • DNA replication • M phase • Regulation of progression through cell cycle • Immune response • Cell division • Spindle organization and biogenesis • Immune system process • Programmed cell death • Regulation of programmed cell death • Apoptosis • Regulation of apoptosis • Cell death • Death	LYMPHOID THYMOCYTE CD8 LYMPHOCYTE ANAPHASE B-CELL LYMPHOMA T-CELL MITOSIS SPINDLE TCR CHECKPOINT CD3 THYMUS CD4 CYTOKINESIS NK IL-2 LEUKEMIA INTERLEUKIN-2 INTERFERING MITOTIC KINETOCHORE PROLIFERATING	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs2b cont'd)			• DNA metabolic process • Cell cycle checkpoint • Microtubule-based process • Antigen processing and presentation • Regulation of mitosis • Positive regulation of apoptosis • Regulation of biological process • Positive regulation of programmed cell death • Cell development • Negative regulation of biological process • Biological regulation • Regulation of cellular process • Induction of apoptosis • Mitotic sister chromatid segregation • Negative regulation of cellular process • Induction of programmed cell death • Sister chromatid segregation • Spindle checkpoint • Negative regulation of programmed cell death • DNA-dependent DNA replication • Chromosome segregation • Negative regulation of apoptosis • Cell differentiation • Cellular developmental process	KILLER INTERPHASE LYMPH PROMETAPHASE LINEAGE VIRUS TUMOR INDUCTION MIDBODY MAB NODE PROLIFERATION JURKAT GERMINAL ENTER HEMATOPOIETIC	
hs2c	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 3-1.	Transcripts enriched in both ITTP and DP by more than 3-fold. • DNA replication, recombination and repair • Cell cycle regulation, progression, mitosis • Lipid/glycolipid-presenting CD1 family • Transcriptional regulation • Regulation of apoptosis	• Mitotic cell cycle • Mitosis • M phase of mitotic cell cycle • Cell cycle phase • M phase • Cell cycle • Cell cycle process • Cell division • Spindle organization and biogenesis • Microtubule-based process • Microtubule cytoskeleton organization and biogenesis • Regulation of mitosis • Cytoskeleton organization and biogenesis • Regulation of progression through cell cycle • Regulation of cell cycle • Organelle organization and biogenesis	MITOTIC SPINDLE MITOSIS ANAPHASE CHECKPOINT KINETOCHORE DIVISION CONGRESSION PROMETAPHASE CENTROMERE PROLIFERATING MICROTUBULE G1 AURORA HELA CHROMATID ANEUPLOIDY CYCLE CENP-A	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs2c cont'd)			• Phosphoinositide-mediated signaling • Second-messenger-mediated signaling • Chromosome segregation • DNA metabolic process • Spindle checkpoint • Microtubule-based movement • DNA replication • Mitotic sister chromatid segregation • Sister chromatid segregation • Cytoskeleton-dependent intracellular transport • Mitotic spindle organization and biogenesis		
hs2d	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 3-2.	Transcripts enriched in more mature cells (SP4, CB4, and AB4) by more than 3-fold. • Intracellular communication • Cell surface receptors • Peptide-presenting MHC antigens • Transcriptional regulation • Regulation of apoptosis	• Antigen processing and presentation of peptide antigen via MHC class I • Antigen processing and presentation of peptide antigen • Immune response • Antigen processing and presentation • Immune system process	HLA-CLASS HLA-A OR-C PERIPHERAL CD8 HLA-C HLA-B HLA-G IFN-GAMMA	O
hs2e	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 3-3.	Transcripts enriched by more than 3-fold in ITTP compared to other lymphocytes. • Immune function	• Nitric oxide mediated signal transduction • CGMP biosynthetic process • CGMP metabolic process	INTERLEUKIN-2	O
hs2f	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 3-4.	Transcripts enriched by more than 3-fold in DP compared to other lymphocytes. • Thymocyte differentiation • Thymocyte survival and positive selection • Modulation of Th1 and Th2 response • CD1 family proteins • T cell co-stimulation	• Lymphocyte activation • Leukocyte activation • Cell activation • T cell activation • Antigen processing and presentation	LYMPHOPROLIFERA - TION MYCOBACTERIAL MYCOBACTERIA INTERACTION	O
hs2i	PMID: 15210650 Title: Gene expression profiles during human CD4+ T cell differentiation. Extracted from Supplementary Table 4-1.	Transcripts showing SP4>CB4>AB4 pattern. • Plasma membrane proteins	[No hits found]	HMG-BOX MMP-2	O
hs3a	PMID: 15897907	Genes which best discriminate apocrine vs	• Alcohol metabolic process • Monocarboxylic acid	ACETOXYMETHYL METABOLISM	O

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs3a cont'd)	Title: Identification of molecular apocrine breast tumours by microarray analysis. Extracted from Supplementary Table Sheet 2.	luminal (AL).	metabolic process • Carboxylic acid metabolic process • Organic acid metabolic process • Steroid biosynthetic process • Lipid metabolic process • Sterol biosynthetic process • Aldehyde metabolic process • Fatty acid metabolic process • Cholesterol biosynthetic process	ANTIANDROGEN ANDROGEN	
hs3b	PMID: 15897907 Title: Identification of molecular apocrine breast tumours by microarray analysis. Extracted from Supplementary Table Sheet 2.	Genes which best discriminate apocrine vs basal (AB).	• Lipid metabolic process • Carboxylic acid metabolic process • Organic acid metabolic process • Monocarboxylic acid metabolic process • Alcohol metabolic process • Sterol biosynthetic process • Cellular lipid metabolic process • Lipid biosynthetic process • Steroid biosynthetic process • Fatty acid metabolic process • Cholesterol biosynthetic process • Sterol metabolic process	DESATURASE FATTY	O
hs4a	PMID: 16260967 Title: Effects of aerobic training on gene expression in skeletal muscle of elderly men. Extracted from Table S2 in the main paper.	Genes whose expression increased after training. • Energy metabolism or mitochondrion • Lipid metabolism • Proton pumps • Collagen • Protein, amino acid dephosphorylation • Heme biosynthesis	• Acetyl-CoA metabolic process • Tricarboxylic acid cycle intermediate metabolic process • Tricarboxylic acid cycle • Acetyl-CoA catabolic process • Cofactor metabolic process • Coenzyme catabolic process • Cellular catabolic process • Cofactor catabolic process	MITOCHONDRIAL MITOCHONDRIA CATALYSIS BURY	O
hs4b	PMID: 16260967 Title: Effects of aerobic training on gene expression in skeletal muscle of elderly men. Extracted from Table S3 in the main paper.	Genes whose expression decreased after training. • Ribosome and protein catabolism • Muscle degradation	• Translation • Macromolecule biosynthetic process • Biosynthetic process • Cellular protein metabolic process • Cellular macromolecule metabolic process • Protein metabolic process	RIBOSOMAL S14 R-PROTEIN RRNA U14	O
hs5a	PMID: 12958056 Title: Gene expression profiling of bronchoalveolar lavage	Genes that are up-regulated in gene expression in acute rejection vs. no rejection (False Discovery Rate =	• Immune system process • Immune response • Leukocyte activation • Response to stimulus	THYMOCYTE CD8 NK TCR	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs5a cont'd)	cells in acute lung rejection. Extracted from Supplementary Table E1.	0.94%). • Acute rejection response • Immune response • Inflammatory response • Transcriptional regulation • TGF-beta signalling • Apoptosis • Nucleotide GPCR receptors • Peptide GPCR receptors • Wnt signalling • Cytokine-CXC chemokine pathways	• Cell activation • Lymphocyte activation • Positive regulation of lymphocyte activation • Regulation of lymphocyte activation • Regulation of cell activation • Positive regulation of isotype switching to IgG isotypes • Lymphocyte mediated immunity • Cellular defense response • Leukocyte mediated immunity • Adaptive immune response • Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains • Regulation of immune system process • Regulation of immune response • Isotype switching to IgG isotypes • Regulation of isotype switching to IgG isotypes • Immunoglobulin mediated immune response • B cell mediated immunity • Positive regulation of B cell activation • Immune effector process • Positive regulation of isotype switching • Immune response-activating cell surface receptor signaling pathway • Immune response-activating signal transduction • Immune response-regulating signal transduction • Immune response-regulating cell surface receptor signaling pathway • Antigen receptor-mediated signaling pathway • Positive regulation of mononuclear cell proliferation • Positive regulation of lymphocyte proliferation • Regulation of B cell activation • Signal transduction	CD4 CD3 KILLER IL-2 T-CELL IMMUNE LYMPHOCYTE MAB LYMPHOCYTIC LYMPHOMA JURKAT PRE-TCR B-CELL ANTIGEN IMMUNOGLOBULIN	
hs5b	PMID: 12958056 Title: Gene expression	Genes with significant changes in gene expression in acute	• Immune system process • Immune response	CD4 NK	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs5b cont'd)	profiling of bronchoalveolar lavage cells in acute lung rejection. Extracted from Supplementary Table E2.	rejection vs. no rejection (False Discovery Rate = 4.63 %). • Acute rejection response • Immune response • Inflammatory response • Transcriptional regulation • TGF-beta signalling • Apoptosis • Nucleotide GPCR receptors • Peptide GPCR receptors • Wnt signalling • Cytokine-CXC chemokine pathways	• Response to stimulus • Leukocyte activation • Lymphocyte activation • Cell activation • Cell surface receptor linked signal transduction • T cell activation • Regulation of lymphocyte activation • Regulation of cell activation • Cellular defense response • Signal transduction • Positive regulation of lymphocyte activation • Cell communication • Regulation of T cell activation • Defense response • Regulation of multicellular organismal process	IL-2 CD3 CD8 KILLER THYMOCYTE ENGAGEMENT TCR T-CELL CYTOLYTIC IMMUNE MAB LYMPHOCYTE CD16 CD56 IMMUNOGLOBULIN IL-12 INTERLEUKIN CD2 PBL LIGATION ALLOGENEIC CTL CYTOTOXICITY PERIPHERAL JURKAT MONOCYTE CYTOKINE EFFECTOR RAFT TH1 MONONUCLEAR IFN-GAMMA BLOOD ANTI-CD3 PHYTOHEMAGGLUTININ GRANZYME LYMPHOID MEMORY NATURAL SURFACE	
hs6a	PMID: 16319128 Title: Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. Extracted from Supplementary Table 1.	Down-regulated genes in FH mutant relative to FH wild-type fibroids. • Extracellular matrix • Cell mobility • Muscle contraction • Organogenesis • Muscle development • Cell adhesion • Plasma membrane	• Anatomical structure development	MICROFIBRIL TRANSFORMING SMOOTH	O
hs6b	PMID: 16319128 Title: Distinct expression profile in fumarate-hydratase-deficient uterine fibroids.	Up-regulated genes in FH mutant relative to FH wild-type fibroids. Extracted from Supplementary Table 1. • Glycolysis • Carbohydrate metabolism	• Glucose catabolic process • Glycolysis • Hexose catabolic process • Monosaccharide catabolic process • Alcohol catabolic process • Glucose metabolic process • Cellular carbohydrate	GLYCOLYTIC ALDO-KETO PEROXIDATION NICOTINAMIDE	O

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs6b cont'd)	Extracted from Supplementary Table 1.	• Hexose metabolism • Iron ion homeostasis • Oxidoreductase activity • Membrane lipid catabolism • Integral to endoplasmic reticulum membrane • Electron transporter activity	catabolic process • Hexose metabolic process • Monosaccharide metabolic process • Carbohydrate catabolic process • Cellular catabolic process • Cellular carbohydrate metabolic process • Carbohydrate metabolic process • Catabolic process • Cellular macromolecule catabolic process • Alcohol metabolic process • Phospholipid catabolic process • Macromolecule catabolic process • Cellular iron ion homeostasis • Iron ion homeostasis		
hs6d	PMID: 16319128 Title: Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. Extracted from Supplementary Table 3.	Up-regulated genes in FH mutant relative to normal myometrium. • Carbohydrate metabolism • Glycolysis	• Glucose catabolic process • Hexose catabolic process • Monosaccharide catabolic process • Alcohol catabolic process • Cellular carbohydrate catabolic process • Carbohydrate catabolic process • Glycolysis • Glucose metabolic process • Cellular carbohydrate metabolic process • Hexose metabolic process • Monosaccharide metabolic process • Carbohydrate metabolic process • Cellular catabolic process • Macromolecule catabolic process • Alcohol metabolic process • Catabolic process • Cellular macromolecule catabolic process • NADP metabolic process • Nicotinamide metabolic process • Cellular iron ion homeostasis • Iron ion homeostasis	GLYCOLYTIC NONSPHEROCYTIC NADP HEMOLYSIS APOFERRITIN RESOLUTION ISOENZYME	O
hs7	PMID: 15817885 Title: Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: expression of a ventricular-like genomic signature.	Genes differentially expressed in atrial fibrillation. • Transcriptional processes and activities • Calcium-dependent signalling pathway • CaMK pathway	• Enzyme linked receptor protein signaling pathway • Developmental process	OVEREXPRESSION	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs7 cont'd)	Extracted from Supplementary Table 3.	• MAPK pathway • Extracellular matrix composition and turnover			
hs8	PMID: 15971941 Title: Derivation of multipotent mesenchymal precursors from human embryonic stem cells. Extracted from Supplementary Table S2.	Genes shared between primary and hESC-derived mesenchymal precursors but significantly different from undifferentiated hESCs. MSC markers including mesenchymal stem cell protein DSC54, hepatocyte growth factor, neuropilin I, forkhead box D1, notch homolog.	• Organ development • Hemostasis • Cell cycle arrest • Blood coagulation • Coagulation • Wound healing • Response to wounding • Regulation of body fluids	COLLAGEN ECM STROMAL	O
hs10	PMID: 16203770 Title: Molecular alterations in primary prostate cancer after androgen ablation therapy. Extracted from Supplementary Data.	Unabridged list of genes differentially expressed between AD and AI prostate cancer.	• Translation • Macromolecule biosynthetic process • Biosynthetic process	RIBOSOMAL S16	O
hs11a	PMID: 16049480 Title: Genes that mediate breast cancer metastasis to lung. Extracted from Supplementary Table 2.	Genes differentially expressed between parental MDA-MB-231 and LM2 cell lines selected to be highly metastatic to lung. • Lung metastatic activity • Growth and survival factors • Chemokines • Cell adhesion receptors • Extracellular proteases • Intracellular enzymes • Transcriptional regulators	• Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II • Immune response • Immune system process	METALLOPROTEIN-ASE-2 SBT	O
hs11b	PMID: 16049480 Title: Genes that mediate breast cancer metastasis to lung. Extracted from Supplementary Table 4.	Lung metastasis candidate genes. Extracellular proteins (eg. SPARC, MMP2) act as virulence genes that may allow tumours to invade, colonise and grow in the lungs.	[No hits found]	METALLOPROTEIN-ASE-2 LINE	O
hs12b	PMID: 16089502 Title: Functional analysis of human hematopoietic stem cell gene expression using zebrafish. Extracted from Supplementary Table S2.	Probesets differentially expressed between adult bone marrow derived Rho-lo and Rho-hi cells. • Cell cycle control	• Cell cycle • Mitotic cell cycle • Cell cycle process • Cell cycle phase • DNA metabolic process • M phase of mitotic cell cycle • Mitosis • M phase • DNA replication • Cellular metabolic process • Cell division • Primary metabolic process	CHECKPOINT MITOSIS ANAPHASE MITOTIC RIBOSOMAL INTERPHASE CYTOKINESIS CYCLE G1 REPLICATION KINETOCHORE PROLIFERATING CDK2 ERYTHROID	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs12b cont'd)			• Metabolic process • Regulation of cell cycle • Macromolecule metabolic process • Regulation of progression through cell cycle • Chromosome organization and biogenesis • Translation • Chromosome organization and biogenesis (sensu Eukaryota) • Cell cycle checkpoint • Chromosome segregation • Response to DNA damage stimulus • DNA-dependent DNA replication • Response to endogenous stimulus • Nucleosome assembly • Mitotic sister chromatid segregation • Spindle organization and biogenesis • Interphase • Macromolecule biosynthetic process • Sister chromatid segregation • DNA repair • Organelle organization and biogenesis • Biopolymer metabolic process • Chromatin assembly • Biosynthetic process • Mitotic spindle organization and biogenesis • Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	DOUBLE-STRAND G2 KI-67 CHROMATID ANEUPLOIDY	
hs12c	PMID: 16089502 Title: Functional analysis of human hematopoietic stem cell gene expression using zebrafish. Extracted from Supplementary Table S3.	Probesets differentially expressed between Rho-lo and Rho-hi cells from both umbilical cord blood and adult bone marrow. • Hematopoietic differentiation and development • Cell cycle control • Wnt signalling • Germ cell development • Globins	• Oxygen transport • Gas transport • Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II • DNA replication • Antigen processing and presentation • DNA metabolic process	THAL BETA-CHAIN DELTA-GLOBIN HBA2 THALASSEMIA ANODE	O
hs15a	PMID: 12756304 Title: A global view of the selectivity of zinc deprivation and excess on genes expressed in human THP-1 mononuclear cells. Extracted from	Group 1 zinc responsive genes. • Nucleic acid binding • Apoptosis • Metabolism • Cell growth and development	• RNA metabolic process • Transcription from RNA polymerase II promoter	SCLEROTOME CD4 RNAP EFFECTOR	O

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs15a cont'd)	Supplementary Table 3.	• Signal transduction • Immune, cytokine • Cytoskeleton			
hs17b	PMID: 16804116 Title: Gene-expression profiling of Waldenstrom macroglobulinemia reveals a phenotype more similar to chronic lymphocytic leukemia than multiple myeloma. Extracted from Supplementary Table S1: MM Unique Genes.	Genes that displayed distinct expression profile in MM compared to CLL and WM. • Signal transduction and intracellular signalling • Cell-surface receptor-linked signalling e.g. AKT, IGF-1R and Wnt signalling • Prostacyclin synthesis • angiopoientin signalling • Integrin-mediated cell adhesion • Early B-cell receptor signalling	• Regulation of biological process • Regulation of cellular process • Biological regulation	MB-1	O
hs17c	PMID: 16804116 Title: Gene-expression profiling of Waldenstrom macroglobulinemia reveals a phenotype more similar to chronic lymphocytic leukemia than multiple myeloma. Extracted from Supplementary Table S1: CLL Unique Genes.	Genes that displayed distinct expression profile in CLL compared to WM and MM. • Apoptosis regulation • Immune response • Cell cycle regulation	[No hits found]	APC	O
hs17d	PMID: 16804116 Title: Gene-expression profiling of Waldenstrom macroglobulinemia reveals a phenotype more similar to chronic lymphocytic leukemia than multiple myeloma. Extracted from Supplementary Table S1: WM CLL B-cell cluster.	A cluster of genes that were over-expressed in B-cell, WM and CLL. • Cell cycle regulation	• Immune system process • Immune response • Cell communication • Signal transduction • Lymphocyte activation • Leukocyte activation • Cell activation • B cell activation • Defense response • Immune response-activating cell surface receptor signaling pathway • Immune response-activating signal transduction • Immune response-regulating signal transduction • Immune response-regulating cell surface receptor signaling pathway • Antigen receptor-mediated signaling pathway • Immune system development • T cell activation	LYMPHOID B-CELL HEMATOPOIETIC LYMPHOCYTE LINEAGE ENGAGEMENT CD8 CD19 TCR LYMPHOMA IMMUNE CD4 PRE-B SRC MONOCYTE NAIVE THYMOCYTE BCR IL-4 CD21 IMMUNOGLOBULIN BURKITT GERMINAL JURKAT LYMPHOCYTIC EXTRANODAL CD3 HISTOCOMPATIBILITY	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs17d cont'd)				ZAP70 NK PRO-B RESTING EFFECTOR KILLER T-CELL F-ACTIN CYTOSKELETON LEUKOCYTE LEUKEMIA	
hs18a	PMID: 16836768 Title: Signatures of human regulatory T cells: an encounter with old friends and new players. Extracted from Additional file 1.	Up-regulated genes comparing CD4+CD25+ T cells versus CD4+CD25- T cells. All differentially expressed genes can be classified into: • Cytokines/chemokines and their receptors • Cell cycle and proliferation • Apoptosis • Signal transduction • Transcriptional regulation The authors identified 3 signalling modules using Pathway Analysis software: • Genes that control T cell receptor signalling, activation and proliferation • Genes that control differentiation and maintenance • Genes that control survival/apoptosis	• Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II • Antigen processing and presentation • Immune response • Response to stimulus • Immune system process	DPB1 DRB1 DRB DR2 DPA1 HLA-DPB1 DQB1*0602 HLA-DRB1 DQW1 AND-DQ DQB1*0302 PCR-SSOP HLA-DR HLA-D OLIGOTYPING DQA1 SBT CD4	O
hs18c	PMID: 16836768 Title: Signatures of human regulatory T cells: an encounter with old friends and new players. Extracted from Additional file 4.	Genes differentially expressed in Foxp3 over-expressing CD4+ Th cell lines cells relative to the GFP transduced CD4+ Th controls. • TNF receptor superfamily • Activation of signal transduction pathways eg. NFkB, JNK, P38, ERK and PI3K • Immune response	[Only top 50 are shown] • Immune system process • Immune response • Response to stimulus • Signal transduction • Cell communication • Lymphocyte activation • Leukocyte activation • T cell activation • Cell death • Death • Biological regulation • Apoptosis • Programmed cell death • Cell activation • Defense response • Cell development • Regulation of cellular process	[Only top 50 are shown] LYMPHOID T-CELL CD4 CD3 CD8 IL-2 LYMPHOCYTE ANTI-CD3 CD25 NAIVE TCR HELPER ENGAGEMENT CD28 IMMUNODEFICIENCY INFECT JURKAT TH1 IMMUNE	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs18c cont'd)			• Regulation of biological process • Positive regulation of biological process • Regulation of lymphocyte activation • Positive regulation of cellular process • Regulation of cell activation • Cell differentiation • Cellular developmental process • Cell surface receptor linked signal transduction • Regulation of apoptosis • Regulation of programmed cell death • Negative regulation of biological process • Negative regulation of cellular process • Positive regulation of lymphocyte activation • Elevation of cytosolic calcium ion concentration • Cytosolic calcium ion homeostasis • Cytokine biosynthetic process • Cytokine metabolic process • Developmental process • Regulation of cytokine biosynthetic process • Cell proliferation • Regulation of T cell activation • Cellular di-, tri-valent inorganic cation homeostasis • Di-, tri-valent inorganic cation homeostasis • Cellular cation homeostasis • Cation homeostasis • Cellular calcium ion homeostasis • Calcium ion homeostasis • Inflammatory response • Response to external stimulus • Cellular ion homeostasis • Cellular chemical homeostasis • Somatic recombination of immunoglobulin genes during immune response • Somatic diversification of immunoglobulins during immune response	NK THYMOCYTE B-CELL PBMC REJECTION IL-10 LYMPHOCYTIC FOXP3 KILLER MONOCYTE UNINFECT IMMUNITY LYMPHOMA COSTIMULATORY HIV-1 T-LYMPHOCYTE IL-4 CYTOMETRY INFECTION INTERLEUKIN CYTOKINE HIV INTERLEUKIN-2 VIRAL STIMULATION ALLOGENEIC IFN-GAMMA MAB PROLIFERATE MONONUCLEAR IL-6	
hs19a	PMID: 15869706 Title: Clinical and	Up-regulated genes in FGFR3 mutated tumors relative to FGFR3 wildtype	• Multicellular organismal process	COMMENSURATE	O

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs19a cont'd)	biological characteristics of cervical neoplasias with FGFR3 mutation. Extracted from Additional file 2: Positive Significant Genes.	tumors. • Genes involved in transcriptional regulation			
hs19b	PMID: 15869706 Title: Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. Extracted from Additional file 2: Negative Significant Genes.	Down-regulated genes in FGFR3 mutated tumours relative to FGFR3 wild type tumours. • Genes involved in transcriptional regulation	• Immune response • Immune system process • Response to stimulus • Defense response • Response to wounding • Antigen processing and presentation • Inflammatory response • Response to external stimulus • Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II • Activation of immune response • Positive regulation of immune system process • Positive regulation of immune response • Innate immune response • Cell adhesion • Biological adhesion • Regulation of immune system process • Regulation of immune response • Positive regulation of multicellular organismal process • Immune effector process • Humoral immune response • Chemotaxis • Taxis • Leukocyte mediated immunity • Adaptive immune response • Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains • Activation of plasma proteins during acute inflammatory response • Complement activation • Leukocyte activation • Locomotory behavior • Regulation of multicellular organismal process • Cell activation	IFN-GAMMA IMMUNE MONOCYTE KILLER HISTOCOMPATIBILITY INFLAMMATORY CD8 CD3 NK MHC INFLAMMATION CYTOKINE RHEUMATOID CHEMOTACTIC IMMUNITY MOLECULE DECIDUAL TNF-ALPHA HLA-DR CD4 NATURAL DC SCLEROSIS	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs19b cont'd)			• Lymphocyte mediated immunity • Behavior • Cell motility • Localization of cell • Prostaglandin biosynthetic process • Prostanoid biosynthetic process • Immunoglobulin mediated immune response • B cell mediated immunity • Acute inflammatory response • Response to stress		
hs20	PMID: 15604246 Title: Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells. Extracted from Supplementary Table 1.	Genes differentially expressed between LHSR and LHS. • Androgen receptor signaling	• Mitotic cell cycle • Cell cycle phase • Cell cycle process • Cell cycle • M phase • M phase of mitotic cell cycle • Mitosis • Cell division • Regulation of progression through cell cycle • Regulation of cell cycle • Chromosome segregation • DNA replication • Regulation of mitosis • Mitotic sister chromatid segregation • Regulation of cellular process • Sister chromatid segregation • Cell cycle checkpoint • Spindle organization and biogenesis • Regulation of biological process • Microtubule-based process • Biopolymer metabolic process • Organelle organization and biogenesis • Cell proliferation • Interphase of mitotic cell cycle • DNA metabolic process • Microtubule cytoskeleton organization and biogenesis • Cellular component organization and biogenesis • Interphase • Mitotic cell cycle checkpoint • Cytoskeleton organization and biogenesis	MITOTIC CHECKPOINT SPINDLE ANAPHASE MITOSIS ARREST CYCLE MICROARRAY G2 PROMETAPHASE INTERPHASE CHROMATID CYTOKINESIS KINETOCHORE BREAST OVEREXPRESSION	E

Table 8.3: Significant GO terms versus PubMed tokens (continued)

ID	Source of gene list	Description and biology under studied	Over-represented GO (Biological Process) terms	Over-represented PubMed tokens	O/E
(hs20 cont'd)			• Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process		

O/E: Text-based ORA method that produced the results shown in 'Over-represented PubMed tokens' column (O = OutlierDM; E = ExtendedHG). Abbreviation "cont'd" means continued from the results for that particular gene list.

8.3 Performance in different organisms

The examples presented have so far been focused on human datasets. Since the human system has been extensively researched upon, literature in this organism is more extensive and most genes have some publications documenting their functions. Therefore the proposed text-based approaches work relatively well in datasets derived from the human system. For most other organisms, however, there is a relative paucity of literature the scope of which can often be very narrow. For example, literature and studies on *Xenopus laevis* are largely focused on developmental biology, while investigations of molecular biology and biochemistry are relatively limited.

OutlierDM and ExtendedHG can be readily extended to other species for which an associated corpus of PubMed abstracts is available. However, their power will depend on the availability and quality of literature. To get an indication of the extent of annotation available for different organisms, literature indices for genes in human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, *Xenopus* and zebrafish were assembled based on the Entrez Gene ID (EGID) and PubMed ID (PMID) mappings provided by NCBI; the amount of genes with at least one PubMed article as of October 2007 were extracted from the 'gene2pubmed' file[1]. For each organism, the number of PubMed articles (PMID) per gene and the number of gene associated with each PMID were determined. The descriptive statistics about the literature index for each of the eight species are shown in Table 8.4.

[1] ftp://ftp.ncbi.nih.gov/gene/DATA; time stamp: 25 Oct 2007.

Table 8.4: Summary of literature index across eight model organisms

Organism	Gene	PMID	PMID per gene			Gene per PMID		
			mean	median	max	mean	median	max
Human	25689	168156	13.44	5	1973	2.05	1	18007
Mouse	42005	113463	12.21	7	1467	4.52	1	24752
Rat	13675	32420	4.12	2	300	1.74	1	5810
Arabidopsis	13521	6787	2.45	2	39	4.89	1	2883
Drosophila	14793	21991	14.82	7	2652	9.97	3	14118
C. elegans	15424	2381	2.85	2	142	18.47	1	10495
Xenopus	11947	2631	2.12	2	263	9.61	1	10755
Zebrafish	13461	3189	2.22	1	217	9.38	2	11117

It can be seen that human, mouse and *Drosophila* have substantially more PMIDs linked to their genes than other species. On average, each gene in these species was cited by more than 13 articles, while less well-studied species such as *Arabidopsis*, *C. elegans*, *Xenopus* and zebrafish have on average less than 3 articles per gene. There is also a noticeable difference in the mean number of genes per PMID across different species. For example, each PMID referred to a mean of 2.05 and 4.52 genes, respectively, in human and mouse, while the mean number of genes cited per article is over 9 for *Drosophila*, *Xenopus* and zebrafish, and 18.47 for *C. elegans*.

