A NOVEL METHOD FOR INTEGRATIVE BIOLOGICAL STUDIES

Abdullatif Sulaiman Al Watban

submitted to the University of Exeter
as a thesis for the degree of

Doctor of Philosophy in Biological sciences

April 2016

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signed)…… Abdullatif Al Watban

Abstract

DNA microarray technology has been extensively utilized in the biomedical field, becoming a standard in identifying gene expression signatures for disease diagnosis/prognosis and pharmaceutical practices. Although cancer research has benefited from this technology, challenges such as large-scale data size, few replicates and complex heterogeneous data types remain; thus the biomarkers identified by various studies have a small proportion of overlap because of molecular heterogeneity. However, it is desirable in cancer research to consider robust and consistent biomarkers for drug development as well as diagnosis/prognosis. Although cancer is a highly heterogeneous disease, some mechanism common to developing cancers is believed to exist; integrating datasets from multiple experiments increases the accuracy of predictions because increasing the sample size improves and enhances biomarkers detection. Therefore, integrative study is required for compiling multiple cancer data sets when searching for the common mechanism leading to cancers.

Some critical challenges of integration analysis remain despite many successful methods introduced. Few is able to work on data sets with different dimensionalities. More seriously, when the replicate number is small, most existing algorithms cannot deliver robust predictions through an integrative study. In fact, as modern high-throughput technology matures to provide increasingly precise data, and with well-designed experiments, variance across replicates is believed to be small for us to consider a mean pattern model. This model assumes that all the genes (or metabolites, proteins or DNA copies) are random samples of a hidden (mean pattern) model. The study implements this model using a hierarchical modelling structure. As the primary component of the system, a multi-scale Gaussian (MSG) model, designed to identify robust differentially-expressed genes to be integrated, was developed for predicting differentially expressed genes from microarray expression data of small replicate numbers. To assure the validity of the mean pattern hypothesis, a bimodality detection method that was a revision of the Bimodality index was proposed.

Acknowledgments

Throughout my PhD studies I am fortunate to have benefited from the help and support of many people. First and foremost, I thank my supervisor Dr Zheng Ron Yang, of the School of Biosciences at the University of Exeter for his inspiration, guidance and insight, and for his continuous support and encouragement through all the stages of my research. His inspired teaching has helped me to achieve this thesis, and my four years of study in the field of Bioinformatics at Exeter University have been the most rewarding and enjoyable in my student life. Also I express my thanks to Dr Nicholas Harmer for his role as a second supervisor.

I would especially like to thank Dr Zihua Yang of the Wolfson Institute of Preventive Medicine at Queen Mary University of London, for the valuable discussion and ideas she gave me, especially for the MSG and HIM development processes. I also thank Professor Richard Everson in the Department of Computer Science at Exeter University for his help and suggestions for MSG design.

I acknowledge with gratitude Saudi Food and Drug Authority, who funded/sponsored me during the period of my PhD study.

Finally, I am deeply grateful to my family. My father Sulaiman Alwatban and my mother Mankiaah Alnowifa have always understood, encouraged and supported me in my endeavours, as have my brothers and sisters who have great confidence in my studies. To my wife, Reufe Alhogbani, and my precious sons Abdulrahman and Nawaf, I give my heartfelt thanks for their endless patience and care at various critical points when I have had to spend less time with them. This project could not have succeeded without their understanding and tolerance and words are inadequate to express how much I appreciate all that they have done during the past four years

Table of Contents

List of Abbreviations

GEO	Gene Expression Omnibus
DEGs	Differentially Expressed Genes
SAM	Significance Analysis of Microarrays
SVM	Support Vector Machine
SOM	Self-Organizing Map
AML	Acute Myeloid Leukemia
ALL	Acute Lymphoblastic Leukemia
PCA	Principle Component Analysis
NMF	Non-negative Matrix Factorization
MSG	Multi-Scale Gaussian
***h*BI**	Heterogeneous Bimodality Index
HIM	Hierarchical Integration Model
CSMSG	Cross-Species Multi-Scale Gaussian
MVMSG	Multivariate Multi-Scale Gaussian
MDI	Multiple Data Integration
BCC	Bayesian Consensus Clustering
ROC	Receiver Operating Characteristic
AUC	Area Under ROC Curve
FDR	False Discovery Rate
BI	Bimodality Index
LHR	Likelihood Ratio test
COPA	Cancer Outlier Profile Analysis
OS	Outlier Sum statistics
ORT	Outlier Robust T-test
PACK	Profile Analysis using Clustering and Kurtosis
IQR	interquartile
EM	Expectation Maximisation
BIC	Bayesian Information Criterion
AIC	Akaike's Information Criterion

BC	Besag's sequential Monte Carlo (Besag and Clifford algorithm)
sep	specificity
sen	sensitivity
DE	Differential Expression
FPD	Full Pattern Discovery
PPD	Partial Pattern Discovery

List of Tables

List of Figures

Chapter 1

Introduction and Overview

Abstract

This chapter briefly introduces the methods used to analyse gene expression microarrays and analyses their advantages as well as their limitations. In particular, I introduce three statistical tools for data mining that I have developed during my PhD study. These tools constitute a complete system for an integrative study of microarrays. Further, I outline the thesis and show my contributions to the area.

1.1 Microarrays and Cancer Research

This section gives basic information that is important to understand the rest of this thesis. I begin with a brief introduction of gene expression and the technologies used to generate expression data. Then I highlight its applications in cancer research.

1.1.1 The Basics of Microarray

The central dogma of molecular biology [1] explains the flow of genetic information from DNA into RNA (transcription) and from RNA into protein (translation). Gene expression is the quantitative measure of the amount of transcribed mRNA (level of transcript) produced by the gene's of interest. The study of gene expression can reveal the function of genes associated with a disease. Many technologies have been used to measure/analyse gene expression. In the past, Northern Blot analysis was used to quantify mRNA levels [2], as was differential display [3]. However, these methods can only analyse small number of genes at a time and do not provide much information pertaining to gene-gene interaction. To avoid these limitations, researchers have moved to using DNA microarrays [4, 5] and Serial Analysis of Gene Expression (SAGE) [6]. Unlike northern blots, microarrays and SAGE technologies allow measurement of the abundance of tens of thousands of transcripts in a tissue sample concurrently. The major difference between the two technologies is that microarrays require *a prior* knowledge of the sequence. Generally, a DNA microarray is a collection of ordered sets of DNA spots/fragments of known sequences on a solid surface [7]. The DNA microarrays can be classified into two main types based on the DNA fragments designed.

1.1.1.1 Spotted microarrays

Spotted microarrays or complementary DNA arrays (cDNA) [5] consist of long cDNA molecules prepared from a cDNA library and attached to a solid surface (i.e. glass slide) by robotic arm [7]. This platform has two channels. It measures the expression profile of genes from two different cell samples simultaneously (e.g., control and disease).The extracted RNAs from a pair of samples are reverse transcribed into cDNA by a polymerisation reaction and labelled using two fluorescent dyes (Red and Green) as illustrated in **Figure 1.1.a.** Then RNAs from the two samples are combined and placed onto the glass slide (array) to hybridise to their complementary probes. After this process, the microarrays are washed and dried to remove any non-specific hybridization. Finally the hybridised microarray is scanned to determine fluorescent dye intensities using two laser sources with two different wavelengths to stimulate each fluorescent dye, to illuminate each probe and both targets (control and disease condition) will fluoresce. The ratio of green and red fluorescence indicates the expression level of the gene corresponding to that probe (i.e. relative gene expression). This done by different imaging processes on the two images produced. First, identify the local background intensity and the spot intensity by taking either mean or median intensity of pixels, which represents the intensity of fluorescence determined by focusing the laser beam on that point of the array.

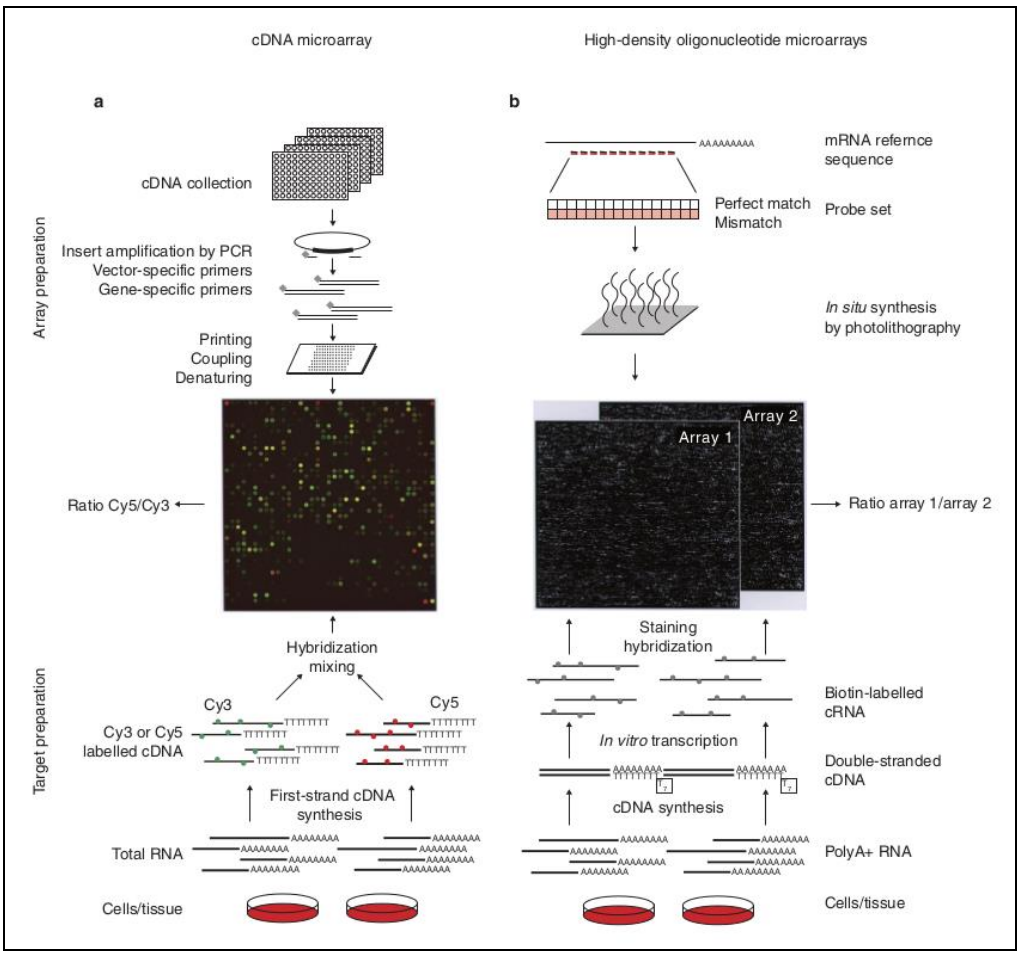


Figure1.1. Schematic illustrates the microarray workflow. (**a)** cDNA microarray, and (**b**) oligo microarray[8].

1.1.1.1 Oligonucleotide microarrays

The other major type of microarray is oligonucleotide arrays or single channel arrays which have been used widely [4] - commonly known as Affymetrix gene chip (**Figure1.1 b**). This is slightly different than the previous one where it uses a single colour designed to detect the expression of one tissue (i.e cancer) at a time. The probes of this type are made from short (segments of cDNA) oligonucleotides (20~25 nucleotides) that are synthesized in situ using photolithography on the microarray plate [7, 8]. In oligonucleotide arrays, each gene is represented by a group of probes called a probe-set in order to eliminate the systemic errors. There are usually 11or 16 (or 10-20) probe-pairs. One is a perfect match (PM) that is identical to the reference sequence and the other is a mismatch (MM) which differs from the PM only in the middle base

(base 13), in order to quantify the non-specific hybridization. The principle of this technology is slightly different to that described above - in the preparation phase the side with 11-20 short oligos with (25 mers) are synthesized in situ using light into the array/surface. Target preparation is done on the opposite side, and differs in that RNA extracted from the tissue is converted to cDNA then labeled and hybridized to the array. The binding between target and prepared is detected by a single fluorescent dye to calculate the absolute abundance. The final step - the image processing – functions as per the spotted array. The expression levels are estimated for each pair of probes as the difference between PM and MM. Finally, the gene expression is the average of the k^{th} expression values obtained from the different pairs. It's worth noting that there are other technologies that been employed in Agilent microarrays, such as ink-jet technology.

1.1.1.3 Normalisation

The most important step after obtaining raw data, from the above mentioned technologies, is the normalisation. Normalisation is applied to microarrays in order to remove any variations caused by microarray technology itself rather than the biological difference between different experimental conditions, thus allowing fair comparison between different experimental conditions. In this section the most common approaches will be described.

The MAS5 algorithm was developed by Affymetrix, it normalises individual arrays independently and sequentially. MM probes are taken into account to estimate average expression by subtracting MM probe value from the PM probe value (i.e. PM-MM).

RMA is a popular normalisation technique, which uses multi chip. In contrast to

MAS5, RMA does not uses MM probe values in its normalisation step because most often the MM probes values are higher than that of PM probe values, which signifies that MM probes as an unreliable measure for non-specific binding.

1.1.2 The Applications of Microarrays

Gene expression has been widely used for research in the biomedical field and many studies have employed it to identify biomarkers in complex diseases such as cancer. Because microarrays can measure expression levels of mRNA for tens of thousands of genes across samples simultaneously, the technology has contributed significantly to the identification of signatures that distinguish between normal and cancerous tissues, between different drug responses and more. The most obvious applications for microarrays include gene discovery, sample classification and stress analysis. In the next few paragraphs, the importance of microarrays in cancer research will be highlighted with a few examples from the literature pertaining to classification, diagnosis and future prediction.

Golub et al. demonstrated a good example of the importance of microarrays in cancer research [9]. They analysed 38 acute leukaemia cancer samples and divided them into two types; acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) using self-organising map (SOM). Also they identified 50 genes, as predictors, which showed a significant expression difference between the two types. They used these genes to assign classes to 34 unclassified samples and they achieved a high accuracy using the cross validation techniques [9]. Another study also used microarray data to distinguish between BRCA1 and BRCA2 mutations in breast cancer. They identified a

subset of genes that can accurately differentiate between the two BRCA types [10].

In a similar study, Ramaswamy et al. identified 128 genes that are differentially expressed between the primary and metastatic tumours [11]. However, they used them as predictor for the outcomes/future of different primary tumours where the tumours that have similar expression pattern genes as the metastatic genes are likely to become metastatic and therefore have worse outcomes.

1.1.3 Microarray databases

The abundance of data resulting from microarrays requires automated approaches to analyse and store. As a result, there have been many projects to develop microarray repositories and tools that enable researchers to communicate their research results [12]. Public microarray data repositories such as GEO [13, 14], Array Express [15] and others hold hundreds of thousands of arrays from multiple species and have expanded rapidly in the last ten years as seen in **Figure1.2**. For example, GEO has collected 1,661,882 samples from 64,516 series that performed on 15,318 platforms (http://www.ncbi.nlm.nih.gov/geo/summary/ **;** accessed on 05/01/2016). The most widely known of these platforms are affymetrix and illumine arrays. The huge available data has enormous potential on researches in finding solid results by integrating different data coherent of existing gene expression data.

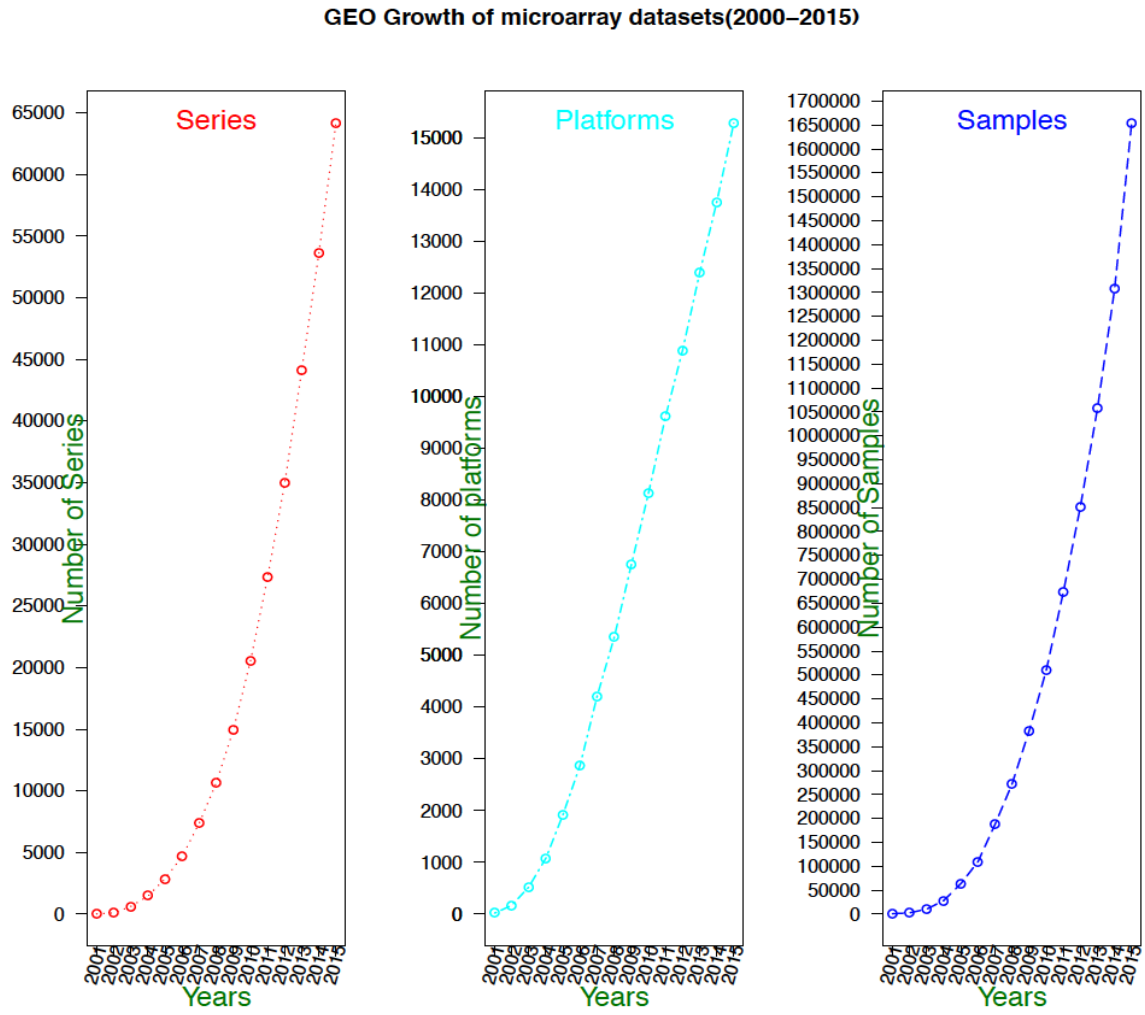


Figure 1.2: GEO Microarray data growth 2000-2015; Data for series, samples and platforms.

1.1.4 Example of gene expression matrix

Gene expression data are usually presented in a matrix where each column represents the gene expression level of a single experiment (microarrays) and the rows represent the genes. Table 1.1 shows an example of a real gene expression data set that downloaded from GEO with accession number (GDS3138). In this example the first column represent the probe id (ID_ref) while the second is the gene symbol and the first two columns after are the expression level in the reference samples and the last two for the testing sample.

Table 1.1: Example of a dataset the database from GEO with accession number (GDS3138) that contains 2 control samples and corresponding samples of a homo species; in GPL570 array platform.

ID_REF	IDENTIFIER	GSM 242136	GSM 242196	GSM 242207	GSM 242208
1007_s_at	DDR1	1147.65	1101.6	645.111	529.475
1053_at	RFC2	1168.96	1294.58	2091.24	2057.69
117_at	HSPA6	101.501	113.92	112.763	159.533
121_at	PAX8	1719.46	1444.92	1481.32	1385.93
1255_g_at	GUCA1A	40.9801	41.6835	31.2603	15.6592
1294_at	UBA7	311.364	336.589	270.467	292.845
1316_at	THRA	135.231	119.512	139.73	114.26
1320_at	PTPN21	172.231	214.62	105.47	117.146
1405_i_at	CCL5	2.89906	2.74331	7.4646	3.73164
1431_at	CYP2E1	166.517	149.555	86.142	95.9834
1438_at	EPHB3	77.2017	75.2635	44.0401	32.7759
1487_at	ESRRA	850.326	720.911	690.522	694.261
1494_f_at	CYP2A6	383.847	296.31	208.898	226.698
1552256_a_at	SCARB1	1047.55	997.821	1005.8	850.118
1552257_a_at	TTLL12	1199.67	1062.9	1392.33	1223.25
1552258_at	LINC00152	375.005	339.095	246.631	309.942
1552261_at	WFDC2	47.5678	64.6362	29.0703	54.8255
1552263_at	MAPK1	363.047	231.802	319.251	392.226

1.1.5 The drawback of Microarrays and the future of gene expression profiling

One of the limitations of microarray technology is the noise obtained from experimental design. The noise results from many sources; PCR and scanners or as result of using different protocols or sample tissues[16]. This noise decreases the quality of expression quantification especially when transcripts are presented in low levels [17]. The low levels may obtained as a result of non-specific binding and/or optical noise. Microarrays are designed with hybridization probes that are based on *a priori* knowledge of sequences, as explained previously. Therefore, they cannot be used to identify novel genes or

transcripts or understand structural variations. Another issue is cross-hybridization, where multiple probes target a single sequence or vice versa and influence the analysis of microarray data [18]. In addition, the comparison of different expression data obtained from different labs requires careful considerations, such as normalization, before analysis.

Although microarrays are still used widely, a new promising technology has emerged that uses next-generation sequencing. Recent developments in high-throughput sequencing have led to the birth of RNA-Sequencing. RNA-Seq is a method for quantifying transcriptomes, the complete set of transcripts in a cell[19]. One benefit of this is that it does not require *a priori* knowledge of probe design as required by microarrays. RNA-Seq is dependent on short read mapping that has single base resolution [19]. RNA-Seq sequences the transcripts directly allowing discovery of novel transcripts and to answer questions regarding the transcriptome[20]. Despite the advantages of RNA-Seq, Microarray technology is still widely used and remains the most popular approach for gene expression analysis. The main reasons for that are the lower cost, the matured analysis techniques and the availability of data. In addition, several studies showed a considerable overlapping of findings between the two platforms [20-23]. On the other hand, RNA-seq has solved most of the technical issues raised by using microarrays. It has the ability to detect the low abundance transcripts, isoforms and genetic variation [20]

1.2 The Analysis Gene Expression Microarrays.

Many data analysis techniques have been developed and applied to gene expression research. One of the most important gene expression analyses is identifying differentially expressed genes (DEGs). A gene is classified as a DEG

if it shows a significant difference in distribution of expression levels between two conditions for instance, normal versus cancer patients. The *t*-test is commonly used to detect DEGs in a gene expression data set between two conditions, while other approaches can also be used - such as ANOVA – where there are more than two conditions. The Wilcoxon test [24] is used in case of multiple conditions for the same participants. More advanced methods have been developed to tackle some common issues in microarray analysis, such as small sample size. For example, Significance Analysis of Microarrays (SAM) [25] has become a favourite choice in microarray analysis studies [26-34]. Other techniques have benefited from using a Bayesian approach, including Cyber-T [35], the random variance model [36], Limma [37], Varmixt [38], SMVar [39] and ROAST [40].