While most articles are scientifically-specific and address individual genes, others may describe the biology of a large number of genes. Therefore the distributions of the number of genes per PMID are extremely skewed. For human genes, a total of 88 articles were found to be linked to more than 100 genes, and 929 referred to more than 10 genes; the highest number of genes referred to by a single abstract was 18007 (as of Oct 2007). Such inequalities in the published literature need to be taken into consideration during enrichment analysis. For the text-based ORA approaches developed in this thesis, only PubMed articles that cross-reference to a single gene were used to construct the text corpus. This is a reasonable solution because, as can be seen from Table 8.4, the median of the number of genes per PMID is either one or close to one for all organisms, indicating that a large proportions of the articles are dealing with one gene, and hence the removal of the less specific articles should not cause a marked loss in information.

8.3.1 Performance on 402 literature gene lists

To assess the performance of OutlierDM and ExtendedHG across different species, 402 gene lists collected from published literature spanning human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, *Xenopus* and zebrafish were analysed. These gene lists were based on 10 major Affymetrix arrays including HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Xenopus laevis, Celegans and Zebrafish (see Section 2.2.5 and Appendix A for details of these gene lists). The text corpora used in this analysis were built according to the procedures described in Section 2.3.

As shown in Figures 8.2 and 8.3, the number of tokens identified as over-represented by OutlierDM and ExtendedHG varies substantially between species. This appears to be related to the amount of literature available to each species in the text corpus used. Those species with a higher amount of overall annotation per gene (i.e. mean number of PMIDs per gene) tend to produce, on average, more significant tokens per gene list tested (Figure 8.4). As can be seen from Table 8.5, gene lists derived from experiments using the human-based arrays (i.e. HG-U133A, HG-U133 Plus 2.0) show the strongest performance; it has over 100,000 articles and an approximately 10:1 ratio of PMIDs to genes. *C. elegans*, *Xenopus* and zebrafish, on the other hand, each have a relatively small corpus with 1~2 % the number of articles of the human literature index, and less than 2 PMIDs per gene. The amount of significant tokens obtained for these organisms is low, and less than one in most cases.

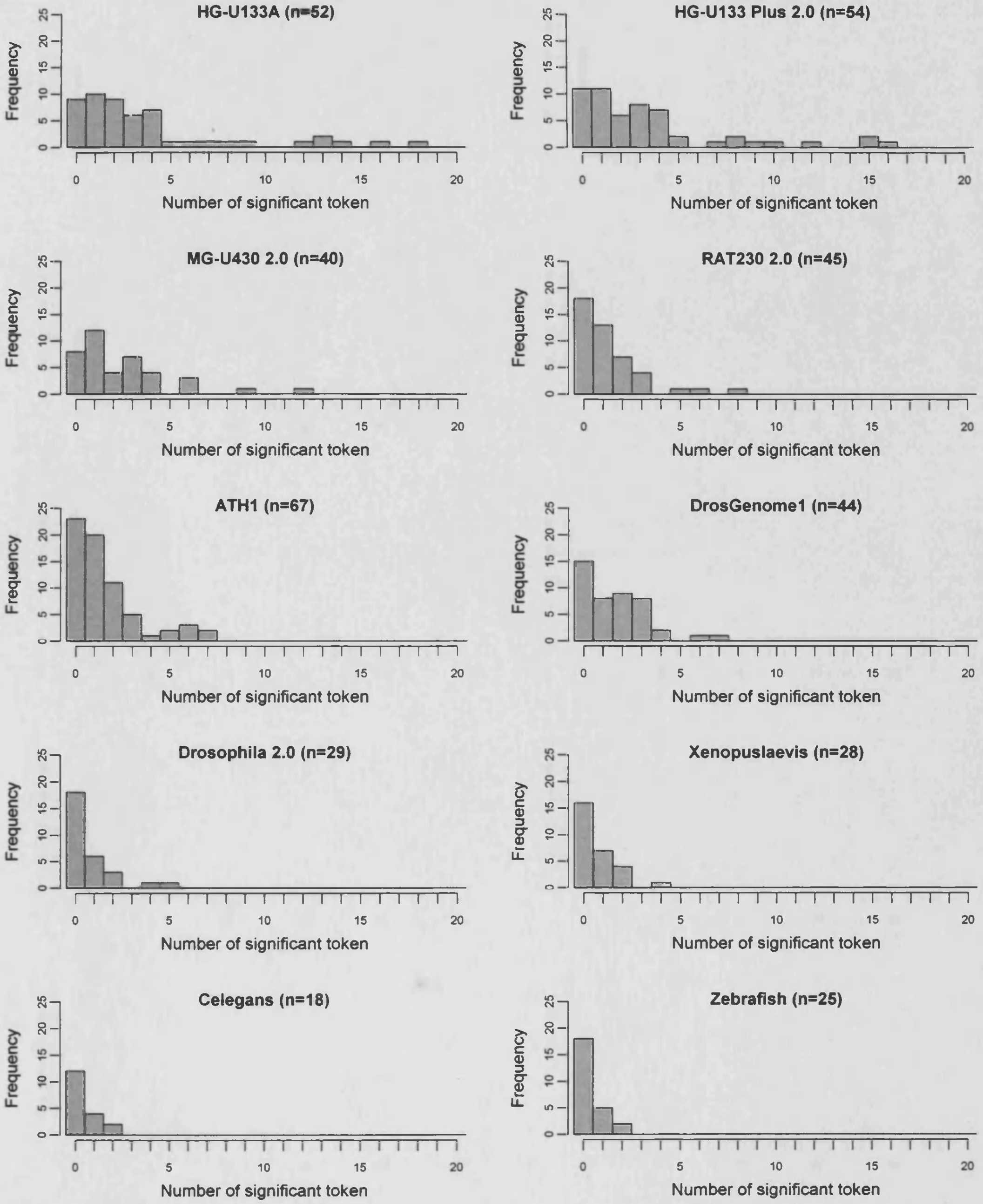


Figure 8.2: Histograms of the number of tokens identified as over-represented by OutlierDM in different species

The number of tokens identified as significantly over-represented by OutlierDM in gene lists derived from experiments performed on 10 Affymetrix platforms. *n* is the number of gene lists available for each platform.

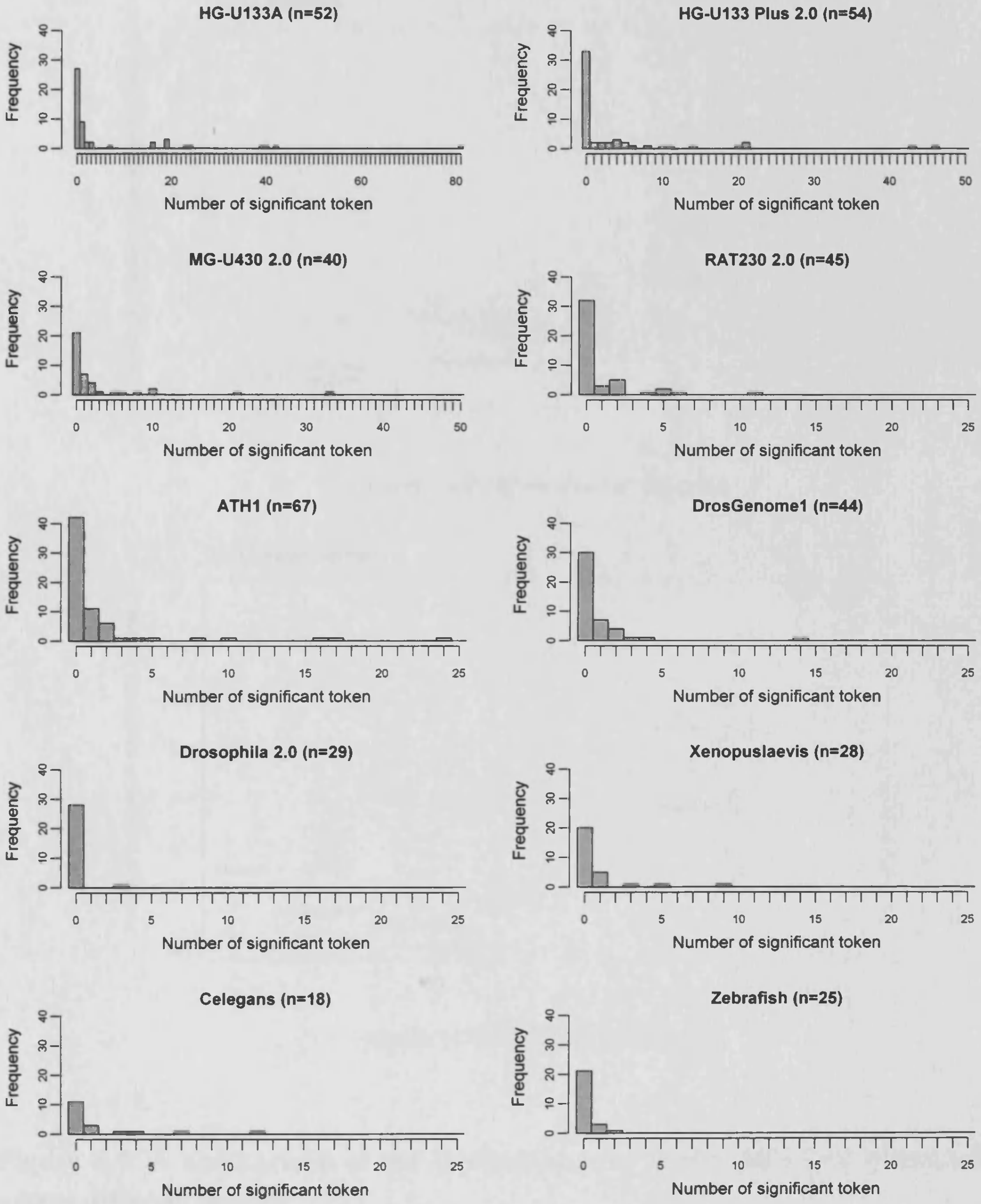


Figure 8.3: Histograms of the number of tokens identified as over-represented by ExtendedHG in different species

The number of tokens identified as significantly over-represented by ExtendedHG in gene lists derived from experiments performed on 10 Affymetrix platforms. *n* is the number of gene lists available for each platform.

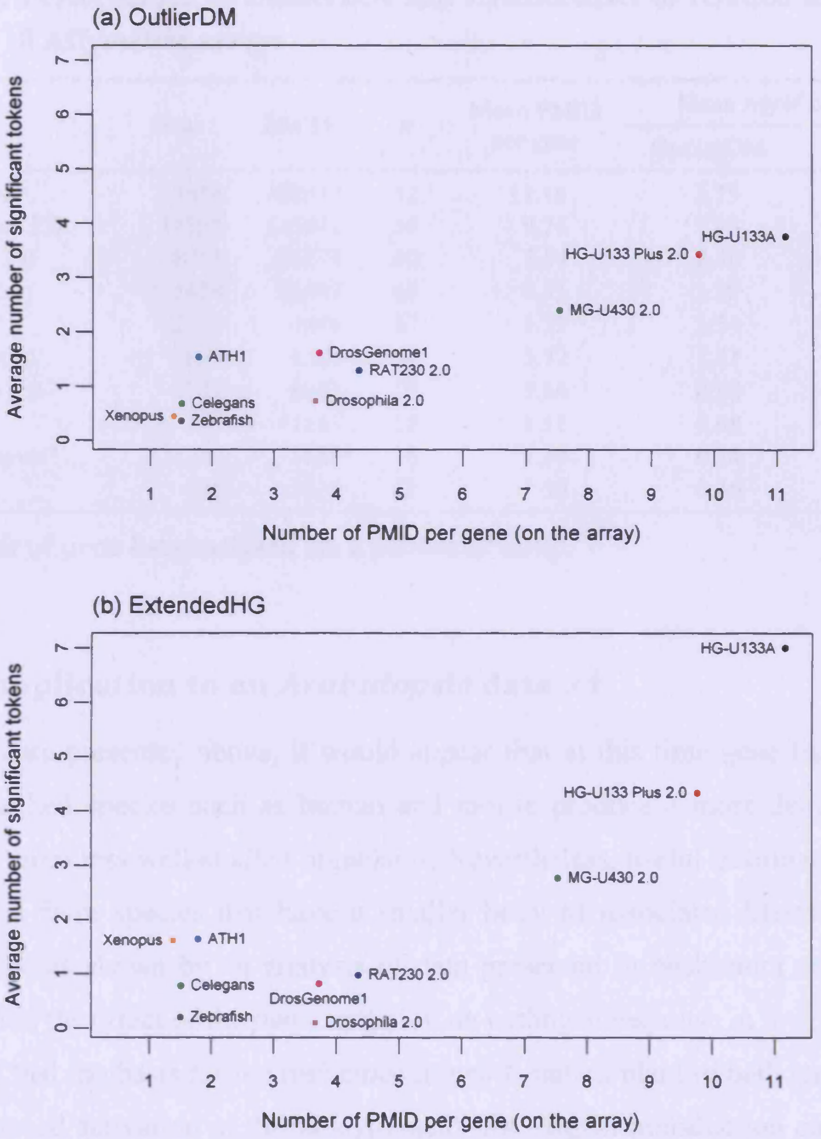


Figure 8.4: A comparison of the performance of OutlierDM and ExtendedHG across different species

The average number of tokens called significant by OutlierDM and ExtendedHG is plotted against the annotation density (number of PMID per gene) for experimentally-derived gene lists that were performed on 10 Affymetrix platforms. These arrays represent 8 different species: human (HG-U133A, HG-U133 Plus 2.0), mouse (MG-U430 2.0), rat (RAT230 2.0), *Arabidopsis* (ATH1), *Drosophila* (DrosGenome1, Drosophila 2.0), *Xenopus* (Xenopus laevis; abbreviated as 'Xenopus' in the plot), *C. elegans* (Celegans) and zebrafish (Zebrafish).

Table 8.5: Performance of OutlierDM and ExtendedHG in relation to literature index for 10 Affymetrix arrays

Array	Gene	PMID	*n*	Mean PMID per gene	Mean significant tokens	
					OutlierDM	ExtendedHG
HG-U133A	9638	107517	52	11.16	3.75	7.00
HG-133 Plus 2.0	11358	110811	54	9.76	3.43	4.33
MG-U430 2.0	8515	64171	40	7.54	2.40	2.78
RAT 230 2.0	5424	23597	45	4.35	1.29	0.98
Ath1	2290	4101	67	1.79	1.54	1.64
DrosGenome1	1694	6303	44	3.72	1.61	0.82
Drosophila 2.0	1722	6303	29	3.66	0.72	0.10
Celegans	767	1161	28	1.51	0.68	0.79
Xenopus laevis	1219	1689	18	1.39	0.44	1.61
Zebrafish	685	1030	25	1.50	0.36	0.20

n = number of gene lists analysed for a particular array.

8.3.2 Application to an *Arabidopsis* dataset

From the data presented above, it would appear that at this time gene lists based on well-researched species such as human and mouse produce a more detailed insight than those from less well-studied organisms. Nevertheless, useful information can still be obtained from species that have a smaller body of associated literature such as *Arabidopsis*, as shown by an analysis of data presented in Nishimura *et al.* (2003). They studied the effect of the *pmr4* mutation on pathogen response in *Arabidopsis* and concluded that the basis for the resistance in *pmr4* mutant plant to pathogens was due to an enhanced activation of the salicylic-acid (SA) signal transduction pathway. The list of differentially expressed genes reported by Nishimura *et al.* was re-analysed using a "trimmed" version of the text corpus built from only those papers published before 2003, so as to mine no more than the knowledge that was available to the authors of the original paper.

As shown in Table 8.6, three tokens, 'salicylic', 'SA' and 'resistance' were identified as over-represented in the gene list; 'salicylic', 'SA' were called significant by both OutlierDM and ExtendedHG, hence recapitulating the key conclusions by Nishimura *et al.* This demonstrates that despite the lower level of available literature, biologically-plausible results can still be obtained for less well-annotated species by using the text-based ORA approaches proposed in this work.

Table 8.6: Significant abstract terms in the Nishimura gene list

(a) OutlierDM

Term	*Chip* frequency	*List* frequency	*Z*-score	*p*-value	Bonferroni *p*-value	Rank
SALICYLIC	43	11	-4.1935	1.37E-05	0.0095	1
SA	22	8	-4.1774	1.47E-05	0.0102	2

(b) ExtendedHG

Term	*Chip* frequency	*List* frequency	Odds ratio	*p*-value	Bonferroni *p*-value	Rank
SA	22	8	1.7724	2.11E-05	0.0146	1
SALICYLIC	43	11	1.2529	4.23E-05	0.0292	2
RESISTANCE	117	16	1.0064	7.17E-05	0.0495	3

Over-represented terms were defined as having Bonferroni p-value ≤ 0.05. The results were ordered by increasing p-values. The text corpus used in this analysis is a trimmed version that contains only articles that were published before 2003. The gene universe used is that based on the Ath1 chip and contains 21,566 unique genes, of which 1360 have abstracts in the trimmed text corpus. A total of 690 tokens were analysed.

8.4 Behaviour of OutlierDM and ExtendedHG

Despite the difference in the underlying mechanisms for assessing over-representation, the results produced by OutlierDM and ExtendedHG generally show a good concordance in most analyses that have been performed (see Figure 7.4 for an example). However, close examination of the significant terms reported by the two methods reveal an interesting characteristic: tokens that were reported as significant by OutlierDM but not by ExtendedHG tend to be terms with relatively low *Chip* and *List* frequencies, whereas terms that were uniquely identified by ExtendedHG as significant are mostly associated with high *Chip* and *List* frequencies. This phenomenon is illustrate in Figure 8.5, in which the literature gene lists from the human arrays, that is HG-U133A and HG-U133 Plus 2.0, were analysed with both

OutlierDM and ExtendedHG, and the tokens that were uniquely identified as significantly enriched by the two approaches were compared. OutlierDM appears to have greater sensitivity in smaller gene lists, since significant tokens that were unique to this method tend to be terms that are supported by few genes in small gene lists. In contrast, ExtendedHG has better power than OutlierDM in detecting enriched terms supported by lots more genes in larger gene lists. A similar pattern recurs for datasets from other organisms besides human, as illustrated in Figure 8.6. Possible explanations for this behaviour are suggested as follows.

Recall that in OutlierDM, all tokens associated with a given gene list are first divided into groups according to their *List* frequency, then local mean and standard deviation of the *Chip* frequencies are calculated for each such group, based on which a *Z*-score is derived for each token as a means for scoring over-representation. Since there are always more observations in groups corresponding to low *List* frequencies (see the scatter plot of *List* versus *Chip* frequencies in Figure 6.1 for an example), therefore the estimation of mean and standard deviation are more accurate for these tokens than tokens corresponding to moderate to high *List* frequencies, for which considerably less observations are available for the estimation of mean and standard deviation. Another factor contributing to the reduced sensitivity of OutlierDM at moderate to high *List* frequency regions is that tokens falling in these regions are susceptible to the masking effect described in Section 6.4.4, preventing potential outliers to be detected.

ExtendedHG scores significance based on the calculation of hypergeometric *p*-values. The odds ratio, which is required for the calculation of *p*-value, is determined as a function of the mean; this is predicted by fitting polynomial regression model to the token frequency data. As discussed in Section 7.3.3, for tokens associated with low *List* and *Chip* frequencies, the odds ratio can be over-estimated due to a poorly predicted mean. Although an adjustment has been implemented to mitigate this problem, there can still be loss of sensitivity in some cases, especially in smaller gene lists.

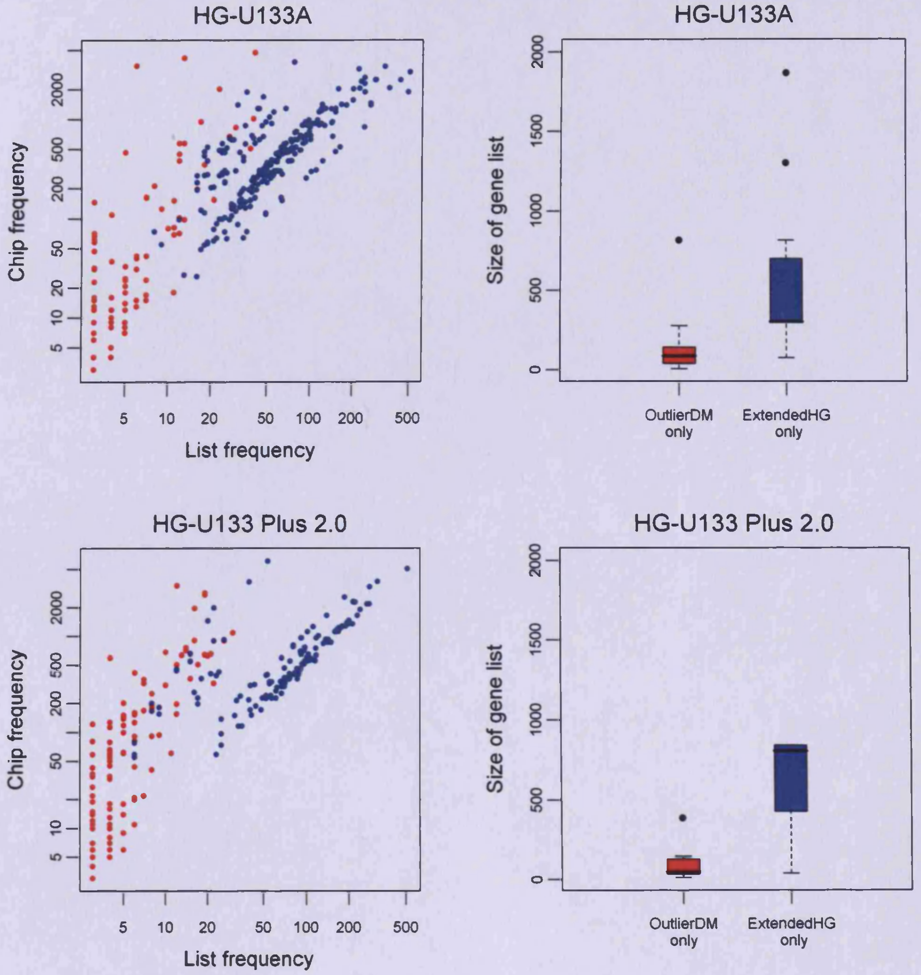


Figure 8.5: Characteristics of significant tokens unique to either OutlierDM or ExtendedHG in human datasets

The red points corresponding to tokens that were identified as over-represented by OutlierDM only, while the blue points corresponding to tokens that were identified as over-represented by ExtendedHG only.

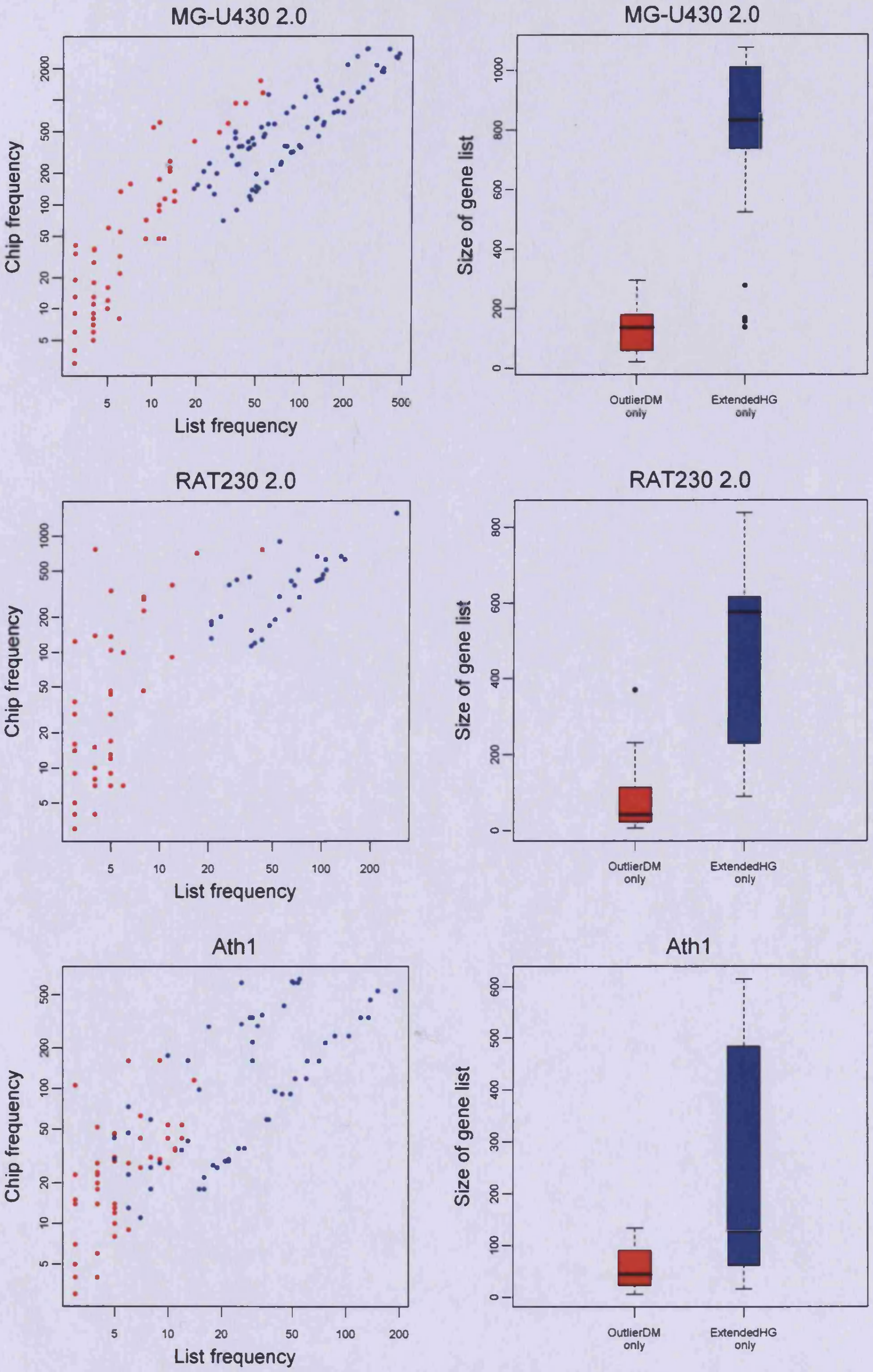


Figure 8.6: Characteristics of significant tokens unique to either OutlierDM or ExtendedHG in mouse, rat and *Arabidopsis* datasets

The red points corresponding to tokens that were identified as over-represented by OutlierDM only, while the blue points corresponding to tokens that were identified as over-represented by ExtendedHG only.

8.5 Discussion

Evaluating the performance of any exploratory approach such as the text-based ORA approaches proposed in this thesis is a challenging task because it is difficult to find datasets for which "ground truth" is known. Therefore a focused approach was undertaken to assess the performance of the proposed methods. Specifically, the evaluation focused on gene lists based on the HG-U133A array, and compared the outputs from OutlierDM and ExtendedHG with those obtained from (i) existing literature-based tools, and (ii) standard ORA approaches that mines GO terms. Such comparative analyses show that the results produced by the proposed text-based ORA approaches (OutlierDM and ExtendedHG) are both biologically relevance and plausible, and in most cases are in accordance with the manually determined annotations (see examples given in Table 8.3). In addition, the proposed text-based methods appear to provide distinct insights into the biological themes over-represented in a gene list compared to the results from undertaking ORA using GO terms, suggesting that the proposed approaches can be used to complement and extend existing ontology-based functional analysis tools for guiding the biological interpretation of microarray data.

Although the proposed text-based methods appear somewhat less effective for non-human species (cf. Table 8.5, Figures 8.2 and 8.3), apparently biologically-meaningful conclusions can still be obtained for dataset derived from certain less well-annotated species like *Arabidopsis*, as illustrated by the analysis based on the Nishimura gene list (Section 8.3.2). It is anticipated that as the biological knowledge and scientific publications accumulate over time, the quality of the results obtained for the less well-studied species such as *C. elegans*, *Xenopus* and zebrafish will also improve.

Chapter 9

PAKORA: a web application for interpreting microarray gene lists using text mining

9.1 Introduction

Having demonstrated the potential of text-based ORA in guiding the biological interpretation of microarray data in the previous Chapters, the focus of this Chapter is on the implementation of a graphical user interface for the proposed text mining algorithms. As pointed out by Alterovitz *et al.* (2008), approximately 97% of the bioinformatics applications hosted in SourceForge[1] do not have a graphical user interface, and these applications might not be utilised to their full potential simply because they are not accessible to a biologist with limited computational skills. A web-based application (named PAKORA) was created during the course of this work, with the aim of providing an easily accessible web interface from which the text-based ORA results can be analysed and visualised. The hope is that this will allow for more widespread dissemination of the proposed text mining algorithms within the biomedical community. The two computationally tractable approaches for performing ORA mining on PubMed abstract texts, based on the use of detection of outliers and the extended hypergeometric test, were implemented within PAKORA. The permutation test-based approach was not implemented because it is extremely computer-intensive, and hence is not suitable for routine analysis via a web service. PAKORA is publicly available at http://www.pakora.cf.ac.uk/pakora.php. Its main features and functionality are described in the following sections.

[1] SourceForge is a web-based open source software repository (http://sourceforge.net). As of July 2009, around 2097 software in SourceForge were categorised as "Bioinformatics" applications.

9.2 Using PAKORA to find significantly over-represented terms in a gene list

The core functionality of PAKORA is to calculate term over-representation for a list of differentially expressed genes, and offers an informative visual representation of the analysis results. Currently, PAKORA supports analysis based on 10 Affymetrix microarray platforms (HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Xenopus laevis, Celegans and Zebrafish), and 8 species (human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, *Xenopus* and zebrafish). Mappings of Affymetrix probeset IDs to Entrez Gene IDs and the relevant PubMed abstracts were retrieved from various databases as described in Section 2.3.1; the text corpora required were constructed according to the text processing procedures outlined in Section 2.3.2.