Another field of microarray analysis uses supervised and or unsupervised learning. Classification and prediction are types of supervised learning. The aim of classification is to benefit from the class label of samples and to identify gene expression pattern that form classification. There are many algorithms proposed for classification, such as support vector machine (SVM) and k-nearest neighbours, where their performance for this kind of analysis depends on the quality of training data. A further important tool for microarray analysis is clustering, which is unsupervised. It aims to assemble genes or samples into different groups based on their similarities. There are also many algorithms proposed for clustering, including hierarchical clustering [41], the k-means algorithm [42], model-based clustering and self-organizing maps [43]. Clustering allows the identification of complex structures or relations between objects (i.e. genes) and does not require *a priori* knowledge of them. However, a microarray data set has thousands of genes between which many

similarities/relationships are possible. Thus clustering does not always give easily interpretable information. For example, including a whole genome for clustering in a study where about 22,000 genes are clustered into either a large number of clusters or large number of genes in each cluster. This makes the analysis of such data tedious and requires more manual work. In addition, it is difficult to estimate cluster number - a well-known issue in clustering and has become a hot topic in machine learning. Despite these issues it's still a useful tool for many applications, including gene expression data, and has been used widely.

More recently, researchers have integrated data to enhance the results obtained by the methods mentioned above. One reason for this is the different results achieved by different methods. The integration analysis has helped the cross-species studies. The cross-species study compares multiple data sets from biological or medical experiments, revealing how genes are conserved among distantly related species [44]. Integration analysis uses either univariate or multivariate statistical analysis methods. The univariate methods include correlation analysis, mutual information and the Fisher combined probability test, as well as a modified *t* test. For instance, correlation analysis and mutual information have been used to examine how a single gene shows differential expression across species [45-49]. When the number of species is large, the Fisher combined probability test [50] has been used [51-53], as has one of the modified *t*-test algorithms such, eBayes [54-56].

The multivariate methods used in integrative studies include non-negative matrix factorization [57, 58], the k-means algorithm [42], mixture models [59] and self-organizing maps [43]. When a multivariate analysis method is used, the focus is mainly on how genes/samples from different species can be clustered

to form meta-knowledge (such as meta-genes or sub-populations) for further examination [60-72].

1.3. Motivation

DNA microarray technology has been extensively utilized in the biomedical field and has become a standard practice for identifying gene expression signatures for disease diagnosis and prognosis as well as pharmaceutical practices. Although cancer research benefits from this technology [73-80], challenges such as large data size, few replicates and the complexity of heterogeneous data types are still present. The biomarkers identified by different studies have therefore a small proportion of overlap due to molecular heterogeneity [81-83]. However, in cancer research, it is desirable to consider robust and consistent biomarkers for drug development as well as diagnosis and prognosis [84-86]. Although cancers are highly heterogeneous diseases, it is believed that some mechanism exists common to developing cancers [87] and that integrating datasets from multiple experiments can reduce prediction bias [88].Integrative approaches were therefore used for comparing multiple cancer data sets in the search for the common mechanism leading to cancers [89].

Many methods have been introduced for integration-analysis. One category works mainly on statistics acquired from individual analysis. First, combining *p* values which has been widely used due to its simplicity - the *p* values derived from two different technologies can simply be combined [90, 91]. Weighted Fisher's method has also been used for integrative studies [90-92]. One issue of applying Fisher's method is that it is biased if one *p* value is extremely small while others are not. Second, a considerable number of studies have focused on a more effective system of meta-analysis that combines effective sizes in their multiple data set analysis [93-95]. The effective size is the standardized

mean between two phenotypes (i.e. cancer/normal) in each set and then combining them as an overall mean. This method has improved the capability to model inter-study variation [93-96]. Third, ranked gene lists, has attracted some researchers for its ability to overcome the effect of outliers; a problem for the previous two methods [74, 97-99].

Another category that has recently received attention is clustering and or decomposing. Methods include principal component analysis (PCA) and non-negative matrix factorization (NMF) [100, 101]. Comparatively speaking, NMF performs better than PCA. NMF [57, 58] is a very successful tool for dimension reduction and visualization [102]. It decomposes an expression matrix into two matrices, one of which contains meta-gene expression while the other is a coefficient matrix that captures the relationship between genes and meta-genes. The main disadvantage of NMF is that it fails to deal with complex patterns across multiple datasets since it is a linear algorithm; in addition all entries must be positive. Consequently, it is not applicable to a differential expression matrix which contains negative values.

It has also been proposed that iCluster [103] could cluster samples of different genomic data by using the joint latent variable/factor analysis of a combined matrix. The main idea is that the latent variable model attempts to establish a linear correlation between the data and the latent variables. The latent variables are regarded as the cluster labels and can explain the co-clusters among integrated sets. However, this is designed to cluster samples requiring the same sample size across sets. More recently, an extension of PCA, called joint and individual variations explained, has been suggested, which breaks down the data into three components; one contains the structural information across datasets, with another for each dataset and the noise [104]. Despite the

successes of these methods they are still sensitive to excessive noise. Bayesian methods obviously make the integration of biological data robust as this can be done through a proper design of prior densities [105]. Bayesian models can also be used to integrate multiple data sources for gene function prediction [106] and for differential-expressed gene prediction [101, 107, 108], as well as for identifying biomarkers associated with clinical outcomes [109].

Even though many successful methods have been introduced, various critical challenges remain. For example, none of the described approaches is able to work on data sets with different dimensionalities. The dimensionalities include number of genes and number of samples. For example, directly merging two datasets requires the same genes/probes. Other studies require the same sample sizes among the different data to be integrated, i.e., subtype discoveries. More seriously, when the replicate number is small, most existing algorithms are unable to deliver robust predictions through an integrative study. In fact, with the gradual maturing of modern high-throughput technology which is providing increasing volumes of precise data, and when experiments are well designed, it is believed that variance across replicates is sufficiently small.

This has led me to consider a mean pattern model, through which I assume that all the genes (or metabolites, proteins or DNA copies) are in fact random samples of a hidden (mean pattern) model. In this thesis, I have implemented a mean pattern model using a hierarchical modelling structure for integration. However, it is important to remove bimodal genes from the basic idea of my proposed method, to ensure that a valid hypothesis is applied.

The majority of gene expression patterns follow a normal distribution [110]. However, a large number of genes, especially in cancer, appear to have an

expression pattern that follows a two or more component mixture distribution [111]. The definition of bimodality is a continuous probability distribution with two distinct modes/peaks; these modes/peaks would usually be high and low expressing values. Several algorithms have been proposed [110, 112, 113] to predict bimodal genes with success. Occasionally, however, their performance is not very satisfactory. As a result, I have developed an algorithm to identify such genes with bimodal characteristics.

Another important component in my integration system is the identification of differentially expressed genes, since this will enhance the analysis by integrating the most important genes for further investigation. In addition, it will increase the power of clustering and make the results interpretable. To detect differential expressed genes in a gene expression data set for a medical/biological investigation across two experiments, the *t*-test is commonly used. Other approaches, e.g., ANOVA in cases of multiple conditions. Alternatively, Significance Analysis of Microarrays (SAM) [25] is widely used for microarray analysis [26-34]. Rather than using non-parametric approaches, theoretical modelling using a Bayesian approach was also applied ([36, 37]; [38]; [39]

The success of variance-adjustment methods such as SAM depends largely on an accurate identification of a subset of similar genes from the same or different data sets. However, there is inherent selection bias with the variance-adjustment approach [114]. When genes have different variances, pinpointing similar genes can become tricky, thus affecting differential expressed gene-detection performance. The performance of model-based algorithms, such as eBayes, is also insufficient when distribution of the variances within and across conditions is markedly different. Thus, I devised a new method to overcome

such limitations and also to make it possible to work with small samples of replicated data, which is common in gene expression data.

1.4. Objectives and Contribution

This thesis aimed to investigate and develop algorithms, which would address the limitations of current microarray data analysis from differential gene expression and bimodal gene identification to comprehensive analysis through integration. The objectives of this work were to develop analytical tools towards integrating data of different sizes (i.e. number of genes/number of sample) and the ability to discover the relationships between different data types (i.e. gene expression and DNA copies).

My research focuses therefore on developing algorithms and tools towards building a comprehensive system for integration by tackling some challenging and interesting computational problems. My principal contributions include:

- The introduction of Multi-Scale Gaussian (**MSG**) model for differential expressed gene identification based on the observation that most microarray expression data show a fat-tailed distribution of differential expressions. A differential expression is the distance between the raw expressions of two different experimental groups for a gene. I modelled the fat-tailed differential expressions using a mixture of null and alternative densities, where the null density captured the minor differential expressions tightly centred near zero and alternative density models captured the major differential expressions located at the tails.
- The introduction of a Heterogeneous Bimodality Index (***h*BI**) to detect bimodal genes. It is based on the assumption that the bimodality is related to the gap between two consecutive sorted expressions. In addition a

comprehensive investigation of bimodality was undertaken in large-scale cancer data obtained from different platforms/studies for seven different cancer types.

- A comprehensive review of the available integration methods, along with the introduction of a Hierarchical Integration Model (**HIM**), which uses mean pattern model for integration.
- Extensions of MSG to work across species (**CSMSG**) and a multivariate MSG (**MVMSG**).

1.5. Thesis Overview

The thesis sets out to present a novel and complete system for integrative study based on gene expression using microarray. This starts with differential gene identification followed by bimodal gene and outlier detection. Finally, a new integrated statistical tool was devised that could be presented and applied to various cancers as well as to cross-species.

The thesis is organised as follows. After this introductory explanation of gene expression analysis methods and applications, *Chapter 2* presents a survey of benchmark methods for differential expressed gene identification and a discussion of their limitations, especially in small replicate samples. The proposed algorithm is discussed in detail, and a comparison is made with the benchmarking algorithms in both simulated and real data. *Chapter 3* describes the bimodality phenomena, along with a technical review of current methods and a discussion of their limitations. It was these limitations that motivated me to introduce a new algorithm to relax the strain by way of the Bimodality Index. The algorithm was also assessed with others in this chapter, based on simulated data.

Chapter 4 extends the evaluation of the proposed method and others based on three sets of real cancer data. Also in this chapter, comprehensive investigation of bimodality on 70 data sets using the proposed method to discover if the bimodality related to a platform or to a specific type or was common across cancers. *Chapter 5* introduces a survey of current advances in the study of integrative genomics. The algorithm developed for integration is also presented with full details while, based on simulated data, the new algorithm is compared to MDI and BCC. *Chapter 6* includes comprehensive details of where all developed algorithms have been used. *Chapter 7* describes a development of MSG for extracting both homogeneous and heterogeneous expression patterns across species as well as other enhancements and the thesis concludes with a summary of the findings and possible directions for future research in *Chapter 8.*

Chapter 2

Predicting Differential Expressed Genes using multi-scale Gaussians

Abstract

When predicting differentially expressed genes from microarray gene expression data with small replicate number, conventional algorithms model the differential expression distribution to adjust for commonly under-estimated variance. I proposed a Multi-Scale Gaussian (MSG) model based on the observation that most microarray expression data show a fat-tailed distribution of differential expressions. A differential expression is the distance between the raw expressions of two different experimental groups. I modelled the fat-tailed differential expressions using a mixture of null and alternative densities, where the null density captured the minor differential expressions tightly centered near zero and alternative density modelled the major differential expressions located at the tails. I used simulated and real data to evaluate the algorithm performance with several benchmark algorithms.

2.1. Introduction

To detect differentially expressed genes in a microarray expression data set across two experiments, the *t* test is commonly used. However the *t* test statistic is often biased when applied to a small number of samples (often called small replicate) of microarray gene expression data due to an underestimated variance. Various algorithms have been developed to provide reliable prediction of differentially expressed genes from microarray gene expression data with small replicate number [25, 35, 36, 54, 114-118]. The majority of these can be thought of as variance-adjustment methods, which rely on borrowing information from all genes or closely related genes to modify the variance of a particular gene. A typical algorithm is SAM (significance analysis of microarrays), which increases the variance of a gene by adding to it a regularization parameter inferred from expressions of other genes [25]. This is done by assuming that the variance across the other genes is representative of the variance of the gene under consideration. This assumption does not necessarily hold in practice. Nevertheless, SAM has been widely used for differential expressed gene detection [26-34].

Bayesian algorithms such as Cyber-T [35], the random variance model [36], Limma [116], Varmixt [38], SMVar [39], and ROAST [40] focus on estimating the posterior variance [119, 120], using a subset of gene expressions to adjust the variance of one gene. The *t* test is then carried out using either analytically derived degrees of freedom, such as Cyber-T and Limma [35, 121] or via the Monte Carlo approach, such as ROAST. Cyber-T and Limma are similar in principle. They model gene expressions in two treatments separately as two Gaussians in order to obtain two adjusted variances for each individual gene. The genes are first ranked according to their expressions. The variance of a

gene is then adjusted depending on the chosen window of similar genes and the degrees of freedom. Cyber-T sets a default window size 101 so as to ensure 50 genes are used for the adjustment of variance. Cyber-T and Limma have been widely used for differential gene identification [26, 27, 29, 34, 122-126].

Expressions of relevant genes in the public database gene expression omnibus (GEO) have also been used to aid inference [115, 127]. This means that the variance of a gene in a data set under investigation can be amended using the variance estimated from the same gene from the data sets in the GEO database.

2.2. Issues related to current methods

The success of variance-adjustment methods largely depends on an accurate identification of a subset of similar genes from the same or different data sets. However, there is inherent selection bias with the variance-adjustment approach [114]. When genes have different variances, pinpointing similar genes can become tricky, thus affecting differentially expressed gene detection performance. **Figure 2.1** illustrates this scenario using a breast cancer data set from GEO with accession GDS3116 [128]. The data contains 58 baseline (control) samples and 58 letrozole (experimental) samples from which 1,000,000 random samples of gene pairs were taken. The Euclidean distance between two genes and the absolute distance between the standard deviations of the two gene expression vectors from which the two members of the pair were drawn was calculated for each pair. The relationship between the two distances is clearly not informative, so in this case it would not be appropriate to use similarly expressed genes to infer the variance of a gene.

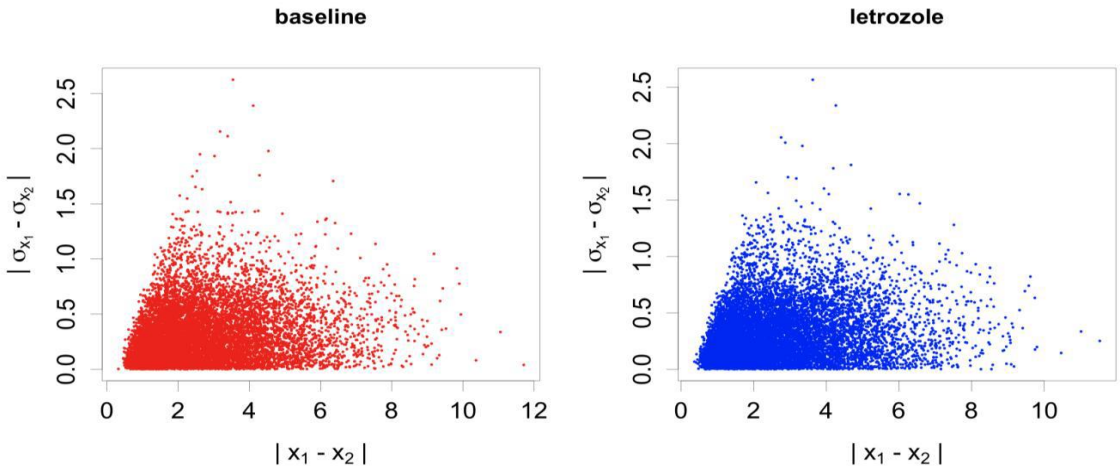


Figure 2.1. Absolute distances between the standard deviations versus gene differential expressions are for the data set (GDS3116) for the baseline samples (left) and the letrozole samples (right): x_1 and x_2 represent two expression vectors, while σ_{x_1} and σ_{x_2} represent their standard deviations. The expressions were log2 normalized.

Efron *et al* developed eBayes [54] by modeling differential expressions as mixtures of non-differential expressions (around mean zero) and differential expressions (away from zero), i.e., the null density and the alternative density. The final testing is done by controlling the local False Discovery Rate (FDR) - the ratio of the null over the full density (which is a mixture of null density and alternative density) of a gene is used to determine whether the gene was differential. In order to estimate the null density, Efron *et al* assumed that the distribution of expression distances in the same condition (control condition) for all genes should be similar to the expression distances across treatments of non-differential genes; this assumption is not necessarily valid in practice. However, for the same data used in Figure 2.1, the distributions of the variances for the within and across conditions are markedly different (**Figure 2.2**). The within condition distance was calculated by randomly selecting two 58-dimensional vectors for the same gene in the baseline samples.

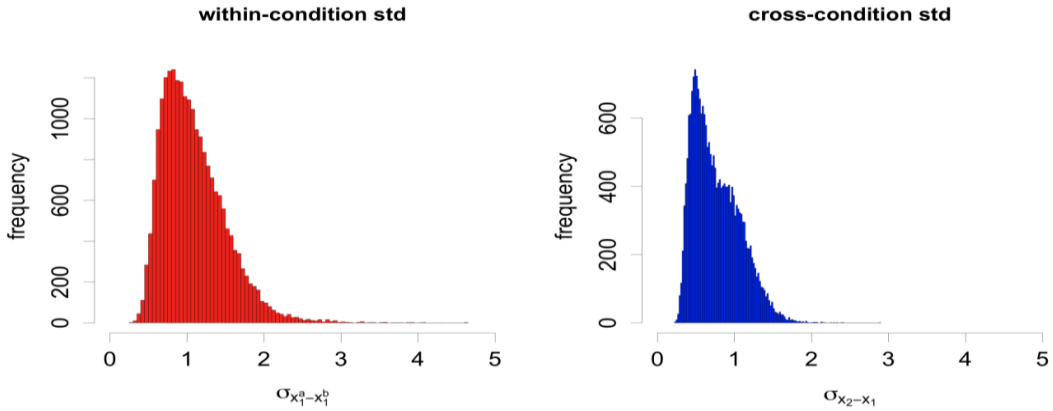


Figure 2.2. Distributions of within-condition standard deviations (left) and cross-condition standard deviations (right). The analysis was based on the same data used in Figure 2.1. The within-condition standard deviation was estimated based on the differential expressions between the baseline samples. The cross-condition standard deviation was estimated based on the differential expressions between the letrozole and baseline samples. $\sigma_{x_1^a - x_1^b}$ denotes the standard deviation of x_1^a and x_1^b, two expression vectors of baseline samples. $\sigma_{x_2 - x_1}$ where x_1 is expression vector of baseline samples and x_2 vector of letrozole samples.

2.3. The proposed method

2.3.1 Multi-Scale Gaussians (MSG)

Here I present an extension to eBayes by modeling expressions via a Bayesian approach. Rather than inferring the alternative density from the empirically estimated null and full densities, I estimate parametrically the null and alternative densities directly from expression distances. This is equivalent to considering the expressions from each condition as a mixture of null and differential genes. In practice, the density of expression distances often has heavier tails than that of a normal distribution (**Figure 2.3**). The heavy tails on both sides of zero correspond to the differential expressions of differential genes, i.e., genes, which show up or down differentiation between non-disease samples and disease samples.

In summary, MSG is based on a probabilistic model as eBayes but relies on less restrictive assumptions. MSG further extends eBayes by adopting a parametric Bayesian approach and models the typical fat-tailed distribution of expression difference across conditions as a mixture of two Gaussians – one describing non-differential expressions (low variance) and the other differential expressions (high variance). This improves upon eBayes, which treats null gene expressions and differential gene expressions across control as being equivalent.

In case of comparing two conditions against each other (single hypothesis), ANOVA is similar to *t* test. The main issue with the *t* test is the variance estimation in the case of the small replicates. ANOVA and MSG addressed variance but MSG addressed variance issue differently. MSG share information across genes to estimate variance which gives more reliable results.

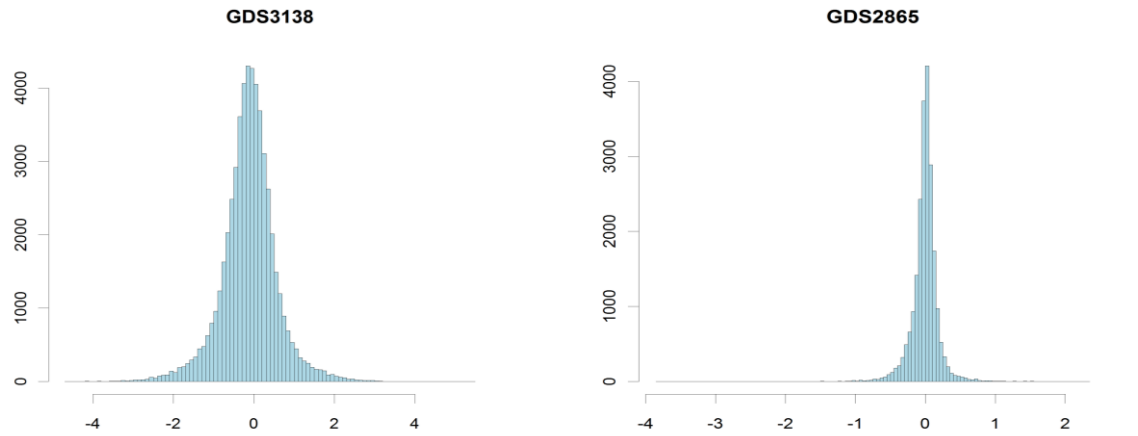


Figure 2.3. The histograms are of differences of expressions across two treatments (cancer versus non-cancer samples or metastasis cancer versus primary cancer samples) for two real data sets discussed later on in this chapter. Details of these datasets are given in Table 2.1. The expressions are on a logarithm 2 scale.

2.3.3. The property of DE

DE is commonly used for testing the null hypothesis that a gene shows no

difference of expression between two experimental conditions. Studies on the robustness of biological systems have indicated that a biological system normally dispatches a small number of genes to respond to a stress [129, 130]. This means that a majority of DEs distribute densely around the origin (zero) and their null hypotheses cannot be rejected. The other subset of DEs distributes sparsely away from the origin and their null hypotheses are rejected. In a one-dimensional case, a sharp peak is seen around zero (a Gaussian distribution with a very small variance) and two long tails (a Gaussian distribution with a very large variance or a uniform distribution).

In terms of hypothesis test, we have three categories of hypotheses. First, those DEs which are around zero are statistically non-differential. Second, those DEs which are positive and away from the origin are statistically up-regulated. Third, those DEs which are negative and away from the origin are statistically down-regulated. It can be seen that all DEs are tested against the origin for the null hypothesis, $H_0 : \mu = 0$. This makes it a natural consideration to use MSG, as described in the next section, Based on this observation, it was proposed that a Multi-Scale Gaussian (MSG) would be used [131-136] to fit the distribution of differential expressions **(Figure 2.4)**. More specifically, minor differential expressions were modelled using a Gaussian with a small variance, while major differential expressions were modelled using a Gaussian with a large variance.

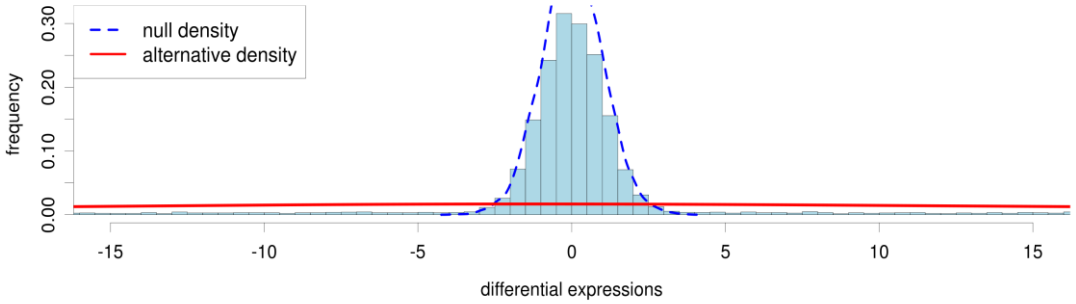


Figure 2.4: Modelling differential expressions using MSG. The broken line represents the null Gaussian density (with a small variance) and the solid line represents the alternative Gaussian density (with a large variance) which captures the fat tails on both sides of the null difference. Here "expression distance" means differential expressions.

In order to demonstrate that using two Gaussians is good to fit the differential expression data, an evaluation analysis being conducted for this purpose. Here I used the regression analysis (R^2) to calculate the fitness as well as *p* value. Firstly, the empirical density was identified and noted by $\mathbf{x}_i$. Secondly, the corresponding MSG density was identified and noted by $\mathbf{y}_i$. Finally, the fitness R^2 was calculated form the two pairs of values. The R-squared is 0.8317 which indicates that MSG model fit the data well.

2.3.4. Algorithm

The control gene expressions and the experimental gene expressions are denoted by $\mathbf{X}_1 = \{\mathbf{x}_{1n}\}_{n=1}^{N}$ and $\mathbf{X}_2 = \{\mathbf{x}_{2n}\}_{n=1}^{N}$, where $\mathbf{X}_{1/2} \in \Re^d$ respectively. Both matrices have *N* rows (number of genes) and *d* (number of replicates) columns. The algorithm also works for data with different replicate numbers across treatments. A mean differential expression vector is generated, denoted as $\mathcal{Z}$, which directly measures the biological significance, similar to eBayes. A multi-scale (full) Gaussian density function $f(z \mid \theta)$ given model parameters has the form:

$$f(z \mid \theta) = w_0 \mathcal{G}(z \mid \mu_0, \sigma_0^2) + w_1 \mathcal{G}(z \mid \mu_1, \sigma_1^2) \qquad z \in \mathcal{Z} \tag{2.1}$$

where $\theta = \{\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, w_0, w_1\}$ is the model parameter set, $\mathcal{G}(z \mid \mu_m, \sigma_m^2)$ is the Gaussian distribution centred at μ_m with standard deviation σ_m (m = 0, 1); μ_0 and μ_1 are the centres of the two Gaussians; w_0 and w_1 are two related mixing coefficients or weights ($w_0 + w_1 = 1$); σ_0^2 is the variance parameter of the null density, $\sigma_1^2 >> \sigma_0^2$ is the variance parameter of the alternative density describing down and up regulated genes. The likelihood function is defined as

$$P(\mathcal{Z} \mid \theta) = \prod_{n=1}^{N} f(z_n \mid \theta) \tag{2.2}$$

where $z_n \in \mathcal{Z}$. Inverse gamma distributed priors are placed (conjugated) on the variances:

$$IG(\sigma_m^2 \mid a_m, b_m) = \frac{b_m^{a_m} \sigma_m^{-2(a_m+1)}}{\Gamma(a_m)} \exp(-b_m \sigma_m^{-2}) \tag{2.3}$$

The mixing coefficients are modelled using non-informative priors. The prior for μ_0 and μ_1 are assigned a Gaussian prior with zero mean and small deviations τ_m:

$$\mu_m \sim \mathcal{G}(0, \tau_m^2) \tag{2.4}$$

As explained in section 2.3.3, the majority of DEs distribute closely around zero and their null hypotheses cannot be rejected. So I set the prior means to be zero, based on my observation that both centres are close to zero, and set both τ_m = 0.5 which corresponds to 1.41 fold change for base two logged expressions (low biological significance). The posterior is then defined as:

$$P(\theta \mid Z,\ \alpha) = \frac{P(Z \mid \theta)P(\theta \mid\ \alpha)}{P(Z \mid \alpha)} \propto P(Z \mid \theta)P(\theta \mid\ \alpha) \tag{2.5}$$

where $\alpha = \{ a_0, a_1, b_0, b_1, \tau_0, \tau_1 \}$ is the hyper-parameter set. The log-posterior can be written as:

$$\log P(\theta \mid Z,\ \alpha) \propto \sum_{m=0}^{1} \sum_{n=1}^{N} (A_{mn} + B_m). \tag{2.6}$$

Here $A_{mn} = \log(w_m \sqrt{\beta_m} \exp(-0.5\beta_m (z_n - \mu_m)^2)$, $B_m = (a_m + 1)\log \beta_m - b_m \beta_m - 0.5\upsilon_m \mu_m^2$ and $\upsilon_m = \tau_m^{-2}$. The model parameters are estimated by maximising the posterior, giving the iterative variance updates

$$\sigma_m^2 = \frac{\sum \vartheta_{n,m} z_n^2 + 2\, b_m}{\sum \vartheta_{n,m} + 2a_m + 2} \tag{2.7}$$

where $\vartheta_{n,m}$ is defined as:

$$\vartheta_{n,m} = \frac{w_m G(z_n \mid \mu_m, \sigma_m^2)}{f(z_n)} \tag{2.8}$$

The iterative update rule for the mixing coefficients can be written as:

$$w_m = \frac{1}{N} \sum_{n=1}^{N} \vartheta_{n,m} \tag{2.9}$$

and the iterative update rule for the centres is:

$$\mu_m = \frac{\beta_m \sum \vartheta_{n,m} z_n}{\beta_m \sum \vartheta_{n,m} + \upsilon_m} \tag{2.10}$$

The learning process is initiated by assigning 0.5 to w_m, 1 to σ_m, and 0 to μ_m. The hyper-parameters were determined based on the empirical differential expression distribution. For both Gaussians, the variance hyper-parameters a_0

and a_1 were set to one, while b_0 was set to be the standard deviation of 90% of the smallest absolute differential expressions, since it is assumed that the majority of differential expressions correspond to those of null genes. b_1 was the 90th percentile of all differential expressions.

Bayesian learning typically has desirable convergence properties [59, 137]. **Figure 2.5** illustrates MSG's estimation convergence of the standard deviation and mixing coefficient parameters on simulations of 10,000 genes with varying number of differential genes. It has been noticed that MSG converged quickly after a few iterations.

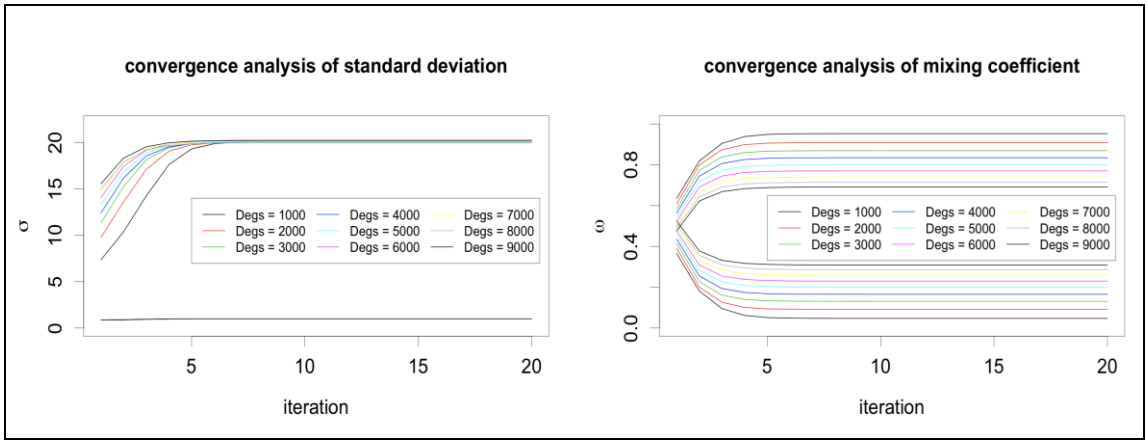


Figure 2.5: MSG's convergence for σ (standard deviation) and ω (mixing coefficient) for simulated gene expressions with varying DEGs (number of differentially expressed genes).

After convergence, the null density and the alternative density are estimated for each gene. Equation (2.1) defines the full density. The Bayes rule is then used to predict whether or not a gene is differential. The null probability is defined of a gene whose mean differential expression is denoted by z, as

$$f_0(z) = \frac{w_0 \mathcal{G}(z \mid \mu_0, \sigma_0^2)}{f(z \mid \theta)} \tag{2.11}$$

The alternative probability that the gene is differentially expressed, is defined as

$$f_1(z) = \frac{w_1 \mathcal{G}(z \mid \mu_1, \sigma_1^2)}{f(z \mid \theta)} \tag{2.12}$$

A gene is predicted as a differential gene if

$$f_1(z) > f_0(z) \tag{2.13}$$

Otherwise a gene is predicted as a non-differential gene.

2.4. Results and Discussion

The proposed algorithm (MSG) was compared against *t*.test, Cyber-T [35], SAM [25], and eBayes [54]. Default parameters were used for these benchmark algorithms.

2.4.1. Synthetic datasets

In order to study the performance of the proposed method as well as the benchmark methods, I used synthetic data. To appropriately play their role, the data must be designed in a way which serve as a mirror for the real microarray data or as closely as possible. Here, I used the flexible microarray data simulation model [138] to generate data with known DEGs. An R package

(*madsim*) implementing their method is available in CRAN. The used model [138] is based on the following:

$$\mathbf{x}_i = \mathbf{a}_i + \mathbf{s}_i + \mathbf{n}_i \tag{2.14}$$

where $\mathbf{x}_i$ is a vector of values generated for gene i. These values come from three sources as per equation (2.14). The values of vector $\mathbf{a}_i$ are expression level for samples. They used beta distribution to obtain values varying between zero and one. Shape parameters of this distribution were chosen to produce more small values than higher ones. This is true for a real data where there are more genes with low intensities than genes with high intensities. The data obtained are scaled to fit real data, which vary between a lower bound and an upper bound. The scaled values represent average expression level for n genes. They assumed that the intensities of each gene are uniformly distributed around the average level. The values of vector $\mathbf{s}_i$ allows one to define DE genes. The values are zero for control samples as well as for test samples in genes that are not differentially expressed, and non-zero for the test samples in genes that are differentially expressed. Finally, the vector $\mathbf{n}_i$ is the independent noise, which obtained from normal distribution with zero mean and 0.4 standard deviation.

2.4.1.1 synthetic dataset 1(2 replicate)

To evaluate the proposed methods and others in presence of noise, I varied the noise levels using *0.1, 0.3,* and *0.5* values for parameter σ_n in *madsim* model. The simulation was repeated ten times corresponding to different initialisations, specifically *rseed* used are *50,100,150,200,250,300,350,400,450,500.* For each noise level the replicate number was two (*m1=m2=2*). A set of 10,000 genes (*N=10000*) was generated of which 1000 (*dgs=0.1*) were chosen as differential genes. The rest of parameter set to their default setting. For each dataset/scenario, corresponding to the noise level, the false discovery rate (FDR) [139-141] associated with treating the top 1,000 genes as being significant was calculated first (**Figure 2.6)**. It can be seen that the mean FDR of MSG is the smallest (0.031). The next smallest mean FDR is Cyber-T (0.032) for a small noise level and this has been switched for medium noise where the average FDR is 0.078, 0.066 for MSG and Cyber-T respectively. For the large noise the MSG is again the best among the other as it gain the smallest FDR(0.08) while this time Cyber-T is 0.13. The FDR of MSG was derived from its posterior error probability (PEP) [54, 142] – i.e., its local false discovery rate.

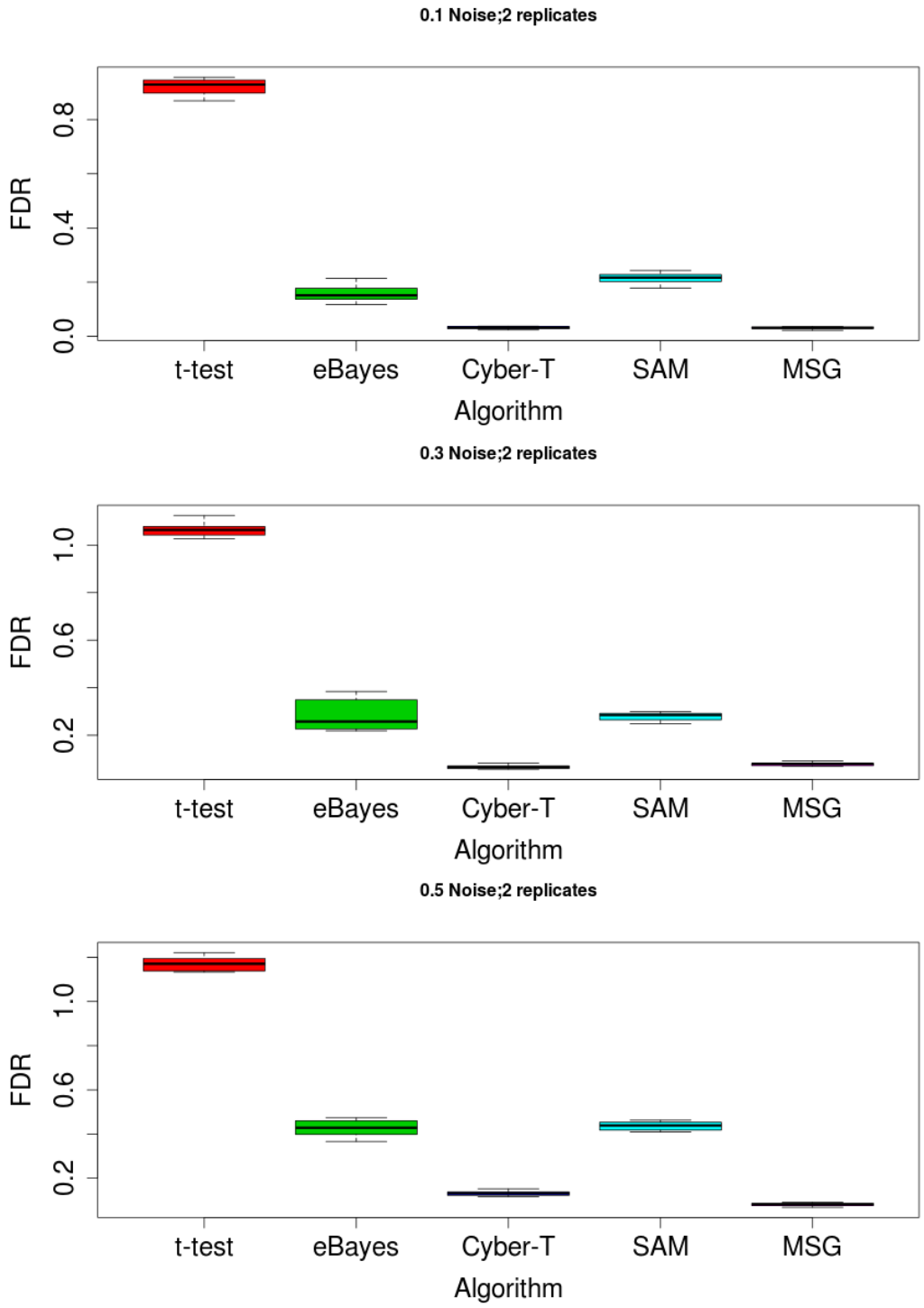


Figure 2.6. Showing FDR of the five algorithms for simulated gene expressions with 2 replicate numbers. The top panel is for 0.1 noise, the middle panel is for 0.2 and the bottom is for 0.5 noise.

The five algorithms were then validated using receiver operating characteristic (ROC) analysis. ROC characterizes a predictor's robustness [143, 144]. Here a predictor is an algorithm used for predicting which subset of genes is significant

in terms of treatment (phenotypic) differentiation. This analysis is different from traditional classification analysis such as discriminant analysis [145], neural network [145], support vector machines [146], or relevance vector machines [147]. In a significance analysis [54] decisions are made based on an arbitrarily-determined significance level (α), such as 0.05 or 0.01. For ROC analysis I varied the significance level between zero and one. **Figure 2.7** shows the ROC curves for this example in the case of σ_n = 0.1. It can be seen that the ROC curves of MSG are the highest (towards to the top-left corner, so that a higher true positive rate can be achieved for a given false positive rate); i.e., in this example, MSG is the most robust with very small number of replicate. Although the others apart of *t* test performed very well. **Figures S2.1 and S2.2** show similar patterns as figure 2.7, for the other different noise level σ_n = 0.3 and 0.5 respectively.

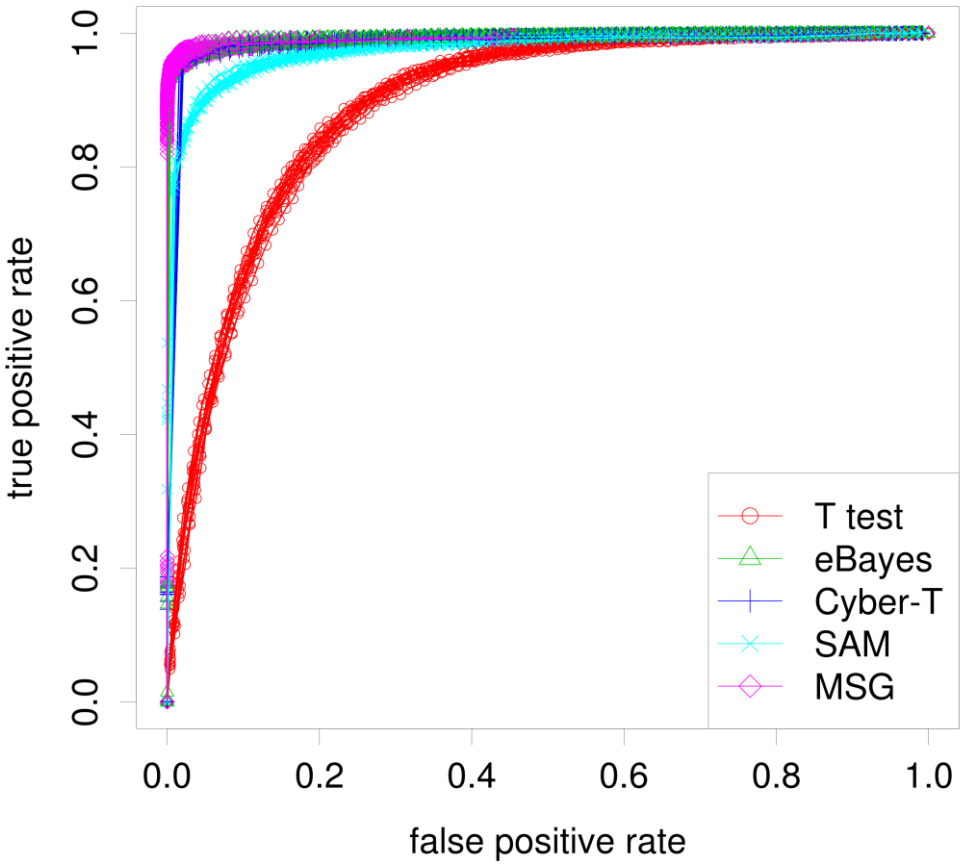


Figure 2.7. Showing ROC curves of the five algorithms for simulated gene expressions with 2 replicates and 0.1 added noise. The horizontal axis represents the false positive rate, which is the rate that non-differential genes are predicted as differential genes. The vertical axis represents the true positive rate (also called sensitivity), which is the rate at which differential genes are correctly predicted. A separate curve is shown for each of the 10 runs.

In order to facilitate the comparison between the five methods, Area under ROC (AUC) was calculated. **Table 2.1** gives the average AUC of 10 runs under different noise levels. It can be seen that MSG for all time has the large AUC while eBayes always is the second best AUC. However, *t* test has the lowest AUC as expected.

Table 2.1. The average AUC for the five algorithms using 3 different value for σ_n. The bold means the best while the italic is the second best.

NOISE	t-test	eBayes	Cyber-T	SAM	MSG
0.1	0.8909	*0.9910*	0.9843	0.9761	**0.9926**
0.2	0.8604	*0.9637*	0.9555	0.9596	**0.9665**
0.3	0.8037	*0.9038*	0.8966	0.8960	**0.9053**

2.4.1.2 Synthetic dataset 2 (3 replicate)

Another data set was also generated with the same parameter as in dataset1 but used 3 replicate instead of 2 (*m1=m2=3*). Ten random samples were generated for this simulation using different seeds as above.

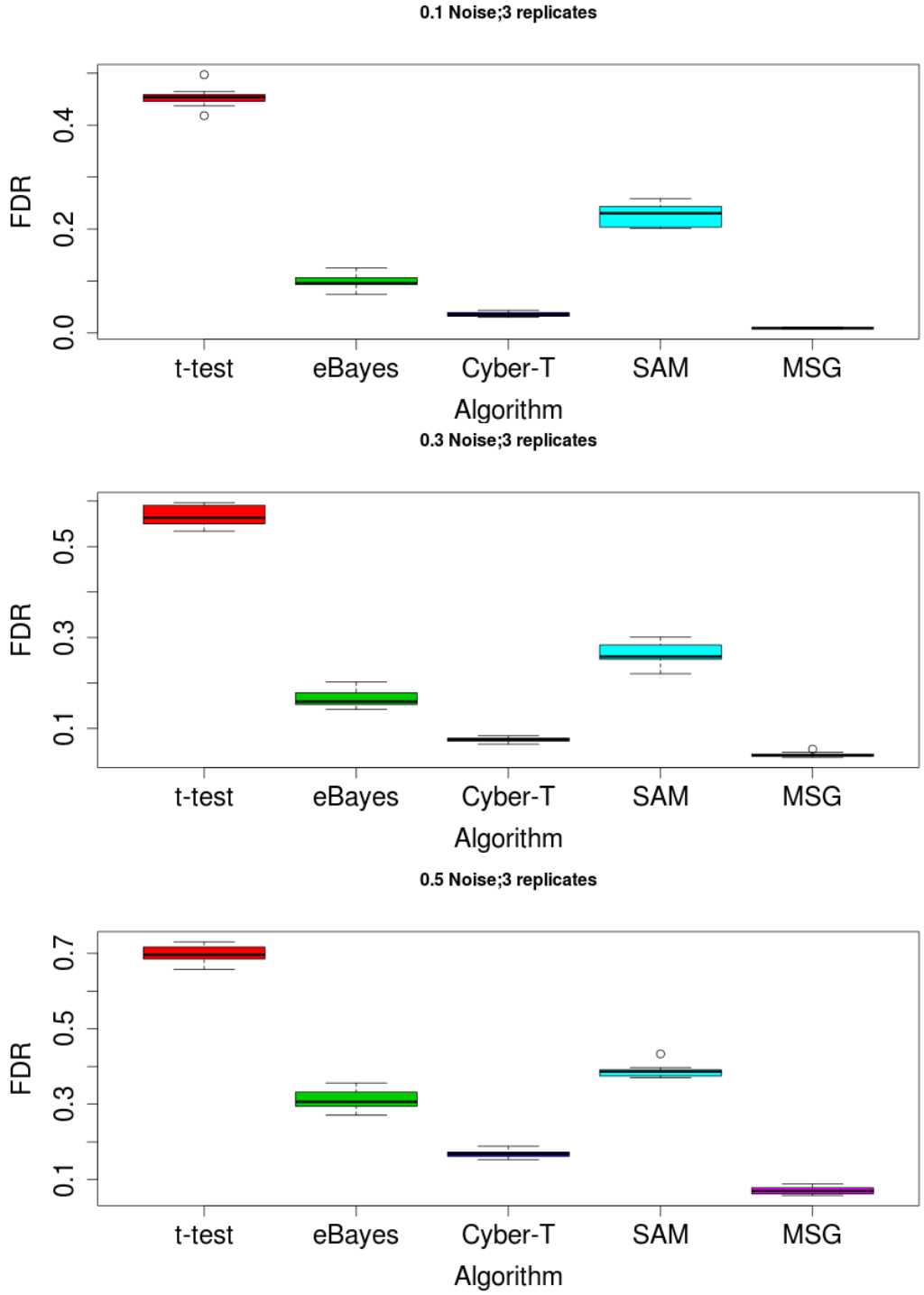


Figure 2.8: FDR of the five algorithms for simulated gene expressions with 3 replicate numbers under different noise levels. The top panel is for small noise level. The middle one is for the medium noise and the last is for large noise.