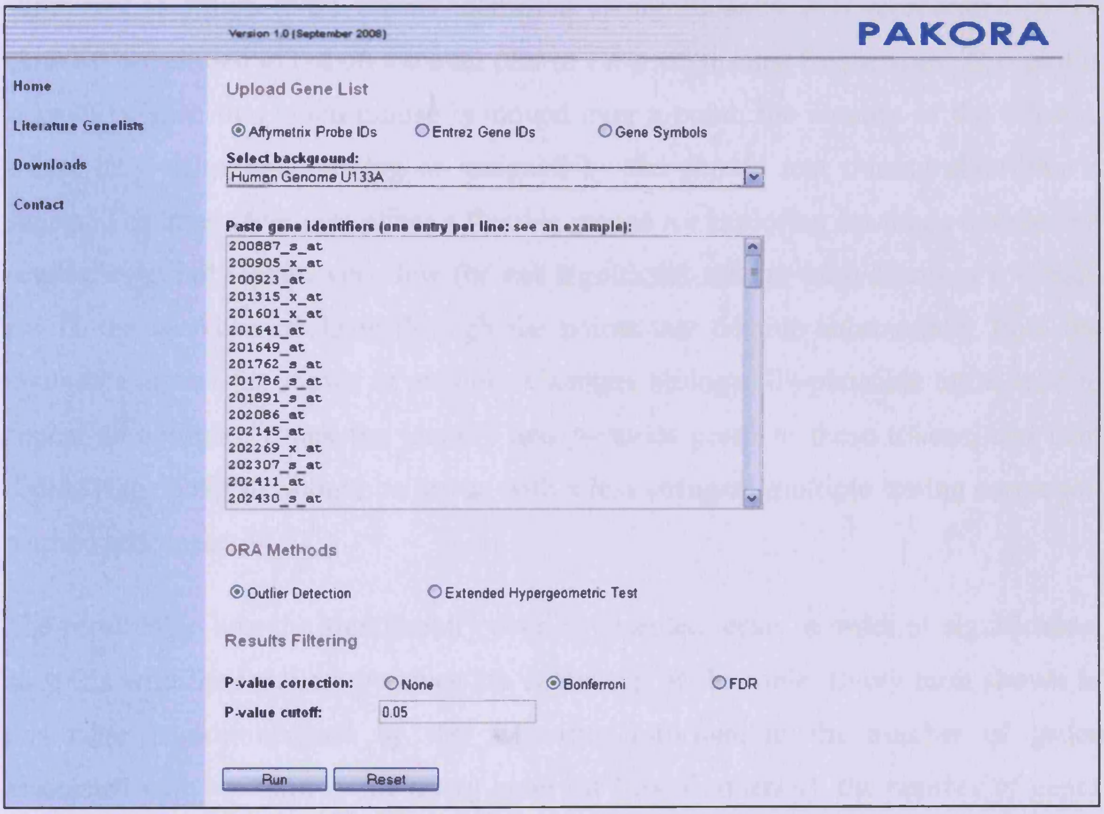
9.2.1 Input to PAKORA

The query interface of PAKORA adopts a simple design (Figure 9.1). An analysis can be initiated in three steps:

Step 1: Upload gene list. The primary input to PAKORA is a list of gene identifiers from any of the currently supported Affymetrix microarray platforms or species (see above for details). Three types of gene identifiers are accepted, including Affymetrix probeset IDs, Entrez Gene IDs and gene symbols. The user can upload the list of gene identifiers either by copy-paste into the web form or as a text file.

Step 2: Select analytical method. After the gene list has been uploaded, the user must specify which of the two text-based ORA algorithms should be used for identifying over-represented terms in the gene list: the outlier detection method or the extended hypergeometric test.

Step 3: Set results filtering options. To identify significantly over-represented terms, the multiple hypothesis testing correction method and the threshold for the *p*-value significance level must be specified. Two multiple testing correction methods are currently implemented within PAKORA, including the method of Bonferroni that controls the family-wise error rate, and the false discovery rate (FDR) method that controls the expected proportion of false discoveries amongst the rejected hypotheses. The default setting is set to the Bonferroni correction and a cutoff of 0.05, i.e. any terms with *p*-values less than or equal to 0.05 after Bonferroni correction will be reported as significantly over-represented in the gene list.



Upload Gene List

● Affymetrix Probe IDs ○ Entrez Gene IDs ○ Gene Symbols

Select background: Human Genome U133A

Paste gene identifiers (one entry per line: see an example):

```
200887_s_at
200905_x_at
200923_at
201315_x_at
201601_x_at
201641_at
201649_at
201762_s_at
201786_s_at
201891_s_at
202086_at
202145_at
202269_x_at
202307_s_at
202411_at
202430_s_at
```

ORA Methods

● Outlier Detection ○ Extended Hypergeometric Test

Results Filtering

P-value correction: ○ None ● Bonferroni ○ FDR

P-value cutoff: 0.05

Figure 9.1: User interface of PAKORA

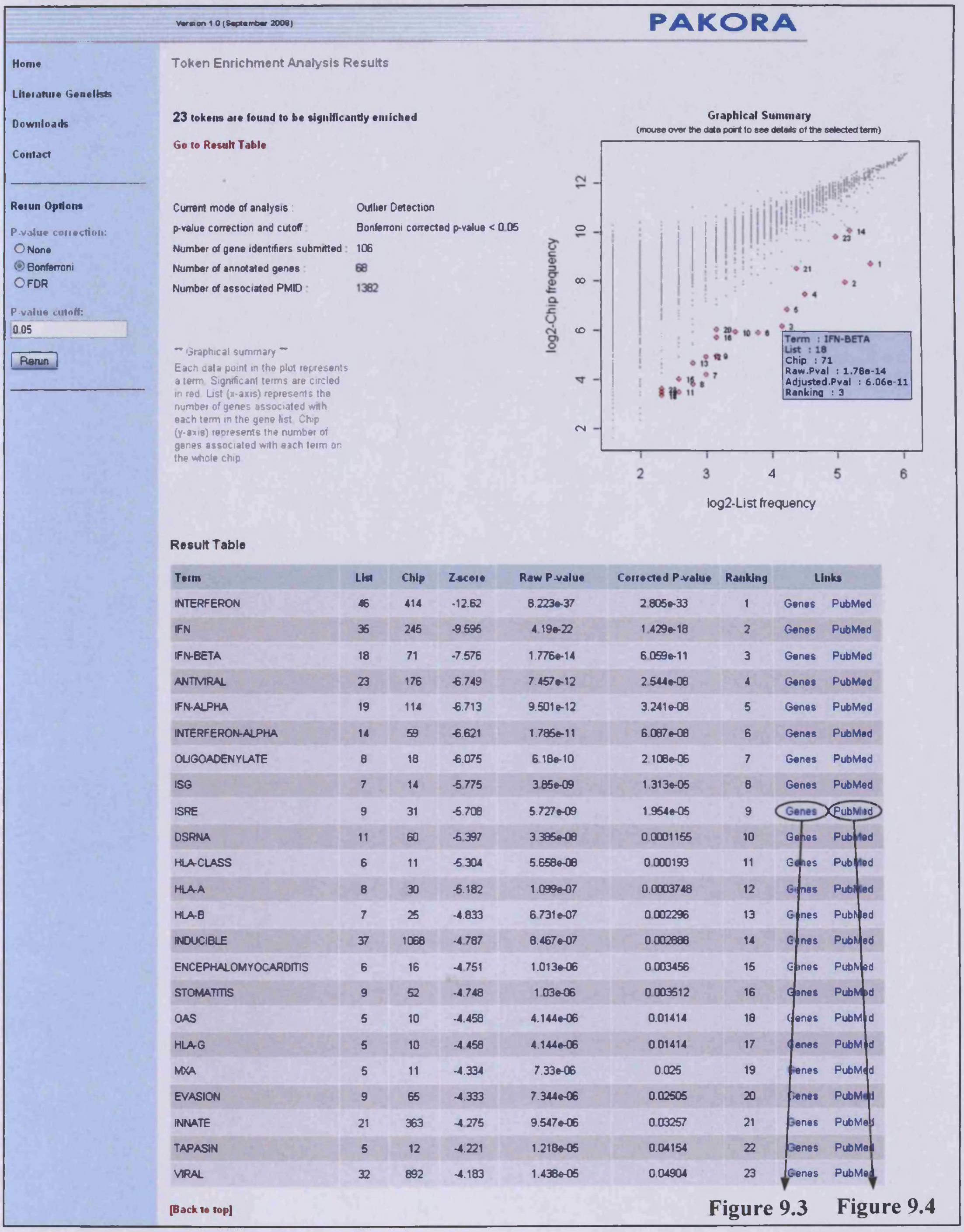
9.2.2 Output and representation of results

The primary output of PAKORA, irrespective of which of the two text-based ORA algorithms was used, is a set of enriched terms from PubMed abstracts relevant to the list of genes uploaded to the system. An example of the result page produced by

PAKORA is shown in Figure 9.2. This set of results was generated by submitting the list of interferon-stimulated genes in the ISG gene list (details of this gene list can be found in Section 2.2.1) to the PAKORA server in the form of Affymetrix probeset IDs, and analysing them with the outlier detection-based method with Bonferroni adjusted p-value ≤ 0.05 as the cutoff.

As can be seen from Figure 9.2, the main result table is preceded by a graphical summary, in which those tokens identified as significantly over-represented in the gene list are circled in red on a scatter plot of *Chip* versus *List* frequencies. This plot is interactive such that when mouse is moved over a point, the identity of the selected token, its p-value and ranking as assigned by the chosen text mining algorithm is shown. The interactive plot offers a flexible means for exploring the token enrichment results; especially when very few (or no) significant tokens were found at a certain cutoff, the user can navigate through the points that deviate substantially from the main data cluster (as shown in previous Chapters biologically-plausible terms tend to appear as outliers), check the identity and p-values given to these tokens, and then decide if the analysis should be re-run with a less stringent multiple testing correction method and/or cutoff.

The result table lists the significantly over-represented terms in order of significance, so terms with the smallest p-values are at the top of the table. Every term shown in this table is accompanied by the following information: the number of genes associated with the term in the query gene list (*List* frequency), the number of genes associated with the term in the background population (*Chip* frequency), the Z-score (for outlier detection method) or the odds ratio (for extended hypergeometric test), the statistical significance (raw and corrected p-values) of such enrichment and the term rank.



Version 1.0 (September 2008)

PAKORA

Rerun Options

P-value correction:

○ None
● Bonferroni
○ FDR

P-value cutoff: 0.05

Token Enrichment Analysis Results

23 tokens are found to be significantly enriched

Go to Result Table

Current mode of analysis :	Outlier Detection
p-value correction and cutoff :	Bonferroni corrected p-value < 0.05
Number of gene identifiers submitted :	106
Number of annotated genes :	68
Number of associated PMID :	1382

** Graphical summary **

Each data point in the plot represents a term. Significant terms are circled in red. List (x-axis) represents the number of genes associated with each term in the gene list. Chip (y-axis) represents the number of genes associated with each term on the whole chip.

Graphical Summary

(mouse over the data point to see details of the selected term)

Term : IFN-BETA
List : 18
Chip : 71
Raw.Pval : 1.78e-14
Adjusted.Pval : 6.06e-11
Ranking : 3

Result Table

Term	List	Chip	Z-score	Raw P-value	Corrected P-value	Ranking	Links
INTERFERON	46	414	-12.62	8.223e-37	2.805e-33	1	Genes PubMed
IFN	36	245	-9.595	4.19e-22	1.429e-18	2	Genes PubMed
IFN-BETA	18	71	-7.576	1.776e-14	6.059e-11	3	Genes PubMed
ANTIVIRAL	23	176	-6.749	7.457e-12	2.544e-08	4	Genes PubMed
IFN-ALPHA	19	114	-6.713	9.501e-12	3.241e-08	5	Genes PubMed
INTERFERON-ALPHA	14	59	-6.621	1.785e-11	6.087e-08	6	Genes PubMed
OLIGOADENYLATE	8	18	-6.075	6.18e-10	2.108e-06	7	Genes PubMed
ISG	7	14	-5.775	3.85e-09	1.313e-05	8	Genes PubMed
ISRE	9	31	-5.708	5.727e-09	1.954e-05	9	Genes PubMed
DSRNA	11	60	-5.397	3.385e-08	0.0001155	10	Genes PubMed
HLA-CLASS	6	11	-5.304	5.658e-08	0.000193	11	Genes PubMed
HLA-A	8	30	-5.182	1.099e-07	0.0003748	12	Genes PubMed
HLA-B	7	25	-4.833	6.731e-07	0.002296	13	Genes PubMed
INDUCIBLE	37	1068	-4.787	8.467e-07	0.002888	14	Genes PubMed
ENCEPHALOMYOCARDITIS	6	16	-4.751	1.013e-06	0.003456	15	Genes PubMed
STOMATITIS	9	52	-4.748	1.03e-06	0.003512	16	Genes PubMed
OAS	5	10	-4.458	4.144e-06	0.01414	18	Genes PubMed
HLA-G	5	10	-4.458	4.144e-06	0.01414	17	Genes PubMed
MXA	5	11	-4.334	7.33e-06	0.025	19	Genes PubMed
EVASION	9	65	-4.333	7.344e-06	0.02505	20	Genes PubMed
INNATE	21	363	-4.275	9.547e-06	0.03257	21	Genes PubMed
TAPASIN	5	12	-4.221	1.218e-05	0.04154	22	Genes PubMed
VIRAL	32	892	-4.183	1.438e-05	0.04904	23	Genes PubMed

[Back to top]

Figure 9.2: Screenshot of PAKORA showing the result page from analysing the ISG gene list using the outlier detection-based approach

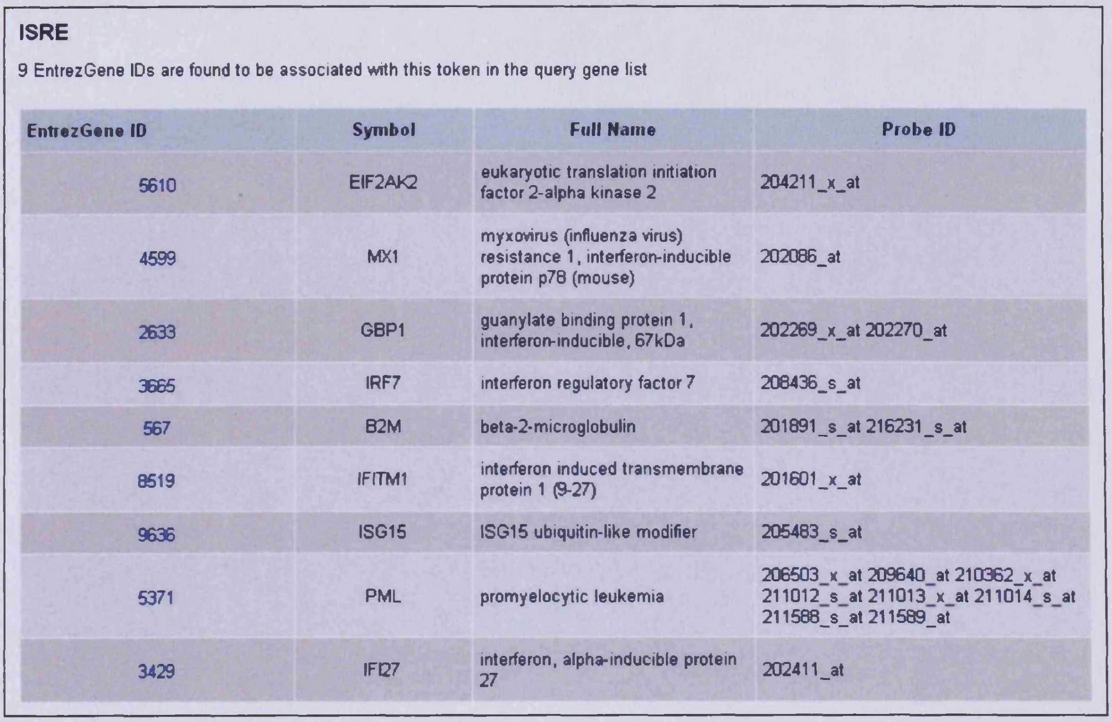
The result page contains two main parts: (i) an interactive plot with the over-represented terms circled in red, and (ii) a result table listing the over-represented terms. Hyperlinks are provided so that user can find out the genes significantly associated with the analysed terms (see Figure 9.3 for an example) and the PubMed abstracts from which they were extracted (see Figure 9.4 for an example).

Context analysis of the over-represented terms

While the over-represented terms listed in the result table provides a first impression of the biological processes or themes related to the gene list, the user can 'drill down' into these results by following the hyperlinks labelled *Gene* and *PubMed* corresponding to these terms. For example, by clicking on the *Gene* hyperlink corresponding to the term 'ISRE' in the result table shown in Figure 9.2, the user will be link to another web page as that shown in Figure 9.3, which lists the genes (including probeset IDs and Entrez Gene IDs, symbols and gene names) that are significantly associated with the selected term. On the other hand, by clicking on the *PubMed* hyperlink, the user can see the PubMed abstracts in which the over-represented term and genes in the query gene list co-occur. These abstracts provide contextual clues to the over-represented terms, allowing the user to explore both the meaning and relationships of the over-represented terms. An example is given in Figure 9.4, in which three PubMed abstracts related to the term 'ISRE' are shown. Consider the following sentences taken from the third abstracts (PMID: 9726442) shown in Figure 9.4 (the over-represented terms are underlined for easy identification):

> *"The double-stranded RNA-dependent protein kinase (PKR) is a serine/threonine kinase that plays an important role in the antiviral activities of interferon (IFN) ... Sequence analysis of the PKR 5'-flanking region identified a canonical IFN-stimulated response element (ISRE), GAAAACGAAACT. Transient transfection of PKR promoter constructs linked to a luciferase reporter gene into human T98G cells indicated that this 5'-flanking region is capable of functioning as a basal promoter that is also inducible by IFN-alpha and IFN-beta but not IFN-gamma."*

We can see from the above examples that the relationships of the over-represented terms (such as 'antiviral', 'interferon', 'ISRE', 'inducible', 'IFN-alpha' and 'IFN-beta') and their interpretation becomes clearer when considered in the original sentences from which they were extracted. This simple example showed that the information contained in the significant terms was of biological relevance; however, their meaning could be better understood by considering the terms in the context of the co-occurring terms and relevant abstracts.

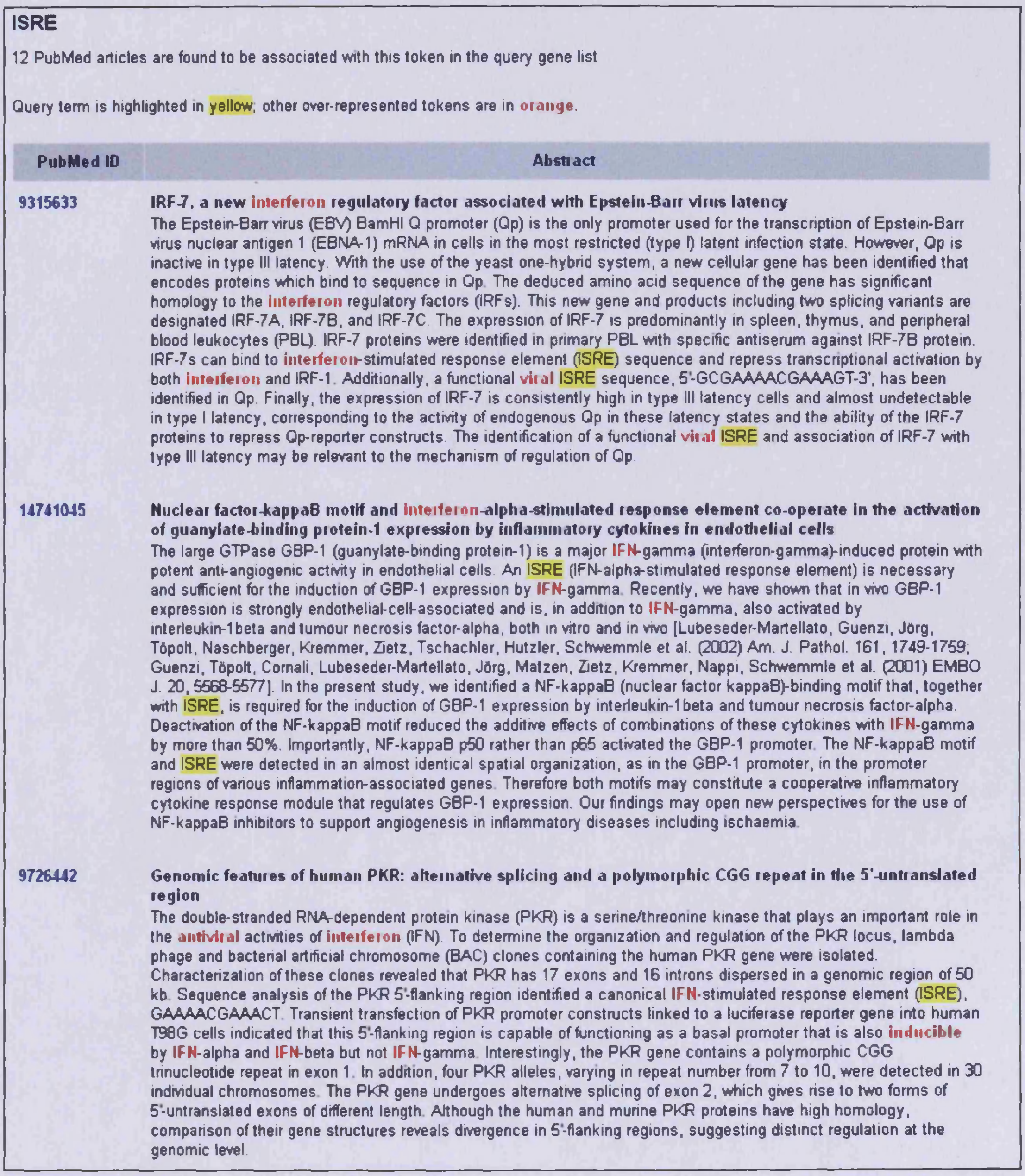


ISRE

9 EntrezGene IDs are found to be associated with this token in the query gene list

EntrezGene ID	Symbol	Full Name	Probe ID
5610	EIF2AK2	eukaryotic translation initiation factor 2-alpha kinase 2	204211_x_at
4599	MX1	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)	202086_at
2633	GBP1	guanylate binding protein 1, interferon-inducible, 67kDa	202269_x_at 202270_at
3665	IRF7	interferon regulatory factor 7	208436_s_at
567	B2M	beta-2-microglobulin	201891_s_at 216231_s_at
8519	IFITM1	interferon induced transmembrane protein 1 (9-27)	201601_x_at
9636	ISG15	ISG15 ubiquitin-like modifier	205483_s_at
5371	PML	promyelocytic leukemia	206503_x_at 209640_at 210362_x_at 211012_s_at 211013_x_at 211014_s_at 211588_s_at 211589_at
3429	IFI27	interferon, alpha-inducible protein 27	202411_at

Figure 9.3: Output from PAKORA showing genes and Affymetrix probeset IDs in the ISG gene list that were significantly associated with the term 'ISRE'



ISRE

12 PubMed articles are found to be associated with this token in the query gene list

Query term is highlighted in yellow; other over-represented tokens are in orange.

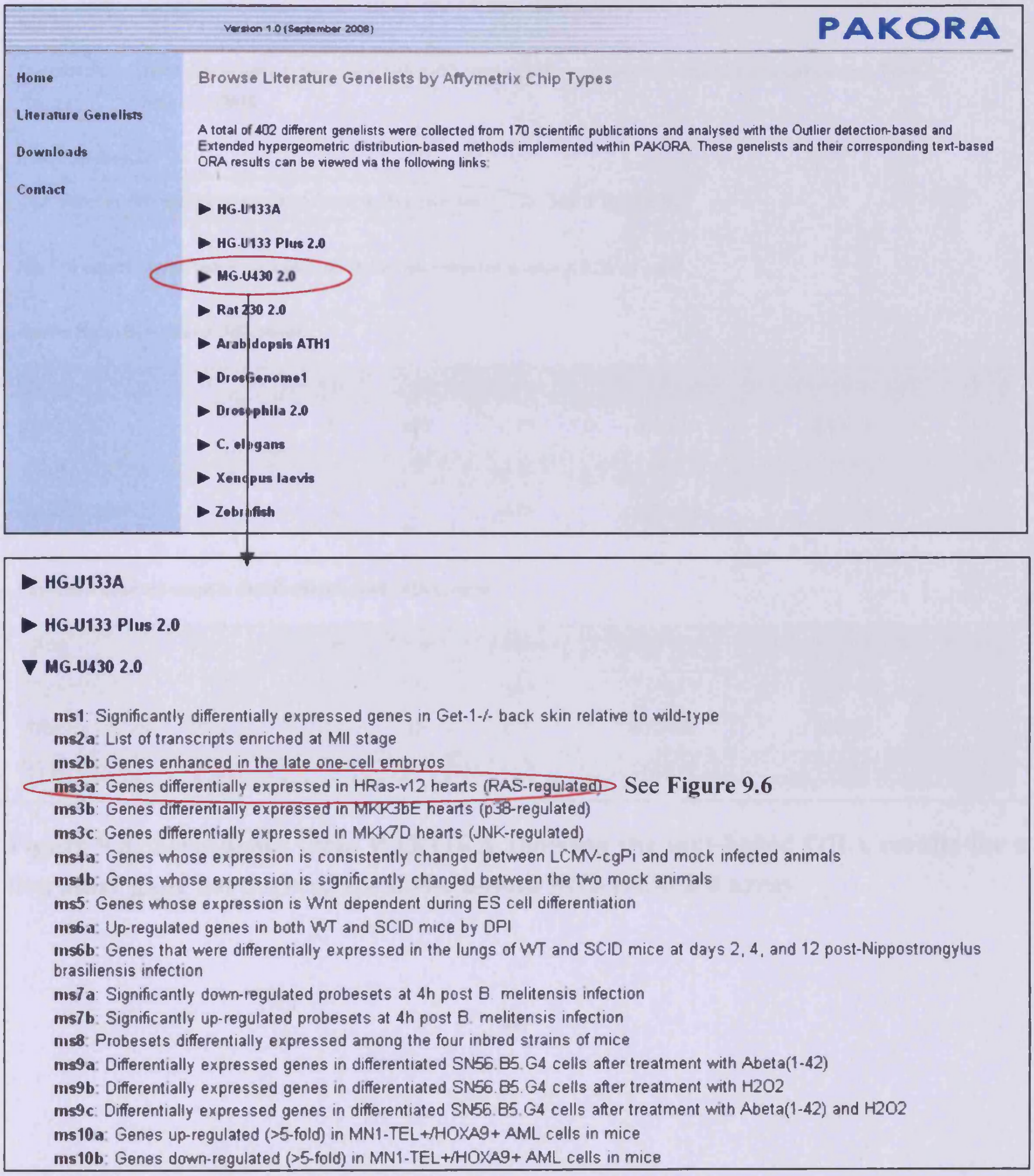
PubMed ID	Abstract
9315633	**IRF-7, a new interferon regulatory factor associated with Epstein-Barr virus latency** The Epstein-Barr virus (EBV) BamHI Q promoter (Qp) is the only promoter used for the transcription of Epstein-Barr virus nuclear antigen 1 (EBNA-1) mRNA in cells in the most restricted (type I) latent infection state. However, Qp is inactive in type III latency. With the use of the yeast one-hybrid system, a new cellular gene has been identified that encodes proteins which bind to sequence in Qp. The deduced amino acid sequence of the gene has significant homology to the interferon regulatory factors (IRFs). This new gene and products including two splicing variants are designated IRF-7A, IRF-7B, and IRF-7C. The expression of IRF-7 is predominantly in spleen, thymus, and peripheral blood leukocytes (PBL). IRF-7 proteins were identified in primary PBL with specific antiserum against IRF-7B protein. IRF-7s can bind to interferon-stimulated response element (ISRE) sequence and repress transcriptional activation by both interferon and IRF-1. Additionally, a functional viral ISRE sequence, 5'-GCGAAAACGAAAGT-3', has been identified in Qp. Finally, the expression of IRF-7 is consistently high in type III latency cells and almost undetectable in type I latency, corresponding to the activity of endogenous Qp in these latency states and the ability of the IRF-7 proteins to repress Qp-reporter constructs. The identification of a functional viral ISRE and association of IRF-7 with type III latency may be relevant to the mechanism of regulation of Qp.
14741045	**Nuclear factor-kappaB motif and interferon-alpha-stimulated response element co-operate in the activation of guanylate-binding protein-1 expression by inflammatory cytokines in endothelial cells** The large GTPase GBP-1 (guanylate-binding protein-1) is a major IFN-gamma (interferon-gamma)-induced protein with potent anti-angiogenic activity in endothelial cells. An ISRE (IFN-alpha-stimulated response element) is necessary and sufficient for the induction of GBP-1 expression by IFN-gamma. Recently, we have shown that in vivo GBP-1 expression is strongly endothelial-cell-associated and is, in addition to IFN-gamma, also activated by interleukin-1beta and tumour necrosis factor-alpha, both in vitro and in vivo [Lubeseder-Martellato, Guenzi, Jörg, Töpolt, Naschberger, Kremmer, Zietz, Tschachler, Hutzler, Schwemmle et al. (2002) Am. J. Pathol. 161, 1749-1759; Guenzi, Töpolt, Cornali, Lubeseder-Martellato, Jörg, Matzen, Zietz, Kremmer, Nappi, Schwemmle et al. (2001) EMBO J. 20, 5568-5577]. In the present study, we identified a NF-kappaB (nuclear factor kappaB)-binding motif that, together with ISRE, is required for the induction of GBP-1 expression by interleukin-1beta and tumour necrosis factor-alpha. Deactivation of the NF-kappaB motif reduced the additive effects of combinations of these cytokines with IFN-gamma by more than 50%. Importantly, NF-kappaB p50 rather than p65 activated the GBP-1 promoter. The NF-kappaB motif and ISRE were detected in an almost identical spatial organization, as in the GBP-1 promoter, in the promoter regions of various inflammation-associated genes. Therefore both motifs may constitute a cooperative inflammatory cytokine response module that regulates GBP-1 expression. Our findings may open new perspectives for the use of NF-kappaB inhibitors to support angiogenesis in inflammatory diseases including ischaemia.
9726442	**Genomic features of human PKR: alternative splicing and a polymorphic CGG repeat in the 5'-untranslated region** The double-stranded RNA-dependent protein kinase (PKR) is a serine/threonine kinase that plays an important role in the antiviral activities of interferon (IFN). To determine the organization and regulation of the PKR locus, lambda phage and bacterial artificial chromosome (BAC) clones containing the human PKR gene were isolated. Characterization of these clones revealed that PKR has 17 exons and 16 introns dispersed in a genomic region of 50 kb. Sequence analysis of the PKR 5'-flanking region identified a canonical IFN-stimulated response element (ISRE), GAAAACGAAACT. Transient transfection of PKR promoter constructs linked to a luciferase reporter gene into human T98G cells indicated that this 5'-flanking region is capable of functioning as a basal promoter that is also inducible by IFN-alpha and IFN-beta but not IFN-gamma. Interestingly, the PKR gene contains a polymorphic CGG trinucleotide repeat in exon 1. In addition, four PKR alleles, varying in repeat number from 7 to 10, were detected in 30 individual chromosomes. The PKR gene undergoes alternative splicing of exon 2, which gives rise to two forms of 5'-untranslated exons of different length. Although the human and murine PKR proteins have high homology, comparison of their gene structures reveals divergence in 5'-flanking regions, suggesting distinct regulation at the genomic level.