Figure2.8 illustrates the comparison between the five algorithms using FDR. It can be seen that MSG had minimal FDR compared to the other algorithms in all noise levels. Among all noise levels, the difference between Cyber-T (mean FDR of 0.03, 0.07, 0.17) and MSG (mean FDR of 0.009, 0.04 ,0.07) is more pronounced compared with the previous example. As same as the previous dataset, the FDR of the proposed method is slightly changed with the increasing of noise while this almost doubled in the others. **Figure 2.9** shows the ROC comparison for 3 replicate numbers at 0.1 noise level, where MSG slightly outperforms the other algorithms. Here, the benchmark algorithms apart of *t* test give better predictions compared with the previous example. **Figures S2.3 and S2.4** show similar patterns as figure 2.9, for the other different noise level σ_n = 0.3 and 0.5 respectively.

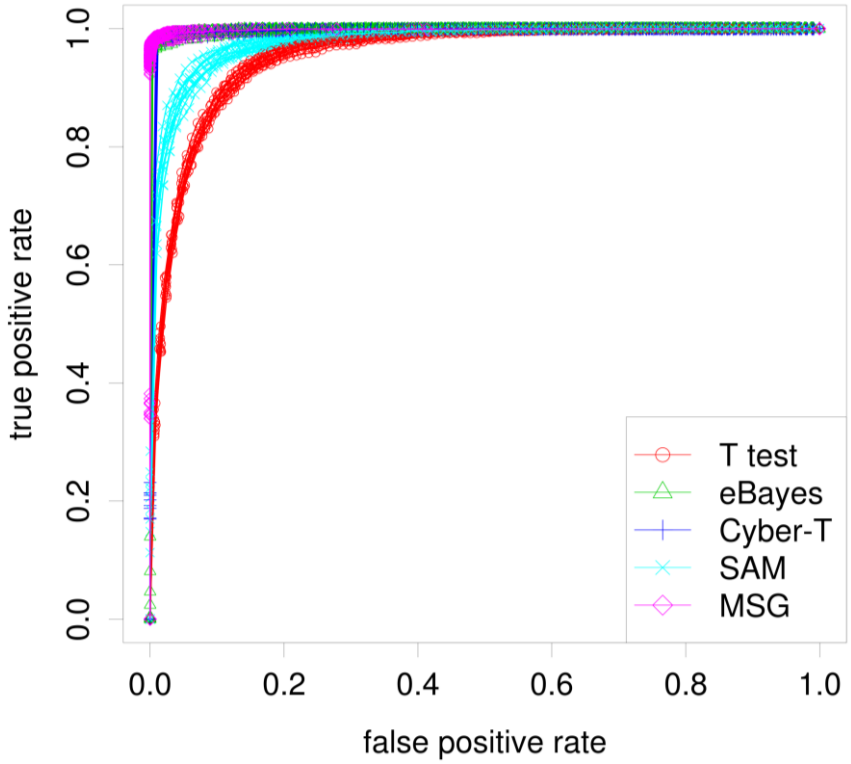


Figure 2.9 ROC curves of four algorithms for simulated gene expressions with three replicates. The horizontal axis represents the false positive rate, which is the rate at which non-differential genes are predicted as differential genes. The vertical axis represents the true positive rate, i.e., the rate that differential genes are correctly predicted. A separate curve is shown for each of the 10 runs.

Further to facilitate the comparison between the five methods, Area under ROC (AUC) was calculated as displayed in **Table 2.2.** It can be seen that the overall AUC is better in this scenario compared to the previous one. MSG has the highest AUC while the eBayes is the second highest.

Table 2.2. shows the average AUC for the five algorithms using 3 different value for σ_n in the three replicate samples datasets. The bold figures mean the best while the italic is the second best.

NOISE	t-test	eBayes	Cyber-T	SAM	MSG
0.1	0.9552	*0.9963*	0.9941	0.9779	**0.9983**
0.2	0.9372	*0.9841*	0.9801	0.9777	**0.9869**
0.3	0.8922	*0.9462*	0.9414	0.8960	**0.9487**

2.4.1.3 Synthetic dataset 3 (5 replicate)

Another data set was also generated with the same parameter as in dataset1 and dataset2 but used 5 replicate instead (*m1=m2=5*). Ten random samples were generated for this simulation using different seeds as above.

Figure2.10 illustrates the comparison between the five algorithms using FDR. It can be seen that MSG had minimal FDR compared to the other algorithms in all noise levels. However, FDR is much lower in all algorithms compared with the previous examples. As same as the previous datasets, the FDR of the proposed method is the smallest among the others in all different noise levels. **Figure 2.11** shows the ROC comparison, in the case of five replicate and σ_n = 0.1 , where all algorithms performed well. This is because of the replicate number is getting large. **Figures S2.5 and S2.6** show similar patterns as figure 2.11, for the other different noise level σ_n = 0.3 and 0.5 respectively.

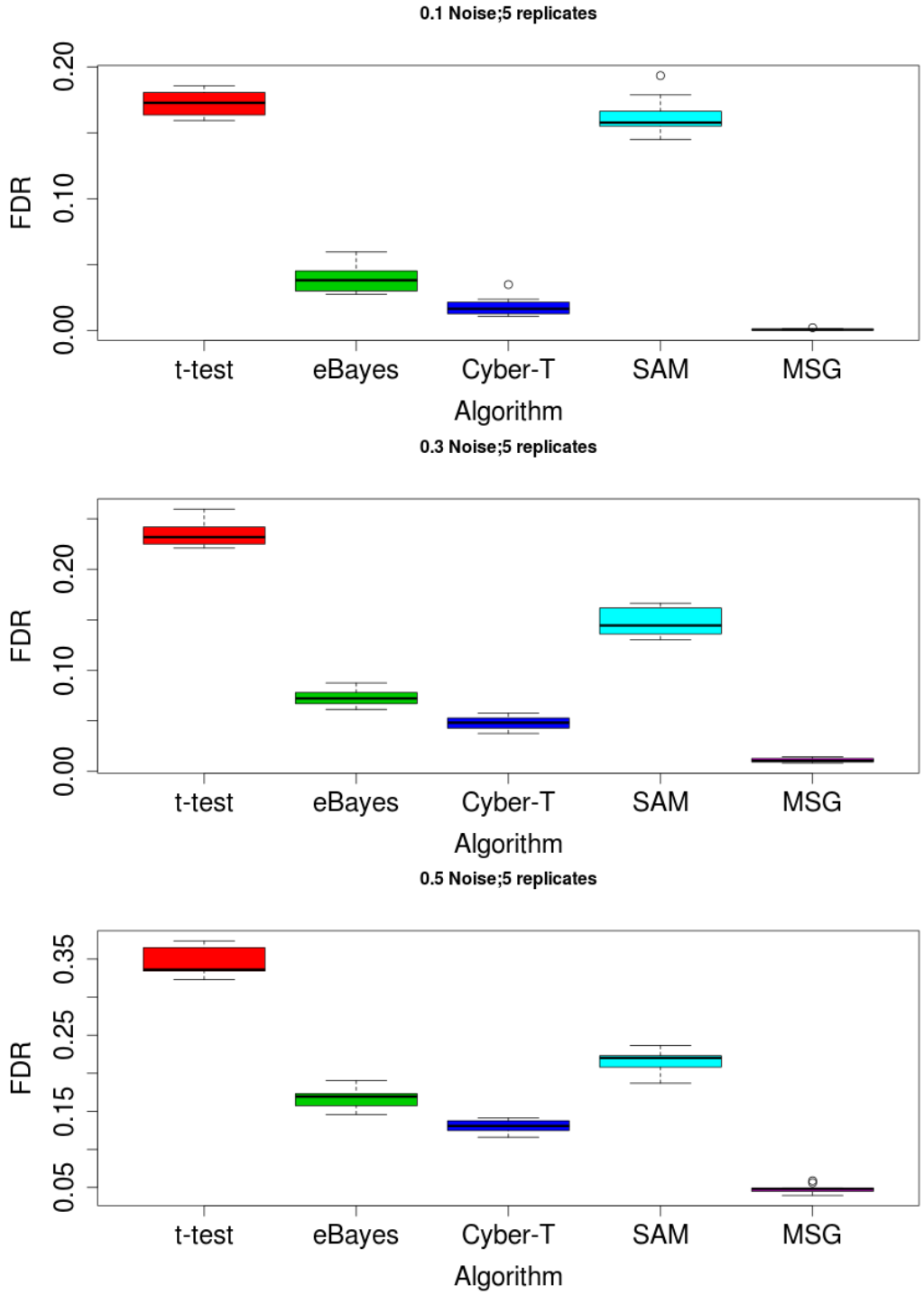


Figure 2.10: FDR of the five algorithms for simulated gene expressions with 5 replicate numbers under different noise levels. The top panel when σ_n=0.1, the middle when σ_n=0.3 and the bottom is when σ_n=0.5.

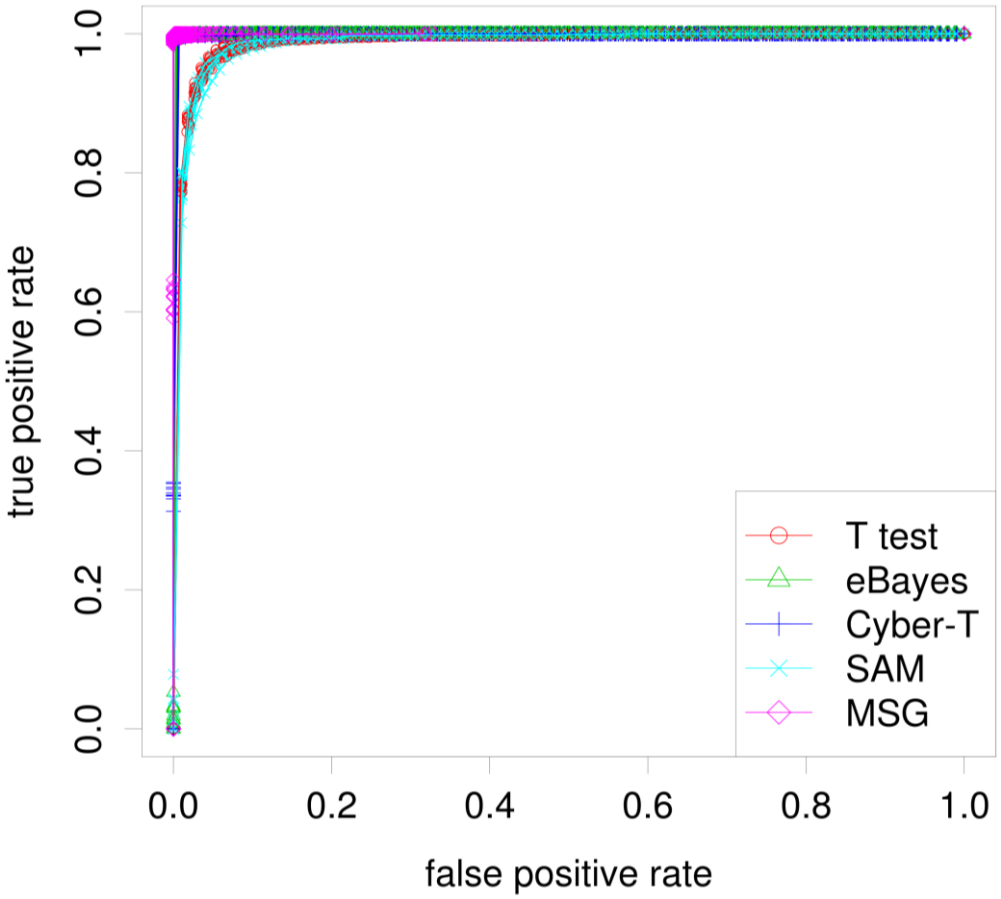


Figure 2.11 ROC curves of four algorithms for simulated gene expressions with five replicate number. The horizontal axis represents the false positive rate, which is the rate at which non-differential genes are predicted as differential genes. The vertical axis represents the true positive rate, i.e., the rate that differential genes are correctly predicted. A separate curve is shown for each of the 10 runs.

The AUC was calculated and summarised in **Table 2.3**. They all have large AUC but MSG is the highest and close to 1.

Table 2.3. shows the average AUC for the five algorithms using 3 different value for σ_n in the five replicate samples datasets. The **bold** means the best while the *italic* is the second best.

NOISE	t-test	eBayes	Cyber-T	SAM	MSG
0.1	0.9884	*0.9979*	0.9977	0.9874	**0.9999**
0.2	0.9814	*0.9947*	0.9932	0.9877	**0.9977**
0.3	0.9577	*0.9793*	0.9774	0.9772	**0.9822**

2.5. Applications on Real Cancer Data

I applied the algorithms to two sets of gene expression data from GEO (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser) **Table 2.4**. Both studies examined cancer metastasis. In the first data set [148], LM2 breast cancer cell line was used with a short hairpin as control and miR-335 expression as treatment. The second data set studied the suppression role of 4.1B in prostate cancer metastasis [149], where poorly metastasised samples were treated as control samples and highly metastasised samples were treated as experimental samples.

Table 2.4. GEO data sets. K is the number of non-cancer samples, M is the number of cancer samples, N is the number of genes.

GEO ID	cancer type	Refs	K	M	N
GDS3138	Breast (metastasises)	[148]	2	2	54677
GDS2865	Prostate	[149]	3	3	22281

2.5.1. Breast cancer data (GDS3138)

The dataset for a breast cancer metastasis study [148] contained two control samples and two experimental samples taken on 54,677 probes. **Figure 2.12** illustrates the Venn diagrams of two analyses. The Venn diagram was generated using the R package VennDiagram [150]. The left panel summarises the prediction results of four algorithms. The significance level was set to 0.05 for eBayes, SAM and Cyber-T. We used 0.95 as the posterior probability threshold for MSG. It can be seen that MSG overlaps 60.9% with SAM, 52.8% with eBayes, and 83.6% with Cyber-T. Cyber-T, MSG, eBayes and SAM have 11,184 (64%), 497 (11%), 324 (5.4%) and one distinct prediction(s) respectively. The right panel shows that only one of the top 20 significantly differential genes

was consistently predicted by all the algorithms. The top 20 predictions were chosen after sorting the *p* values in ascending order and alternative probabilities (see Section **2.3** for definitions) in a descending order. Both eBayes and SAM have three distinct predictions, MSG has nine distinct predictions and Cyber-T has 19 distinct predictions. MSG has 11 overlaps with eBayes and SAM.

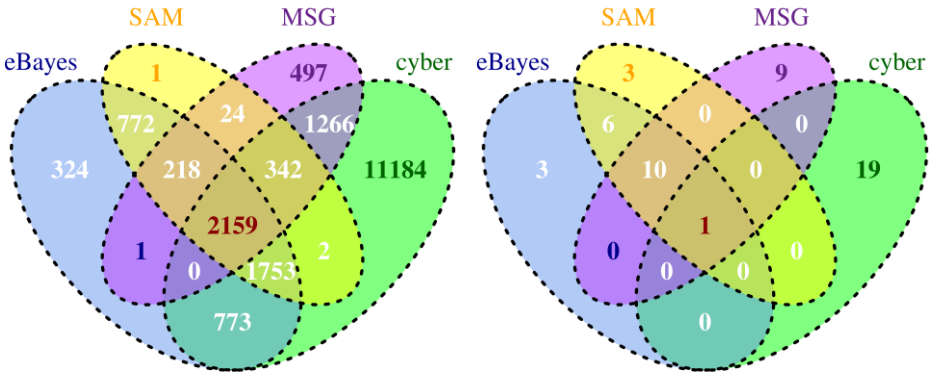


Figure 2.12. Venn diagram analysis for the breast cancer data set (GDS3138), where the left panel summarises the significant genes using fixed thresholds, while the right panel shows the result for the top 20 predictions.

Figure 2.13 shows the ($\log_2$) raw expressions of 13 down regulated genes (left panel) and seven up regulated genes (right panel) among the top 20 predictions of MSG. It can be seen that the predictions are consistent with the raw expression distributions. **Figures 2.14** to **2.16** illustrate the top 20 predictions made by the three benchmark algorithms. MSG is comparable to eBayes (**Figure 2.14**) and SAM (**Figure 2.15**), but Cyber-T (**Figure 2.16**) gives poorer predictions.

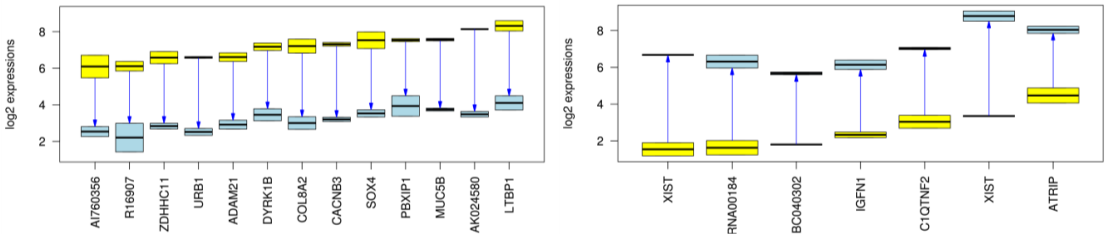


Figure 2.13. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by MSG for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

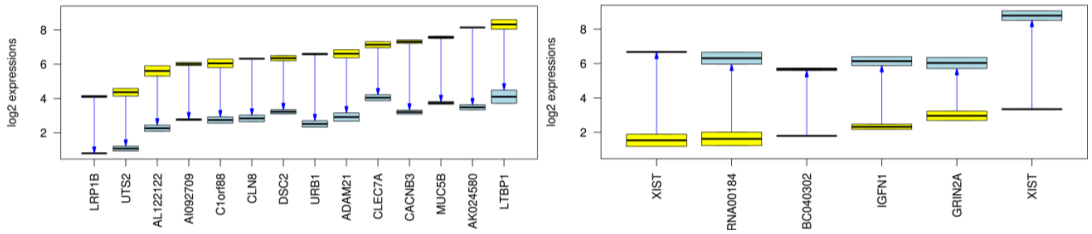


Figure 2.14. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by eBayes for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

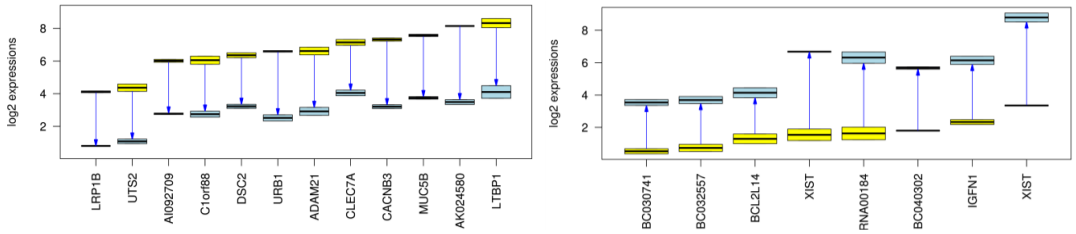


Figure 2.15. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by SAM for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

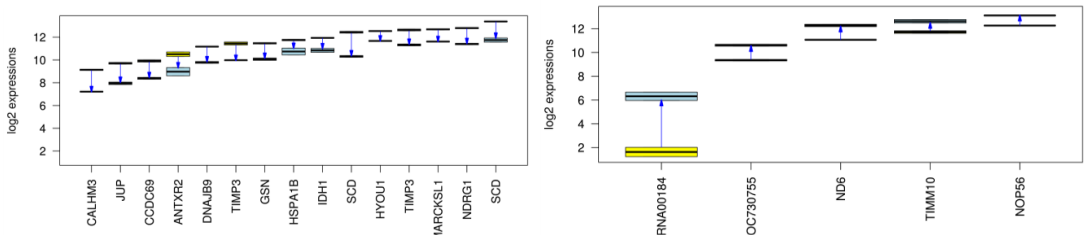


Figure 2.16. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by Cyber-T for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

Figure 2.17 illustrates a selection of the top 10 predictions of MSG and the top two predictions of the benchmark algorithms plotted in terms of their MSG

alternative probability (i.e., that the differential expression corresponds to a differentially expressed gene). The predictions of MSG are located far out in the tails, indicating a high probability that the gene is differentially expressed. The predictions of eBayes are similar to those of MSG. However the SAM and Cyber-T predictions are closer to the null area (the area where the alternative probability is becoming smaller): this agrees with the characteristics of the predictions that were illustrated in **Figures 2.13** to **2.16**.

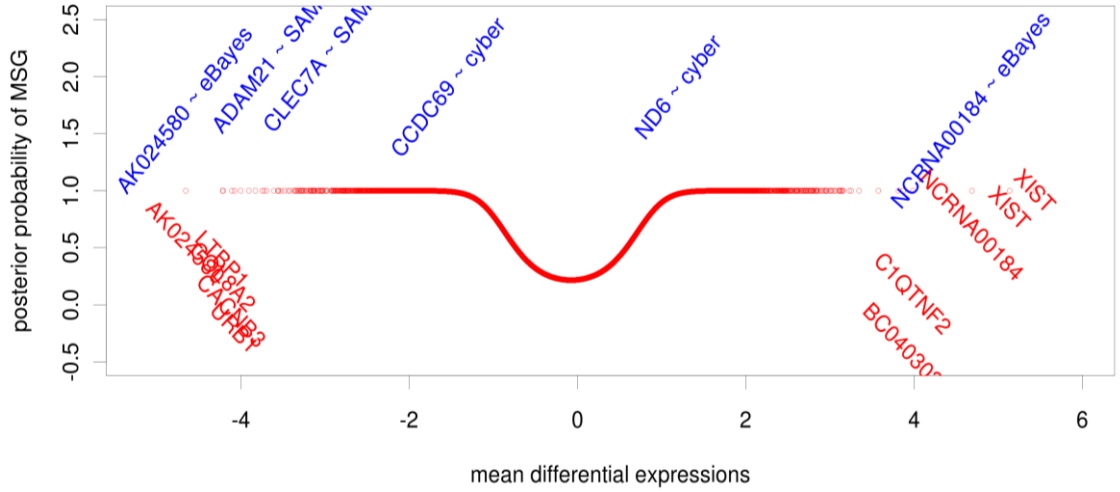


Figure 2.17. The MSG alternative probability of the (log2) mean differential expression with annotations for a selection of the top 10 MSG predictions (bottom) and the top two predictions given by the other algorithms (top) for the breast cancer data set (GDS3138).

The *p* values (for eBayes, Cyber-T and SAM) and the null probabilities (see section **2.3** for definition) for the nine distinct predictions of MSG (among MSG's top 20) are shown in **Table 2.5**, in which a null probability is used since it is synonymous with a small *p* value. The three benchmark algorithms agree with MSG for these nine genes, albeit with very different rankings (the values in brackets in **Table 2.5** are the rankings of the genes for each algorithm).

Table 2.5. The *p* values and the null probabilities of the four algorithms for the nine distinct genes among MSG's top 20 genes for GDS3138 (see right panel of Figure 2.10). The ranking of each gene for a given algorithm is shown in the brackets.

Probe/symbol	eBayes	Cyber-T	SAM	MSG
221900_at/COL8A2	0.00580 (482)	5.55E-16 (208)	0.00168 (201)	2.91E-22 (5)
223749_at/C1QTNF2	0.00133 (31)	1.57E-12 (567)	0.00011 (31)	1.65E-21 (6)
213668_at/SOX4	0.00523 (422)	1.75E-12 (576)	0.00157 (190)	4.31E-20 (10)
1556760_at/R16907	0.01563 (1580)	4.31E-14 (359)	0.00639 (604)	4.06E-19 (12)
232417-at/ZDHHC11	0.00345 (237)	4.94E-14 (366)	0.00109 (135)	1.41E-17 (14)
205875_at/ATRIP	0.00587 (493)	7.01E-10 (1132)	0.00234 (264)	2.13E-17 (15)
217270_at/DYRK1B	0.00420 (307)	4.38E-09 (1396)	0.00137 (172)	2.32E-17 (17)
207838_at/PBXIP1	0.00652 (561)	4.03E-08 (1807)	0.00258 (296)	2.72E-16 (18)
227082_at/AI760356	0.00226 (101)	7.67E-11 (876)	0.00063 (73)	7.56E-16 (20)

C1QTNF2 has been noted as significant predictor of breast cancer [149]. COL8A2 has also been studied for its function in contributing to breast cancer development [151]. SOX4 was found in relation to breast cancer samples [152] and has been used for developing drugs for breast cancer [153]. ZDHHC11 has been found to be frequently amplified in breast cancer cell lines and primary tumour tissues [154], while ATRIP was found to interact with BRCA1 for the checkpoint function of ATR in breast/ovarian cancer [155].

Serine/Theronine kinase MIRK/DYRK1B demonstrates up to 10-fold over-expression in breast tumour cells [156], and PBXIP1 has been recognised as a signature of breast cancer diagnosis [157]. AI760356 is related to the regression of DNA-binding dependent glucocorticoid [158], which is associated with the inhibition of apoptosis in breast epithelia cells [159]. R16907 is not well documented, but it did show a significant down-regulation trend (**Fig 2.13**). **Table S2.1** gives the prediction details of the top 20 genes predicted by four algorithms.

The correlations between the null probabilities of MSG and the *p* values of other algorithms are presented in **Table 2.6**. It can be seen that SAM and eBayes had the largest correlation. The correlation between MSG and other algorithms is smaller (but not insignificant) than the correlations among the three benchmark algorithms.

Table 2.6. Correlation coefficients between the *p* values of eBayes, Cyber-T, SAM and the null probabilities of MSG for GDS3138.

	Cyber-T	SAM	MSG
eBayes	0.84325	0.98947	0.60235
Cyber-T		0.84588	0.52393
SAM			0.66889

2.5.2. Prostate cancer data (GDS2865).

The prostate cancer study data contained 22281 probes with three control and three experimental samples [149]. The Venn diagram (**Figure 2.18**) illustrates all the significant predictions (significance levels were set to 0.05 for the three benchmark algorithms and the posterior probability threshold was again set to 0.95), as well as the top 20 genes for the four algorithms. MSG agreed with eBayes for 90.3% of the predictions and agreed with SAM for 75.8% of the predictions. MSG had only one discordant prediction with Cyber-T, while it was

also evident that MSG agreed with eBayes and SAM for the 11 (=5+3+2+1) top genes. The right panel of the diagram (**Figure 2.18**) shows that Cyber-T, SAM, eBayes and MSG had ten, eight, five and four distinct predictions respectively. The left panel of **Figure 2.19** shows the $\log_2$ raw expressions of nine down regulated genes among the top 20 predictions of MSG, and the right panel shows 11 up regulated genes among the top 20 predictions of MSG: predictions that are consistent with the raw expression distributions.

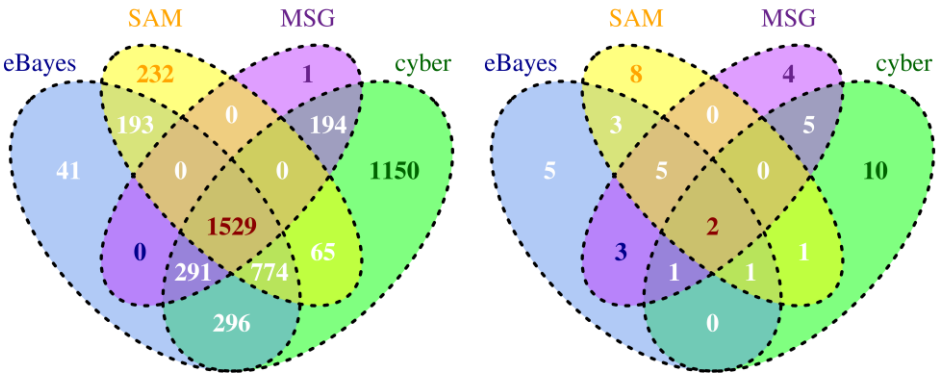


Figure 2.18 Venn diagram analysis for the prostate cancer data set (GDS2865). The left panel shows the result using fixed thresholds to make prediction. The right panel shows the result of top 20 predictions. The diagram was generated using VennDiagram [150]

Figures 2.20 to 2.22 show the raw expression distribution of the top predictions of the three benchmark algorithms. Generally speaking, MSG shows larger distances between the two groups compared with the three benchmark-algorithms.

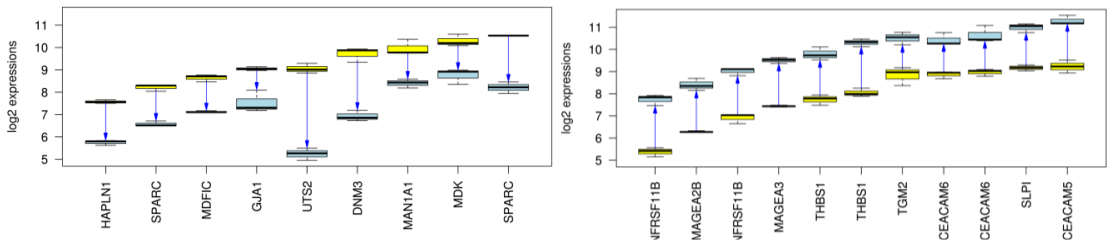


Figure 2.19. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by MSG for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

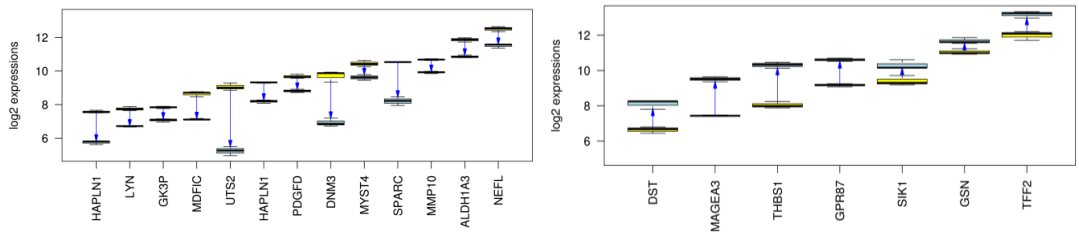


Figure 2.20. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by SAM for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

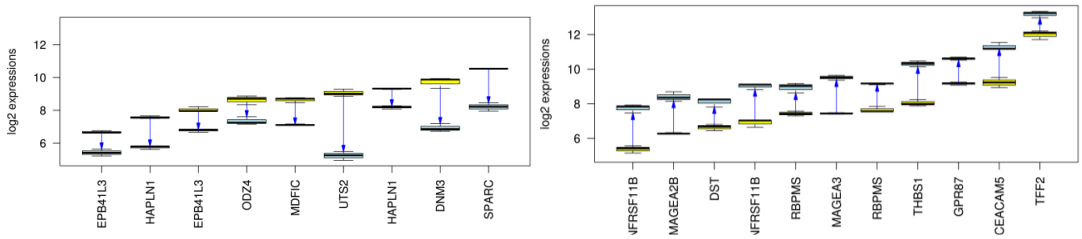


Figure 2.21. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by eBayes for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

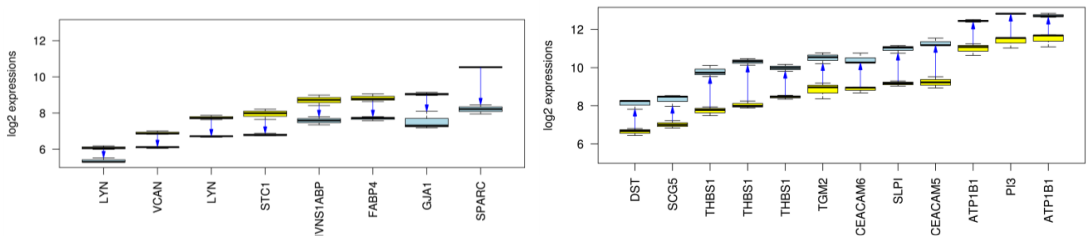


Figure 2.22. The log2 expression distributions of the top predictions identified significantly down (left panel) / up (right panel) regulated genes predicted by Cyber-T for GDS3138. The arrows are used to denote the direction from control expressions to experimental expressions.

Figure 2.23 shows the validation of the top 10 predictions of MSG and the top two predictions by the other three algorithms. The predictions of eBayes are consistent with MSG. The gene GSN, as predicted by SAM, is located in the null area and the gene TGM2, predicted by Cyber-T, is close to the null area.

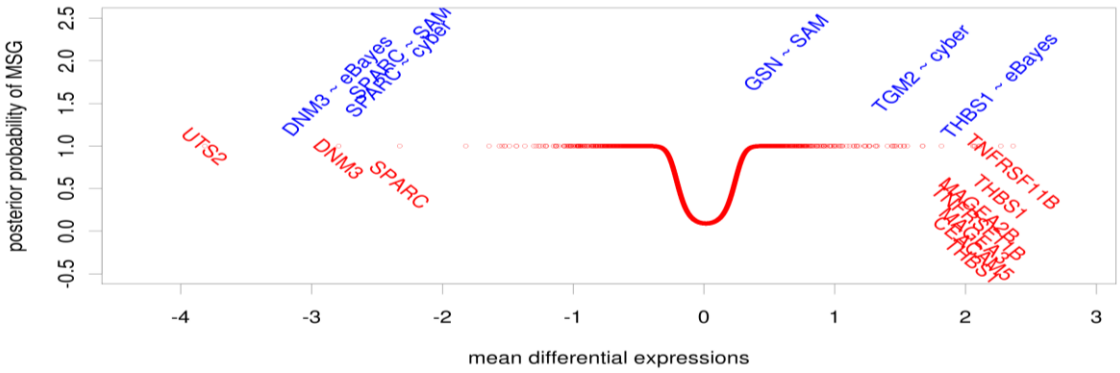


Figure 2.23. The MSG alternative probability of the (log2) mean differential expressions with annotations for a selection of the top 10 MSG predictions (bottom) and the top two predictions given by the other algorithms (top) for the prostate cancer data set (GDS2865).

The unique genes amongst MSG's top 20 are summarised in **Table 2.7**. The rankings given by the algorithms here are more accordant than for the breast cancer data (MSG's unique top 20 are ranked among the top 150 genes of the others). **Table S2.2** gives the *p* values and the null probabilities of the top 20 predictions of the four algorithms.

Table 2.7. The table shows the *p* values and the null probabilities for the four distinct genes among MSG's top 20 predictions for the prostate cancer data set, GDS2865.The ranking of each gene for a given algorithm is given within the brackets, but the ranking for Cyber-T is unavailable due to tied zeroes. Table S2.3 shows 73 genes with zero *p* values given by Cyber-T.

Probe/symbol	eBayes	Cyber-T	SAM	MSG
212667_at/SPARC	1.02E-05 (23)	0	7.85E-05 (42)	2.34E-63 (13)
211657_at/CEACAM6	3.40E-05 (49)	0	0.000135 (100)	3.04E-63 (14)
208116_s_at/MAN1A1	5.87E-05 (61)	0	0.000236 (133)	1.90E-57 (16)
209035_at/MDK	1.74E-05 (31)	0	7.85E-05 (42)	5.64E-56 (17)

SPARC has been widely studied for prostate cancer metastasis progression [160-162]. CEACAM6 has also been found to be a significant contributor to prostate cancer metastasis. For instance, it was found that it does not show any differentiation between prostate cancer and normal tissues, but shows high differentiation in metastasis status [163]. It has also been recognized as a mediator of metastasis [164] and has been found in various metastasising cancer patients, including prostate cancer [165].

MAN1A1 has been found to be differentially expressed between primary and metastasising prostate cancer patients [166, 167]. MDK's function is to promote cell growth, survival, migration and angiogenesis. It has been found to be over-expressed in various human cancers [168] and shows high differentiation status in metastasis prostate cancer patients [169].

Table 2.8 gives the correlations between the four algorithms. The treatment of *p* values and null posterior probabilities is similar to what was used in compiling **Table 2.6**. In this data set, the correlation of MSG with eBayes and Cyber-T is more significant compared with that in the breast cancer data set, although these methods are better correlated amongst themselves than with MSG.

Table 2.8. Correlation coefficients between the *p* values of eBayes, Cyber-T, SAM and the null probabilities of MSG for the GDS2865 data set.

	Cyber-T	SAM	MSG
eBayes	0.95844	0.98239	0.70459
Cyber-T		0.93009	0.68989
SAM			0.68097

2.6. Conclusion

A multi-scale Gaussian (MSG) model was presented for predicting differential expressed genes from microarray expression data of small replicate number. MSG models the typical fat-tailed distribution of differential expressions as a mixture of two Gaussians – one describing minor differential expressions using a small variance and one describing major differential expressions using a large variance. In this respect, MSG generalizes the concept of eBayes which models the null density and the full density empirically. MSG estimates the null density and the alternative density parametrically under the Bayesian framework.

It was found that for simulated datasets, MSG compared favourably to three benchmark algorithms – SAM, Cyber-T and eBayes . When applied to real data sets, MSG offers unique and insightful predictions of differential expressed genes. In summary, multi-scale mixture modelling, which is becoming more and more interesting for statistical purposes, helps capture the characteristics of differential expression data and thus serves as a useful alternative to existing approaches in differentially expressed gene detection.

Chapter 3

Predicting bimodal genes via gap maximisation

Abstract

Bimodality is a common phenomenon that occurs in gene expression data for certain types of studies, including cancer studies and drug effect studies. Several algorithms developed to tackle this problem have met with some success, but since their performance is occasionally unsatisfactory, a new algorithm was suggested that would detect bimodal genes. This new algorithm, is based on the assumption that, as suggested by [110], bimodality is associated with the gap between two consecutive expressions. The performance of this new algorithm has been demonstrated against several widely-used benchmark algorithms, using both real and simulated data sets.

3.1. Introduction

3.1.1. Bimodality

Although mRNA expression profile for most genes is expected to follow a normal distribution across samples [110], it appears that a considerable number of genes have an expression pattern that follows a two or more component mixture distribution [111]. The mixture of two densities with distinct centres is known as bimodality. Furthermore, density distribution can be defined by the presence of two characteristic peaks (**Figure 3.1**). Bimodality is caused by somatic mutations, such as the amplification of the tyrosine kinase proto-oncogene receptor 'erbB2', during the development of cancer [170], or through germ cell mutations, such as SNPs [171]. Some genetic mutations can have a high level of genetic alteration, while the level of alteration in others can be normal [172].

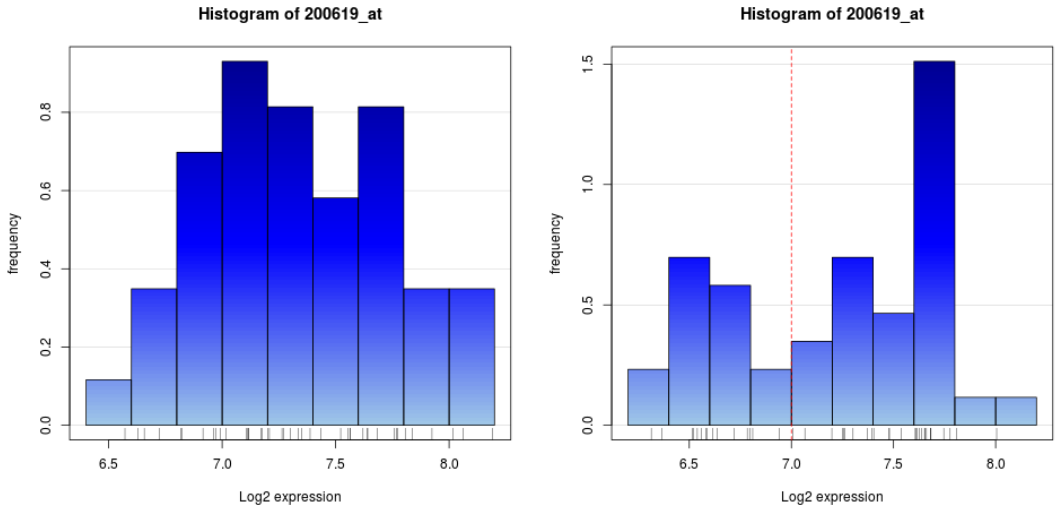


Figure 3.1 Histograms for the SF3B2 gene (200619_at). The left panel is the gene expression in normal samples and shows normal distribution. The right panel is for the same gene in cancer samples and it has bimodal distribution with the broken vertical line representing the classification threshold between the two modes.

Genetic translocations, which commonly occur in cancer cells, are a result of the rearrangement of parts between non-homologous chromosomes [173]. However, these mutations play a main role in cancer progression or in disease development more generally. Furthermore, genomic lesions may affect some samples but not others; this means that bimodal expression patterns have occurred. An example of these recurrent fusions was observed in prostate cancer datasets by Tomlins *et al*., who found ERG and ETV1 genes over-expressed in some of the samples in multiple datasets [174]. Another study showed that oncogene HER2 was over-expressed in 15–20% of breast tumours compared with normal breast tissues [175].

Bimodality can appear in many biological systems as well as in cancer, as noted by Mason and his group [176], who observed that the expression levels for some genes show a distinct bimodal distribution in human skeletal muscle tissue. Bimodal distribution has also been investigated in many blood glucose studies [177, 178]. Interestingly, bimodality can occur in homogenous disease, tissues or patient groups. [176, 179]. It also can be observed in healthy samples but gain attention in cancer or disease state where it believed to play main role in treatment. Thus, Bimodal genes are not always differentially expressed genes. This has led to the discovery of such methods that can identify the bimodal patterns rather than identifying differentially expressed genes. Genes with a bimodal distribution may be useful candidates to distinguish subgroups of disease such as different types of breast cancer. In addition, it has been demonstrated that bimodal gene expression may plays another role in identifying clinical groups (cancer vs normal). In some genes, the bimodal pattern observed in one group and unimodal in the other group. For example, transcription factor genes show a unimodal patterns in normal lung samples

while those genes showed a bimodal patterns in lung cancer samples[180, 181] Cancer as a complex disease [182] has many subtypes that respond to treatment differently [81]. Since tumours with such genetic alterations are not likely to respond to the same drug, finding a genotype only may not be suitable for drug selection. Therefore the critical key for drug design is a full understanding of both genotypes and phenotypes [172], and these differences in response have led to a focus on the testing of individual patients. Other important criteria for personalised medicine also include accurate decisions that correspond to a clear distinction of expression values of a specific gene into two groups. As a result, in the search for a therapeutic target, identifying such bimodal genes makes an important contribution to the process of treatment, and may also help biologists to understand the complexity of cancer or to discover cancer subtypes. Genes whose expression follows bimodal distribution in multiple cancer data are likely to be candidate biomarkers for diagnostics or prognostics, or for predicting therapy response [183].

3.1.2. Heterogeneity and drug resistance

Researchers have put considerable effort into studying the complex disease of cancer, with the aim of understanding its molecular characteristics [184, 185]. Cancer patients with similar tumour characteristics are unlikely to respond to a specific treatment in the same way [186]. In breast cancer, for example, variant responses to drugs such as Tamoxifen and Herceptin are evidence of the heterogeneity in such pathological factors as oestrogen receptor (ER) and HER2 status [187]. Large numbers of patients have gained from using Tamoxifen for hormone receptor-positive, but the same drug has failed in subgroups of patients who carry specific variants in the cytochrome gene P450

2D6 (CYP2D6) [188, 189]. Other evidence provides clear support that HER2 targeted therapies aim to block tumour cell growth. Trastuzumab, as a first drug approved by the US Food and Drug Administration (FDA) for this purpose, has been a beneficial therapy, either alone or in combination with chemotherapy, in about 25% of patients with positive HRE2 cancer [190-193]. However, some patients have not responded to the Trastuzumab treatments, producing the need for an accurate grouping of HRE2-positive patients [189].

Another target for many cancer types was epidermal growth factor receptor (EGFR). The drug Gefitinib (Iressa), which has also been approved by the FDA, suppresses the ATP binding function of EGFR and has resulted in partial remission regression for 10-30% of patients with NSCLC (non-small cell lung cancer) [194-198]. But, as reported by Giaccone *et al* [199], the same drug was not adequate in combination with chemotherapy. Finally, a study by researchers at Harvard established that genetic alterations were associated with drug response [200]. Thus, it has become clear that efficient cancer drugs will be based on understanding the molecular profile of human cancers.

The heterogeneity in cancer makes both diagnosis and treatment difficult. This is clear in cases of breast cancer expression patterns analysis, through which at least five subtypes of tumour were found [81, 201]. Chin and collaborators showed that tumours with the same phenotype and transcriptional factors can result from a different set of genomic alterations [185], and showed that ER-negative, high-grade basal subtypes can be divided into two groups, low and high genetic stability. As a result, drug design based only on the phenotype will not be sufficient if knowledge of the genotype is not taken into account. The heterogeneity observed in cancer has led to personalised medicine being

thought of as offering better care for cancer than standard care, in the way it targets individual patients.

In summary, genetic heterogeneity in cancer is common. This means that different subsets of people within a population that had different underlying factors which cause same kind of disease. These factors either genetic variation or environmental factors. For example, ERG and ETV1 genes over-expressed in some of the samples but not all in many datasets [174]. Thus, identifying such bimodal genes makes an important contribution to the process of treatment, and may also help biologists to understand the complexity of cancer or to discover cancer subtypes. Genes whose expression follows bimodal distribution in multiple cancer data are likely to be candidate biomarkers for diagnostics or prognostics, or for predicting therapy response [183].

3.2. Methods used for bimodality identification

3.2.1. Background

In the last few years there has been a huge number of microarray experiments using genome-wide analysis of gene expression that have helped biologists to investigate gene expression patterns for cancers and other diseases [202]. The first concern was to identify the differentially-expressed genes between two sample conditions, such as cancer vs. normal. Many methods were proposed and have been widely used for this task, such as the *t*-test, SAM [25], CyberT [35] and eBayes [203] that were explained in chapter 2. Methods based on means and variances comparison are powerful tools for detecting the differential expressed genes in the case of all samples in a group sharing a common average, and with more or less equal variation around it. However, some diseases, such as cancer, are following a heterogeneous expression

distribution which indicates that a subset of samples (patients/cell lines) show similar expressions to non-cancer samples. This skewedness lowers the mean of cancer samples towards the mean of non-cancer samples to affect the accuracy of the *t*-test. Thus, its important task to identify those genes with bimodal behaviour for its importance in cancer as seen in some studies[204-206] and further to enhance the power of other tests such DEG identifications and clustering.

It has been noticed that even within a specific subgroups of cancer patients, the same genes are not necessarily expressed in the same way among all samples [207]. This has altered the focus on identifying DEGs to search for a single sample (i.e. outlier) or small samples (i,e. bimodal) that have divergent expressions from the rest of the group.

In relation to the detection of bimodal genes it is important to discover a set of genes that are tightly regulated around two conditions at the transcript level [208]. Genes with bimodal behaviour in multiple cancer datasets are strongly evidenced to be ideal biomarker candidates for the treatment of the cancer or of more general diseases [209]. Many studies have applied biological annotation, such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways [210], or Gene Ontology (GO) [211] to a set of bimodal genes. This shows that they are utilized in cell-cell communication and play a main role in changes in biological function [212, 213].