Figure 9.4: Output from PAKORA showing a subset of PubMed abstracts for which the term 'ISRE' and the genes in the ISG gene list co-occur

The over-represented terms were highlighted in different colours for easy identification. Term selected by the user is highlighted in yellow; whilst other over-represented terms are highlighted in orange. This analysis was performed based on the text corpus constructed as of 25 Oct 2007.

9.3 Using PAKORA to browse and download the literature gene lists

In Section 8.2.2, a comprehensive performance assessment was carried out to compare results produced by the proposed text-based approaches with that from GOstats when applied to the 52 human HG-U133A literature gene lists. Similar analyses were also performed for literature gene lists derived from other arrays, except for those based on the *Arabidopsis* and *Xenopus laevis* chips, because the Bioconductor GO annotation library files as required by GOstats are not available for these two chip types. For chip types where this analysis is possible, their text-based ORA and GOstats results are readily accessible via PAKORA for review by researchers with the relevant biological background. The user can browse through these pre-processed results by following the link "Literature Genelists" featured at the main menu. The 402 literature gene lists were organised according to the microarray platform on which they were based (Figure 9.5). From the expandable menu, the user can view all the gene lists available for a particular chip type and the biological conditions under study. By clicking on the gene list's identifier, the user will be presented with result tables summarising the text-based ORA results associated with that gene list. An example is shown in Figure 9.6, from which we can see that the outlier detection-based and the hypergeometric test-based approaches found three significant tokens each when applied to the literature gene list with identifier 'ms3a'. The results of applying GOstats to the same dataset can be viewed by clicking on the link "View enriched GO (biological process) terms in this gene list"; the user will then get a table listing the GO terms that were reported as significantly enriched by GOstats at a threshold of 0.05 after Bonferroni correction, as shown in Figure 9.7. This kind of comparison would help the user to make a judgement regarding the performance of the proposed text-based ORA methods, as opposed to standard ORA tools that mine GO (such as GOstats), in different species for which the amount and quality of literature data varies substantially.



Version 1.0 (September 2008)

PAKORA

Browse Literature Genelists by Affymetrix Chip Types

A total of 402 different genelists were collected from 170 scientific publications and analysed with the Outlier detection-based and Extended hypergeometric distribution-based methods implemented within PAKORA. These genelists and their corresponding text-based ORA results can be viewed via the following links:

- ▶ HG-U133A
- ▶ HG-U133 Plus 2.0
- ▶ MG-U430 2.0
- ▶ Rat 230 2.0
- ▶ Arabidopsis ATH1
- ▶ DrosGenome1
- ▶ Drosophila 2.0
- ▶ C. elegans
- ▶ Xenopus laevis
- ▶ Zebrafish

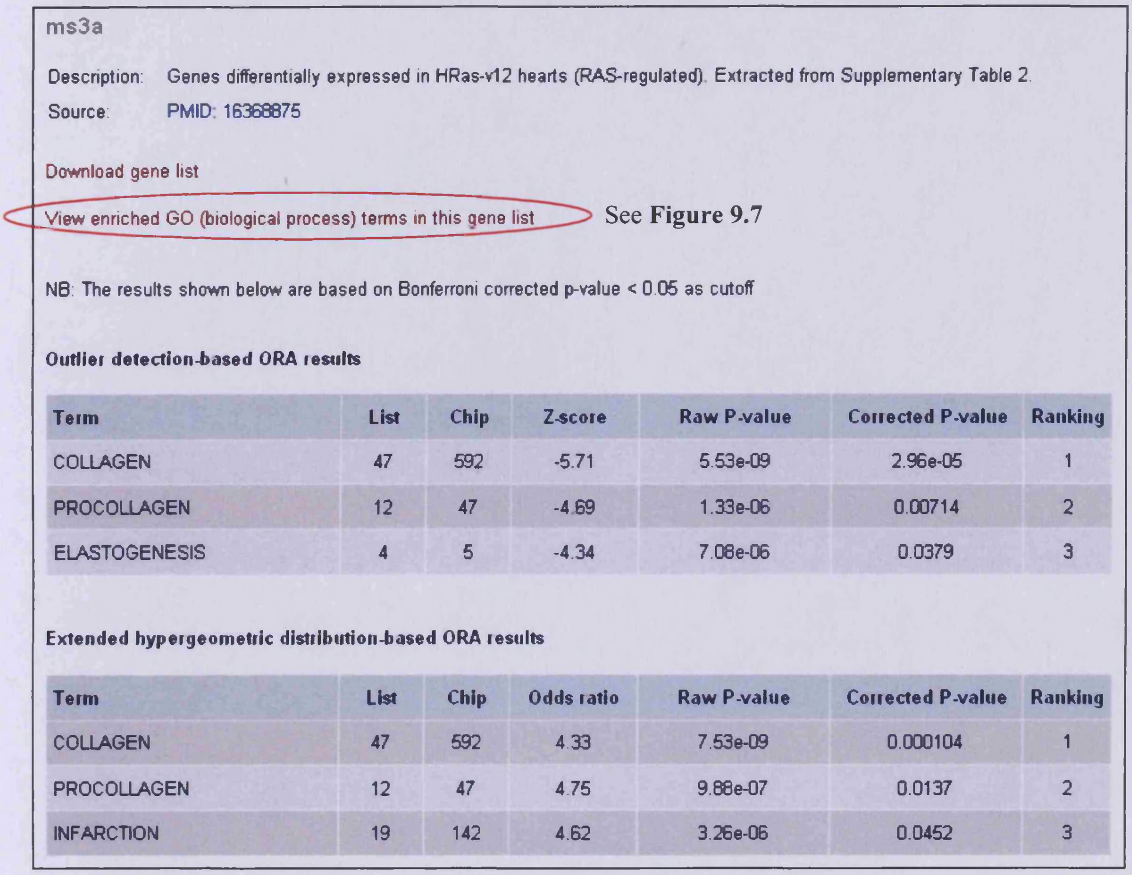
▶ **HG-U133A**

▶ **HG-U133 Plus 2.0**

▼ **MG-U430 2.0**

ms1: Significantly differentially expressed genes in Get-1-/- back skin relative to wild-type
ms2a: List of transcripts enriched at MII stage
ms2b: Genes enhanced in the late one-cell embryos
ms3a: Genes differentially expressed in HRas-v12 hearts (RAS-regulated) See Figure 9.6
ms3b: Genes differentially expressed in MKK3bE hearts (p38-regulated)
ms3c: Genes differentially expressed in MKK7D hearts (JNK-regulated)
ms4a: Genes whose expression is consistently changed between LCMV-cgPi and mock infected animals
ms4b: Genes whose expression is significantly changed between the two mock animals
ms5: Genes whose expression is Wnt dependent during ES cell differentiation
ms6a: Up-regulated genes in both WT and SCID mice by DPI
ms6b: Genes that were differentially expressed in the lungs of WT and SCID mice at days 2, 4, and 12 post-Nippostrongylus brasiliensis infection
ms7a: Significantly down-regulated probesets at 4h post B. melitensis infection
ms7b: Significantly up-regulated probesets at 4h post B. melitensis infection
ms8: Probesets differentially expressed among the four inbred strains of mice
ms9a: Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with Abeta(1-42)
ms9b: Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with H2O2
ms9c: Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with Abeta(1-42) and H2O2
ms10a: Genes up-regulated (>5-fold) in MN1-TEL+/HOXA9+ AML cells in mice
ms10b: Genes down-regulated (>5-fold) in MN1-TEL+/HOXA9+ AML cells in mice

Figure 9.5: Screenshots showing how the text-based ORA results for each of the literature gene lists used in this work can be viewed within PAKORA



ms3a

Description: Genes differentially expressed in HRas-v12 hearts (RAS-regulated). Extracted from Supplementary Table 2.

Source: PMID: 16368875

Download gene list

View enriched GO (biological process) terms in this gene list

See Figure 9.7

NB: The results shown below are based on Bonferroni corrected p-value < 0.05 as cutoff

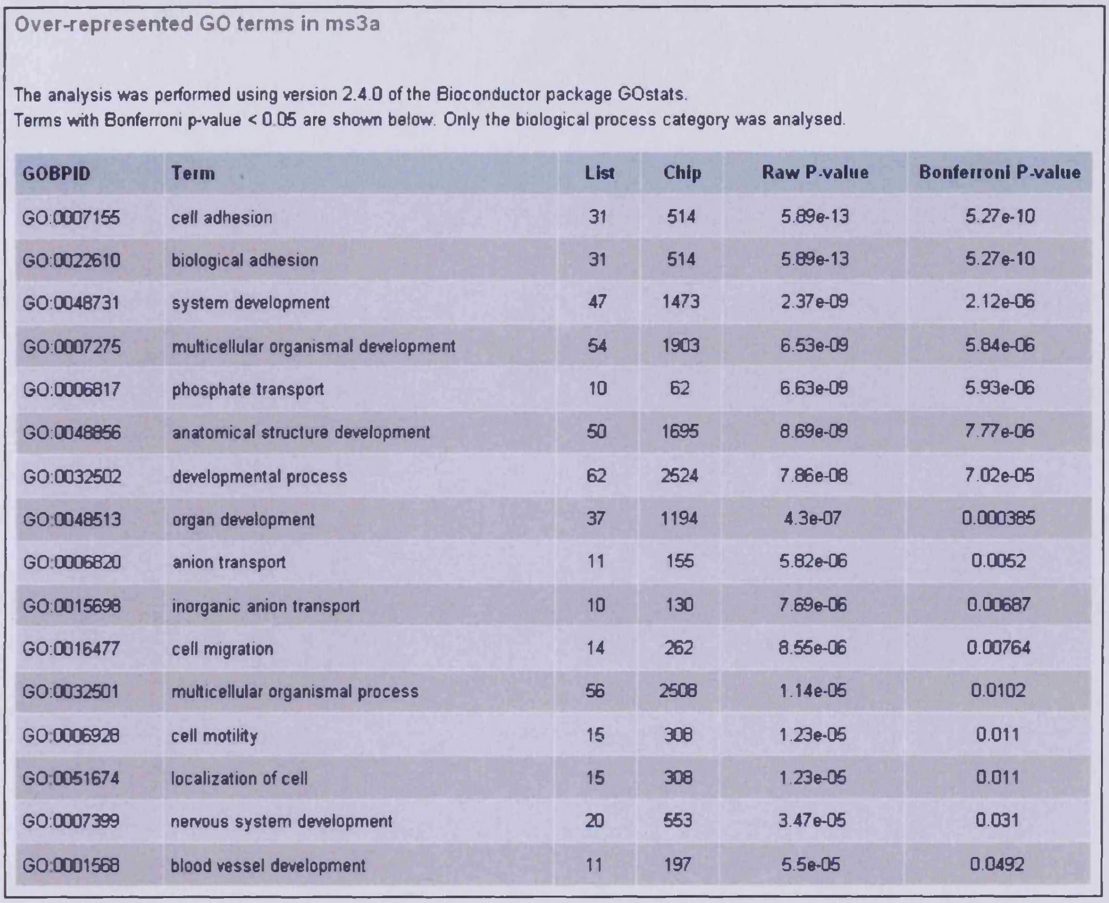
Outlier detection-based ORA results

Term	List	Chip	Z-score	Raw P-value	Corrected P-value	Ranking
COLLAGEN	47	592	-5.71	5.53e-09	2.96e-05	1
PROCOLLAGEN	12	47	-4.69	1.33e-06	0.00714	2
ELASTOGENESIS	4	5	-4.34	7.08e-06	0.0379	3

Extended hypergeometric distribution-based ORA results

Term	List	Chip	Odds ratio	Raw P-value	Corrected P-value	Ranking
COLLAGEN	47	592	4.33	7.53e-09	0.000104	1
PROCOLLAGEN	12	47	4.75	9.88e-07	0.0137	2
INFARCTION	19	142	4.62	3.26e-06	0.0452	3

Figure 9.6: Screenshot from PAKORA showing the text-based ORA results for a literature gene list derived from the mouse MG-U430 2.0 array



Over-represented GO terms in ms3a

The analysis was performed using version 2.4.0 of the Bioconductor package GOstats.
Terms with Bonferroni p-value < 0.05 are shown below. Only the biological process category was analysed.

GOBPID	Term	List	Chip	Raw P-value	Bonferroni P-value
GO:0007155	cell adhesion	31	514	5.89e-13	5.27e-10
GO:0022610	biological adhesion	31	514	5.89e-13	5.27e-10
GO:0048731	system development	47	1473	2.37e-09	2.12e-06
GO:0007275	multicellular organismal development	54	1903	6.53e-09	5.84e-06
GO:0006817	phosphate transport	10	62	6.63e-09	5.93e-06
GO:0048856	anatomical structure development	50	1695	8.69e-09	7.77e-06
GO:0032502	developmental process	62	2524	7.86e-08	7.02e-05
GO:0048513	organ development	37	1194	4.3e-07	0.000385
GO:0006820	anion transport	11	155	5.82e-06	0.0052
GO:0015698	inorganic anion transport	10	130	7.69e-06	0.00687
GO:0016477	cell migration	14	262	8.55e-06	0.00764
GO:0032501	multicellular organismal process	56	2508	1.14e-05	0.0102
GO:0006928	cell motility	15	308	1.23e-05	0.011
GO:0051674	localization of cell	15	308	1.23e-05	0.011
GO:0007399	nervous system development	20	553	3.47e-05	0.031
GO:0001568	blood vessel development	11	197	5.5e-05	0.0492

Figure 9.7: Screenshot of PAKORA showing the GOstats results for a literature gene list derived from the mouse MG-U430 2.0 array

9.4 Conclusions

It has been demonstrated in the previous Chapters that the outlier detection algorithm and extended hypergeometric test can be effectively integrated into a wider statistical framework for mining textual information and discern a coherent picture that exists within complex groups of genes. In this Chapter, a web application named PAKORA is presented. PAKORA was designed to provide an easily accessible interface to the proposed text-based ORA algorithms and offers an intuitive visual representation of the analysis results. With PAKORA, the user gets an overview of the over-represented term shared by genes in a gene list and can quickly follow the links to find the relevant publications. This makes the exploratory analysis of complex microarray gene lists more efficient. At this stage in its development, PAKORA only outputs the results as a web page. Future incarnations of this tool could look into providing textual output that can be imported into Excel spreadsheets or saved automatically via HTTP. In conclusion, PAKORA provides a gateway to explore text-based information associated with a list of differentially expressed genes, with the advantage of obtaining this information consistently and automatically, making it a useful

Chapter 10

Discussion and conclusions

10.1 Accomplishments

The aims of this research were (i) to determine whether existing applications of over-representation analysis (ORA), which are generally performed using GO terms or related controlled vocabularies as the associated annotation resource, can be extended to a wider mining of free-text; and (ii) to develop improved text mining approaches for identifying significantly over-represented terms or biological concepts within a list of differentially expressed genes by mining the associated literature information.

Initial exploration was based on a simple tokenisation of PubMed abstracts (Chapter 2), followed by the identification of over-represented tokens using the classical hypergeometric distribution (Chapter 3). When this approach was tested on 52 gene lists derived from microarray experiments using human arrays (Affymetrix HG-U133A chip), a dramatic and hitherto under-appreciated feature was observed, which is that gene lists generated from a typical microarray experiment tend to have more PubMed articles (as in PMID count) associated with them than equivalently-sized random gene lists. A similar trend was also seen for gene lists based on other popular model organisms such as mouse and rat (Section 4.1.1). These findings suggest that there is an excess of PMID annotation inherent with highly annotated gene lists. Further investigations into the potential causes of annotation bias found that there is a strong trend whereby those gene lists showing an excess of PMID annotation are also those whose constituent genes have been known for longer period of time (Section 4.3). It thus seems that gene lists generated from real-life biological experiments tends to favour groups of genes and areas of biology that have been studied for longer, and for which a greater amount of published literature are available. This may suggest that the (funded) research using microarrays to date has been rather conservative and largely focused on well-established areas of biology (such as immunology) where the genes that are likely to be differentially expressed have been known for a while. When then combined with the accumulated biological knowledge about these genes, the result is the annotation bias effect described above.

Annotation bias has a negative impact on ORA approaches that use the standard hypergeometric distribution to assess over-representation. Failing to account for such bias can lead to many common and apparently uninformative terms to be reported as significantly enriched (Section 3.3.1). This is because there is simply more text associated with the highly-annotated gene list than expected by chance, therefore even a relatively modest increase in frequency of a common word would produce a significant hypergeometric *p*-value. Experimentally-derived gene lists are generally associated with more GO terms than random gene lists as well (Section 4.1.2), suggesting that annotation bias may have a similar influence on other ORA-based functional enrichment tools that mine different annotation resources, such as GO terms or KEGG pathways. The issue of annotation bias has been raised by Blaschke *et al.* (2001), who stated that "It is not clear whether the over-representation of genes or abstracts constitutes a real problem, because they represent true biological or editorial biases; but certainly they bring a different type of information which should be taken into account during analysis". Similar concerns were also voiced by Khatri and Draghici (2005) and Krallinger *et al.* (2005). However, no solution has been proposed to address this problem in the context of enrichment analysis (be it using free-text or controlled vocabularies), and it has so far been overlooked by existing ORA tools.

Three different text-based ORA approaches have been developed during the course of this research to address the effect due to annotation bias. The first approach is based on the use of a permutation test (Chapter 5); this method produces biologically-plausible results and no longer considers those common and apparently uninformative terms as significant. However, the usefulness of this approach is hampered by being extremely computationally intensive, and therefore not suitable for routine analysis.

Two computationally tractable approaches were subsequently developed, which are based on the detection of outliers (OutlierDM) and the extended hypergeometric distribution (ExtendedHG). Analyses based on selected datasets showed that OutlierDM and ExtendedHG are able to identify tokens that are of biological relevance whilst compensating for the effect of annotation bias. Although they differ in the underlying statistical algorithms used to assess over-representation, the results produced by OutlierDM and ExtendedHG generally show a good concordance in most datasets that have been analysed (Chapters 6 and 7). In addition, the results are very similar to those generated by the permutation test-based method. These methods can be applied not only to well-studied species but also to less well-annotated species such as *Arabidopsis* (Section 8.3.2).

Due to the lack of a "gold standard" and appropriate evaluation metrics, it is not possible to formally evaluate the performance the approaches proposed herein against related ontology-based or text mining approaches. An alternative strategy has been adopted instead, which focused on the 52 literature gene lists using the HG-U133A array, and compared the outcome from OutlierDM and ExtendedHG with those obtained from a standard ORA approach that mines GO terms when applied to these gene lists. The biological relevance and plausibility of the over-represented tokens and GO terms were assessed against the perceived biology of the original publication. This focused performance review showed that the proposed approaches not only provide a similar but also distinct insight into the themes over-represented in a gene list compared to the results from undertaking ORA using GO terms (Section 8.2.2).

Several groups have undertaken the challenge of incorporating literature-based information into data mining algorithms to interpret the underlying biological significance of a list of differentially expressed genes (Blaschke *et al.* 2001; Chaussabel and Sher 2002; Frijters *et al.* 2008; Glenisson *et al.* 2004; Jelier *et al.* 2007; Jenssen *et al.* 2001; Shatkay *et al.* 2000). Their approaches are briefly reviewed in Section 1.4.4. The majority of these methods aim at finding functional associations between genes and terms, based on their co-occurrences in literature. The two methods that carry out text mining in a similar spirit to the approaches proposed here are GEISHA (Blaschke *et al.* 2001) and CoPub (Frijters *et al.* 2008). GEISHA evaluates the significance of terms associated with a gene cluster by comparing their frequency of abstracts with the frequency of abstracts containing these terms in different gene clusters. This system typically requires *more than one* gene cluster (gene list) to calculate the test statistic for testing over-representation, which is lacking in flexibility. Based on the work examples presented in Blaschke *et al.* (2001), it is clear that GEISHA is also susceptible to the effect of annotation bias. The CoPub system, on the other hand, calculates keyword enrichment for a list of differentially expressed genes using the Fisher's exact test. Since in this system CoPub maps the terms in abstracts to thesauri concepts, it is not apparent if its performance is affected by annotation bias as conventional tools implementing Fisher's exact test might.

The text-based ORA approaches developed in this project can be viewed as a complement to, and extension of, the existing functional enrichment tools. A web application PAKORA was implemented to provide an easily assessable interface to the proposed text-based ORA algorithms (Chapter 9). Taken together, the work presented here is a contribution to the more general need currently experienced in high-throughput genomics analysis, where data mining methods that access the literature efficiently and effectively are in demand. Inevitably, there are many unresolved issues and future possibilities raised by the work presented in this thesis. In the following sections, I will provide a general discussion on aspects where additional opportunities exist to improve their potential, and more general future research directions will also be suggested.

10.2 Limitations, open problems and future work

Conceptually, the text-based ORA mining framework proposed in this thesis can be viewed as having three major components: (i) an annotation database, which contains the textual information; (ii) statistical algorithms for assessing term over-representation; and (iii) output and result presentation. Each of these can affect the performance of the proposed methods and the comprehensiveness of the analytical results. Specific issues and possible extensions related to these components are discussed below.

10.2.1 Issues concerning the selection of PubMed articles

The methods described here depend on a corpus of articles relevant to the genes being studied (e.g. all genes appearing on an array), and an index that links the articles to the appropriate genes. Currently, the manually-curated citations provided by NCBI were used to retrieve the relevant gene-related PubMed abstracts. Although such curation provides high quality literature index, this process, together with the volume of research activity in different areas, means that the coverage of less heavily studied species is still limited, and this has a direct effect on the power of the proposed methods. Incorporation of additional gene-citation links, perhaps from species-specific databases, would increase the amount of textual information in the corpus and improve the power of the text mining methods proposed in this work. To improve the comprehensiveness of the results, the annotation database could be expanded to include textual information extracted from other specialised resources such as OMIM (Amberger *et al.* 2009), which provides a high level of details about genes and disease phenotypes. With full-text articles become increasingly accessible, future explorations should also tap into these resources. It has been reported that information density is highest in abstracts, but the information coverage in full texts is much greater than in abstracts (Schuemie *et al.* 2004). So it would be interesting to investigate if an expansion to mine full-text articles would add values or reduce the specificity of the proposed text-based methods.

As mentioned previously, articles that deal with large-scale sequencing, nomenclature, or protein family characterisation studies are typically associated with a large number of genes (Section 2.3.2). The inclusion of these articles in the text corpus is undesirable because their abstracts typically contain little explicit information about gene function. Tokens extracted from these articles are more likely to appear as significant simply because the articles are linked to most of the genes found in a given gene list (which is possible for articles such as sequencing reports, which are typically tagged with >1000 genes). To address this issue, the text corpus used in the current system was filtered to retain only those PMIDs that cross-reference to one EGID. While maximising the number of abstracts that deal specifically with the biological role of a given gene, this strategy inevitably leads to loss of information. For example, only 70% of all the articles indexed for the human genes were retained after filtering (Table 2.5). A more elaborate document retrieval approach that weighs the information content in an abstract according to the context could be adopted in future to improve the quality and quantity of the text corpus. For example, one might use Gene Reference Into Function (GeneRIF) (Mitchell *et al.* 2003) as the document selection criterion. GeneRIF is a curated resource that provides annotated links between PubMed articles containing functional information about the gene (or protein) and the corresponding Entrez Gene record. A GeneRIF entry is a short statement (up to 225 characters in length) summarising the function related to a specific gene as reported in the article. Therefore, by considering only articles that are supported by at least one GeneRIF entry, it is possible to identify a set of articles that contain functional information about genes. Further rounds of the document retrieval process could be performed by using those articles tagged with GeneRIFs as "seeds", followed by a search of the literature database for articles related or most relevant to the seeds. Document similarity measures described in Weiss *et al.* (2004), or the probabilistic algorithms proposed by Shatkay *et al.* (2000), could be explored in the future for their applicability to this task.

10.2.2 Issues concerning corpus manipulation and text processing

Much of the work presented in this thesis has focused on the development of statistical techniques for accurate identification of over-represented terms. So far, only very basic text processing such as simple tokenisation and stemming have been performed on the text in the corpus. There are two reasons to support the choice of performing as little by way of corpus manipulation at this stage. First, it enables the issue of annotation bias to be explored, and solutions to be developed, without the potentially confounding effect of a heavily pre-manipulated corpus. Since only minimal modifications have been done on the text in the corpus, any effects observed can then be confidently categorised as a real effect of annotation bias, and not some side-effect due to text manipulation. Second, it defines a baseline state for which to assess the improvements made by applying more elaborate natural language processing (NLP) on free-text.

There are several areas concerning text processing that could be made more sophisticated and complex in the future. These include the removal of stopwords and the use of name entity recognition techniques in combination with established thesauri (e.g. UMLS metathesaurus, MeSH and GO) to allow for the identification of multi-word biological concepts and synonyms, as well as to provide mapping to the same gene. These steps should reduce the noise caused by natural language variation and improve the information content of the over-represented tokens. On the other hand, popular term-weighting strategy such as the TF-IDF scheme (where the term frequency is multiplied by the inverse document frequency; see Weiss *et al.* 2004 for more details) could be used to weigh terms according to their relevance in the abstracts. This weight can also be used as a guide to remove common words with less semantic values from the corpus prior to conducting downstream statistical analysis. This will reduce the token space and minimise the problem of multiple hypothesis testing described in Section 3.2.5.

10.2.3 Issues concerning ORA's threshold-based strategy

Like other ORA approaches, the methods proposed in this thesis require an initial selection of differentially expressed genes by an arbitrarily chosen cutoff threshold. A major criticism to such "threshold-based" approach is that different choices of the cutoff value will produce different lists of differentially expressed genes and alter the result of the enrichment analysis. Moreover, many genes with moderate but meaningful expression changes may be discarded by the selected threshold regardless of their relative position in the ranked list, leading to a loss in statistical power. In recent years, an alternative mode of analysis that does not involve an initial gene selection step has been proposed. Examples of these include Gene Set Enrichment Analysis (GSEA) (Mootha *et al.* 2003; Subramanian *et al.* 2005; Tian *et al.* 2005) and Functional Class Scoring (FCS) (Pavlidis *et al.* 2004). These methods consider the distribution of a functionally-defined group of genes in the ranked list of genes and allow adjustments for their correlation structure. While a few studies have shown that such threshold-free approach enables the detection of more subtle functional categories that were overlooked by ORA (Ben-Shaul *et al.* 2005; Pavlidis *et al.* 2004), Manoli and coworkers (Manoli *et al.* 2006) found that ORA produced more consistent results than GSEA with respect to the concordance between analyses on differentially expressed genes obtained by different statistical methods from three prostate cancer data sets. Although it would be computationally challenging in scale, it may be possible to develop threshold-free methods that can accommodate annotation bias and thus be applied to the mining of PubMed tokens; works are currently underway to explore this question.

10.2.4 Issues concerning presentation of results

In many cases, the information contained in the over-represented terms can provide a first impression of the biological processes or themes related to a gene list. However, the interpretation of these significant tokens would become clearer when considered in the context of related terms and their associations with genes in the gene list. A simple illustration is given in Figure 10.1, where the 23 terms that were identified as significantly enriched in the ISG gene list by using OutlierDM are represented in the form of a dendrogram, in which the distances between terms are proportional to the number of genes the terms shared such that terms that have many genes in common are grouped together (the genes associated with each of the significant term are shown in Figure 10.2). We can see from the dendrogram that terms corresponding to interferon-inducible antiviral proteins such as 'OAS', 'oligoadenylate', and 'MxA' are close to each other, whereas terms related to antigen presentation by major histocompatibility complex (MHC) molecules such as 'HLA-A', 'HLA-B', 'HLA-G', form another cluster. MHC-dependent antigen presentation is another level of interferon action that augments the adaptive and acquired immune responses, which is different from the direct inhibition of viral replication by ISGs (such as MxA or 2'-5' OAS). By organising the output in this way, the relation between the significant terms becomes more specific and clearer.

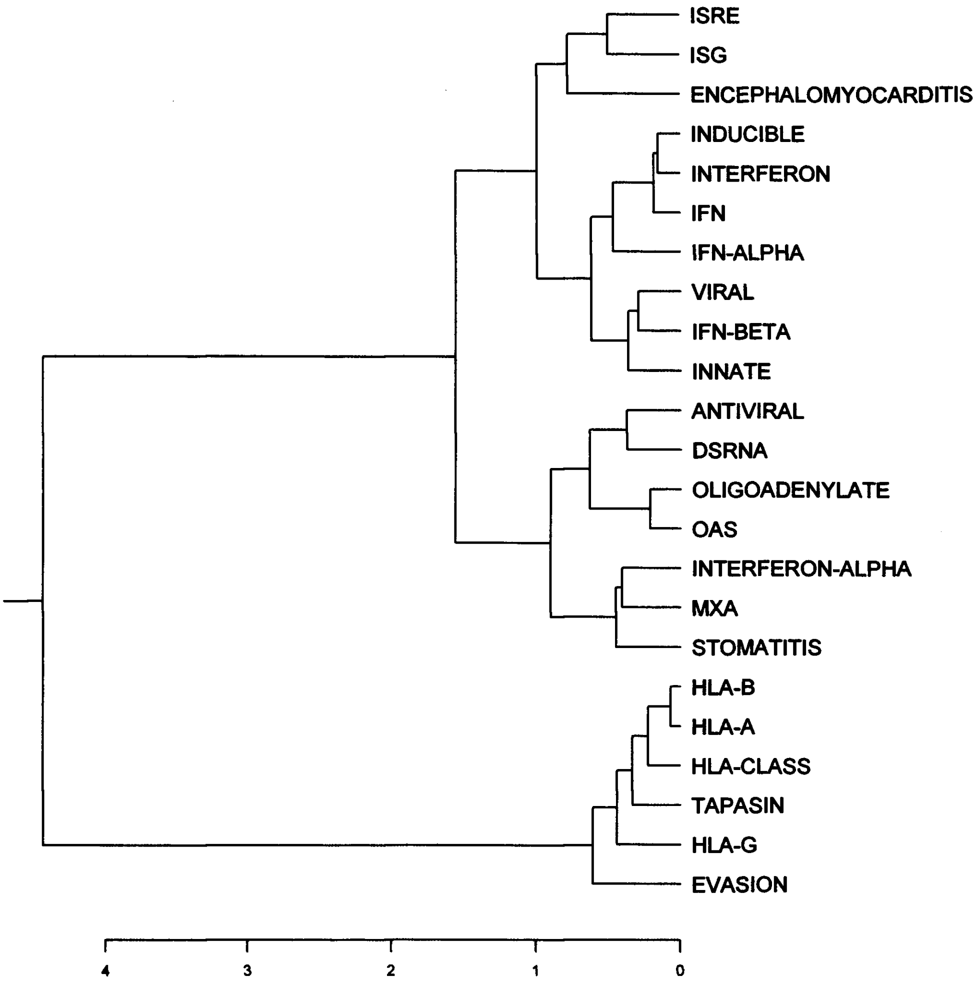


Figure 10.1: Dendrogram showing association between terms

The associations between 23 significantly over-represented terms in the ISG gene list as identified by using OutlierDM is represented as a dendrogram. The distances between the terms are proportional to the number of genes they have in common. Specifically, cosine similarity was calculated for each pair of terms based on the corresponding gene vectors.

	EVASION	HLA-G	TAPASIN	HLA-CLASS	HLA-A	HLA-B	STOMATITIS	MXA	INTERFERON-ALPHA	OAS	OLIGOADENYLATE	DSRNA	ANTIVIRAL	INNATE	IFN-BETA	VIRAL	IFN-ALPHA	IFN	INTERFERON	INDUCIBLE	ENCEPHALOMYOCARDITIS	ISG	ISRE
ADAR	0	0	0	0	0	0	1	0	1	0	0	1	1	1	0	1	1	1	1	1	0	0	0
B2M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1
C1R	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C1S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0
C3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
DDX58	1	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0
EIF2AK2	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1
GBP1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1
HERC5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	0	0	0
HLA-A	1	0	1	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0
HLA-B	1	0	1	1	1	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0
HLA-C	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
HLA-E	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0
HLA-F	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HLA-G	1	1	0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	1	1	1	0	0	0
HLA-J	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IFI16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	0	0	0
IFI27	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1	0	1	1
IFI35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
IFI44	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0
IFI6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
IFIH1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0
IFIT1	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	0
IFIT3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
IFIT5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
IFITM1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	1	1	1	0	1	1
IFITM3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
IRF7	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1
IRF9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	0
ISG15	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1
ISG20	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0
LGALS3BP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
MX1	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1
MX2	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	1	1	0	0	0
MYD88	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0
NMI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
OAS1	0	0	0	0	0	0	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	0	0
OAS2	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0	0	0
OAS3	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	0	0	0
OASL	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
OGFR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
PLSCR1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0
PML	1	0	0	0	0	0	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1
PSMB8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0
PSME2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
PYHIN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
SP100	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0
SP110	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	0
STAT1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
TAP1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
TREX1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0
TRIM21	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
TRIM22	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
TRIM34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
USP18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	0	0

Figure 10.2: A binary matrix of the significant tokens and genes

The columns of the matrix correspond to the significantly over-represented terms in the ISG gene list (based on OutlierDM), and the rows are genes associated with these term. The values 1 and 0 signify 'present' and 'absent', respectively.

To access contextual clues that go beyond the isolated meaning of individual terms, it is necessary to examine the text from which these terms are extracted. Methods that select sentences or abstracts containing the maximum concentration of significant terms have been found to be particularly useful in this respect and facilitate biological interpretation by human experts (Blaschke *et al.* 2001). On the other hand, literature profiling approaches that clustered genes based on keyword (or significant term) co-occurrence in abstracts have also shown promise in guiding the interpretation of large and heterogeneous gene lists (Alako *et al.* 2005; Chaussabel and Sher 2002; Jelier *et al.* 2007), and could be evaluated in future work.

10.2.5 Other considerations

In addition to the annotation data and statistical algorithms used, another factor that can influence the performance of the proposed text-based approaches is the quality of the gene list. Intuitively, the proposed methods will perform better on gene lists containing a notable portion of up- or down-regulated genes that are participated in certain interesting biological processes than on less specific gene lists whose constituent genes are spread throughout all possible biological processes. Other factors such as the size of gene list, the amount of annotation associated with the genes and the dependency among genes will also affect the sensitivity and specificity of the proposed approaches. For this reason, the enrichment *p*-values are not directly comparable across gene lists, and should only be treated as a scoring system that plays an advisory role rather than decision-making role (Huang *et al.* 2008). Functional enrichment analysis such as that presented herein is therefore considered more of an exploratory procedure rather than a pure statistical solution.

While the peer-reviewed scientific literature will continue to be the prime resource for accessing biological knowledge, in the future it may become increasingly necessary to integrate other sources of information (such as protein interactions, pathways information and annotated data from previous gene expression analysis) to provide a more comprehensive data-mining environment, because results jointly learnt from more than one type of data are likely to produce insights that might not be apparent from mining one type of data in isolation. The construction of a gold standard and standard evaluation procedure will facilitate the growth and development of the field.

10.3 Conclusions

To conclude, the aims and objectives of this research have been broadly achieved. I have described the problems and challenges associated with existing ORA methods when adapting them for mining text-based information, and three novel approaches have been proposed to address some of these problems. Analysis performed on several independent datasets show that the proposed methods produce biologically-meaningful results that are in good agreement with the manually determined annotations. These examples also demonstrate that a coherent picture that exists within complex group of genes can be discerned by incorporating textual information embedded in literature as a knowledge source into the analysis of gene expression data. Collectively, the text-based ORA approaches presented in this thesis can be used to complement and extend existing ontology-based functional analysis tools for guiding the biological interpretation of complex microarray data.

Appendix A

Details of literature gene lists

402 different gene lists were collected from 170 scientific papers for testing the performance of the text-based ORA methods presented in this thesis. These gene lists were derived from microarray experiments performed on 10 major Affymetrix platforms, including HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Celegans, Xenopus laevis and Zebrafish arrays. The gene identifiers in these gene lists are in the form of Affymetrix probeset IDs; these gene lists are included on the CD-ROM attached to this thesis. Details of each of these gene lists, including its size and the PubMed article from which it was extracted, are detailed as follow.

Table A.1: Literature gene lists based on the HG-U133A array

List ID	PubMed ID	First author	Year	# Probe sets	Description
hs1a	16531451	Provenzani	2006	867	Probesets differentially expressed in the total RNA sample. Extracted from Supplementary Table I.
hs1b	16531451	Provenzani	2006	2018	Probesets changed in the polysomal RNA sample. Extracted from Supplementary Table II.
hs1c	16531451	Provenzani	2006	397	Probesets common to polysomal and total RNA sample. Extracted from Supplementary Table III.
hs2a	15210650	Lee	2004	269	Transcripts with greater than 3 fold enrichment in every T cell subpopulation compared to TSC. Extracted from Supplementary Table 1.
hs2b	15210650	Lee	2004	1521	Transcripts whose expression changed by more than 3 fold during T cell differentiation. Extracted from Supplementary Table 2.
hs2c	15210650	Lee	2004	134	Transcripts enriched in both ITTP and DP by more than 3 fold. Extracted from Supplementary Table 3-1.
hs2d	15210650	Lee	2004	76	Transcripts enriched in more mature cells (SP4, CB4, and AB4) by more than 3 fold. Extracted from Supplementary Table 3-2.
hs2e	15210650	Lee	2004	28	Transcripts enriched by more than 3 fold in ITTP compared to other lymphocytes. Extracted from Supplementary Table 3-3.
hs2f	15210650	Lee	2004	28	Transcripts enriched by more than 3 fold in DP compared to other lymphocytes. Extracted from Supplementary Table 3-4.
hs2g	15210650	Lee	2004	16	Transcripts enriched by more than 3 fold in SP4 compared to other lymphocytes. Extracted from Supplementary Table 3-5.
hs2h	15210650	Lee	2004	25	Transcripts enriched in naïve CD4 T cells (CB4, and AB4) by more than 3 fold. Extracted from Supplementary Table 3-6.
hs2i	15210650	Lee	2004	32	Transcripts showing SP4>CB4>AB4 pattern. Extracted from Supplementary Table 4-1.
hs2j	15210650	Lee	2004	240	Transcripts showing more than 2 fold higher expression in CB4 than in AB4. Extracted from Supplementary Table 4-2.
hs3a	15897907	Farmer	2005	400	Genes which best discriminate apocrine vs. luminal (AL). Extracted from Supplementary Table Sheet 2.
hs3b	15897907	Farmer	2005	400	Genes which best discriminate porcine vs. basal (AB). Extracted from Supplementary Table Sheet 2.
hs3c	15897907	Farmer	2005	400	Genes which best discriminate basal vs. luminal (BL). Extracted from Supplementary Table Sheet 2.
hs4a	16260967	Radom-Aizik	2005	181	Genes whose expression increased after training. Extracted from Table S2 in the main paper.
hs4b	16260967	Radom-Aizik	2005	216	Genes whose expression decreased after training. Extracted from Table S3 in the main paper.
hs5a	12958056	Gimino	2003	135	Genes that are up-regulated in gene expression in acute rejection vs. no rejection. Extracted from Supplementary Table E1.
hs5b	12958056	Gimino	2003	858	Genes with significant changes in gene expression in acute rejection vs. no rejection. Extracted from Supplementary Table E2.

Table A.1: Literature gene lists based on the HG-U133A array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
hs6a	16319128	Vanharanta	2006	181	Down-regulated genes in FH mutant relative to FH wild-type fibroids. Extracted from Supplementary Table 1.
hs6b	16319128	Vanharanta	2006	179	Up-regulated genes in FH mutant relative to FH wild-type fibroids. Extracted from Supplementary Table 1.
hs6c	16319128	Vanharanta	2006	228	Down-regulated genes in FH mutant relative to normal myometrium. Extracted from Supplementary Table 3.
hs6d	16319128	Vanharanta	2006	199	Up-regulated genes in FH mutant relative to normal myometrium. Extracted from Supplementary Table 3.
hs7	15817885	Barth	2005	1434	Genes differentially expressed in atrial fibrillation. Extracted from Supplementary Table 3.
hs8	15971941	Barberi	2005	412	Genes shared between primary and hESC-derived mesenchymal precursors but significantly different from undifferentiated hESCs. Extracted from Supplementary Table S2.
hs9a	15558013	O'Donnell	2005	26	Genes down-regulated in N+ primary tumours. Extracted from Supplementary Figure 1.
hs9b	15558013	O'Donnell	2005	93	Genes up-regulated in N+ primary tumours. Extracted from Supplementary Figure 1.
hs10	16203770	Best	2005	256	Unabridged list of genes differentially expressed between AD and AI prostate cancer. Extracted from Supplementary Data.
hs11a	16049480	Minn	2005	113	Genes differentially expressed between parental MDA-MB-231 and LM2 cell lines selected to be highly metastatic to lung. Extracted from Supplementary Table 2.
hs11b	16049480	Minn	2005	65	Lung metastasis candidate genes. Extracted from Supplementary Table 4.
hs12a	16089502	Eckfeldt	2005	2707	Probesets differentially expressed between umbilical cord blood derived Rho-lo and Rho-hi cells. Extracted from Supplementary Table S1.
hs12b	16089502	Eckfeldt	2005	4677	Probesets differentially expressed between adult bone marrow derived Rho-lo and Rho-hi cells. Extracted from Supplementary Table S2.
hs12c	16089502	Eckfeldt	2005	304	Probesets differentially expressed between Rho-lo and Rho-hi cells from both umbilical cord blood and adult bone marrow. Extracted from Supplementary Table S3.
hu13	16205643	Liu	2006	235	Genes induced by c-Myb and v-Myb in MCF-7 cells. Extracted from Supplementary Table 1.
hs14a	14872006	Hall	2004	299	Genes differentially expressed in 19 paired human samples comparing pre and post mechanical unloading with a LVAD. Extracted from Supplementary Table S1.
hs14b	14872006	Hall	2004	144	Genes differentially expressed in non-ischemic cohort. Extracted from Supplementary Table S3.
hs14c	14872006	Hall	2004	97	Genes differentially expressed in ischemic cohort. Extracted from Supplementary Table 4.
hs14d	14872006	Hall	2004	22	Genes differentially expressed in acute MI cohort. Extracted from Supplementary Table 5.

Table A.1: Literature gene lists based on the HG-U133A array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
hs15a	12756304	Cousins	2003	104	Group 1 zinc responsive genes. Extracted from Supplementary Table 3.
hs15b	12756304	Cousins	2003	86	Group 4 zinc responsive genes. Extracted from Supplementary Table 4.
hs16	16116475	Chow	2006	57	Genes differentially expressed in C666-1 RASSF1A-transfected clones. Extracted from Table 1 in the original paper.
hs17a	16804116	Chng	2006	73	Genes that displayed distinct expression profile in WM compared to CLL and MM. Extracted from Supplementary Table S1: WM Unique Genes.
hs17b	16804116	Chng	2006	1247	Genes that displayed distinct expression profile in MM compared to CLL and WM. Extracted from Supplementary Table S1: MM Unique Genes
hs17c	16804116	Chng	2006	396	Genes that displayed distinct expression profile in CLL compared to WM and MM. Extracted from Supplementary Table S1: CLL Unique Genes.
hs17d	16804116	Chng	2006	314	A cluster of genes that were over-expressed in B-cell, WM and CLL. Extracted from Supplementary Table S1: WM CLL B-cell cluster.
hs18a	16836768	Pfoertner	2006	46	Up-regulated genes comparing CD4+CD25+ T cells vs. CD4+CD25- T cells. Extracted from Additional file 1.
hs18b	16836768	Pfoertner	2006	21	Down-regulated genes comparing CD4+CD25+ T cells vs. CD4+CD25- T cells. Extracted from Additional file 1.
hs18c	16836768	Pfoertner	2006	313	Genes differentially expressed in Foxp3 over-expressing CD4+ Th cell lines cells relative to the GFP transduced CD4+ Th controls. Extracted from Additional file 4.
hs19a	15869706	Rosty	2005	262	Up-regulated genes in FGFR3 mutated tumours relative to FGFR3 wild-type tumours. Extracted from Additional file 2: Positive Significant Genes.
hs19b	15869706	Rosty	2005	552	Down-regulated genes in FGFR3 mutated tumours relative to FGFR3 wild-type tumours. Extracted from Additional file 2: Negative Significant Genes
hs20	15604246	Berger	2004	1207	Genes differentially expressed between LHSR and LHS. Extracted from Supplementary Table 1.

Table A.2: Literature gene lists based on the HG-U133 Plus 2.0 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
hu1	15361855	Donninger	2004	1191	Differentially regulated genes identified in advanced papillary serous tumour specimens. Extracted from Supplementary Table S1.
hu2	16205643	Liu	2006	1527	Genes induced by c-Myb and v-Myb in human monocytes. Extracted from Supplementary Table 3.
hu3a	16982809	Szatmari	2006	81	Genes regulated by PPAR-gamma agonist, RAR-alpha agonists and the RAR-alpha antagonist (referred to as cluster 3 in the main text). Extracted from Supplementary Table S1.
hu3b	16982809	Szatmari	2006	72	Genes down-regulated by PPAR-gamma ligand (referred to as cluster 6 in the main text). Extracted from Supplementary Table S2.
hu4a	16926187	Vendelin	2006	195	Genes differentially expressed in contrasts 1 and 2 (NPS stimulated vs. unstimulated NPSR1-A cells, and NPS stimulated NPSR1-A vs. NPS stimulated HEK-293H, respectively). Extracted from Supplementary Table S1.
hu4b	16926187	Vendelin	2006	43	Co-regulated genes. Extracted from Supplementary Table 3.
hu5a	16210406	Jaatinen	2006	690	Differentially expressed genes in CD133+ relative to CD133–. Extracted from Supplementary Table 2.
hu5b	16210406	Jaatinen	2006	257	Genes expressed only in CD133+ and absent in CD133- samples. Extracted from Supplementary Table 3.
hu6	16242812	Dunckley	2006	225	Genes whose expression differed significantly when comparing ND non-NFT neurons to AD non-NFT neurons and then to AD NFT neurons. Data downloaded from http://www.tgen.org/research/index.cfm?pageid=502
hu7	16949412	Hassan	2006	50	Top 50 probesets that displayed differential expression between TL samples and TNL samples. Extracted from Table II in the main paper.
hu8a	16772347	Lampron	2006	724	Probesets related to GIP-dependent AIMAH. Extracted from Supplementary Table 2.
hu8b	16772347	Lampron	2006	94	Probesets with intensity levels linked to the presence of a GIP-dependent nodule. Extracted from Supplementary Table 4.
hu9a	16728703	Zhan	2006	51	Top 50 over-expressed genes unique to PR subgroup. Extracted from Supplementary Table S2.
hu9b	16728703	Zhan	2006	56	Top 50 over-expressed genes unique to LB subgroup. Extracted from Supplementary Table S2.
hu9c	16728703	Zhan	2006	52	Top 50 over-expressed genes unique to MS subgroup. Extracted from Supplementary Table S2.
hu9d	16728703	Zhan	2006	53	Top 50 over-expressed genes unique to HY subgroup. Extracted from Supplementary Table S2.
hu9e	16728703	Zhan	2006	53	Top 50 over-expressed genes unique to CD-1 subgroup. Extracted from Supplementary Table S2.
hu9f	16728703	Zhan	2006	57	Top 50 over-expressed genes unique to CD-2 subgroup. Extracted from Supplementary Table S2.
hu9g	16728703	Zhan	2006	52	Top 50 over-expressed genes unique to MF subgroup. Extracted from Supplementary Table S2.

Table A.2: Literature gene lists based on the HG-U133 Plus 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
hu9h	16728703	Zhan	2006	52	Top 50 under-expressed genes unique to PR subgroup. Extracted from Supplementary Table S3.
hu9i	16728703	Zhan	2006	58	Top 50 under-expressed genes unique to LB subgroup. Extracted from Supplementary Table S3.
hu9j	16728703	Zhan	2006	55	Top 50 under-expressed genes unique to MS subgroup. Extracted from Supplementary Table S3.
hu9k	16728703	Zhan	2006	54	Top 50 under-expressed genes unique to HY subgroup. Extracted from Supplementary Table S3.
hu9l	16728703	Zhan	2006	55	Top 50 under-expressed genes unique to CD-1 subgroup. Extracted from Supplementary Table S3.
hu9m	16728703	Zhan	2006	54	Top 50 under-expressed genes unique to CD-2 subgroup. Extracted from Supplementary Table S3.
hu9n	16728703	Zhan	2006	58	Top 50 under-expressed genes unique to MF subgroup. Extracted from Supplementary Table S3.
hu9o	16728703	Zhan	2006	172	Genes commonly dysregulated in CD-1 and CD-2 groups. Extracted from Supplementary Table S5.
hu9p	16728703	Zhan	2006	131	Genes that were differentially expression between CD-1 and CD-2 subgroups. Extracted from Supplementary Table S6.
hu10	16670265	Radmacher	2006	157	Bullinger Validation Signature that separated AML from normal karyotype. Extracted from Supplementary Table S1.
hu11	16953664	Maier	2006	70	Genes differentially expressed in response to retrovirally mediated MDR1 over-expression. Extracted from Table 1 in the original paper.
hu12a	16638148	Oudes	2006	4176	Genes detected in sorted cells but not in the whole tissue. Extracted from Supplementary Table 2.
hu12b	16638148	Oudes	2006	239	Genes detected in whole prostate but not in sorted cells. Extracted from Supplementary Table 4.
hu12c	16638148	Oudes	2006	197	Genes detected only in prostate luminal cells. Extracted from Supplementary Table 5.
hu12d	16638148	Oudes	2006	150	Genes detected only in prostate basal cells. Extracted from Supplementary Table 6.
hu12e	16638148	Oudes	2006	632	Genes detected only in prostate stromal cells. Extracted from Supplementary Table 7.
hu13	16863911	Fruehauf	2006	123	Genes differentially expressed in AMD3100 + G-CSF-mobilized PBPC compared to G-CSF-mobilized cells. Extracted from Table 1 in the original paper.
hu14	16474848	Hooi	2006	361	Genes present 2 fold expression changes by ST7-1a expression in PC-3 cells. Extracted from Supplementary table.
hu15a	16682435	Peddada	2006	183	Genes significantly increased upon differentiation. Extracted from Supplementary Table 1.
hu15b	16682435	Peddada	2006	12	Genes significantly increased upon differentiation and significantly increased in D-UT vs. D-MT comparison. Extracted from Supplementary Table 2.
hu15c	16682435	Peddada	2006	8	Genes significantly increased upon differentiation and significantly decreased in D-UT vs. D-MT comparison. Extracted from Supplementary Table 3.

Table A.2: Literature gene lists based on the HG-U133 Plus 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
hu15d	16682435	Peddada	2006	45	Genes significantly decreased upon differentiation. Extracted from Supplementary Table 4.
hu15e	16682435	Peddada	2006	4	Genes significantly decreased upon differentiation and significantly increased in D-UT vs. D-MT comparison. Extracted from Supplementary Table 5.
hu15f	16682435	Peddada	2006	26	Genes significantly changed upon differentiation in D-MD vs. D-CD comparison. Extracted from Supplementary Table 6.
hu16	16507782	Mense	2006	133	Hypoxia-regulated genes in both astrocytes and HeLa cells. Extracted from Table S1.
hu17	16785517	Fulcher	2006	1521	Genes differentially expressed between LPS and galectin-1-treated MDDCs. Extracted from Supplementary Table 1.
hu18	16379004	DeFilippis	2006	859	Differentially regulated genes between HCMV-infected versus uninfected human fibroblasts treated with control siRNA at both 4h and 8h post-infection. Extracted from Supplementary Table 1.
hu19a	16288205	Charafe-Jauffret	2006	1233	Probesets significantly differentially expressed between luminal cell lines and basal cell lines. Supplementary. Extracted from Supplementary Table 2.
hu19b	16288205	Charafe-Jauffret	2006	1309	Probesets significantly differentially expressed between luminal cell lines and mesenchymal cell lines. Extracted from Supplementary Table 3.
hu19c	16288205	Charafe-Jauffret	2006	227	Probesets significantly differentially expressed between basal cell lines and mesenchymal cell lines. Extracted from Supplementary Table 4.
hu20a	16644866	Armstrong	2006	205	hESC-specific transcripts. Extracted from Supplementary Table S1.
hu20b	16644866	Armstrong	2006	84	Highly expressed transcripts in both hES-NCL1 and H1 cell lines. Extracted from Supplementary Table S2.
hu20c	16644866	Armstrong	2006	61	Genes which were unique to hES-NCL1. Extracted from Supplementary Table S3.
hu20d	16644866	Armstrong	2006	49	Genes which were unique to H1. Extracted from Supplementary Table S3.
hu20e	16644866	Armstrong	2006	110	New hESC markers. Extracted from Supplementary Table S4.

Table A.3: Literature gene lists based on the MG-U430 2.0 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
ms1	16949565	Yu	2006	294	Significantly differentially expressed genes in Get-1-/- back skin relative to wild-type. Extracted from Supplementary Table 4.
ms2a	16860309	Potireddy	2006	490	List of transcripts enriched at MII stage. Extracted from Supplementary Table 2.
ms2b	16860309	Potireddy	2006	1808	Genes enhanced in the late one-cell embryos. Extracted from Supplementary Table 3.
ms3a	16368875	Mitchell	2006	185	Genes differentially expressed in HRas-v12 hearts (RAS-regulated). Extracted from Supplementary Table 2.
ms3b	16368875	Mitchell	2006	251	Genes differentially expressed in MKK3bE hearts (p38-regulated). Extracted from Supplementary Table 2.
ms3c	16368875	Mitchell	2006	257	Genes differentially expressed in MKK7D hearts (JNK-regulated). Extracted from Supplementary Table 2.
ms4a	16940520	Kunz	2006	98	Genes whose expression is consistently changed between LCMV-cgPi and mock infected animals. Extracted from Supplementary Table S1.
ms4b	16940520	Kunz	2006	77	Genes whose expression is significantly changed between the two mock animals. Extracted from Supplementary Table S2.
ms5	16943279	Lindsley	2006	1583	Genes whose expression is Wnt dependent during ES cell differentiation. Extracted from Supplementary Table 1.
ms6a	16926388	Reece	2006	123	Up-regulated genes in both WT and SCID mice by DPI. Extracted from Supplementary Table S2.
ms6b	16926388	Reece	2006	220	Genes that were differentially expressed in the lungs of WT and SCID mice at days 2, 4, and 12 post-Nippostrongylus brasiliensis infection. Extracted from Supplementary Table S3.
ms7a	16926395	He	2006	1138	Significantly down-regulated probesets at 4h post B. melitensis infection. Extracted from Supplementary Table 1.
ms7b	16926395	He	2006	288	Significantly up-regulated probesets at 4h post B. melitensis infection. Extracted from Supplementary Table 2.
ms8	16772024	Korostynski	2006	1528	Probesets differentially expressed among the four inbred strains of mice. Extracted from Supplementary Table S2.
ms9a	16945109	Heinitz	2006	75	Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with Abeta(1-42). Extracted from Table 1 in the main paper.
ms9b	16945109	Heinitz	2006	14	Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with H2O2. Extracted from Table 2 in the main paper.
ms9c	16945109	Heinitz	2006	15	Differentially expressed genes in differentiated SN56.B5.G4 cells after treatment with Abeta(1-42) and H2O2. Extracted from Table 3 in the main paper.
ms10a	16105979	Kawagoe	2005	36	Genes up-regulated by more than 5 fold in MN1-TEL+/HOXA9+ AML cells in mice. Extracted from Supplementary Table S1.

Table A.3: Literature gene lists based on the MG-U430 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ms10b	16105979	Kawagoe	2005	32	Genes down-regulated by more than 5 fold in MN1-TEL+/HOXA9+ AML cells in mice. Extracted from Supplementary Table S2.
ms11a	17227850	Jankovic	2007	388	CMP-associated transcript. Extracted from Supplementary Dataset 1.
ms11b	17227850	Jankovic	2007	2038	HSC-associated transcript. Extracted from Supplementary Dataset 2.
ms12a	17183314	Niedernhofer	2006	1674	Genes differentially expressed in Ercc1-/- compared with wild-type liver. Extracted from Supplementary Table III.
ms12b	17183314	Niedernhofer	2006	1665	Genes differentially expressed in aged (130-wk-old) wild-type mouse liver compared to young (2-mo-old) controls. Extracted from Supplementary Table V.
ms12c	17183314	Niedernhofer	2006	1773	Genes differentially expressed in adult (4-mo-old) wild-type mouse livers compared to young (2-mo-old) controls. Extracted from Supplementary Table VI.
ms13a	15967997	Rossi	2005	595	Age-up-regulated genes in LT-HSC. Extracted from Supplementary Table 4.
ms13b	15967997	Rossi	2005	383	Age-down-regulated genes in LT-HSC. Extracted from Supplementary Table 4.
ms14a	16110338	Corbo	2005	252	Genes up-regulated in rd7 mutant retina at P21. Extracted from Supplementary Table S4.
ms14b	16110338	Corbo	2005	138	Genes down-regulated in rd7 mutant retina at P21. Extracted from Supplementary Table S5.
ms15a	16373508	Shiina	2005	767	Down-regulated genes in 8-week-old AR-/- ovaries. Extracted from Supplementary Table 1.
ms15b	16373508	Shiina	2005	346	Up-regulated genes in 8-week-old AR-/- ovaries. Extracted from Supplementary Table 2.
ms15c	16373508	Shiina	2005	323	Down-regulated genes in 3-week-old AR-/- ovaries. Extracted from Supplementary Table 3.
ms15d	16373508	Shiina	2005	516	Up-regulated genes in 3-week-old AR-/- ovaries. Extracted from Supplementary Table 4.
ms16a	16399799	Beverdam	2006	266	Genes specifically up-regulated in 11.5 dpc male somatic cells. Extracted from Supplementary Table S1.
ms16b	16399799	Beverdam	2006	50	Genes specifically down-regulated in 11.5 dpc male somatic gonad cells. Extracted from Supplementary Table S3.
ms16c	16399799	Beverdam	2006	243	Genes specifically up-regulated in 11.5 dpc female somatic cells. Extracted from Supplementary Table S5.
ms17a	16166195	Denolet	2005	509	Transcripts down-regulated in SCARKO as compared to control mice at day 10. Extracted from Supplementary Table 1.
ms17b	16166195	Denolet	2005	342	Transcripts up-regulated in SCARKO as compared to control mice at day 10. Extracted from Supplementary Table 1.
ms18	16423883	Jeong	2006	492	Genes differentially expressed in the liver of SRC-2-/- mice as compared to wild-type mice. Extracted from Supplementary Table 2.