Ertel and Tozeren evaluated the expression of bimodal genes in diabetes types I and II in order to investigate the role of bimodal genes in cell communication and immune response, expanding their study to transcript level regulation using the available bioinformatics repositories [214]. The incidence of bimodality was

also mentioned in a number of cancer and other disease studies [179, 215, 216]. The bimodality could be calculated using F-statistic [179], outlier detection methods [174, 217, 218], or a combination of mixture modelling and special coefficients [110, 112, 113, 216].

Many clustering methods are used to group genes based on their expression values. These methods vary as some are classified and grouped based on the use of different distance metrics [41], while others can be gathered around by the data structure, such as k-means and mixture model [145, 219, 220], with the latter being widely used in bimodality pattern discoveries. A number of studies have used the previous clustering methods to group genes into two groups and have then analysed the bimodality by using Kurtosis value [113], the Likelihood ratio test [112], or the bimodality index [110].

Another approach to investigate bimodality is through outlier detection methods, which are mainly used to identify the low or high expression values in a fraction of the sample. These methods group the genes expression using a specific quantile to detect the outlier sample expressions such as a Cancer Outlier Profile Analysis "COPA" [174], outlier sum statistics "OS" [217], the outlier robust *t*-test (ORT), MOST [221], LSOSS [222] and distribution-based outlier sum statistics [223]; these are some of the typical algorithms in the literature.

3.2.2. Technical review

Two approaches have been applied in identifying such bimodal genes. One approach is based on outlier detection methods, since the biologists include the outliers in the scope of bimodality definition. The second is designed for bimodality identification. For the former several methods have been proposed for handling gene expression data with heterogeneous distributions. Cancer

outlier profile analysis algorithms COPA [174], outlier sum statistics (OS) [217], the outlier robust *t*-test (ORT), MOST [221], LSOSS [222] and distribution-based outlier sum statistics [223], which are some typical algorithms in the literature. For the latter approach many algorithms have been suggested such as PACK [113], Likelihood ratio test (LHR) [112], and the bimodality index (BI) [110].

3.2.2.1 Outlier detection methods

Many methods have been proposed to identify genes with bimodal distributions for the first group. They were derived from the *t*-test by replacing the mean using the median, and using the median's absolute deviation instead of standard error used by the standard *t*-test. One of these methods is COPA (Cancer Outlier Profile Analysis) [174], which is based on the median of gene expression instead of the mean. COPA consist of three steps. First, the data are centred to the median and each gene's median expression set to zero. Second the data are scaled by dividing each gene's expression value to their median average difference. Finally, different percentiles are used in order to filter the data. Tomlins (2005) and MacDonald (2006) used the 75th, 90th, and 95th.

Let x_{ij} be the observed expression value of a non-cancer sample $i = 1, \ldots, m$ and let y_{ij} be the observed expression values of cancer samples $i = 1, \ldots, n$. The gene index $j = 1, \ldots, g$. *T*-test can be calculated as the following equation:

$$t_j = \frac{\bar{\mathbf{y}}_{\mathrm{j}} - \bar{\mathbf{x}}_{\mathrm{j}}}{s_j} \tag{3.1}$$

whereas $\bar{\mathbf{y}}_{\mathrm{j}}$ and $\bar{\mathbf{x}}_{\mathrm{j}}$ are the means value for gene *j* in the cancer group and control group respectively and s_j is the pooled standard deviation of gene *j* and defined as:

$$s_j^2 = \frac{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2 + \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_{ij}} - \bar{\mathrm{y}}_{\mathrm{j}})^2}{(m+n)-2} \tag{3.2}$$

The *t*-test fails to detect heterogeneous differential genes. Thus, COPA was proposed as the following:

$$COPA_j = \frac{\mathrm{q_r(y_{ij})} - med_j}{MAD_j} \tag{3.3}$$

where $\mathrm{q_r}$ is the percentile of cancer data, med_j is the median of all normal values and cancer data for gene *j* , while MAD_j is the median absolute difference of all values of gene *j*

Another method, named the outlier sum statistic (OS), was proposed by Tibshirani and Hastie, in order to detect genes that over expressed or down expressed in subsets of the sample [217]. This was motivated by the limitation of COPA, which is that using a fixed r^{th} sample percentile may not work well for detecting more than one outlier [218]. The idea behind this method is that it sums up the expression of the outlier cancer samples, which are identified by quintiles as:

$$\hat{\mathrm{y}}_{\mathrm{ij}} = \frac{\sum_{i=1}^{n}\mathrm{y_{ij}} - med_j}{MAD_j} \tag{3.4}$$

$$OS_j = \hat{\mathrm{y}}_{\mathrm{ij}}.I[\hat{\mathrm{y}}_{\mathrm{ij}} > q_{75}(j) + IQR(j)]$$

where $IQR(j) = q_{75}(j) - q_{25}(j)$ is the interquartile of all expression data for gene j, and q^{75} is the 75 percentile of expressions. This can also be used to find the down regulated outliers as they used the 25th as percentile minus the IQR in eqn 3.4.

A common problem of COPA and OS is the use of the overall median because this is not the appropriate replacement of the normal sample mean, and might overestimate the non-cancer sample mean and affect the outlier detection. A slight modification to OS has been proposed by Wu which is Outlier Robust T-statistic (ORT) [218] and defined as:

$$ORT_j = \frac{\sum_{i \in R_j} (\mathrm{y}_{ij} - med_j^N)}{median\{|x_{ij} - med_j^N|_{i \le m}, |x_{ij} - med_j^C|_{i \le n}\}} \quad (3.5)$$

where med_j^N is the median of Normal expression sample and med_j^C is the median obtained from Cancer expression values, *m* and *n* represent the number of non-cancer samples and cancer samples respectively, and

$$R_j = \{i \le n : y_{ij} > q_{75}(x_{ij}, 1 \le i \le m) + IQR(x_{ij} \le 1 \le m)\} \quad (3.6)$$

The only difference between OS and ORT is that ORT uses only the control (non-cancer) sample to centralise the data, and scales the control and cancer separately, whereas OS uses them both [173]. Although OS and ORT are designed for control and disease data, specifically normal vs. cancer, ORT has been modified to deal only with tumour samples, as proposed by Hellwig and her group [183]. In addition to this, all outlier methods mentioned are designed to tackle the outliers based on the quintiles of the gene expression from all the samples. MOST statistic was then proposed to overcome the arbitrary setting used by OS and ORT with the aim of considering all possible numbers of outliers [221]. MOST arranged the cancer samples from high to low in order to

find the outliers. However, this method does not work perfectly in some cases where the number of cancer samples is small, as the standard deviation σ_k for one is 0,

$$MOST_j = \underset{i<k<J}{MAX}\left(\frac{\sum_{1\le i\le k}(\mathrm{y_{ij}} - med_j^N)}{1.4826\times median\{|x_{ij} - med_j^N|_{i\le m}, |x_{ij} - med_j^C|_{i\le n}\}} - \mu_k\right)/\sigma_k \qquad (3.7)$$

where *k* varies from one to the number of cancer samples.

More recently, LSOSS has been proposed; based on detecting change points in the ordered gene expression of cancer samples it uses these points to identify the outliers in cancer samples [222].

$$LSOSS_j = K\frac{\bar{\mathrm{y}}_{\mathrm{s_{ik1}}} - \bar{x}_j}{s_j} \qquad (3.8)$$

where *k* is obtained by minimizing the pooled sum of squares for cancer samples only and s_j is calculated as:

$$s_j^2 = \frac{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2 + \sum_{\mathrm{i=1}}^{\mathrm{k}}(\mathrm{y_{ij}} - \bar{\mathrm{y}}_\mathrm{j})^2 + \sum_{\mathrm{i=k+1}}^{\mathrm{n}}(\mathrm{y_{ij}} - \bar{\mathrm{y}}_\mathrm{j})^2}{(m+n)-2} \qquad (3.9)$$

An improvement has also been made to the previous two methods by using an explicit form for the *p* value [223] .

3.2.2.2 Bimodality detection methods/ combination of mixture modelling and special coefficients

In this category, PACK (Profile Analysis using Clustering and Kurtosis) is a method proposed for finding the outliers or genes with bimodal distribution [113]. PACK's general approach is based on two steps. The first is clustering, using an EM algorithm with BIC as a model for the selection criterion in order to

find the optimal number of clusters. The second uses kurtosis to find relevant classifiers as:

$$Kurtosis(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^4}{(n-1)(n-2)(n-3)\sigma^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (3.10)$$

The positive kurtosis means there are outlier subgroups in the data, while the negative shows a major bimodal pattern. The researchers validated their methodology in six independent cancer datasets, three prostate cancer datasets and three breast cancers. They also successfully extended the application of their suggested method to other breast cancer data [224]. Although previously applied successfully, important genes could be missed in the case of an 80 to 20 percent split of the samples into two groups, because the kurtosis is close to zero [110]. The disadvantage of PACK is that it failed to detect the bimodal distribution with unbalanced sizes of peaks.

Another method is the Likelihood ratio test (LHR) which Ertel and Tozeren used to identify bimodal genes [112], and their importance in disease progression [112, 225]. The hypothesis that gene expression distribution is bimodal, against the null hypothesis that the expression follows a normal distribution, was tested and has the form:

$$LHR = \frac{LH_{mixture}}{LH_{normal}} \quad (3.11)$$

Where $LH_{mixture}$ is the likelihood of a normal mixture model with two components and LH_{normal} is the likelihood of a normal model with one component. Small LHRs indicate that the distribution of data(gene expression) does not have bimodal patterns, while the large LHRs suggest bimodal distribution. This method was inherited from glucose bimodality studies [177, 178]. First Ertel and

Tozeren used the Box-Cox transformation to eliminate the skewedness and then applied the EM algorithm. Secondly, they used the log Likelihood ratio with the *p* value obtained by using chi-square with six degrees of freedom aiming to find more accurate *p* values. Finally, they used $p<0.001$ as an ad hoc *p*-value to target genes with bimodality function. In this regard another technique was developed and applied by Bessarabova and colleagues to estimate bimodality by maximizing *t*-statistic like τ [216].

More recently, Wang *et al* (2009) have proposed a new approach which they have called the Bimodality Index (BI) [110]. The bimodality index method was motivated by the limitations of using BIC or AIC as a model for criteria selection, so these were replaced by a criterion that gives a better discrimination between the bimodal distributions. The method assumes that the distribution of a gene with bimodal expression can be expressed as a mixture of two distributions with an equal variance. The bimodality index method is based on the distance between two clusters means and defined as:

$$BI = \left[\pi(1-\pi)\right]^{\frac{1}{2}}\delta \tag{3.12}$$

where π is the proportion of samples in cancer data and δ is the distance between the two subgroups. Although there is merit in using BI and as it has also outweighed the others because of its capability to identify the bimodal genes and rank them, it does however have some limitations as the two subgroups were assumed have the same variance and this is not true in all cases.

3.3. The proposed method

3.3.1. Motivation

It is unrealistic to assume that two clusters of a bimodal gene should have the same variance. The examination of various data sets has clearly shown that one of two clusters, either being of lowly expressed samples or of highly expressed samples is very likely to demonstrate a comparatively flat distribution while the other shows a tight cluster. Thus, I deliberately removed the constraint of homogeneous variance across genes because this constraint is certainly not valid in reality. The main differences between BI and hBI are summarised in Table 3.1. Finally, it worth noting that since significance analysis is critical to real biological/medical applications, I enhanced the Bimodality Index by using Besag's sequential Monte Carlo approach to deliver significance analysis.

Table 3.1: summary of the differences between BI and hBI

Factor	**BI**	**hBI**
model	parametric	Non-parametric
cross-cluster variance	homogeneous	heterogeneous
variance across genes	homogeneous	heterogeneous
Output	Index	Index and P value
Outlier sensitivity	No	Yes
Sample size	Large	Large

3.3.2. Algorithms

My proposed algorithm is a revision of the Bimodal Index (BI) [110], which is defined in Equation (3.12). The use of this definition implies a homogeneous variance for two clusters of samples. A one-side *t*-statistic of the i^{th} gene can be defined as

$$t_i = \frac{\mu_{H,i} - \mu_{L,i}}{\sqrt{\frac{\sigma_{L,i}^2}{n_{L,i}} + \frac{\sigma_{H,i}^2}{n_{H,i}}}} \tag{3.13}$$

where $\sigma^2_{L,i}$ is the variance of lowly expressed samples; $\sigma^2_{H,i}$ is the variance of highly expressed samples; $n_{L,i}$ is the number of lowly expressed samples; $n_{H,i}$ is the number of highly expressed samples; $\mu_{L,i}$ is the mean of lowly expressed samples; and $\mu_{H,i}$ is the mean of highly expressed samples of the *i*th gene. If $\sigma^2_{L,i} = \sigma^2_{H,i} = \sigma^2_i$, this one side *t*-statistic becomes

$$t_i = \sqrt{\frac{n}{\sigma_i^2}}\delta_i\sqrt{\pi_{H,i}(1-\pi_{H,i})} \tag{3.14}$$

where $\pi_{H,i}$ is the proportion of highly expressed samples of the *i*th gene, if the sample size is fixed for all genes.

$$t_i \propto \sigma_i^{-1}\delta\sqrt{\pi_{H,i}(1-\pi_{H,i})} \tag{3.15}$$

It can be seen that if homogeneous exists across subgroups and genes, BI is equivalent to one side *t*-statistic. However this can hardly be true in real applications. I therefore revised BI by employing heterogeneous variance. In the one side *t*-statistic, I use percentile estimations to replace the parametric estimation of means and variances as shown below

$$t_i = \frac{q^{25}_{H,i} - q^{75}_{L,i}}{\sqrt{\frac{\sigma^2_{L,i}}{n_{L,i}} + \frac{\sigma^2_{H,i}}{n_{H,i}}}} \tag{3.16}$$

Here q^{25}_H is the 25th percentile of highly expressed samples, q^{75}_L is the 75th percentile of lowly expressed samples, and the variances are calculated using

$$\sigma = \frac{\mathrm{IQR}}{1.34896} \tag{3.17}$$

where IQR is the interquartile of expression data. I assume that the separation between lowly expressed samples and highly expressed samples occurs at one of the largest gaps between consecutively sorted samples. Therefore I introduce the gap between lowly expressed samples and highly expressed samples to enhance the bimodality test. My heterogeneous bimodal index (*h*BI) is then defined below:

$$h\mathrm{BI}_{\mathrm{i}} = \alpha\left(m_{H,i} - M_{L,i}\right) + (1-\alpha)t_i \tag{3.18}$$

where $m_{H,i}$ is the minimum of highly expressed samples, $M_{L,i}$ is the maximum of lowly expressed samples of the i^{th} gene and $\alpha > 0$ is a trade-off between the gap effect and the *t*-statistic. In this thesis, $\alpha = 0.75$.

BI uses an arbitrary threshold to make decision based on the indexes. I employed the sequential Monte Carlo approach [226] Besag and Clifford, 1996) to deliver significance analysis. The procedure for the algorithm is shown below:

Step 1. BI calculation for each gene

1.1. sort expressions

1.2. calculate the distance between every consecutive expression and record them as a gap list.

1.3. sort the gap list

1.4. calculate the revised BI for the top ten gaps and record them in a bimodality list

1.5. maximise the bimodality list

Step 2. Apply BC algorithm to obtain *p* values

3.4. Evaluation and comparison of the proposed method and others

To evaluate the proposed algorithm by making a comparison with the Likelihood ratio test, Kurtosis test and BI test, I calculated sensitivity (Sen) and specificity (Spe), and used receiver operator characteristic (ROC) analysis [144]. Sensitivity is the ratio of correctly-predicted bimodal genes. Specificity is the ratio of correctly-predicted non-bimodal genes. I specifically calculated the area under the ROC curve (AUC) for comparison, and in all scenarios used the critical threshold of p=0.05.

3.4.1. Evaluation of control data (simulated data)

For the evaluation, five scenarios were chosen with different combinations of parameters to represent a wide range of bimodal shapes. Hellwig et al.(2010) in their comprehensive study of bimodality have identified an impression of major types of bimodality distributions [183]. They employed many bimodality scores to identify those shapes based on a real cancer data. In this thesis I have used all possible shapes that identified by the reference paper [183]. In conclusion, all shapes are applicable for the identification of genes with characteristic distributions that the benchmark algorithms designed for it.

For all five scenarios, 950 genes were designed as unimodal and 50 genes were designed as bimodal. Each gene had 40 replicates. Of these, 30 replicates were designed with low expressions, and ten with high expressions, and each simulation was repeated ten times.

3.4.1.1. Scenario 1

Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were

drawn from a normal distribution of mean ten and standard deviation one. Highly expressed samples of a bimodal gene were drawn from a normal distribution of mean 12 with variable standard deviation drawn from a uniform distribution between one and five. This scenario is in favour of kurtosis whereas only small samples or different variance between the two subgroups exist. **Table 3.2** shows the comparison based on the mean values among ten simulations for four algorithms using specificity, sensitivity, and AUC at the critical threshold p=0.05. It can be seen that hBI and Kurtosis have similar performances and hBI slightly outperforms Kurtosis analysis. The likelihood test shows the worst performance with the sensitivity as 0.06 although its specificity is 1.

Table 3.2. The averaged measurements for scenario 1; where LHR is likelihood ratio, K is stand for kurtosis, BI is Bimodality index and hBI is for the proposed method, spe stands for specificity, sen is sensitivity and auc is Area under ROC

	LR	K	BI	hBI
spe	1	0.983	0.975	0.992
sen	0.062	0.858	0.532	0.84
auc	0.995	0.964	0.852	0.992

3.4.1.2. Scenario 2

Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were drawn from a normal distribution of mean ten and standard deviation one. Highly expressed replicates of a bimodal gene followed a uniform distribution in the interval between zero and five in addition to maximum of low expressions. The averaged measurements are shown in **Table 3.3**. In this scenario, kurtosis has shown the worst accuracy (36%) while the other relatively similar and higher, 99.9%. In addition, the result has shown that the Likelihood test has very low sensitivity while BI and *h*BI perform equally well.

Table 3.3. The averaged measurements for scenario 2 ; where LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method, spe is specificity, sen is sensitivity and auc is Area under ROC

	LR	K	BI	hBI
Spe	1	0.986	0.997	0.9963
Sen	0.058	0	0.954	0.924
Auc	0.999	0.36	0.999	0.9985

3.4.1.3. Scenario 3

Samples of unimodal genes were drawn from a uniform distribution in the interval between ten and 12. Lowly expressed replicates of a bimodal gene were drawn from the same low expression distribution as bimodal genes and highly expressed replicates of a bimodal gene were drawn from a normal distribution with two units added to the maximum of the low expressions. **Table 3.4** shows the summary of the simulations for this scenario. This scenario indicates that Kurtosis again failed to have a sensible accuracy (14%). Meanwhile hBI shows the highest AUC (0.999), similar to LR (0.996) and BI (0.993); hBI also outweighs LR and BI in term of sensitivity. The sensitivities of BI and LR are 0.81 and 0.77 respectively, while *h*BI's sensitivity is 0.94.

Table 3.4. The averaged measurements for scenario 3; where LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method, spe is specificity, sen is sensitivity and auc is Area under ROC

	LR	K	BI	hBI
Spe	0.997	0.9519	0.9898	0.996
Sen	0.774	0	0.812	0.94
Auc	0.996	0.1413	0.9933	0.999

3.4.1.4. Scenario 4.

Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed replicates of a bimodal gene were drawn from a mixture of a normal distribution of mean ten and a normal distribution of mean 12. The standard deviation of the former was designed as one and that of the latter was designed as three. Highly expressed replicates of

a bimodal gene were drawn from the low expressions plus white noise, with two units above the maximum low expression. **Table 3.5** shows the summary of ten simulations on random samples for this scenario. All performed very well in terms of AUC. This means that there are some suitable statistical significance levels by which perfect separation between unimodal and bimodal genes can be found.

Table 3.5. The averaged measurements for scenario 4; where LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method, spe is specificity, sen is sensitivity and auc is Area under ROC

	LR	K	BI	hBI
spe	1	0.9861	0.9963	0.997
sen	0.468	0	0.93	0.952
auc	0.998	0.9572	0.9984	0.9988

3.4.1.5. Scenario 5

Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed replicates of a bimodal gene were organised as a mixture of three normal distributions with mean values as ten, 11 and 12, as well as standard deviation values as three, two and one. Highly expressed replicates of a bimodal gene were drawn in the same way as Scenario 4. Based on ten random simulations for this scenario, it was observed that although LR and BI showed reasonably good values of AUC, their sensitivities were not acceptable. **Table 3.6** shows that these two algorithms have the same problem that was encountered in Scenario 4: their *p* values tend to be large, which leads to the difficulty of using command significance levels to make decisions. Kurtosis analysis does not work well because its AUC value drops to 0.66, not very far from 0.5, hence occurs by chance. In this scenario *h*BI performs the best in all measurements while BI has 69% sensitivity.

Table 3.6. The averaged measurements for scenario 5; where LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method, spe is specificity, sen is sensitivity and auc is Area under ROC

	LR	K	BI	hBI
spe	1	0.9844	0.9845	0.9873
sen	0	0	0.698	0.766
auc	0.9267	0.6676	0.9593	0.9867

3.4.2 Evaluation of Real Data

In this section I applied all four methods to a three real cancer data sets. This is to make the comparison more realistic. Also, all data were normalised to base two logarithm scale. All evaluation measurements were based on using three different p values.

3.4.2.1. The GSE11121 dataset

The data set was downloaded from GEO (Gene Expression Omnibus) where the raw expression was deposited under accession number GSE11121. It contained 200 lymph node-negative breast cancer patients who were not treated by systemic therapy after surgery. The data was for a derivation study to find prognostic motifs [227]. Gene expression profiling of patients was done using the Affymetrix HG-U133A microarray platform compromising 22283 probe sets.

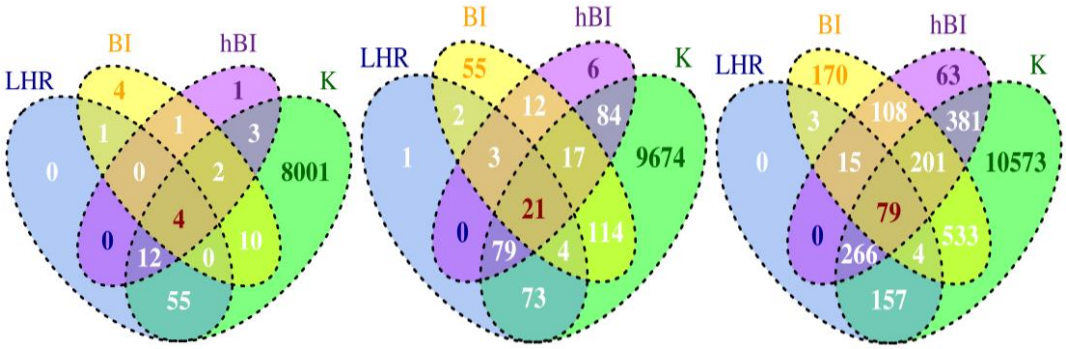
I transformed the expression using a base two logarithm before analysis. I used three significance levels of critical *p* values (0.001, 0.01 and 0.05) to predict bimodal genes, and the predicted bimodal genes using these three significance levels are shown in **Table 3.7**. The Likelihood test predicted from 0.3% to 2.3% bimodal genes, BI predicted from 0.01% to 5% bimodal genes, and *h*BI predicted bimodal genes from 0.01% to 5% as well. However Kurtosis analysis ended up with too many predictions up to 54.7%, which was unreasonable.

Even for the significance level of 0.001, it still predicted 36.3% bimodal genes, which is far beyond a realistic level.