Table A.3: Literature gene lists based on the MG-U430 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ms19	16641092	Costinean	2006	429	Genes differentially expressed in transgenic relative to wild-type mice. Extracted from Supplementary Table 3.
ms20	16847333	Sun	2006	248	Genes differentially expressed in day 5 postpartum Mkl1 KO vs. wild-type mammary glands. Extracted from Supplementary Table S1.

Table A.4: Literature gene lists based on the RAT230 2.0 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
rat1a	16854511	Mainwaring	2006	1215	Genes with altered expression post paraquat dosing compared to control lung. Extracted from Supplementary Table 2.
rat1b	16854511	Mainwaring	2006	543	This is a subset of genes in rat1a, selected based on a more stringent statistical test. Extracted from Supplementary Table 3.
rat2	16770850	Bardag-Gorce	2006	95	Differentially expressed genes between the peaks of UAL (group 3) and controls (group 1). Extracted from Table 2 in the main paper.
rat3	16574311	Chen	2006	257	Genes regulated by PPARγ in RIE cells. Extracted from Supplementary data.
rat4a	16202214	Jaster	2005	22	Genes up-regulated in PPARγ1-overexpressing LTC cells. Extracted from Table 1A in the main paper.
rat4b	16202214	Jaster	2005	19	Genes down-regulated in PPARγ1-overexpressing LTC cells. Extracted from Table 1B in the main paper.
rat5a	16715494	Ko	2006	18	Genes up-regulated by haloperidol and clozapine. Extracted from Table IA in the main paper.
rat5b	16715494	Ko	2006	14	Genes down-regulated by haloperidol and clozapine. Extracted from Table IB in the main paper.
rat5c	16715494	Ko	2006	17	Genes up-regulated by amphetamine. Extracted from Table IIA in the main paper.
rat5d	16715494	Ko	2006	5	Genes down-regulated by amphetamine. Extracted from Table IIB in the main paper.
rat5e	16715494	Ko	2006	15	Genes up-regulated by amphetamine, but down-regulated by haloperidol and clozapine. Extracted from Table IIIA in the main paper.
rat5f	16715494	Ko	2006	6	Genes down-regulated by amphetamine, but up-regulated by haloperidol and clozapine. Extracted from Table IIIB in the main paper.
rat5g	16715494	Ko	2006	23	Genes up-regulated by haloperidol, clozapine, and amphetamine. Extracted from Table IVA in the main paper.
rat5h	16715494	Ko	2006	12	Genes down-regulated by haloperidol, clozapine, and amphetamine. Extracted from Table IVB in the main paper.
rat6a	16809437	Lahousse	2006	1777	Genes that displayed a significant expression change within 12 h of MEHP exposure. Extracted from Supplementary Table 1.
rat6b	16809437	Lahousse	2006	536	Genes altered significantly at 1, 3, 6, or 12 h following fetal DBP exposure. Extracted from Supplementary Table 2.
rat6c	16809437	Lahousse	2006	176	Genes showed altered expression at fetal and prepubertal ages. Extracted from Supplementary Table 3.
rat7a	17082646	Geoffrey	2006	2410	Genes over-represented in DRlyp/lyp. Extracted from Supplementary Table A.
rat7b	17082646	Geoffrey	2006	2660	Gene over-represented in DR+/+ . Extracted from Supplementary Table B.

Table A.4: Literature gene lists based on the RAT230 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
rat8a	16401727	Luyendyk	2006	315	Genes differentially expressed in LPS/Veh/RAN-treated rats relative to LPS/Veh/FAM-treated rats. Extracted from Supplementary Table 1.
rat8b	16401727	Luyendyk	2006	145	Heparin-responsive probesets in LPS/RAN-treated rats. Extracted from Supplementary Table 2.
rat8c	16401727	Luyendyk	2006	29	Subset AB: Genes differentially expressed in LPS/Veh/RAN-treated rats relative to LPS/Veh/FAM-treated rats, and also by heparin. Extracted from Supplementary Table 3.
rat9a	17069981	Coyle	2007	113	Genes significantly up-regulated in Allodynic rats. Extracted from Supplementary Table 3.
rat9b	17069981	Coyle	2007	20	Genes significantly down-regulated in Allodynic rats. Extracted from Supplementary Table 3.
rat10	17187413	Yovchev	2007	30	Genes up-regulated in the oval cell enriched fractions. Extracted from Table 1 in the main paper.
rat11a	17332525	Xia	2007	436	Transcripts with more than 2 fold changes in their expression by day 4 after Adjudin treatment. Extracted from Supplementary Table 1.
rat11b	17332525	Xia	2007	1466	Transcripts that were significantly altered following Adjudin treatment (P < 0.05 by ANOVA). Extracted from Supplementary Table 2.
rat12a	18158353	Frank	2008	164	Genes significantly up-regulated by biaxial stretch in neonatal rat ventricular cardiomyocytes. Extracted from Supplementary Table S2.
rat12b	18158353	Frank	2008	21	Genes significantly down-regulated by biaxial stretch in neonatal rat ventricular cardiomyocytes. Extracted from Supplementary Table S2.
rat12c	18158353	Frank	2008	238	Genes significantly up-regulated by phenylephrine in neonatal rat ventricular cardiomyocytes. Extracted from Supplementary Table S3.
rat12d	18158353	Frank	2008	211	Genes significantly down-regulated by phenylephrine in neonatal rat ventricular cardiomyocytes. Extracted from Supplementary Table S3.
rat13a	18366630	Yukhananov	2008	512	Genes differentially expressed with p < 0.01 at 24h following injection of carrageenan (24h vs. control). Extracted from Additional file 1.
rat13b	18366630	Yukhananov	2008	629	Genes differentially expressed with p < 0.01 at 28d following injection of carrageenan (28d vs. control). Extracted from Additional file 1.
rat14a	17540011	Brouillette	2007	42	Genes differentially expressed between trained and naive saline treated rats. Extracted from Supplemental Table S1.
rat14b	17540011	Brouillette	2007	32	Genes differentially expressed between scopolamine and naive saline treated rats. Extracted from Supplemental Table S2.
rat15a	18436381	Pedersen	2008	663	Genes affected by traumatic brain injury (TBI) only. Extracted from Supplementary data file 2.
rat15b	18436381	Pedersen	2008	82	Genes affected by both FGL treatment and TBI. Extracted from Supplementary data file 3.

Table A.4: Literature gene lists based on the RAT230 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
rat16a	18095365	Dihal	2008	42	Genes significantly up-regulated by quercetin in the distal colon mucosa of rats. Extracted from Supplementary Table 1.
rat16b	18095365	Dihal	2008	165	Genes significantly down-regulated by quercetin in the distal colon mucosa of rats. Extracted from Supplementary Table 1.
rat17	18355885	Hirode	2008	78	Genes significantly differentially expressed in response to 5 drugs that induce hepatic phospholipidosis in rats (including amiodarone, amitriptyline, clomipramine, imipramine, and ketoconazole). Extracted from Table 3 in the main paper.
rat18a	18405950	Rodd	2008	220	Genes that were differentially expressed in the nucleus accumbens (ACB) of iP rats between the ethanol and water groups. Extracted from Supplemental Table A.
rat18b	18405950	Rodd	2008	253	Genes that were differentially expressed in the nucleus accumbens (ACB) of iP rats between the ethanol and saccharin groups. Extracted from Supplemental Table B.
rat19a	17693601	Schmidt-Ott	2007	854	Probesets that were significantly up-regulated during epithelial differentiation when compared with freshly isolated metanephric mesenchyme (baseline). Extracted from Supplemental Table S1.
rat19b	17693601	Schmidt-Ott	2007	552	Probesets that were significantly down-regulated during epithelial differentiation when compared with freshly isolated metanephric mesenchyme (baseline). Extracted from Supplemental Table S2.
rat20	17033635	Conti	2007	272	Genes affected in two out of three antidepressant treatments: deprivation (SD), and fluoxetine (FLX). Extracted from Table 4 of the main paper.

Table A.5: Literature gene lists based on the Ath1 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
ath1	12920300	Nishimura	2003	685	Genes response to pathogen attack. Extracted from Supplementary Table S2.
ath2a	16243906	Vanneste	2005	3110	Differentially expressed genes in the wild type versus the slr1 mutant (denoted as the 3110 significant genes in the main text). Extracted from Supplementary Table 1.
ath2b	16243906	Vanneste	2005	913	Lateral root initiation genes (denoted as the 913 LRI genes in the main text). Extracted from Supplementary Table 2.
ath2c	16243906	Vanneste	2005	99	LRI genes with arf7 arf19–dependent auxin inducibility. Extracted from Supplementary Table 6.
ath3a	16299223	Bläsing	2005	222	3h glucose-repressed genes. Extracted from Supplementary Table 6.
ath3b	16299223	Bläsing	2005	218	3h glucose-induced genes. Extracted from Supplementary Table 6.
ath3c	16299223	Bläsing	2005	223	3h sucrose-repressed genes. Extracted from Supplementary Table 6.
ath3d	16299223	Bläsing	2005	224	3h sucrose-induced genes. Extracted from Supplementary Table 6.
ath3e	16299223	Bläsing	2005	223	4h photomorphogenesis repressed genes. Extracted from Supplementary Table 6.
ath3f	16299223	Bläsing	2005	219	4h photomorphogenesis induced genes. Extracted from Supplementary Table 6.
ath3g	16299223	Bläsing	2005	220	30min NO3 repressed genes. Extracted from Supplementary Table 6.
ath3h	16299223	Bläsing	2005	219	30min NO3 induced genes. Extracted from Supplementary Table 6.
ath3i	16299223	Bläsing	2005	235	3h NO3 repressed genes. Extracted from Supplementary Table 6.
ath3j	16299223	Bläsing	2005	209	3h NO3 induced genes. Extracted from Supplementary Table 6.
ath3k	16299223	Bläsing	2005	217	3h mannitol repressed genes. Extracted from Supplementary Table 6.
ath3l	16299223	Bläsing	2005	228	3h mannitol induced genes. Extracted from Supplementary Table 6.
ath3m	16299223	Bläsing	2005	232	4h carbon fixation repressed genes. Extracted from Supplementary Table 6.
ath3n	16299223	Bläsing	2005	218	4h carbon fixation induced genes. Extracted from Supplementary Table 6.
ath3o	16299223	Bläsing	2005	221	4h light repressed genes. Extracted from Supplementary Table 6.
ath3p	16299223	Bläsing	2005	228	4h light induced genes. Extracted from Supplementary Table 6.
ath3q	16299223	Bläsing	2005	643	Genes responsive to carbon fixation and glucose. Extracted from Supplementary Table 8.
ath4a	17006513	Mouchel	2006	189	Gene regulated due to residual Uk-1 loci other than brx. Extracted from Supplementary Table 1: root introgression drag bg.

Table A.5: Literature gene lists based on the Ath1 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ath4b	17006513	Mouchel	2006	4006	Genes strongly regulated in root of brx. Extracted from Supplementary Table 1: >2x, brx vs. Sav+resc., root.
ath4c	17006513	Mouchel	2006	3072	Up-regulated genes in root and seedling of brx relative to Sav-0. Extracted from Supplementary Table 1: >2x, vs. Sav+resc., root+seedling.
ath4d	17006513	Mouchel	2006	26	Genes not rescued by BL treatment. Extracted from Supplementary Table 1.
ath5a	16805732	Mandaokar	2006	19	Jasmonate-responsive genes in stamens at 0.5h. Extracted from Supplementary Table S1.
ath5b	16805732	Mandaokar	2006	198	Jasmonate-responsive genes in stamens at 2h. Extracted from Supplementary Table S1.
ath5c	16805732	Mandaokar	2006	491	Jasmonate-responsive genes in stamens at 8h. Extracted from Supplementary Table S1.
ath5d	16805732	Mandaokar	2006	811	Jasmonate-responsive genes in stamens at 22h. Extracted from Supplementary Table S1.
ath6a	16107481	Nagpal	2005	911	Genes whose expression changes during development and peaks at stages 13-14. Extracted from Supplementary Table S1.
ath6b	16107481	Nagpal	2005	417	Genes whose expression changes during development and peaks at stages 11-12. Extracted from Supplementary Table S1.
ath6c	16107481	Nagpal	2005	387	Genes whose expression changes during development and peaks pre-stage 11. Extracted from Supplementary Table S1.
ath6d	16107481	Nagpal	2005	79	Genes whose expression in stage 1-10 flowers is greater in Columbia flowers than in the arf6 arf8 flowers. Extracted from Supplementary Table S2.
ath6e	16107481	Nagpal	2005	472	Genes whose expression in stage 11-12 flowers is greater in Columbia flowers than in the arf6 arf8 flowers. Extracted from Supplementary Table S2.
ath6f	16107481	Nagpal	2005	692	Genes whose expression in stage 13-14 flowers is greater in Columbia flowers than in the arf6 arf8 flowers. Extracted from Supplementary Table S2.
ath6g	16107481	Nagpal	2005	35	Genes induced by auxin-treatment. Extracted from Supplementary Table S3.
ath7a	15908603	Tatematsu	2005	1592	Down-regulated genes in axillary bud. Extracted from Supplementary Table 1b.
ath7b	15908603	Tatematsu	2005	1184	Up-regulated gene in axillary bud. Extracted from Supplementary Table 1c.
ath7c	15908603	Tatematsu	2005	272	SRE-containing down-regulated genes in axillary bud. From Supplementary Table 2a.
ath7d	15908603	Tatematsu	2005	162	Up1-containing up-regulated genes in axillary bud. From Supplementary Table 2b.
ath7e	15908603	Tatematsu	2005	193	Up2-containing up-regulated genes in axillary bud. From Supplementary Table 2c.
ath8a	15894741	Tung	2005	115	Genes regulated in the stigmas relative to wild-type stigmas. Extracted from Supplementary Table 1.
ath8b	15894741	Tung	2005	33	Genes regulated in ablated ovary samples. Extracted from Supplementary Table 2.

Table A.5: Literature gene lists based on the Ath1 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ath9	15937231	Sibout	2005	547	Genes with altered expression in the double mutant. Extracted from Supplementary Table 10.
ath10a	16021403	Zhang	2005	104	Genes preferentially expressed in young inflorescences. Extracted from Supplementary Table 6.
ath10b	16021403	Zhang	2005	105	Genes preferentially expressed in all three reproductive organs (i.e. the young inflorescences, stage-12 floral buds and developing siliques). Extracted from Supplementary Table 7.
ath10c	16021403	Zhang	2005	82	Genes preferentially expressed in young inflorescences and stage 12 floral buds. Extracted from Supplementary Table 8.
ath10d	16021403	Zhang	2005	92	Genes preferentially expressed in stage 12 floral buds and silique. Extracted from Supplementary Table 9.
ath11a	16299169	Suh	2005	614	Genes up-regulated in the epidermis of the base of the stem and the epidermis of the top of the stem. Extracted from Supplementary Data: Up regulated top and base epid.
ath11b	16299169	Suh	2005	630	Genes up-regulated in the epidermis of the top of the stem only. Extracted from Supplementary Data: Up-regulated top epidermis only.
ath11c	16299169	Suh	2005	657	Genes found to be up regulated in the epidermis of the base of the stem only. Extracted from Supplementary Data: Up-regulated base epidermis only.
ath12	16492731	Braybrook	2006	718	Genes induced by LEC2. Extracted from Supplementary Table 2.
ath13	16299182	Lin	2005	1122	Differentially expressed genes under At5PTase13 deficiency. Extracted from Supplementary Table 1.
ath14	16299171	Umbach	2005	203	Genes differentially expressed in antisense versus wild-type leaves. Extracted from Supplementary Table II.
ath15	16330762	Brown	2005	72	Genes induced by UV-B in wild type but show much reduced UV-B induction in uvr8-1 mutant. Extracted from Supplementary Table 2.
ath16a	16372013	Leibfried	2005	104	Genes up-regulated by WUS but not to STM or LFY induction. Extracted from Supplementary Table S1.
ath16b	16372013	Leibfried	2005	44	Genes down-regulated by WUS but not to STM or LFY induction. Extracted from Supplementary Table S1.
ath17a	16183833	Davletova	2005	668	Transcripts significantly changed by more than 2 fold in wild type plants following treatment with hydrogen peroxide. Extracted from Supplementary Table 2.
ath17b	16183833	Davletova	2005	8	Transcripts significantly enhanced in Zat12 over-expressing plants more than in WT plant. Extracted from Supplementary Table 3.
ath17c	16183833	Davletova	2005	90	Transcripts significantly enhanced by more than 2 fold in Zat12 over-expressing plants in response to hydrogen peroxide. Extracted from Supplementary Table 4.

Table A.5: Literature gene lists based on the <u>Ath1</u> array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ath17d	16183833	Davletova	2005	637	All transcripts significantly elevated in Zat12-overexpressing plants in response to hydrogene peroxide stress. Extracted from Supplementary Table 5.
ath18	16648214	Kolbe	2006	44	Genes responsive to short-term DTT treatment. Extracted from Supplementary Table S2.
ath19a	15608331	Yang	2005	64	Genes down-regulated by MSBP1. Extracted from Supplementary Table 1.
ath19b	15608331	Yang	2005	116	Genes up-regulated by MSBP1. Extracted from Supplementary Table 1.
ath20a	15505214	Monte	2004	579	Genes response to 1-h continuous red light (Rc) in wild-type (WT) Arabidopsis seedlings. Extracted from Supplementary Table 3.
ath20b	15505214	Monte	2004	370	Genes that were induced in the WT by 1-h Rc. Extracted from Supplementary Table 4
ath20c	15505214	Monte	2004	210	Genes that were repressed in the WT by 1-h Rc. Extracted from Supplementary Table 5.

Table A.6: Literature gene lists based on the DrosGenome1 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
dros1	16277749	Wertheim	2005	159	Probesets differentially expressed between parasitised and control larvae. Extracted from Additional data file 1.
dros2	15345053	Wang	2004	1454	Genes up-regulated in tubule. Extracted from Additional data file 1.
dros3a	15777795	Johansson	2005	150	Genes induced in Drosophila mbn-2 cells 6 h post-challenge with crude LPS. Extracted from Supplementary Table 2.
dros3b	15777795	Johansson	2005	518	Differentially expressed genes in Drosophila mbn-2 cells 6 h post-challenge with crude LPS or E. coli. Extracted from Supplementary Table 3.
dros4a	16357214	Goodliffe	2005	272	Genes up-regulated in response to Myc. Extracted from Supplementary data 1.
dros4b	16357214	Goodliffe	2005	214	Genes were elevated upon reduction of Pc. Extracted from Supplementary data 2.
dros4c	16357214	Goodliffe	2005	129	Genes that were repressed by a factor of 0.533 or more by the ectopic accumulation of dMyc. Extracted from Supplementary data 3.
dros5	12777520	Michalak	2003	51	Transcripts differed significantly in F1 hybrids from both D. simulans and D. mauritiana. Extracted from Supplementary Table 1.
dros6a	16356271	Beckstead	2005	4188	Genes that change their expression by more than 1.5 fold in at least one time point in *EcRi* animals. Extracted from Additional data file 1.
dros6b	16356271	Beckstead	2005	743	20E-regulated and 20E primary-response genes. Extracted from Additional data file 2.
dros7	16264191	Badenhorst	2005	286	Genes differentially expressed in Nurf301 mutants compared with wild-type larvae. Extracted from supplementary data downloaded from http://home.ccr.cancer.gov/badenhorst.
dros8a	12586708	Asha	2003	1286	Genes that were up-regulated in Ras-act hemocytes. Extracted from Supplementary Table 1.
dros8b	12586708	Asha	2003	260	Genes that were down-regulated in Ras-act hemocytes. Extracted from Supplementary Table 2.
dros9	16616121	Terry	2006	403	Candidate stem cell genes: genes enriched in Os+ bgcn- compared to bgcn-. Extracted from Supplementary Table 1.
dros10a	16333985	Sørensen	2005	34	Genes early up-regulated by stress. Extracted from Table 3A in the original paper.
dros10b	16333985	Sørensen	2005	123	Genes early down-regulated by stress. Extracted from Table 3B in the original paper.
dros10c	16333985	Sørensen	2005	34	Genes late up-regulated by stress. Extracted from Table 3C in the original paper.
dros11a	16204451	Yang	2005	2308	Genes under-represented in fly photoreceptor cells relative to whole head. Extracted from Supplementary Table 1.
dros11b	16204451	Yang	2005	1499	Genes enriched in fly photoreceptor cells relative to whole head. Extracted from Supplementary Table 2.
dros12a	16584578	Girardot	2006	4503	Genes that showed age-dependent expression in body parts. Extracted from Additional file 1.

Table A.6: Literature gene lists based on the DrosGenome1 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
dros12b	16584578	Girardot	2006	126	Most head- or thorax-enriched genes. Extracted from Additional file 2.
dros12c	16584578	Girardot	2006	112	Common age-responsive genes. Extracted from Additional file 4.
dros13	16798875	Mack	2006	539	Genes whose expression levels differed significantly in the reproductive tract of unmated vs. mated females at 0, 3, 6, or 24 h postmating. Extracted from Supplementary Table 3.
dros14a	15136717	Landis	2004	168	Genes up-regulated with age. Extracted from Supplementary Table 1: old up.
dros14b	15136717	Landis	2004	494	Genes down-regulated with age. Extracted from Supplementary Table 1: old down.
dros14c	15136717	Landis	2004	108	Genes up-regulated with oxidative stress. Extracted from Supplementary Table 1: O2 up.
dros14d	15136717	Landis	2004	234	Genes down-regulated with oxidative stress. Extracted from Supplementary Table 1: O2 down
dros14e	15136717	Landis	2004	97	Genes up-regulated with age and oxidative stress. Extracted from Supplementary Table 1: old O2 up
dros14f	15136717	Landis	2004	154	Genes down-regulated with age and oxidative stress. Extracted from Supplementary Table 1: old O2 down
dros14g	15136717	Landis	2004	8	Genes down-regulated with age but up-regulated with oxidative stress. Extracted from Supplementary Table 1: old down O2 up
dros14h	15136717	Landis	2004	6	Genes up-regulated with age but down-regulated with oxidative stress. Extracted from Supplementary Table 1: old up O2 down
dros15	16754642	Osada	2006	353	Probesets differentially expressed in MMM and ZZZ . Extracted from Supplementary Table 1.
dros16	15090076	Loop	2004	321	Genes differentially expressed between brat/k06028 and wild type control. Extracted from Additional file 3.
dros17a	16938865	Kadener	2006	552	Genes down-regulated in TIM-MJD fly heads. Extracted from Supporting Table 5.
dros17b	16938865	Kadener	2006	368	Genes up-regulated in TIM-MJD fly heads. Extracted from Supporting Table 6.
dros17c	16938865	Kadener	2006	597	Genes down-regulated in TIM-MJD fly brains. Extracted from Supporting Table 7.
dros17d	16938865	Kadener	2006	739	Genes up-regulated in TIM-MJD fly brains. Extracted from Supporting Table 8.
dros18	14749722	Roxström-Lindquist	2004	427	Genes that were significantly up-regulated following infection with virus, bacteria or fungi in parallel. Extracted from Supplementary Table 2.
dros19a	15695583	Apidianakis	2005	133	Genes up-regulated by P. aeruginosa strain CF5. Extracted from Supplementary Table 1.
dros19b	15695583	Apidianakis	2005	80	Genes down-regulated by P. aeruginosa strain CF5. Extracted from Supplementary Table 1.
dros19c	15695583	Apidianakis	2005	16	Genes up-regulated by P. aeruginosa strain PA14. Extracted from Supplementary Table 2.

Table A.6: Literature gene lists based on the DrosGenome1 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
dros19d	15695583	Apidianakis	2005	12	Genes down-regulated by P. aeruginosa strain PA14. Extracted from Supplementary Table 2.
dros19e	15695583	Apidianakis	2005	241	Genes differentially expressed in CF5-infected vs. PA14-infected flies. Extracted from Supplementary Table 3.
dros20	15458575	Girardot	2004	1368	Stress responsive genes. Extracted from Supplementary Table S1.

Table A.7: Literature gene lists based on the Drosophila 2.0 array

List ID	PubMed ID	First author	Year	# Probe sets	Description
dm1	15851659	Mackay	2005	3727	Genes differentially expressed between fast and slow selection lines. Extracted from Supplementary Table 5.
dm2a	16269142	Barmina	2005	16	Genes differentially expressed between Male T1-Male T2. Extracted from Supplementary Table 1.
dm2b	16269142	Barmina	2005	23	Genes differentially expressed between Female T1-Female T2. Extracted from Supplementary Table 2.
dm2c	16269142	Barmina	2005	143	Genes differentially expressed between Male T1-Female T1. Extracted from Supplementary Table 3.
dm2d	16269142	Barmina	2005	35	Genes differentially expressed between Male T2-Female T2. Extracted from Supplementary Table 4.
dm3a	16581772	Rehwinkel	2006	187	Genes regulated in PIWI and AUB knockdowns. Extracted from Supplementary Table I.
dm3b	16581772	Rehwinkel	2006	472	Genes changed by at least 1.5 fold in expression levels in cells depleted of Drosha. Extracted from Supplementary Table II.
dm3c	16581772	Rehwinkel	2006	1081	Genes regulated in AGO1-depleted cells. Extracted from Supplementary Table III.
dm3d	16581772	Rehwinkel	2006	372	Genes regulated in AGO2-depleted cells. Extracted from Supplementary Table IV.
dm3e	16581772	Rehwinkel	2006	137	Genes up-regulated in cells depleted of Drosha and AGO1. Extracted from Supplementary Table VI.
dm4	16907832	Sun	2006	33	Genes that were differentially expressed after a 24 h exposure to phenobarbital in the diet of third-instar larvae as compared with larvae reared on control diet. Extracted from Table 1 in the main paper.
dm5	16624921	Hughes	2006	2329	Genes that displayed significant variation among isogenic lines. Extracted from Supplementary Table S1.
dm6a	17044737	Edwards	2006	1593	Probesets differing significantly for the main effect of selection line. Extracted from Supplementary Table S2.
dm6b	17044737	Edwards	2006	1539	Probesets that displayed significant differences in contrast statements. Extracted from Supplementary Table S3.
dm6c	17044737	Edwards	2006	12	Probesets that displayed sexually antagonistic expression. Extracted from Supplementary Table S4.
dm7	17054780	Morozova	2006	582	Probesets differentially expressed following alcohol exposure. Extracted from Additional data file 2.
dm8a	16258543	Liu	2005	127	Genes commonly affected in both males and females after depletion of HP1. Extracted from Supplementary Table 1.
dm8b	16258543	Liu	2005	203	Genes specifically affected in males after depletion of HP1. Extracted from Supplementary Table 2.
dm8c	16258543	Liu	2005	119	Genes specifically affected in females after depletion of HP1. Extracted from Supplementary Table 3.
dm9a	18408154	Wang	2008	141	Probesets that were up-regulated in single- vs. group-housing condition. Extracted from Supplemental dataset S1.

Table A.7: Literature gene lists based on the Drosophila 2.0 array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
dm9b	18408154	Wang	2008	48	Probesets that were down-regulated in single- vs. group-housing condition. Extracted from Supplemental dataset S1.
dm10a	18367466	Rand	2008	74	Probesets that were up-regulated in response to methylmercury treatment. Extracted from Figure 3 in the main paper.
dm10b	18367466	Rand	2008	213	Probesets that were down-regulated in response to methylmercury treatment. Extracted from Supplementary Figure 2.
dm11	18628398	Gilchrist	2008	243	Genes affected by the negative elongation factor (NELF)-depletion. Extracted from Supplementary Table S1.
dm12a	18296696	Kopp	2008	102	OR, OBP, and GR genes that were detected in the third antennal segment and did not show evidence of sex-biased expression. Extracted from Supplementary Table 2.
dm12b	18296696	Kopp	2008	22	OR, OBP, and GR genes that were detected in the third antennal segment and showed evidence of sex-biased expression. Extracted from Supplementary Table 3.
dm13a	17578907	Kadener	2007	73	Direct CLK target genes from heads. Extracted from Supplementary file TargetHeads.xls.
dm13b	17578907	Kadener	2007	46	Direct CLK target genes from S2 cells. Extracted from Supplementary file TargetS2cells.xls.
dm14	17448252	Qin	2007	500	Top 500 genes that showed the most significant changes of their ribosomal occupancy between the development periods. Extracted from Additional data file 2.