Table 3.7 Predicted bimodal genes for 3 significance levels for dataset GDS11121. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

Significance levels	LHR	K	BI	hBI
0.001	72	8087	22	23
0.01	182	10065	227	221
0.05	523	12193	1112	1112

The overlap analysis between four algorithms based on significance level 0.001, using the VennDiagram package in R [228] is shown in **Figure 3.2**. I found that *h*BI was the most similar to BI. The overlap percentage between these two algorithms is 31.8%, i.e., 100*7/(7+14+1). The overlap degree between LHR and *h*BI is 5.1%, and the overlap degree between LHR and BI is 5.6% (Figure 4.1a). It seems that 91.3% of predicted bimodal genes of *h*BI are also predicted by Kurtosis. This percentage drops to 69.6% between *h*BI and LHR, and to 30% between *h*BI and BI. Further the overlap percentage between BI and *h*BI is 23.3% for significance level 0.01 and 55.6% between the *h*BI and LHR and 16.5% between BI and LHR (Figure 4.1.b). Here 90.9% of predicted bimodal genes of *h*BI are predicted by Kurtosis as well. This percentage drops to 46.6% between *h*BI and LHR, and to 24% between *h*BI and BI. For significance level 0.05, the overlap between hBI and BI was found to be 36.2%, between *h*BI and LHR it was 68.8%, and between BI and LHR was 19.3% (Figure 2.1 c). In addition, 83.3% (32.3%, 36.2%) of *h*BI's predictions are consistent with Kurtosis (LHR, BI).



(a) (b) (c)

Figure 3.2. Venn diagram illustrates the overlap between the methods for GSE11121 with the significance levels 0.001(a), 0.01(b), and 0.05(c).

Figure 3.3 shows the top five bimodal genes predicted based on the significance level 0.001, where A-D was predicted by all and E was predicted by *h*BI only. It can be seen that they showed different types of distributions. Both *FOXA1* (Fig 3.3A) and *GATA3* (Fig.3.3B) show a pattern in which the high expressions form a tight cluster. However the low expressions demonstrate a flatter distribution or form smaller clusters. *TDRD12* (Fig.3.3C) and *GRIA2* (Fig, 3.3D) have tight clusters formed by low expressions and their high expressions display flat distributions. *SH3GL3* (Fig.3.3E) shows a different pattern from the other four. It is composed of two more tightly formed clusters, one small and one large, and the gap between the two clusters is large. Analysis of these patterns proves one important notion – that the use of restrictive assumptions of data distribution may not be sufficient for accurate prediction of bimodal genes in real applications, where distribution can be varied over many different formats.

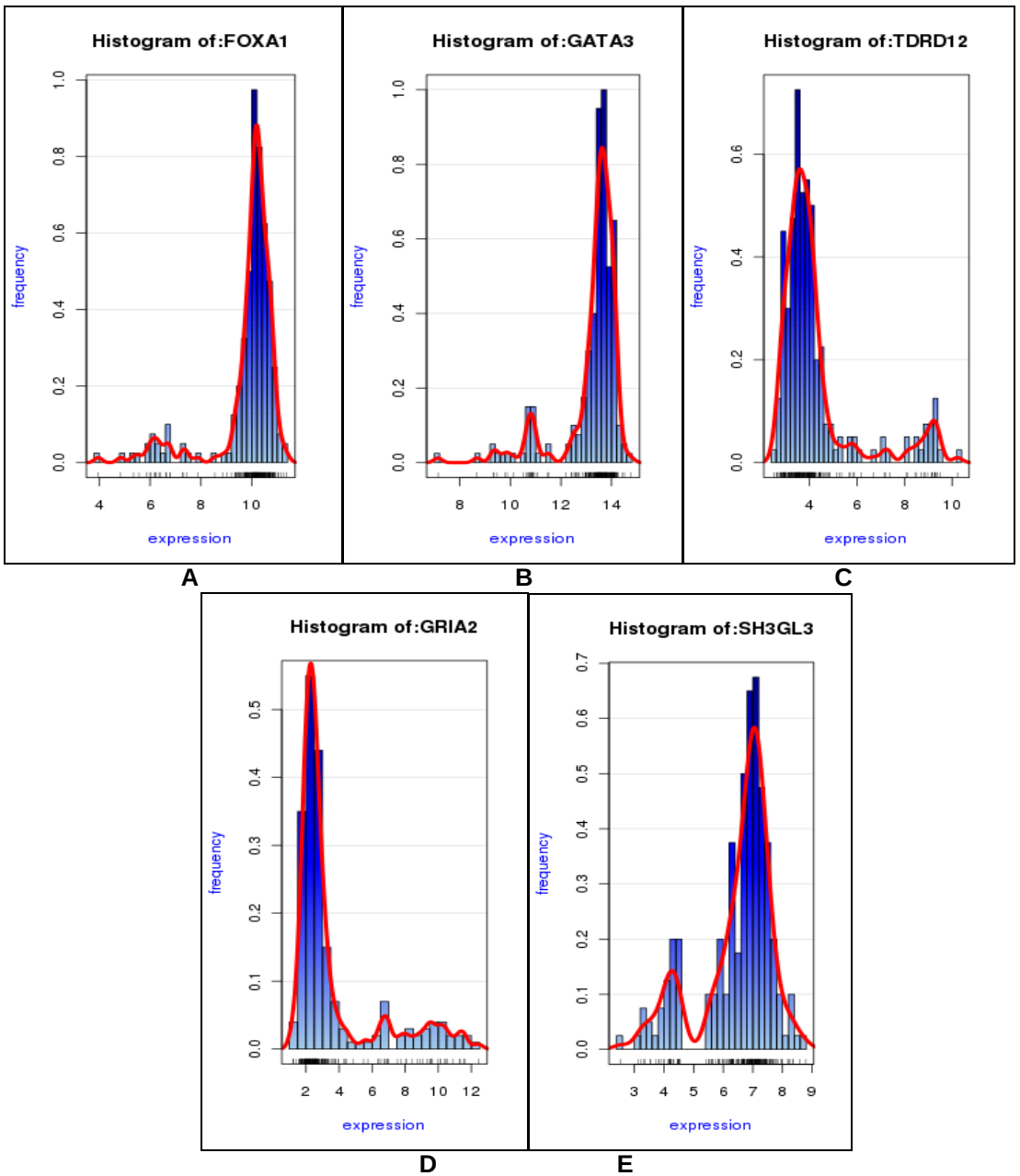


Figure 3.3. Density analysis of four bimodal genes (**A-D**) predicted by all four algorithms at the significance level 0.001, and **E** predicted only by *h*BI at the same significance level. The horizontal axes represent log2 expressions and the vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

The *p* values of four algorithms for the genes uniquely predicted by *h*BI at the significance level 0.01 are given in **Table 3.8**. The data shows that for those

bimodal genes predicted by *h*BI, the ranking by other algorithms is far low. For instance, the Kurtosis rank of C6orf64 is 18643 and the Kurtosis rank of *GULP1* is 22101.

Table 3.8 The *p* values of bimodal genes predicted ONLY by *h*BI at significance level 0.01 for GDS11121. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

Row	ID_Ref	Gene symbol	LHR	K	BI	hBI
4575	205048_s_at	PSPH	0.27(2990)	0.1(13507)	0.058(1293)	0.002(66)
21574	222214_at	unknown	0.065(652)	0.03(11438)	0.039(864)	0.004(97)
20803	221440_s_at	RBBP9	0.052(540)	0.36(17000)	0.043(974)	0.005(113)
15000	215627_at	unknown	0.31(3584)	0.02(11311)	0.012(265)	0.008(195)
18148	218784_s_at	C6orf64	0.07(713)	0.54(18643)	0.018(416)	0.008(199)
15288	215915_at	GULP1	0.19(1879)	0.97(22101)	0.026(582)	0.009(226)

Figure 3.4 shows four of them which have gene symbols and are indeed bimodal genes. However three other algorithms failed to predict them. For instance, PSPH was ranked by *h*BI at the 66th position (p = 0.002). Likelihood, Kurtosis and BI respectively ranked it at the 2990th position (p = 0.2), 13507th (p = 0.1), and the 1293rd (p = 0.05). This gene is highly expressed in African-American colorectal cancer patients compared with European- Americans [229]. PSPH is also expressed at a higher level in responding patients versus non-responding groups, which supports its importance as a therapeutic target for non-small-cell lung cancer [230]. RBBP5 was ranked at the 113th position (p = 0.005), but was ranked at the 540th position (p = 0.52), the 17000th position (p = 0.36), and the 974th position (p = 0.043) by the Likelihood, Kurtosis and BI tests. RBBP5 was found to show only 40% of Pancreatic ductal adenocarcinomas (PDAs) [231].

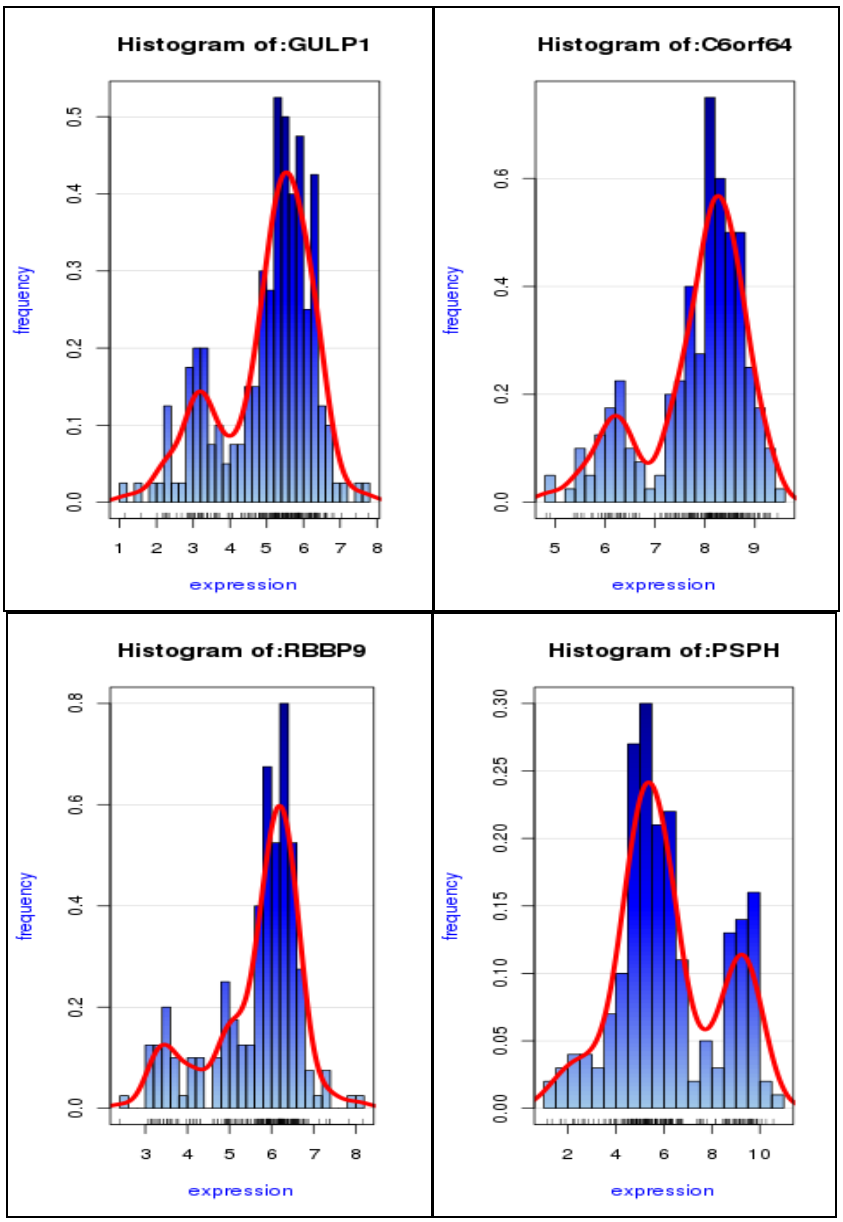


Figure 3.4 Density analysis of four bimodal genes predicted only by *h*BI at the significance level 0.01. The horizontal axes represent $\log_2$ expressions and vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

3.4.2.2. The GSE2034 data set

The data set was downloaded from GEO (Gene Expression Omnibus). It contained 286 lymph node-negative patients who were not treated specifically by adjuvant systemic therapy, 180 lymph node–negative relapse-free patients, and 106 lymph node–negative patients who had developed a distinct metastasis. The data was subject to derivation analysis by Wang *et al* to predict risk to patients based on gene expression profiles [232]. Gene expression

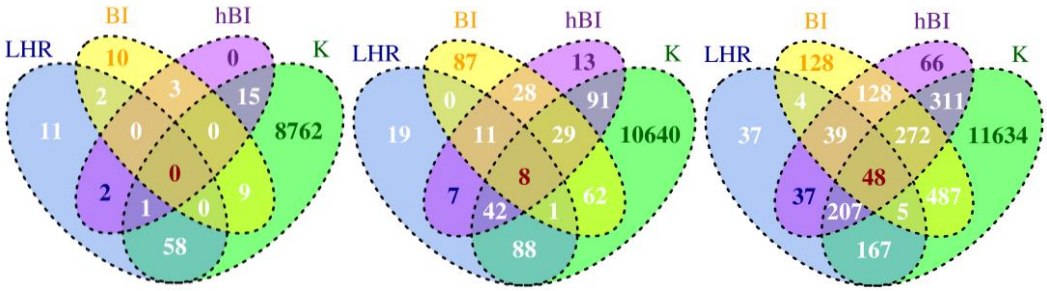
profiling of patients was done using the Affymetrix HG-U133A microarray platform, compromising 22283 probe sets. The raw expression was deposited at the NCBI GEO data repository under accession number GSE2034.

For the purposes of this study, the expressions have been transformed onto a common log2-scale. **Table 3.9** shows the predicted bimodal genes using these three significance levels. The Likelihood test predicted from 0.3% to 2.44% bimodal genes, BI predicted from 0.09% to 5% bimodal genes and *h*BI predicted bimodal genes from 0.1% to 5% as well. However the Kurtosis analysis ended up with too many predictions up to 58.9%, which is unreasonable. Even for the significance level 0.001, it still predicted 39.7% bimodal genes, which is far beyond a realistic level.

Table 3.9 The number of predicted bimodal genes is shown for three significance levels for data set GDS2034. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

Significance level	LHR	K	BI	hBI
0.001	74	8845	24	21
0.01	176	10961	226	229
0.05	544	13131	1111	1108

Figure 3.5 shows the overlap analysis between four algorithms at three different significance levels using the Venn Diagram [228]. It was found that *h*BI was most similar to BI. The overlap percentage between these two algorithms is 39.6%. The overlap degree between LHR and *h*BI is 38.6%, while the overlap degree between LHR and BI is 11.3% (Fig 3.5b). Kurtosis as well as *h*BI predicted 74.2% of the bimodal genes. This percentage drops to 29.7% between *h*BI and LHR, as well as to 33.2% between *h*BI and BI. Figures 3.5a and 3.5c show the overlap analysis at the significance levels of 0.001 and 0.05.



(a) (b) (c)

Figure 3.5. The Venn diagram illustrates the overlap between the methods at significance levels 0.001(a), 0.01(b) and 0.05(c) (GDS2034).

The unique predictions of *h*BI at the significance level 0.01 are shown in **Table 3.10**, which shows that most rankings of other algorithms are far behind. The three genes listed in **Table 3.10**, which all show typical bimodal genes are also shown in **Figure 3.6**.

ERBB2 was identified only by *h*BI at the 114th position (p = 0.005), whereas it was ranked at the 372th position (p = 0.035), the 11301st position (p = 0.198), and the 1564th position (p = 0.07) respectively, by the Likelihood, Kurtosis and BI tests. It has been reported that over-expressed in 30% of breast cancer patients are linked to poor prognosis [233, 234]. *FGFR3* was ranked by *h*BI at the 122nd position (p = 0.005). The Likelihood, Kurtosis and BI tests ranked it at the 3609th position (p = 0.267), the 11594th position (p = 0.198), and the 1564th position (p = 0.07), respectively. It was highly expressed in Tamoxifen resistance breast tumours [235]. *WHSC1* was ranked by *h*BI at the 155th position (p = 0.006), but ranked respectively at the 2194th position (p = 0.18), the 12468th position (p = 0.276) and the 775th position (p = 0.034) by the Likelihood, Kurtosis and BI tests. It is over-expressed in approximately 75% of neuroblastomas and is associated with aggressiveness [236].

Table 3.10 The *p* values of top bimodal genes predicted only by *h*BI at significance level 0.01 for data set GDS2034. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

ID_Ref	Gene symbol	LH	K	BI	hBI
210246_s_at	ABCC8	0.07(730)	0.56(14986)	0.015(336)	0.004(95)
216836_s_at	ERBB2	0.035(372)	0.17(11301)	0.019(433)	0.005(114)
204379_s_at	FGFR3	0.267(3609)	0.198(11594)	0.07(1564)	0.005(122)
203163_at	KATNB1	0.044(452)	0.844(17028)	0.016(361)	0.006(152)
220559_at	EN1	0.084(908)	1.3E-05(18580)	0.134(2987)	0.006(154)
209052_s_at	WHSC1	0.180(2194)	0.276(12468)	0.034(775)	0.006(155)
221195_at	RNFT1	0.301(4169)	6.2E-05(21324)	0.011(245)	0.007(160)
209644_x_at	CDKN2A	0.032(344)	7.1E-09(21579)	0.101(2251)	0.007(166)
210990_s_at	LAMA4	0.051(526)	3.9E-06(20503)	0.077(1717)	0.008(181)
215867_x_at	CA12	0.112(1234)	0.595(15222)	0.028(633)	0.008(186)
210066_s_at	AQP4	0.059(618)	0.193(11527)	0.015(342)	0.009(217)
207302_at	SGCG	0.071(770)	2.2E-05(19451)	0.095(2133)	0.009(220)
213174_at	TTC9	0.042(432)	0.576(15072)	0.015(333)	0.01(227)

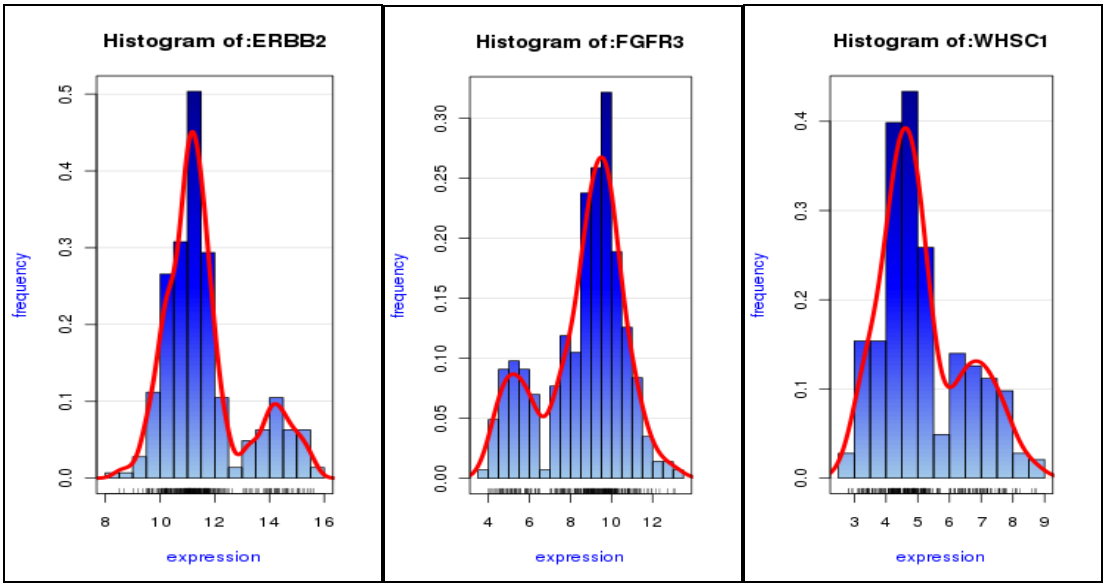


Figure 3.6. Density analysis of three bimodal genes only predicted by *h*BI at the significance level 0.01. The horizontal axes represent $\log_2$ expressions and vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

3.4.2.3. The GSE1456 dataset

The data set was downloaded from GEO (Gene Expression Omnibus). It contained 159 patients with breast cancer, 126 of whom received adjuvant systemic therapy and 33 who were not being treated with it. The data was derivation research to predict the impact of adjuvant therapies on breast cancer patients using gene expression profiling [237]. Gene expression profiling of

patients was done using the Affymetrix HG-U133B microarray platform, compromising 22283 probe sets. The row expression was deposited at the NCBI GEO data repository under accession number GSE1456. **Table 3.11** shows the numbers of genes identified within different thresholds. The overlap between the four methods at the significance level 0.01 is indicated in **Figure 3.7**. The Likelihood test predicted from 1.16% to 8.4% bimodal genes, BI predicted from 0.1% to 5% bimodal genes and *h*BI predicted bimodal genes from 0.1% to 5% as well (Figure 4.6a). However the Kurtosis analysis ends up with too many predictions (up to 57.9%), which is unreasonable. Even for the significance level 0.001, it still predicts 36.6% bimodal genes, which is a far more realistic level. Figures 4.6b and 4.6c show the overlap analysis for the significance levels at 0.001 and 0.05.

Table 3.11. Number of predicted bimodal genes for three critical *p* values for data set GDS1456. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

Significance level	LHR	K	BI	hBI
0.001	260	8152	23	24
0.01	701	10299	226	227
0.05	1878	12909	1129	1136

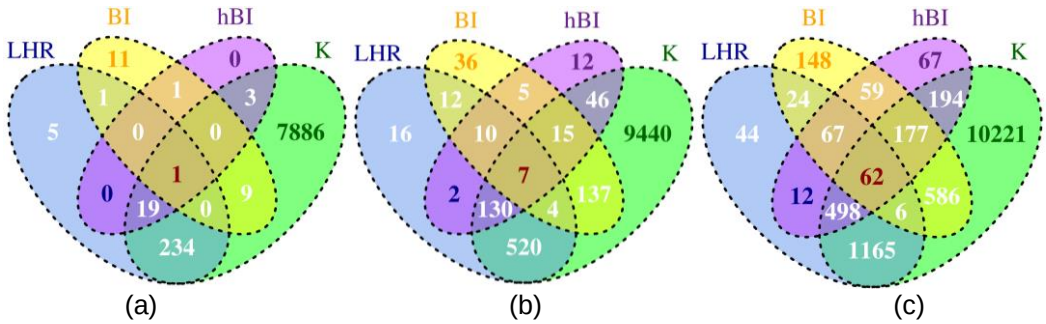


Figure 3.7. Venn diagram illustrating the overlap between methods at significance levels 0.001(a), 0.01(b), and 0.5(c).

Table 3.12 gives the bimodal genes uniquely predicted by *h*BI. It again shows that the rankings of other algorithms sometimes are very behind. **Figure 3.8** shows six of them with typical bimodality. RBP5 was ranked at the 181th position (p = 0.008), but ranked at the 779th position (p = 0.011), the 13344th position (p = 0.062) and the 533th position (p = 0.023) by the Likelihood, Kurtosis and BI tests. It is reported to be associated with axillary lymph node metastases in breast cancer and to play an important role in regulating the migration of breast cancer tumours[238].

Table 3.12 The *p* values of top bimodal genes predicted by *h*BI at critical *p* value 0.01 for data set GDS1456. LHR is likelihood ratio, K is for kurtosis, BI is Bimodality index and hBI is for the proposed method.