Table A.8: Literature gene lists based on the Celegans array

List ID	PubMed ID	First author	Year	# Probe sets	Description
ce1a	16626960	Vartiainen	2006	433	Up-regulated genes in transgenic C. elegans over-expressing alpha-synuclein. Extracted from Supplementary Table S7.
ce1b	16626960	Vartiainen	2006	67	Down-regulated genes in transgenic C. elegans over-expressing alpha-synuclein. Extracted from Supplementary Table S3.
ce2a	16480708	Wang	2006	167	Genes that are up-regulated upon over-expression of both CeTwist and CeE/DA. Extracted from Supplementary data S1.
ce2b	16480708	Wang	2006	34	Genes that are down-regulated upon over-expression of both CeTwist and CeE/DA. Extracted from Supplementary data S2.
ce3	16184190	Shen	2005	816	Genes differentially regulated by ire-1 and xbp-1. Extracted from Supplementary Table S10.
ce4a	15256590	Huffman	2004	1092	Genes that showed significant regulation by Cry5B. Extracted from Supplementary Table 2.
ce4b	15256590	Huffman	2004	1083	Genes that showed significant regulation by cadmium. Extracted from Supplementary Table 3.
ce5a	15308663	McElwee	2004	1348	Genes that were up-regulated in daf-2 compared with daf-16;daf-2. Extracted from Supplementary data <- Final gene lists.xls <- worksheet: Daf-2 final gene list.
ce5b	15308663	McElwee	2004	926	Genes that were down-regulated in daf-2 compared with daf-16;daf-2. Extracted from Supplementary data <- Final gene lists.xls <- worksheet: Daf-2 final gene list.
ce6a	16809667	O'Rourke	2006	71	Genes up-regulated following infection of C. elegans with M. nematophilum. Extracted from Supplementary data S1.
ce6b	16809667	O'Rourke	2006	22	Genes down-regulated following infection of C. elegans with M. nematophilum. Extracted from Supplementary data S1.
ce7a	17096597	Troemel	2006	144	Genes differentially expressed between daf-2 and daf-2;pmk-1. Extracted from Supplementary Table S2.
ce7b	17096597	Troemel	2006	449	Genes differentially expressed between exposure to E. coli strain OP50 and wild-type P. aeruginosa strain PA14. Extracted from Supplementary Table S4.
ce7c	17096597	Troemel	2006	287	Genes differentially expressed between exposure to gacA mutant and wild-type P. aeruginosa strain PA14. Extracted from Supplementary Table S5.
ce8a	16962739	Towers	2006	44	Up-regulated genes in the two mutant alleles cx35 and cx18 as compared to the wild-type. Extracted from Supplementary Table 1.
ce8b	16962739	Towers	2006	71	Down-regulated genes in the two mutant alleles cx35 and cx18 as compared to the wild-type. Extracted from Supplementary Table 2.
ce8c	16962739	Towers	2006	124	Differentially expressed genes between wild-type and dys-1 mutant. These genes were selected based on transcript presence calls and a statistical confidence level of 0.05. Extracted from Supplementary Table 3.

Table A.8: Literature gene lists based on the Celegans array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
ce9	17187676	Chen	2006	466	Genes that show a down regulation of 2 fold or higher in daf-19(-) animals compared to the daf-19(+) control animals. Extracted from Additional data file 3.
ce10a	15620651	Colosimo	2004	167	Genes differentially expressed between AFD and AWB. Extracted from Supplementary Table S1.
ce10b	15620651	Colosimo	2004	1513	Genes differentially expressed between AFD and AWB and the unsorted whole embryonic cells. Extracted from Supplementary Table S2.
ce11a	15780142	Fox	2005	1012	unc-4::GFP enriched genes. Extracted from Additional data file 9.
ce11b	15780142	Fox	2005	1586	N2 enriched genes. Extracted from Additional data file 10.
ce12	18627611	Greiss	2008	190	Genes up-regulated by at least 1.5 fold 2 hours after X-ray treatment. Extracted from Additional file 2.
ce13	17368442	Kirienko	2007	1949	Genes that showed altered expression in the lin-35 mutant background compared with wild type. Extracted from Supplementary file 1.
ce14a	17612406	Von Stetina	2007	1637	Embryonic Pan-neural enriched genes. Extracted from Additional data file 1.
ce14b	17612406	Von Stetina	2007	1562	Larval Pan-neural enriched genes. Extracted from Additional data file 1.
ce14c	17612406	Von Stetina	2007	995	Embryonic A-class enriched genes. Extracted from Additional data file 1.
ce14d	17612406	Von Stetina	2007	412	Larval A-class enriched genes. Extracted from Additional data file 1.

Table A.9: Literature gene lists based on the Xenopus laevis array

List ID	PubMed ID	First author	Year	# Probe sets	Description
xp1	16872594	Urban	2006	463	Genes regulated by hedgehog. Extracted from Supplementary material Part I.
xp2	16651540	Sinner	2006	276	Endoderm-enriched genes that were differentially expressed between vegetal and animal cap regions. Extracted from Supplementary Table S1.
xp3a	16756679	Hufton	2006	188	Genes regulated by ectopic organizer signalling. Extracted from Additional file S2.
xp3b	16756679	Hufton	2006	220	Genes that showed similar regulation in the full organizer conditions. Extracted from Additional file S3.
xp4a	17024524	Malone	2006	25	Top 25 male biased differentially expressed genes in both species. Extracted from Supplementary Table 4.
xp4b	17024524	Malone	2006	25	Top 25 female biased differentially expressed genes in both species. Extracted from Supplementary Table 4.
xp4c	17024524	Malone	2006	25	Top 25 differentially expressed genes that showed unbiased expression in both species. Extracted from Supplementary Table 4.
xp4d	17024524	Malone	2006	25	Top 25 male biased genes that showed no difference in expression between species. Extracted from Supplementary Table 4.
xp4e	17024524	Malone	2006	25	Top 25 female biased genes that showed no difference in expression between species. Extracted from Supplementary Table 4.
xp5a	15901671	Gurvich	2005	666	Genes most stimulated by both VPA and TSA in xenopus. Extracted from Supplementary data xenopus.stim key.
xp5b	15901671	Gurvich	2005	599	Genes most inhibited by both VPA and TSA in xenopus. Extracted from Supplementary data xenopus.inhib key.
xp6a	16871633	Grow	2006	1108	Comparison A: St53 1dPA Versus st57 1dPA. Genes differentially expressed between st53 blastemas and st57 pseudoblastemas at day 1 postamputation. Extracted from Supplementary Table using P (Welch's T-test) < 0.05.
xp6b	16871633	Grow	2006	2428	Comparison B: St53 5dPA Versus st57 5dPA. Genes differentially expressed between st53 blastemas and st57 pseudoblastemas during regeneration of the limb (st53) or during spike formation (st57) at 5 days postamputation. Extracted from Supplementary Table using P (Welch's T-test) < 0.05.
xp6c	16871633	Grow	2006	1677	Comparison C: St53 1dPA Versus st53 5dPA. Genes differentially expressed from day 1 to day 5 during the regeneration of the st53 limb. Extracted from Supplementary Table using P (Welch's T-test) < 0.05.
xp6d	16871633	Grow	2006	115	Comparison D: St57 1dPA Versus st57 5dPA. Genes differentially expressed from day 1 to day 5 during st57 spike formation.Extracted from Supplementary Table using P (Welch's T-test) < 0.05.

Table A.9: Literature gene lists based on the Xenopus laevis array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
xp7	18650498	Tonge	2008	172	Genes up-regulated by at least 5 fold in 3-day in-vitro conditioned compared to the primary DRG. Extracted from Supplementary Table S1.
xp8a	17705306	Cha	2007	20	Top 20 genes that showed up-regulated expression levels in the presence of FoxC1 morpholino. Extracted from Table 1A in the main paper.
xp8b	17705306	Cha	2007	20	Top 20 genes that showed down-regulated expression levels in the presence of FoxC1 morpholino. Extracted from Table 1A in the main paper.

Table A.10: Literature gene lists based on the Zebrafish array

List ID	PubMed ID	First author	Year	# Probe sets	Description
zf1	16631158	Cheng	2006	295	Liver-enriched genes. Extracted from Supplementary Table 1.
zf2	16869712	Lien	2006	662	Genes that were differentially expressed during heart regeneration process. Extracted from dataset S1.
zf3a	16443690	Andreasen	2006	67	Transcripts enhanced at least 2 fold by TCDD exposure in comparison with their time matched vehicle control. Extracted from Table 2 in original paper.
zf3b	16443690	Andreasen	2006	132	Transcripts repressed at least 2 fold by TCDD exposure in comparison with their time matched vehicle control. Extracted from Table 3 in original paper.
zf4	15827125	Weber	2005	387	Genes differentially regulated between clo and WT siblings. Extracted from Supplementary Table S1.
zf5a	16322560	Chen	2005	141	Transcript down-regulated between def mutant and WT. Extracted from Supplementary material Table 1.
zf5b	16322560	Chen	2005	23	Transcript differentially up-regulated between def mutant and WT. Extracted from Supplementary material Table 2.
zf6	16714409	Carney	2006	163	TCDD-induced genes in the zebrafish heart at 73, 74, 76 and 84 hpf. Extracted from Supplementary data S1.
zf7a	17251491	Leung	2007	78	Over-expressed genes in the RPE Compared with the Retina at 52 hpf. Extracted from Table 2 from the main paper.
zf7b	17251491	Leung	2007	988	Under-expressed genes in the RPE Compared with the Retina at 52 hpf. Extracted from Supplementary Table S3.
zf8	16484454	Giraldez	2006	811	Genes that were up-regulated in MZdicer compared to wild type and MZdicer+miR-430 at 9 hpf. Extracted from Supplementary Table S1.
zf9	16638810	Link	2006	220	Differentially expressed genes between ectodermal and mesendodermal cells. Extracted from Supplementary Table S2.
zf10a	15901671	Gurvich	2005	600	Genes most stimulated by both VPA and TSA in zebrafish. Extracted from Supplementary data zebrafish.stim key.
zf10b	15901671	Gurvich	2005	617	Genes most inhibited by both VPA and TSA in zebrafish. Extracted from Supplementary data zebrafish.inhib key.
zf11a	18495758	Mathew	2008	66	Genes enhanced at least 1.7 fold by TCDD exposure in regenerating fins of larvae. Extracted from Supplemental Table S1.
zf11b	18495758	Mathew	2008	69	Genes repressed at least 1.7 fold by TCDD exposure in regenerating fins of larvae. Extracted from Supplemental Table S2.
zf11c	18495758	Mathew	2008	30	Genes enhanced at least 1.7 fold by TCDD exposure in regenerating fins of larvae and adults. Extracted from Supplemental Table S3.
zf11d	18495758	Mathew	2008	58	Genes repressed at least 1.7 fold by TCDD exposure in regenerating fins of larvae and adults. Extracted from Supplemental Table S4.

Table A.10: Literature gene lists based on the Zebrafish array (continued)

List ID	PubMed ID	First author	Year	# Probe sets	Description
zf12a	17698971	Bahary	2007	302	Genes differentially regulated in VegfAa morphants. Extracted from Supplementary Table S1.
zf12b	17698971	Bahary	2007	301	Genes differentially regulated in VegfAb morphants. Extracted from Supplementary Table S2.
zf12c	17698971	Bahary	2007	39	Genes differentially regulated in both VegfAa and VegfAb morphants. Extracted from Supplementary Table S3.
zf13a	17699609	Maves	2007	188	Genes regulated in control MO versus pbx2-MO; pbx4-MO at 10-somite stage. Extracted from Supplementary Table S1.
zf13b	17699609	Maves	2007	258	Genes regulated in control MO versus pbx2-MO; pbx4-MO at 18-somite stage. Extracted from Supplementary Table S2.
zf14a	17369489	Liu	2007	87	Genes down-regulated in udu-sq1 mutant. Extracted from Supplementary Table S3.
zf14b	17369489	Liu	2007	56	Genes up-regulated in udu-sq1 mutant. Extracted from Supplementary Table S4.

Appendix B

Computer codes and scripts

Three text-based ORA approaches have been developed for mining the abstract text associated with a list of differentially expressed genes and to search within them for terms or biological concepts that are significantly over-represented. The first approach is based on the use of a permutation test; the second approach is based on the detection of outliers (OutlierDM); the third approach uses the extended hypergeometric distribution to assess over-representation (ExtendedHG). The source code for these methods is written in R and Perl, and has been tested on a Microsoft Windows platform with a 2.8 GHz processor and 2 GB of RAM. The R scripts were developed under R-2.6 and BioConductor-2.1, while the Perl scripts were developed under Perl v5.8.7.

Besides the list of differentially expressed genes, the proposed text-based ORA methods required three additional pre-processed data files, which link the gene identifiers in the query gene list with the relevant abstracts and token frequency data.

- array_annfile.data: provides mappings between the Affymetrix probeset IDs, Entrez Gene IDs and PubMed IDs.
- array_termfile.data: contains pre-processed tokens extracted from the abstracts.
- array_chiphits.data: records the *Chip* frequencies, that is the number of genes containing a certain token of interest on a given chip type.

Pre-processed data files for the 10 Affymetrix GeneChip® arrays used in this work are provided on the CD-ROM attached to this thesis. The scripts for creating these data files are also included in the CD-ROM; the procedures for creating the associated text corpus and processing the abstract text can be found in Chapter 2.

B.1 R functions for performing text-based ORA

This section provides an overview of the three R functions - `run.Permutation`, `run.OutlierDM` and `run.ExtendedHG` – that were developed for performing ORA using tokens extracted from PubMed abstracts.

Function `run.Permutation`

Description

A method for identifying significantly over-represented abstract terms within a list of differentially expressed genes based on the use of a permutation test, as described in Chapter 5 of this thesis.

Usage

```
run.permutation(x, input.file=NULL, runname="myRun", chip="hgu133a",
idType="affy", jackknife=1, adj.method="bonferroni", cutoff=0.05,
nperm=100000, data.dir)
```

Arguments

`x` A vector of gene identifiers. This argument can be ignored when the input is read from a file.

`input.file` File that contains the list of genes to be analysed; one entry per line. Set to `"NULL"` when the input is a vector of gene identifiers.

`runname` Name of the analysis. Default to `"myRun"`.

`chip` Name of the chip type on which the gene list was based. This should be one of: `"hgu133a"`, `"hgu132plus2"`, `"mouse4302"`, `"rat2302"`, `"ath1121501"`, `"drosgenome1"`, `"drosophila2"`, `"celegans"`, `"xenopuslaevis"`, `"zebrafish"`, which correspond to the Affymetrix arrays HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Celegans, Xenopus laevis, and Zebrafish, respectively.

`idType` The type of gene identifier in the query gene list. This can be in the form of Affymetrix probeset ID or EntrezGene ID (EGID). For probeset ID, set `idType = "affy"`. For EGID, set `idType = "egid"`. Default to `"affy"`.

`jackknife` Jackknife score (integer). The greater the value, the more conservative the *p*-value will be. Default to 1.

`adj.method` Methods used for adjusting the raw *p*-values for multiple hypothesis testing. This is done by a call to `p.adjust`, so the argument should be one of: `"holm"`, `"hochberg"`, `"hommel"`, `"bonferroni"`, `"BH"`, `"BY"` or `"fdr"`. Default to `"bonferroni"`.

`cutoff` Cutoff value for the adjusted *p*-values. Default to 0.05.

`nperm` Number of permutations to be performed. Default to 100000.

`data.dir` Directory in which the pre-processed data files are located.

Values

Three files are generated by this function:

- `runname_EmpiricalCount.txt` This file records the empirical counts for the test statistics based on the randomised data.
- `runname_ResultTable.txt` This file contains all the tokens and their *p*-values as calculated by the permutation test.
- `runname_EnrichedTerms.txt` This file contains the significant tokens and their *p*-values as determined using the specified cutoff value.

Examples

```
# Assume that the pre-processed data files are stored in the
# directory ./Data, and that the differentially expressed genes
# are stored in a file named “sampleGenelist.txt” in the form of
# Affymetrix probeset ID.

# Example 1: Input gene identifiers as vector

testDEG <- scan("sampleGenelist.txt", strip.white=TRUE, sep="\n",
what="character")

run.permutation(x=testDEG, input.file=NULL, runname="Testrun",
chip="hgu133a", idType="affy", jackknife=1, adj.method="bonferroni",
cutoff=0.05, nperm=100000, data.dir="./Data")

# Example 2: Read the gene list directly from file

run.permutation(input.file="sampleGenelist.txt", runname="Testrun",
chip="hgu133a", idType="affy", jackknife=1, adj.method="bonferroni",
cutoff=0.05, nperm=100000, data.dir="./Data")
```

Source code

The source code is included in the CD-ROM attached to this thesis.

Function `run.OutlierDM`

Description

A method for identifying significantly over-represented abstract terms within a list of differentially expressed genes based on the detection of outliers, as described in Chapter 6 of this thesis.

Usage

```
run.OutlierDM(x, input.file=NULL, runname="myRun", chip="hgu133a",
idType="affy", jackknife=1, adj.method="bonferroni", cutoff=0.05,
windowSize=10, data.dir)
```

Arguments

`x` A vector of gene identifiers. This argument can be ignored when the input is read from a file.

`input.file` File that contains the list of genes to be analysed; one entry per line. Set to `“NULL”` when the input is a vector of gene identifiers.

`runname` Name of the analysis. Default to `“myRun”`.

`chip` Name of the chip type on which the gene list was based. This should be one of: `“hgu133a”`, `“hgu132plus2”`, `“mouse4302”`, `“rat2302”`, `“ath1121501”`, `“drosgenome1”`, `“drosophila2”`, `“celegans”`, `“xenopuslaevis”`, `“zebrafish”`, which correspond to the Affymetrix arrays HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Celegans, Xenopus laevis, and Zebrafish, respectively.

`idType` The type of gene identifier in the query gene list. This can be in the form of Affymetrix probeset ID or EntrezGene ID (EGID). For probeset ID, set `idType = “affy”`. For EGID, set `idType = “egid”`. Default to `“affy”`.

`jackknife` Jackknife score (integer). The greater the value, the more conservative the p-value will be. Default to 1.

`adj.method` Methods used for adjusting the raw p-values for multiple hypothesis testing. This is done by a call to `p.adjust`, so the argument should be one of: `“holm”`, `“hochberg”`, `“hommel”`, `“bonferroni”`, `“BH”`, `“BY”` or `“fdr”`. Default to `“bonferroni”`.

`cutoff` Cutoff value for the adjusted p-values. Default to 0.05.

`windowSize` The minimum number of observations used to estimate local mean and standard deviation. Default to 10.

`data.dir` Directory in which the pre-processed data files are located.

Values

The result is a list with the following names-values:

`EnrichedTerms` This contains the significantly over-represented terms whose adjusted *p*-values satisfy the cutoff threshold after correcting for multiple hypothesis testing. The results are presented as a table with the following columns:

`Token`: Significantly enriched terms

`Chip`: Total number of genes associated with a particular token in the background

`List`: Total number of genes associated with a particular token in the input gene list

`Z.score`: Note that a negative sign associated with a *Z*-score implies over-representation

`Raw.pval`: Raw *p*-values

`Adjusted.pval`: Adjusted *p*-values

`Ranking`: Rank of the terms according to their raw *p*-values

`ResultTable` This contains all tokens that were subjected for testing. The results are presented as a table with columns same as those for `EnrichedTerms`.

Examples

```
# Assume that the pre-processed data files are stored in the
# directory ./Data.
# Example 1: Input gene identifiers as vector

testDEG <- scan("sampleGenelist.txt", strip.white=TRUE, sep="\n",
what="character")

out1 <- run.OutlierDM(x=testDEG, input.file=NULL, runname="Testrun",
chip="hgu133a", idType="affy", jackknife=1, adj.method="bonferroni",
cutoff=0.05, windowSize=10, data.dir="./Data")

# Example 2: Read the gene list directly from file

out2 <- run.OutlierDM(input.file="sampleGenelist.txt", runname=
"Testrun", chip="hgu133a", idType="affy", jackknife=1, adj.method=
"bonferroni", cutoff=0.01, windowSize=10, data.dir="./Data")
```

Source code

The R script implementing this approach is shown in Section B.2, and the Perl script called by this function is shown in Section B.4. The source code is included in the CD-ROM attached to this thesis.

Function `run.ExtendedHG`

Description

A method for identifying significantly over-represented abstract terms within a list of differentially expressed genes based on the extended hypergeometric distribution, as described in Chapter 7 of this thesis.

Usage

```
run.ExtendedHG(x, input.file=NULL, runname="myRun", chip="hgu133a",
idType="affy", jackknife=1, adj.method="bonferroni", cutoff=0.05,
data.dir)
```

Arguments

`x` A vector of gene identifiers. This argument can be ignored when the input is read from a file.

`input.file` File that contains the list of genes to be analysed; one entry per line. Set to "NULL" when the input is a vector of gene identifiers.

`runname` Name of the analysis. Default to "myRun".

`chip` Name of the chip type on which the gene list was based. This should be one of: "hgu133a", "hgu132plus2", "mouse4302", "rat2302", "ath1121501", "drosgenome1", "drosophila2", "celegans", "xenopuslaevis", "zebrafish", which correspond to the Affymetrix arrays HG-U133A, HG-U133 Plus 2.0, MG-U430 2.0, RAT230 2.0, Ath1, DrosGenome1, Drosophila 2.0, Celegans, Xenopus laevis, and Zebrafish, respectively.

`idType` The type of gene identifier in the query gene list. This can be in the form of Affymetrix probeset ID or EntrezGene ID (EGID). For probeset ID, set `idType = "affy"`. For EGID, set `idType = "egid"`. Default to "affy".

`jackknife` Jackknife score (integer). The greater the value, the more conservative the p-value will be. Default to 1.

`adj.method` Methods used for adjusting the raw p-values for multiple hypothesis testing. This is done by a call to `p.adjust`, so the argument should be one of: "holm", "hochberg", "hommel", "bonferroni", "BH", "BY" or "fdr". Default to "bonferroni".

`cutoff` Cutoff value for the adjusted p-values. Default to 0.05.

`data.dir` Directory in which the pre-processed data files are located.

Values

`EnrichedTerms` This contains the significantly over-represented terms whose adjusted *p*-values satisfy the cutoff threshold after correcting for multiple hypothesis testing. The results are presented as a table with the following columns:

`Token`: Significantly enriched terms

`Chip`: Total number of genes associated with a particular token in the background

`List`: Total number of genes associated with a particular token in the input gene list

`Odds.ratio`: Odds ratios associated with the tokens

`Raw.pval`: Raw *p*-values

`Adjusted.pval`: Adjusted *p*-values

`Ranking`: Ranks of the terms according to their raw *p*-values

`ResultTable` This contains all tokens that were subjected for testing. The results are presented as a table with columns same as those for `EnrichedTerms`.

Dependencies

This method depends on the `BiasedUrn` R package.

Examples

```
# Assume that the pre-processed data files are stored in the
# directory ./Data.
# Example 1: Input gene identifiers as vector

testDEG <- scan("sampleGenelist.txt", strip.white=TRUE, sep="\n",
what="character")

out1 <- run.ExtendedHG(x=testDEG, input.file=NULL, runname="Testrun",
chip="hgu133a", idType="affy", jackknife=1, adj.method="bonferroni",
cutoff=0.05, data.dir="./Data")

# Example 2: Read the gene list directly from file
out2 <- run.ExtendedHG(input.file="sampleGenelist.txt", runname=
"Testrun", chip="hgu133a", idType="affy", jackknife=1, adj.method=
"bonferroni", cutoff=0.05, data.dir="./Data")
```

Source code

The R script implementing this approach is shown in Section B.3, and the Perl script called by this function is shown in Section B.4. The source code is included in the CD-ROM attached to this thesis.

B.2 Source code of OutlierDM

```
run.OutlierDM <- function(x, input.file=NULL, runname="myRun", chip="hgu133a",
idType="affy", windowSize=10, jackknife=1, adj.method="bonferroni", cutoff=0.05,
data.dir)
{
      # Check if the input.genelist is a vector or a file
      if(is.null(input.file)){
         myGenelist <- "tmpGL.txt"
         write.table(x, file=myGenelist, sep="\n", quote=FALSE, row.names=FALSE,
                     col.names=FALSE)
      }else{
         myGenelist <- input.file
      }

      # Locate the pre-processed data files
      data.dir <- paste(data.dir, chip, sep="/")
      annotationFile = paste(chip, "_annfile.data", sep="")
      termFile = paste(chip, "_termfile.data", sep="")
      chipFile = paste(chip, "_chiphits.data", sep="")

      # Determine List frequency associated with the tokens in Perl
      working.dir <- getwd()
      perl.dir <- paste(working.dir, "Perlcodes", sep="/")
      cat(myGenelist, idType, data.dir, annotationFile, termFile, chipFile,
          chip, runname, file="param.txt", sep="\t")
      perl.command <- paste('perl "', perl.dir, '/getListhits.pl"', sep="")
      xx <- as.vector(system(paste(perl.command), show.output.on.console=TRUE,
            invisible=TRUE, intern=TRUE))

      inputID.count <- paste(xx[1])
      Chip.totalEGID <- as.numeric(xx[2])
      List.totalEGID <- as.numeric(xx[3])
      Chip.annEGID <- as.numeric(xx[4])
      List.annEGID <- as.numeric(xx[5])
      List.totalPMID <- as.numeric(xx[6])
      Count.file <- paste(runname, "_PmFreq.txt", sep="")

      # Token frequency data filtering and pre-processing
      dat <- read.table(file=Count.file, header=TRUE, sep="\t", na.string="")
      colnames(dat) <- c("Token", "Chip", "List")
      dat$List <- dat$List - jackknife
      dat <- dat[dat$List > 1, ]
      set.seed(1234)
      idx <- sample(1:nrow(dat))
      dat <- dat[idx, ]
      dat[ , c(2:3)] <- log2(dat[ , c(2:3)])
      dat <- dat[order(dat$List, decreasing=TRUE), ]

      # Initialise some variables
      smChipMean <- vector(length=nrow(dat))
      smChipSD <- vector(length=nrow(dat))
      uniqChipMean <- sort(unique(dat$List))
      uniqChipSD <- sort(unique(dat$List))
      countPoints <- sort(unique(dat$List))
```

```
# Local mean and SD estimation
for(n in 1:length(countPoints)) {
    i <- dat$List == countPoints[n]
    if(sum(i) >= windowSize) {
       myData <- dat[i, "Chip"]
       uniqChipMean[n] <- mean(myData)
       uniqChipSD[n] <- sd(myData)
    }else{
       weHaveGot <- sum(i)
       ni <- dat$List < countPoints[n]
       resDat <- dat[ni, ]
       myData <- c(dat[i, "Chip"], resDat[1:(windowSize-weHaveGot), "Chip"])
       uniqChipMean[n] <- mean(myData)
       uniqChipSD[n] <- sd(myData)
    }
}

#  Local mean and SD smoothing
chipMean.model <- lm(uniqChipMean ~ countPoints + I(countPoints^2) +
                     I(countPoints^3))
chip.mean <- chipMean.model$coef %*% rbind(1, countPoints, countPoints^2,
             countPoints^3)
chipSD.model <- lm(uniqChipSD ~ countPoints + I(countPoints^2) +
                   I(countPoints^3))
chip.sd <- chipSD.model$coef %*% rbind(1, countPoints, countPoints^2,
           countPoints^3)

for(n in 1:length(countPoints)) {
    i <- dat$List == countPoints[n]
    smChipMean[i] <- chip.mean[n]
    smChipSD[i] <- chip.sd[n]
}

# Z-score and p-value calculation
mcDat <- cbind(dat,
               "Z.score"=vector(length=nrow(dat)),
               "Raw.pval"=vector(length=nrow(dat)),
               "Adjusted.pval"=vector(length=nrow(dat)),
               "Ranking"=vector(length=nrow(dat))
         )
mcDat[ ,c(2:3)] <- 2^mcDat[ ,c(2:3)]
mcDat$List <- mcDat$List + jackknife
mcDat$Z.score <- (dat$Chip - smChipMean)/smChipSD
mcDat$Raw.pval <- pnorm(mcDat$Z.score)
mcDat$Adjusted.pval <- p.adjust(mcDat$Raw.pval, method=adj.method)
mcDat$Ranking <- rank(mcDat$Raw.pval, ties.method="first")
sorted.res <- mcDat[order(mcDat$Raw.pval, decreasing=FALSE, na.last=NA), ]
hits <- sorted.res[sorted.res$Adjusted.pval < cutoff, ]


# Clean up
unlink(c("param.txt", Count.file, "tmpGL.txt"))
```

```
    # Output
    cat("\n//------  Text-based ORA results for ", runname, "  ------//\n",
        sep="")
    cat("Current mode of analysis: OutlierDM", "\n", sep="")
    cat("Current analysis setting: ", adj.method, " corrected p-value < ",
        cutoff, "\n",  sep="")
    cat("Gene identifiers submitted for analysis: ", inputID.count, "\n",
        sep="")
    cat("Annotated genes in the gene list: ", List.annEGID, "\n", sep="")
    cat("Annotated genes in the background: ", Chip.annEGID, sep="", "\n")
    cat("Total PMID associated with the gene list: ", List.totalPMID, "\n", sep="")
    cat(nrow(hits), " tokens are found to be significantly over-represented",
        "\n\n", sep="")
    print(hits)

    results <- (list(EnrichedTerms=hits, ResultTable=sorted.res))
    invisible(results)

}
```

B.3 Source code of ExtendedHG

```
run.ExtendedHG <- function(x, input.file=NULL, runname="myRun", chip="hgu133a",
idType="affy", jackknife=1, adj.method="bonferroni", cutoff=0.05, data.dir)
{
      require("BiasedUrn")

      # Check if the input.genelist is a vector or a file
      if(is.null(input.file)){
         myGenelist <- "tmpGL.txt"
         write.table(x, file=myGenelist, sep="\n", quote=FALSE, row.names=FALSE,
                     col.names=FALSE)
      }else{
         myGenelist <- input.file
      }

      # Locate the pre-processed data files
      data.dir <- paste(data.dir, chip, sep="/")
      annotationFile = paste(chip, "_annfile.data", sep="")
      termFile = paste(chip, "_termfile.data", sep="")
      chipFile = paste(chip, "_chiphits.data", sep="")

      # Determine List frequency associated with the tokens in Perl
      working.dir <- getwd()
      perl.dir <- paste(working.dir, "Perlcodes", sep="/")
      cat(myGenelist, idType, data.dir, annotationFile, termFile, chipFile,
          chip, runname, file="param.txt", sep="\t")
      perl.command <- paste(‘perl "’, perl.dir, ‘/getListhits.pl"’, sep="")
      xx <- as.vector(system(paste(perl.command), show.output.on.console=TRUE,
            invisible=TRUE, intern=TRUE))

      # Get the output from Perl
      inputID.count <- paste(xx[1])
      Chip.totalEGID <- as.numeric(xx[2])
      List.totalEGID <- as.numeric(xx[3])
      Chip.annEGID <- as.numeric(xx[4])
      List.annEGID <- as.numeric(xx[5])
      List.totalPMID <- as.numeric(xx[6])
      Count.file <- paste(runname, "_PmFreq.txt", sep="")

      # Token frequency data filtering and pre-processing
      dat <- read.table(file=Count.file, header=TRUE, sep="\t", na.string="")
      colnames(dat) <- c("Token", "Chip", "List")
      lm.mu <- lm(dat$List ~ dat$Chip + I(dat$Chip^2) + I(dat$Chip^3) +
                  I(dat$Chip^4) + I(dat$Chip^5) + I(dat$Chip^6) + I(dat$Chip^7))
      estList <- lm.mu$coef %*% rbind(1, dat$Chip, dat$Chip^2, dat$Chip^3,
                 dat$Chip^4, dat$Chip^5, dat$Chip^6, dat$Chip^7)
      dat <- cbind(dat, "Mu"=as.vector(estList))
      dat <- dat[dat$Chip > 1, ]
      dat <- dat[dat$List - jackknife > 1, ]
```

```
# Function for calculating odds ratio
calOdds <- function(x, mu.index, Chip.Gene, List.Gene)
{
    chip.hit <- as.numeric(x[2])
    list.hit <- as.numeric(x[3])
    myMu <- as.numeric(x[mu.index])

    if(myMu < List.Gene)
        p2 <- myMu

    if(myMu > List.Gene)
        p2 <- List.Gene * 0.999

    p1 <- List.Gene - p2

    if(myMu >= chip.hit)
        myOdds = 1
    else
        myOdds <- (p2/(chip.hit-p2))/(p1/(Chip.Gene-chip.hit-p1))
}

# Function for calculating Fisher non-central hypergeometric P-values
nchyperGPval <- function(x, odds.index, Chip.Gene, List.Gene, jackscore)
{
    chip.hit <- as.numeric(x[2])
    list.hit <- as.numeric(x[3])
    myOdds <- as.numeric(x[odds.index])
    myOdds <- ifelse(myOdds < 0, 1, myOdds)
    pFNCHypergeo(x=list.hit-jackscore-1, m1=chip.hit,
                 m2=Chip.Gene-chip.hit, n=List.Gene-jackscore,
                 odds=myOdds, precision=1E-40, lower.tail=FALSE)
}

# P-value calculation
res <- data.frame(cbind(dat,
                "Odds" = vector(length=nrow(dat)),
                "Adj.odds" = vector(length=nrow(dat)),
                "Raw.pval"=vector(length=nrow(dat)),
                "Adjusted.pval"=vector(length=nrow(dat)),
                "Ranking"=vector(length=nrow(dat))
        ))
res[ ,"Odds"] <- apply(res, 1, calOdds, mu.index=4, Chip.Gene=Chip.annEGID,
                       List.Gene=List.annEGID)
idx <- res$Odds > median(res$Odds) + 3*mad(res$Odds) & res$List < 10
res$Adj.odds <- res$Odds
res[idx, "Adj.odds"] <- median(res[!idx, "Odds"])
res$Raw.pval <- apply(res, 1, nchyperGPval, odds.index=6,
      Chip.Gene=Chip.annEGID, List.Gene=List.annEGID, jackscore=jackknife)
res$Adjusted.pval <- p.adjust(res$Raw.pval, method=adj.method)
res$Ranking <- rank(res$Raw.pval, ties.method="first")
sorted.res <- res[order(res$Raw.pval, decreasing=FALSE, na.last = NA), ]
hits <- sorted.res[sorted.res$Adjusted.pval < cutoff, ]
hits <- hits[ ,-c(4,5)]
colnames(hits) <- c("Token", "Chip", "List", "Odds.ratio", "Raw.pval",
                    "Adjusted.pval", "Ranking")
        # Clean up
        unlink(c("param.txt", Count.file, "tmpGL.txt"))

        # Output
        cat("\n//------  Text-based ORA results for ", runname, "  ------//\n",
            sep="")
        cat("Current mode of analysis: ExtendedHG", "\n", sep="")
        cat("Current analysis setting: ", adj.method, " corrected p-value < ",
            cutoff, "\n",  sep="")
        cat("Gene identifiers submitted for analysis: ", inputID.count, "\n",
            sep="")
        cat("Annotated genes in the gene list: ", List.annEGID, "\n", sep="")
        cat("Annotated genes in the background: ", Chip.annEGID, sep="", "\n")
        cat("Total PMID associated with the gene list: ", List.totalPMID, "\n",
            sep="")
        cat(nrow(hits), " tokens are found to be significantly over-represented",
            "\n\n", sep="")
        print(hits)

        results <- (list(EnrichedTerms=hits, ResultTable=sorted.res))
        invisible(results)

  }
```

B.4 Perl script invoked in OutlierDM and ExtendedHG

```
#!/usr/bin/perl
#
# Name: getListhits.pl
# Description: a script for calculating the List frequency associated with
#  each token in a given gene list



use strict;
use warnings;


# Get parameters from OutlierDM or ExtendedHG

my @params=();
my $paramfile="param.txt";
open(IN, $paramfile) or die "Cannot open $paramfile\n";
while(<IN>){
     chomp $_;
     @params=split(/\t/, $_);
}
close IN;

my $genelist=$params[0];
my $idType=$params[1];
my $dataDir=$params[2];
my $annotationFile=$params[3];
my $termFile=$params[4];
my $chipFile=$params[5];
my $chipType=$params[6];
my $identifier=$params[7];


# Get the mapping between PMID and tokens
my %pmid2term=();
open(IN, "$dataDir\/$termFile") or die "Cannot open $dataDir\/$termFile\n";
while(<IN>){
     chomp $_;
     my @col=split(/\|/, $_);
     my $pmid=$col[0];
     my $terms=$col[1];
     if($pmid and $terms){
        $pmid=~s/\s+//g;
        $pmid2term{$pmid}=$terms;
     }
}
close IN;

# Map Entrez Gene ID to PubMed ID

my %chipLL2PM=();
my %chipAFFY2LL=();
my %annChipEGID=();
```

```
open(IN, "$dataDir\/$annotationFile") or die "Cannot open $dataDir\/$annotationFile\n";
while(<IN>){
     next unless ($. >1);
     chomp $_;
     my @col=split(/\t/,$_);
     my $probesetid=lc $col[0];
     my $geneid=$col[1];
     my $pmid=$col[2];

     if($probesetid=~/^AFFX/i or $geneid eq 'NA'){
        ;
     }else{
        $chipLL2PM{$geneid}=$pmid;
        $chipAFFY2LL{$probesetid}=$geneid;
        if($pmid ne 'NA'){
           $annChipEGID{$geneid}++;
        }
     }
}
close IN;

my $ChipAllGenes = scalar keys %chipLL2PM;
my $ChipAnnotatedGenes = scalar keys %annChipEGID;


# Get Chip frequency from pre-processed data files

my %chipHits=();
open(IN, "$dataDir\/$chipFile") or die "Cannot open $dataDir\/$chipFile\n";
while(<IN>){
     chomp $_;
     my @col=split(/\t/, $_);
     my $term=$col[0];
        $term=~s/\"//g;
     my $count=$col[1];

     if($term and $count){
        $chipHits{$term}=$count;
     }
}
close IN;


# Associate the user input gene identifiers with the corresponding PMIDs

my %mygenelist=();
my $inputIDcounter=0;
open(IN, $genelist) or die "Cannot open $genelist\n";
while(<IN>){
     next if /^[ \t]*$/ ;
     chomp $_;
     my @col=split(/\t/,$_);
     my $query=lc $col[0];
        $query=~s/\s+//g;
     if($idType eq 'affy'){
        if(exists $chipAFFY2LL{$query}){
           my $queryGeneID=$chipAFFY2LL{$query};
           $mygenelist{$queryGeneID}=$chipLL2PM{$queryGeneID};
        }
     }elsif($idType eq 'egid'){
        if(exists $chipLL2PM{$query}){
           $mygenelist{$query}=$chipLL2PM{$query};
        }
     }
     $inputIDcounter++;
}
close IN;


# Build a hash where the keys are term and the values are Entrez Gene ID

my %listHits=();
my %uniquePMID=();
my $ListAllGenes=0;
my $ListAnnotatedGenes=0;

foreach my $gene (keys %mygenelist){
     my %seen=();
     my @pmid_array=split(/\s+/, $mygenelist{$gene});

     if($mygenelist{$gene} ne 'NA'){
       $ListAnnotatedGenes++;
     }

     foreach my $pubmedID (@pmid_array){
        if(exists $pmid2term{$pubmedID}){
           chomp $pmid2term{$pubmedID};
           my @term_array=split("\t",$pmid2term{$pubmedID});
           foreach my $term(@term_array){
              $listHits{uc $term}.=$gene." " unless $seen{$term}++;
           }
       }
        $uniquePMID{$pubmedID}++ if($pubmedID ne 'NA');
     }
     $ListAllGenes++;
}

my $ListPMIDCount = scalar keys %uniquePMID;


# Output token frequencies to file

my $outfile=$identifier."_PmFreq.txt";
open(OUT, ">$outfile") or die "Cannot write to $outfile\n";
print OUT "Token\tChip\tList\n";
```

```
# Loop through the listHits hash and calculate term occurrence

foreach my $word (keys %listHits){
     $word =~s/\"//g;
     $listHits{$word}=~s/\s$//;
     if(exists $chipHits{$word}){
        my $list = [split(" ", $listHits{$word}) ];
        my $wordFrequency = scalar @$list;
        print OUT "\"$word\"\t$chipHits{$word}\t$wordFrequency\n";
     }
}


# Output

print "$inputIDcounter\n";
print "$ChipAllGenes\n";
print "$ListAllGenes\n";
print "$ChipAnnotatedGenes\n";
print "$ListAnnotatedGenes\n";
print "$ListPMIDCount\n";

exit;



### Subroutine

sub getFreq{
    my ($ref,$delim) = @_;
    my @list=split($delim, $ref);
    my @uniqToken=keys %{ {map{$_,1}@list} };
    my $count=scalar @uniqToken;
    return $count;
}
```

Appendix C

Publication list

The work presented in this thesis has been published in:

Leong, H.S. and Kipling, D. (2009). Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res*, 37(11):e79.

Bibliography

Acuna, E. and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper*. Department of Mathematics, University of Puerto Rico at Mayaguez. In proceedings IPSI 2004, Venice.

Affymetrix. (2001). *Array design for the GeneChip human genome U133 set. Technical note*. Affymetrix, Santa Clara, CA.

Affymetrix. (2002). *Statistical algorithms desciption document. Technical report*. Affymetrix, Santa Clara, CA.

Affymetrix. (2005). *GeneChip exon array design. Technical note*. Affymetrix, Santa Clara, CA.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. 1st ed. John Wiley & Sons, Inc.

Alako, B.T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J. and Jenster, G. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, 6:51.

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

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578-580.

Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J.M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006). BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, 34(Web Server issue):W472-476.

Alterovitz, G., Jiwaji, A. and Ramoni, M.F. (2008). Automated programming for bioinformatics algorithm deployment. *Bioinformatics*, 24(3):450-451.

Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*, 74(12):5350-5354.

Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, 37(Database issue):D793-796.

Angelini, C., De Canditiis, D., Mutarelli, M. and Pensky, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. *Stat Appl Genet Mol Biol*, 6:Article24.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25-29.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Muller, R., Meese, E. and Lenhof, H.P. (2007). GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res*, 35(Web Server issue):W186-192.

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509-519.

Bar-Joseph, Z. (2004). Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493-2503.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data.* 3rd ed. John Wiley & Sons.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N. and Edgar, R. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885-890.

Barry, W.T., Nobel, A.B. and Wright, F.A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943-1949.

Beissbarth, T. and Speed, T.P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464-1465.

Beisvag, V., Junge, F.K., Bergum, H., Jolsum, L., Lydersen, S., Gunther, C.C., Ramampiaro, H., Langaas, M., Sandvik, A.K. and Laegreid, A. (2006). GeneTools--application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, 7:470.

Ben-Gal, I. (2005). Outlier detection. In Maimon, O. and Rockach, L. (eds), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, p. 131-146.

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):289-300.

Ben-Shaul, Y., Bergman, H. and Soreq, H. (2005). Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7):1129-1137.

Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003). Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19(18):2502-2504.

Blaschke, C., Oliveros, J.C. and Valencia, A. (2001). Mining functional information associated with expression arrays. *Funct Integr Genomics*, 1(4):256-268.

Blom, E.J., Bosman, D.W., van Hijum, S.A., Breitling, R., Tijsma, L., Silvis, R., Roerdink, J.B. and Kuipers, O.P. (2007). FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics*, 23(9):1161-1163.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267-270.

Bodenreider, O. (2006). Lexical, Terminological, and Ontological Resources for Biological text Mining. In Ananniadou, S. and McNaught, J. (eds), *Text Mining for Biology and Biomedicine*. Artech House, p. 43-66.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185-193.

Boorsma, A., Foat, B.C., Vis, D., Klis, F. and Bussemaker, H.J. (2005). T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res*, 33(Web Server issue):W592-595.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Socienty. Series B (Methodological)*, 26(2):211-252.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710-3715.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS Lett*, 480(1):17-24.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P. and Sansone, S.A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31(1):68-71.

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

Breslin, T., Eden, P. and Krogh, M. (2004). Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, 5:193.

Bruford, E.A., Lush, M.J., Wright, M.W., Sneddon, T.P., Povey, S. and Birney, E. (2008). The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res*, 36(Database issue):D445-448.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*, 8(1):R3.

Castillo-Davis, C.I. and Hartl, D.L. (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891-892.

Chaussabel, D. and Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):RESEARCH0055.

Chen, H. and Sharp, B.M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5:147.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699-705.

Curtis, R.K., Oresic, M. and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol*, 23(8):429-435.

Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 31(1):19-20.

Davies, L. and Gather, U. (1993). The Identification of Multiple Outliers. *Journal of the American Statistical Association*, 88(423):782-792.

Delongchamp, R., Lee, T. and Velasco, C. (2006). A method for computing the overall statistical significance of a treatment effect among a group of genes. *BMC Bioinformatics*, 7 Suppl 2:S11.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(5):P3.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*, 14(4):457-460.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7.

Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*. Chapman and Hall/CRC Press.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S.A. and Tainsky, M.A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, 31(13):3775-3781.

Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C. and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Res*, 17(10):1537-1545.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71-103.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919-929.

Edgar, R., Domrachev, M. and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207-210.

Edgington, E.S. and Onghena, P. (2007). *Randomization Tests*. 4th ed. Chapman & Hall/CRC.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107-129.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863-14868.

Ernst, C., Bureau, A. and Turecki, G. (2008). Application of microarray outlier detection methodology to psychiatric research. *BMC Psychiatry*, 8:29.

Ernst, J. and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7:191.

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

Faraway, J.J. (2002). *Practical Regression and Anova in R*. Available at http://www.maths.bath.ac.uk/~jjf23/book/pra.pdf.

Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fog, A. (2008). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. *Communications in Statistics: Simulation and Computation*, 37(2):241-257.

Fortin, M.-J., Jacquez, G.M. and Shipley, B. (2002). Computer-intensive methods. In El-Shaarawi, A. and Piegorsch, W., W. (eds), *Encyclopedia of Environmetrics*. Wiley, Chichester, p. 399-402.

Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., Polman, J. and Alkema, W. (2008). CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res*, 36(Web Server issue):W406-410.

Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*:707-718.

Gibbons, J.D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*. 4th ed. CRC Press LLC.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y. and De Moor, B. (2003). Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*:391-402.

Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y. and De Moor, B. (2004). TXTGate: profiling gene groups with text-based information. *Genome Biol.*, 5(6):R43.

Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93-99.

Goeman, J.J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980-987.

Gold, D.L., Coombes, K.R., Wang, J. and Mallick, B. (2007). Enrichment analysis in high-throughput genomics - accounting for dependency in the NULL. *Brief Bioinform*, 8(2):71-77.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531-537.

Good, P.I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 3rd ed. Springer.

Grossmann, S., Bauer, S., Robinson, P.N. and Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, 23(22):3024-3031.

Gupta, P., Yoshida, R., Imoto, S. and Miyano, S. (2007). Statistical absolute evaluation of gene ontology terms with gene expression data. In, *Proc. 3rd International Symposium on Bioinformatics Research and Applications, Lecture Note in Bioinformatics*. Springer-Verlag, p. 146-157.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):761-771.

Hampel, F.R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887-1896.

Hardiman, G. (2004). Microarray platforms--comparisons and contrasts. *Pharmacogenomics*, 5(5):487-502.

Harkness, W.L. (1965). Properties of the Extended Hypergeometric Distribution. *The Annals of Mathematical Statistics*, 36(3):938-945.

Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, 17 Suppl 1:S97-106.

Hawkins, D.M. (1980). *Identification of outliers*. Chapman and Hall.

Hoaglin, D.C., Iglewicz, B. and Tukey, J.W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991-999.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800-803.

Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nat Genet*, 36(7):664.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C. and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.*, 4(10):R70.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. and Lempicki, R.A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*, 35(Web Server issue):W169-175.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1-13.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44-57.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109-126.

Iglewicz, B. and Hoaglin, D.C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249-264.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J., Jr., Boguski, M.S., Lashkari, D., Shalon, D., Botstein,

D. and Brown, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83-87.

Jelier, R., Jenster, G., Dorssers, L.C., Wouters, B.J., Hendriksen, P.J., Mons, B., Delwel, R. and Kors, J.A. (2007). Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics*, 8:14.

Jensen, L.J., Saric, J. and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7(2):119-129.

Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21-28.

Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, 23(3):306-313.

John, J.A. and Draper, N.R. (1980). An alternative family of transformations. *Applied Statistics*, 29(2):190-197.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. 3rd ed. Wiley and Sons.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428-432.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27-30.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480-484.

Khatri, P., Draghici, S., Ostermeier, G.C. and Krawetz, S.A. (2002). Profiling gene expression using onto-express. *Genomics*, 79(2):266-270.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587-3595.

Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180-182.

Kim, S.B., Yang, S., Kim, S.K., Kim, S.C., Woo, H.G., Volsky, D.J., Kim, S.Y. and Chu, I.S. (2007). GAzer: gene set analyzer. *Bioinformatics*, 23(13):1697-1699.

Kim, S.Y. and Volsky, D.J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144.

Krallinger, M., Erhardt, R.A. and Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today*, 10(6):439-445.

Krauthammer, M., Rzhetsky, A., Morozov, P. and Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1-2):245-252.

Krauthammer, M., Kaufmann, C.A., Gilliam, T.C. and Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci U S A*, 101(42):15148-15153.

Kruskal, W. (1988). Miracles and Statistics: The Casual Assumption of Independence. *Journal of the American Statistical Association*, 83(404):929-940.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O. and Davis, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A*, 94(24):13057-13062.

Lee, H.K., Braynen, W., Keshav, K. and Pavlidis, P. (2005a). ErmineJ: tool for functional analysis of gene expression data sets. 6:269.

Lee, J.S., Katari, G. and Sachidanandam, R. (2005b). GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, 6:189.

Lee, M.S., Hanspers, K., Barker, C.S., Korn, A.P. and McCune, J.M. (2004). Gene expression profiles during human CD4+ T cell differentiation. *Int Immunol*, 16(8):1109-1124.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*, 2(8):RESEARCH0032.

Liang, P. and Pardee, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257(5072):967-971.

Lindsay, R. and Gordon, M. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 50(7):574-587.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20-24.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675-1680.

MacCallum, R.M., Kelley, L.A. and Sternberg, M.J. (2000). SAWTED: structure assignment with text description--enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16(2):125-129.

Maere, S., Heymans, K. and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448-3449.

Manly, B.F.J. (1976). Exponential data transformation. *The Statistician*, 25(37-42).

Manly, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Third ed. Chapman & Hall/CRC.

Manoli, T., Gretz, N., Grone, H.J., Kenzelmann, M., Eils, R. and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22(20):2500-2506.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D. and Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 5(12):R101.

Marton, M.J., DeRisi, J.L., Bennett, H.A., Iyer, V.R., Meyer, M.R., Roberts, C.J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D.E., Jr., Hartwell, L.H., Brown, P.O. and Friend, S.H. (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med*, 4(11):1293-1301.

Masseroli, M., Galati, O. and Pinciroli, F. (2005). GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*, 33(Web Server issue):W717-723.

Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M. and Ward, J.M. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc*:460-464.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267-273.

Muller, H.M., Kenny, E.E. and Sternberg, P.W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309.

Nam, D. and Kim, S.Y. (2008). Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189-197.

Nelson, S.J., Schopen, M., Savage, A.G., Schulman, J.L. and Arluk, N. (2004). The MeSH translation maintenance system: structure, interface design, and implementation. *Stud Health Technol Inform*, 107(Pt 1):67-69.

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):R81.

Nishimura, M.T., Stein, M., Hou, B.H., Vogel, J.P., Edwards, H. and Somerville, S.C. (2003). Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science*, 301(5635):969-972.

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res*, 37(Web Server issue):W317-322.

Nueda, M.J., Conesa, A., Westerhuis, J.A., Hoefsloot, H.C., Smilde, A.K., Talon, M. and Ferrer, A. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics*, 23(14):1792-1800.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T.F., Rezwan, F., Sharma, A., Williams, E., Bradley, X.Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S.G., Rocca-Serra, P., Sansone, S.A., Sklyar, N., Zhao, M., Sarkans, U. and Brazma, A. (2009). ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868-872.

Pasquier, C., Girardot, F., Jevardat de Fombelle, K. and Christen, R. (2004). THEA: ontology-driven analysis of microarray data. *Bioinformatics*, 20(16):2636-2643.

Pavlidis, P., Qin, J., Arango, V., Mann, J.J. and Sibille, E. (2004). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem.Res.*, 29(6):1213-1222.

Pericchi, L.R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, 68:35-43.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population. *Royal Statistical Society B*, 4:119-130.

Platanias, L.C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat Rev Immunol*, 5(5):375-386.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V.V. and Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform Ser Workshop Genome Inform*, 9:72-80.

Raychaudhuri, S., Schutze, H. and Altman, R.B. (2002). Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12(10):1582-1590.

Raychaudhuri, S., Chang, J.T., Imam, F. and Altman, R.B. (2003). The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res*, 31(15):4553-4560.

Raychaudhuri, S. (2006). *Computational Text Analysis: for functional genomics and bioinformatics*. 1st ed. OUP Oxford.

Reimand, J., Kull, M., Peterson, H., Hansen, J. and Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(Web Server issue):W193-200.

Rhee, S.Y., Wood, V., Dolinski, K. and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 9(7):509-515.

Rindflesch, T.C., Tanabe, L., Weinstein, J.N. and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput*:517-528.

Rivals, I., Personnaz, L., Taing, L. and Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401-407.

Robinson, M.D., Grigull, J., Mohammad, N. and Hughes, T.R. (2002). FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 3:35.

Roff, D.A. (2006). *Introduction to Computer-Intensive Methods of Data Analysis in Biology*. 1st ed. Cambridge University Press.

Rubinstein, R. and Simon, I. (2005). MILANO--custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics*, 6:12.

Sakia, R.M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, 41(2):169-178.

Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J., Conklin, B.R. and Pico, A.R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217.

Sanda, C., Weitzel, P., Tsukahara, T., Schaley, J., Edenberg, H.J., Stephens, M.A., McClintick, J.N., Blatt, L.M., Li, L., Brodsky, L. and Taylor, M.W. (2006). Differential gene induction by type I and type II interferons and their combination. *J Interferon Cytokine Res*, 26(7):462-472.

Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I. and Bork, P. (2006). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22(6):645-650.

Saxena, V., Orgill, D. and Kohane, I. (2006). Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res*, 34(22):e151.

Schadt, E.E., Li, C., Su, C. and Wong, W.H. (2000). Analyzing high-density oligonucleotide gene expression array data. *J Cell Biochem*, 80(2):192-202.

Scheer, M., Klawonn, F., Munch, R., Grote, A., Hiller, K., Choi, C., Koch, I., Schobert, M., Hartig, E., Klages, U. and Jahn, D. (2006). JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res*, 34(Web Server issue):W510-515.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467-470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A*, 93(20):10614-10619.

Schuemie, M.J., Weeber, M., Schijvenaars, B.J., van Mulligen, E.M., van der Eijk, C.C., Jelier, R., Mons, B. and Kors, J.A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597-2604.

Sealfon, R.S., Hibbs, M.A., Huttenhower, C., Myers, C.L. and Troyanskaya, O.G. (2006). GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, 7:443.

Shah, N.H. and Fedoroff, N.V. (2004). CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, 20(7):1196-1197.

Shalon, D., Smith, S.J. and Brown, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 6(7):639-645.

Shatkay, H., Edwards, S., Wilbur, W.J. and Boguski, M. (2000). Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:317-328.

Smalheiser, N.R. and Swanson, D.R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57(3):149-153.

Smid, M. and Dorssers, L.C. (2004). GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics*, 20(16):2618-2625.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.

Somogyi, R., Wen, X., Ma, W. and Barker, J.L. (1995). Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. *J Neurosci*, 15(4):2575-2591.

Srinivasan, P. (2004). Text mining:Generating hypothesis from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55:396-413.

Stapley, B.J., Kelley, L.A. and Sternberg, M.J. (2002). Predicting the sub-cellular location of proteins from text using support vector machines. *Pac Symp Biocomput*:374-385.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc.Natl.Acad.Sci.U.S.A*, 102(43):15545-15550.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, 30(1):7-18.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907-2912.

Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. 27(6):1210-1217.

Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75-82.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281-285.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. *Pac Symp Biocomput*:541-552.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38):13544-13549.

Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2-8.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644-648.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116-5121.

Usadel, B., Nagel, A., Steinhauser, D., Gibon, Y., Blasing, O.E., Redestig, H., Sreenivasulu, N., Krall, L., Hannah, M.A., Poree, F., Fernie, A.R. and Stitt, M. (2006). PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics*, 7:535.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530-536.

Vanharanta, S., Pollard, P.J., Lehtonen, H.J., Laiho, P., Sjoberg, J., Leminen, A., Aittomaki, K., Arola, J., Kruhoffer, M., Orntoft, T.F., Tomlinson, I.P., Kiuru, M., Arango, D. and Aaltonen, L.A. (2006). Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. *Hum Mol Genet*, 15(1):97-103.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484-487.

Vencio, R.Z., Koide, T., Gomes, S.L. and Pereira, C.A. (2006). BayGO: Bayesian analysis of ontology term enrichment in microarray data. *BMC Bioinformatics*, 7:86.

Vencio, R.Z. and Shmulevich, I. (2007). ProbCD: enrichment analysis accounting for categorization uncertainty. *BMC Bioinformatics*, 8:383.

Wain, H.M., Lush, M., Ducluzeau, F. and Povey, S. (2002). Genew: the human gene nomenclature database. *Nucleic Acids Res*, 30(1):169-171.

Waring, J.F., Ciurlionis, R., Jolly, R.A., Heindel, M. and Ulrich, R.G. (2001). Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. *Toxicol Lett*, 120(1-3):359-368.

Weeber, M., Klein, H., Berg, L. and Vos, R. (2001). Using Concepts in Literature-based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. *Journal of the American Society for Information Science*, 52(7):548-557.

Weiss, S.M., Indurkhya, N., Zhang, T. and Damerau, F. (2004). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.

Wheelan, S.J., Martinez Murillo, F. and Boeke, J.D. (2008). The incredible shrinking world of DNA microarrays. *Mol Biosyst*, 4(7):726-732.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Ostell, J., Pruitt, K.D., Schuler, G.D., Shumway, M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 36(Database issue):D13-21.

Wilbur, W.J. and Coffee, L. (1994). The effectiveness of document neighboring in search enhancement. *Inf Process Manag*, 30:253-266.

Wrobel, G., Chalmel, F. and Primig, M. (2005). goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, 21(17):3575-3577.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. and Weinstein, J.N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28.

Zhang, B., Schmoyer, D., Kirov, S. and Snoddy, J. (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16.

Zheng, Q. and Wang, X.J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res*, 36(Web Server issue):W358-363.

Zhou, X. and Su, Z. (2007). EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, 8:246.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K.B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358-375.