Row	ID_Ref	Gene symbol	LH	K	BI	hBI
12745	235059_at	RAB12	0.042(1684)	0.278(16884)	0.016(363)	0.003(76)
9989	232298_at	hCG_1806964	0.038(1568)	0.833(21499)	0.014(333)	0.005(135)
7862	230171_at	unknown	0.05(1961)	0.091(14083)	0.026(608)	0.006(141)
21196	243510_at	unknown	0.728(11372)	0.476(18813)	0.187(4253)	0.006(144)
17140	239454_at	SCARF2	0.069(2454)	0.061(13303)	0.029(665)	0.006(149)
14704	237018_at	unknown	0.048(1843)	0.164(15391)	0.015(351)	0.007(166)
11510	233823_at	FAM184B	0.095(3096)	0.206(15980)	0.01(244)	0.007(172)
1533	223820_at	RBP5	0.011(779)	0.062(13344)	0.023(533)	0.008(181)
17089	239403_at	CCDC120	0.191(5381)	0.014(10857)	0.015(361)	0.008(188)
10932	233243_at	unknown	0.042(1699)	0.215(16123)	0.01(243)	0.009(241)
14467	236781_at	unknown	0.199(5582)	0.016(11068)	0.017(397)	0.009(221)

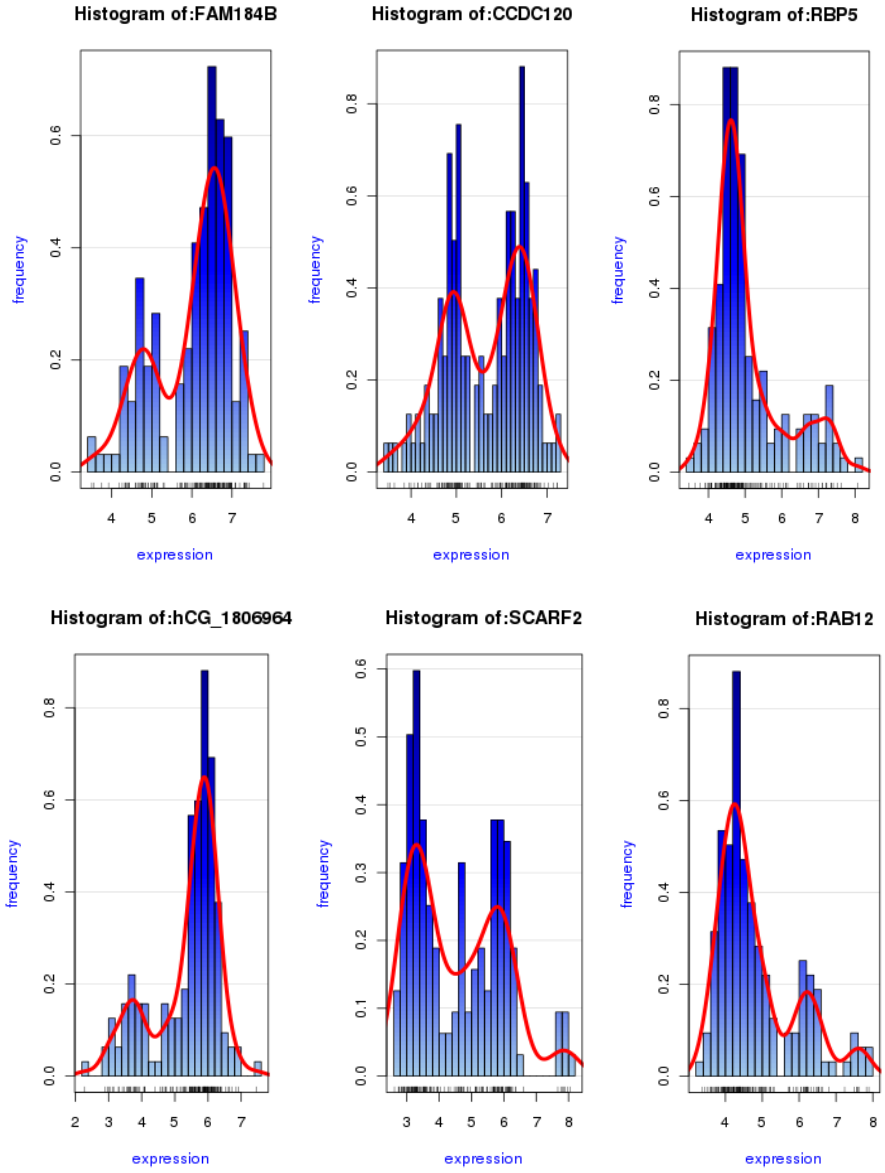


Figure 3.8 Density analysis of six bimodal genes predicted only by *h*BI at the significance level 0.001. The horizontal axes represent $\log_2$ expressions and vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

3.5 Remarks

It is worth to discuss the differences in these methods in term of characteristic distributions. Out of the top 10 genes identified by each method, I have summarised the differences of these algorithms. For example genes with the smallest *p* value from kurtosis method is either of a unimodal with a small outlier group with high expression values except for one case the other way around or bimodal densities with nearly equal sizes in both groups (Fig 3.9,top

two rows). This can be linked to the negative kurtosis (outliers) and positive kurtosis (bimodal). The top 10 genes identified by LHR show clearly a large group with a small variance and a small group with large variance (Fig 3.9, top two rows).

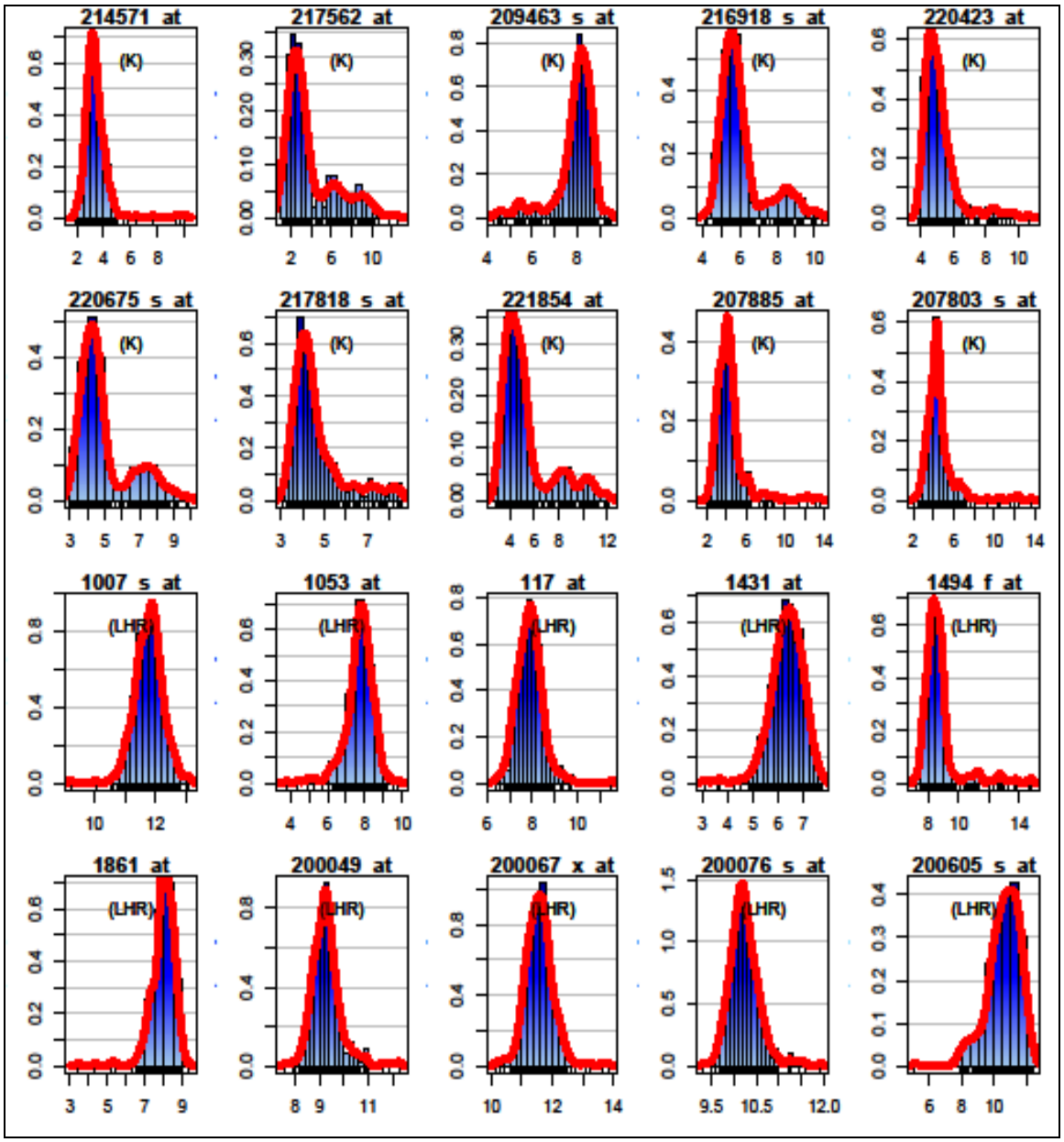


Figure 3.9: The characteristic distributions of the top 10 genes identified by kurtosis (top 2 rows- labelled in brackets with (K)), and LHR(row 3 and 4- labelled in brackets with(LHR)).

Similarly, the top genes recognized by BI clearly show bimodal genes with two peaks (**Fig 3.10**, top two rows). BI was able to identify unbalanced sizes between the two groups. The majority of top ranked genes by BI is bimodal genes with considerable sizes in both groups and there is no outlier identified in the top ones. However, top ten genes predicted by *h*BI show a combination between all cases identified by all algorithms (**Fig 3.10**, bottom two rows). I have admitted that the proposed method is sensitive to outliers as the others, but not BI. Four genes are in the top 10 genes for both hBI and BI and 3 genes common between k and hBI. As noticed in this chapter that each algorithm is designed to target genes with special characteristic distributions as seen in (**Figs 3.9 and 3.10**). Thus, relying on one algorithm will definitely ignore some bimodal genes. However, hBI promised to be a good method for such analysis where it able to identify different shapes.

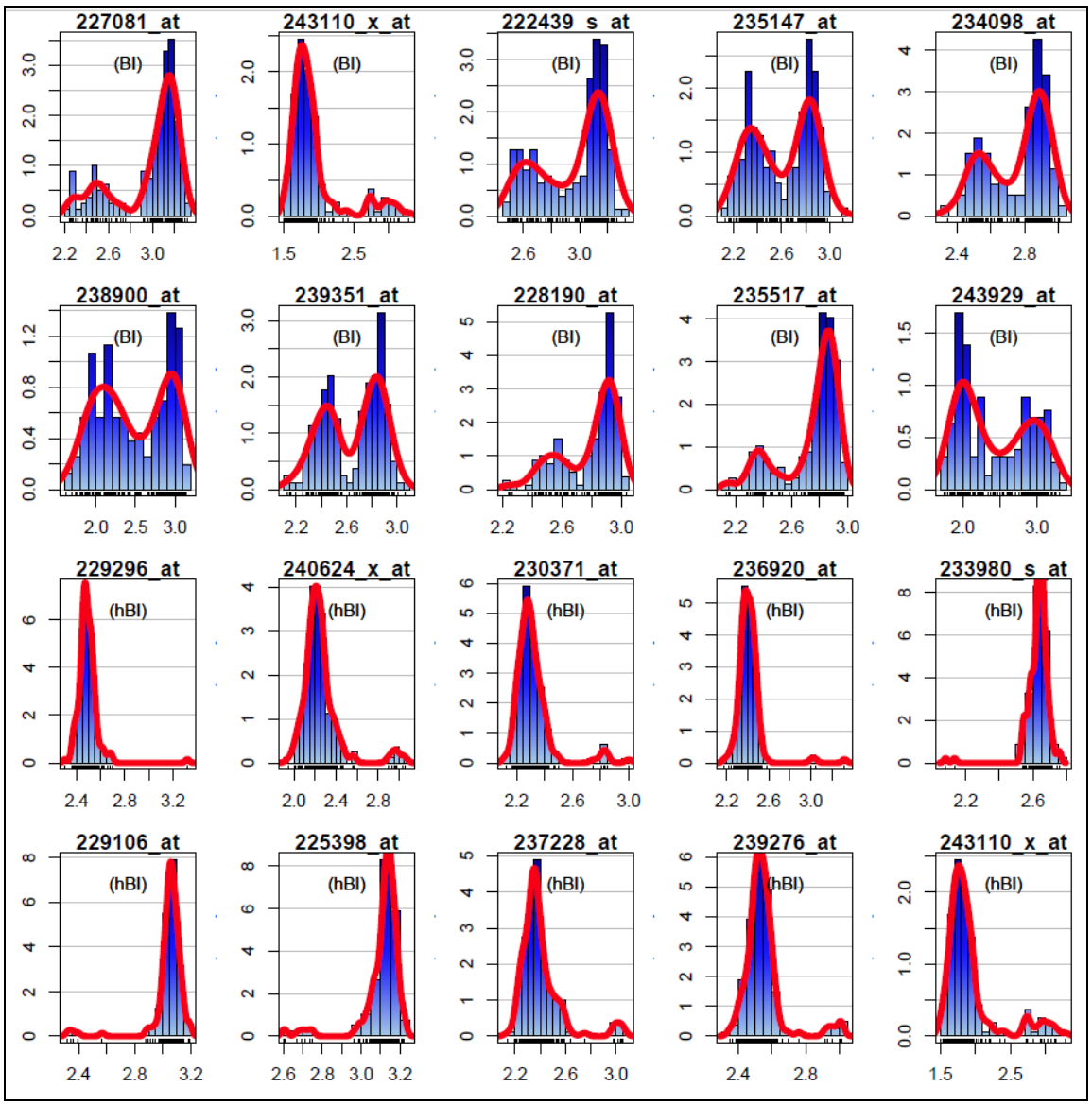


Figure 3.10: The characteristic distributions of the top 10 genes identified by BI (top 2 rows- labelled in brackets with(BI)), and hBI(row 3 and 4- labelled in brackets with(hBI)).

3.6. Conclusion

The importance of bimodality was examined with some examples of its effects in cancer and drug design, and all available methods for bimodality identification were critically discussed. To overcome all the disadvantages of other methods, a novel bimodal gene prediction algorithm (*h*BI) proposed by way of relaxing the constraints of the Bimodal Index algorithm. First, the constraint of cross-cluster homogeneous variance was removed. It is very unrealistic to assume that two

clusters of a bimodal gene should have the same variance. Examination of various real data sets has clearly shown that one of two clusters, either being of lowly expressed samples or of highly expressed samples, is very likely to demonstrate a comparatively flat distribution while the other shows a tight cluster. Secondly, I deliberately removed the constraint of homogeneous variance across genes because this constraint is certainly confusing. Obvious evidence of this is that the variance of unimodal genes and bimodal genes will not show homogeneous variance.

In addition to these two revisions, I also emphasised the impact of gaps between consecutive expressions of sorted samples on bimodal formulation. This is because I have observed in real data sets that lowly expressed samples often demonstrate a tight cluster and highly expressed samples show; i) a comparatively large variance; and ii) distantly depart from the tight cluster of lowly expressed samples -or the other way around. In this case, although I was using percentiles to estimate mean values and standard deviations, the *t*-statistic was still not working well, i.e., it was very likely to be small due to the large variance of the highly expressed samples. A gap impact was therefore introduced into the prediction of bimodal genes. Because significance analysis is critical to real biological/medical applications, I enhanced the Bimodal Index using Besag's sequential Monte Carlo approach to deliver significance analysis for the proposed algorithm as well as for the Bimodal Index algorithm.

In this chapter, all simulations showed that the new algorithm is comparable or better than the benchmark algorithms in simulated data sets. However, when applying to real cancer data sets, the new algorithm is partially consistent with benchmark algorithms and provides better insights into the analysis of bimodal genes. Importantly, most of the bimodal genes predicted by the new algorithm

do show typical bimodality. A small percentage of my predictions are not favoured by benchmark algorithms.

Chapter 4

Investigation of Bimodality Patterns in Cancer gene expression data

Abstract

Bimodality is one of the common phenomena observed in many cancer studies, and the identification of bimodal genes is an important task since these genes can be biomarker candidates. Chapter 3 presented a new algorithm to identify bimodal genes and showed that it performed better in the simulated data than other algorithms and worked well on a real data. Here, I used the proposed method to extend the investigation to real applications, since it was most important to study the phenomena among large-scale gene expressions of cancer. In this chapter I investigate bimodality patterns among different patient cohorts, platforms, and cancer types, and showed that diversity among cancer samples is common.

4.1. Introduction

A comprehensive investigation of the genes with bimodal distribution was carried out in different cancer types and included more than 70 data sets, covering seven cancer types. The data were performed on different platforms. The study designed to test if the bimodality is common phenomenon in cancer gene expression and not influenced by specific platform, type, or study.

The gene expression distribution is assumed to be a normal distribution. As a result, many studies have used differential expression methods to identify biomarkers in cancer by comparing the two states (normal vs. cancer; type I vs. type II). However, this is not always a true assumption since many studies have recently indicated that some genes show a bimodal distribution within the category [179, 205]. Bimodality has been reported in many different cancer types as well as in other diseases, e.g., diabetes, and this might be related to the heterogeneity of cancer.

In this section, I investigate the genes with comprehensive bimodal distribution in different cancer types. I hypothesised that bimodality of gene expression is dependent on the disease and common property of a gene, and not related to a specific platforms or so. However, it is obvious that various groupings of different samples, arrays, and platforms may influence and/or increase certain outcomes with regard to bimodality. Thus, in this chapter an investigation is conducted to validate this assumption, based on a large-scale analysis. It has been reported in a small scale analysis on breast cancer that bimodality is platform-independent [216]. Large numbers of the bimodal genes have been identified as differential expressed genes as reported in some studies [239]. However, the normal samples might have a bimodal distribution. This is linked

to the subgroups in a population (i.e. disease free samples/control). All data used in this chapter were selected from the same physiological stage of cancer (i.e. stage1 or stage 2). In this regards some of the original datasets in this chapter were divided into two or more independent sets. This has been explained in following section.

4.2. Datasets

The datasets used in this project were downloaded from the NCBI GEO data repository [13, 14]. In this study I include seven cancer types with a minimum of ten sets of each type. A simple explanation of each dataset is presented here with information on the design of the experiments, but full details can be found in the reference for each dataset or on the GEO website.

4.2.1. Colon Cancer Data

Table 4.1. Colon cancer data sets used in this study

Dataset	Probe ID	Sample	Microarrays platform
GSE18088	54675	53	HG-U133_Plus_2
GSE31595 _1	54675	20	HG-U133_Plus_2
GSE31595 _2	54675	17	HG-U133_Plus_2
GSE38026	33257	16	HuGene-1_1-st
GSE4107	54675	12	HG-U133_Plus_2
GSE4183 _1	54675	15	HG-U133_Plus_2
GSE4183 _2	54675	15	HG-U133_Plus_2
GSE4183 _3	54675	15	HG-U133_Plus_2
GSE8671	54675	32	HG-U133_Plus_2
GSE10950	22184	24	Illumina BeadChip Human Ref8-v2

Table 4.1 gives a summary of colon cancer data sets used later in this chapter for bimodality investigation. The GSE18088 set profiled 53 primary stage II colon cancer patients treated by elective standard oncological resection, none of whom had received adjuvant chemotherapy [240]. This was profiled on HG-U133_Plus_2 platform. The GSE31595 set contains 37 samples from stage II and III colon cancer patients. Concerning bimodality identification, each stage was used as an independent set, specifically 20 for stage II and 17 for stage III.

These were also performed on HG-U133_Plus_2 platform. The GSE38026 dataset contains 16 colorectal cancer patients with their normal corresponding. This project used only the 16 cancer samples regardless of KRAS-mutation presence or absences, and was profiled on HuGene-1_1-st microarray platform.

The GSE4107 expression data [241] were extracted from the colonic mucosa of healthy controls and patients where only 12 cancer cases were included in the study. This data was analysed using U133-Plus 2.0 Array. The GSE4183 dataset contained 15 patients with colorectal carcinomas (CRC), 15 with adenoma, 15 with inflammatory bowel diseases (IBD) and eight control samples [242, 243]. The analysis used only three different disease states as independent sets, and the data was analysed on HGU133 Plus 2.0 microarrays.

The GSE8671 dataset containing 32 samples from patients with colorectal adenomas and a corresponding normal sample were used as the disease cases in the analysis, which was performed on HG-U133_Plus_2 [244]. The GSE10950 set was designed to compare colon tumours (24 sample) with normal colon mucosa (24 sample) and was analysed using Illumina BeadChip Human Ref8-v2 [245].

4.2. 2. Liver Cancer Data

Table 4.2 gives a summary of liver cancer data sets used in this chapter for bimodality investigation. The GSE24520 set was divided to three separate data sets of 13, 12, and 13 samples respectively. The data from 24526 probes was conducted on the Illumina platform but needs to be removed or replaced as it is too complicated.

Table 4.2. Liver cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE10694_2	121	78	CapitalBio Mammalian miRNA Array Services V1

GSE14520	22268	22	HG-U133A_2
GSE24520_1	24526	13	Illumina HumanRef-8 v3.0 expression beadchip
GSE24520_2	24526	12	Illumina HumanRef-8 v3.0 expression beadchip
GSE24520_3	24526	13	Illumina HumanRef-8 v3.0 expression beadchip
GSE25097	37582	268	Affymetrix 1.0 microarray, Custom CDF
GSE29721	54675	10	HG-U133_Plus_2
GSE31370_2	47323	15	Illumina HumanHT-12 V4.0 expression beadchip
GSE32225	24526	149	Illumina HumanRef-8 WG-DASL v3.0
GSE35306	33297	20	HuGene-1_0-st
GSE38941	54675	17	HG-U133_Plus_2

The GSE29721 dataset consists of 10 cancer samples and 10 normal adjacent tissues obtained from patients with hepatic cellular carcinoma (HCC) [246]. The expression data for this set was generated using HG-U133_Plus_2 platform. The GSE31370 data includes expression profiles of three different type of liver cancer, namely hepatocellular carcinoma (HCC), cholangiocarcinoma (CC) and HCC with fibrous stroma. In respect of bimodality analysis only one type was included, with 15 samples [247]. All data were obtained using Illumina HumanHT-12 V4.0 expression beadchip platform. The GSE32225 dataset was performed on Illumina HumanRef-8 WG-DASL v3.0 for 149 patients with intrahepatic cholangiocarcinoma (ICC). They also provided six samples as control [248], but were dropped from our analysis for the reason mentioned earlier. The GSE35306 dataset contained three different types of disease [249], but only the combined hepatocholangiocarcinomas (cHCC-CC) were included. This type consisted of 20 samples profiled on HuGene-1_0-st.

The GSE38941 data were obtained from four patients with HBV-associated acute liver failure; there were 17 samples which were amplified on HG-U133_Plus_2 [250]. Part of the GSE10694 was included. GSE14520 contains huge gene expression data of human hepatocellular carcinoma (HCC) as well as their normal corresponding tissues. From this only 22 samples were carried out on HG-U133A 2.0 arrays. The GSE25097 data were profiled from human

liver hepatocellular carcinoma profiles on Rosetta/Merck Human RSTA Affymetrix 1.0 microarray, Custom CDF and used only the 268 HCC tumour samples.

4.2.3. Prostate Cancer Data

Table 4.3. Prostate cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE16560_1	6144	281	Human 6k
GSE21034	43419	131	HuEx-1_0-st
GSE29079	17881	47	HuEx-1_0-st
GSE41408_1	17881	48	HuEx-1_0-st
GSE6605-GPL8300	12558	25	HG_U95Av2
GSE6605-GPL92	12553	25	HG_U95B
GSE6605-GPL93	12579	25	HG_U95C
GSE6606-GPL8300	12625	65	HG_U95Av2
GSE6606-GPL92	12553	66	HG_U95B
GSE6606-GPL93	12646	65	HG_U95C

Table 4.3 gives a summary of prostate cancer data sets used later in the chapter for bimodality investigation. The GSE29079 data contains gene expression profiles of 48 normal and 47 prostate tumour tissue samples. Those performed on HuEx-1_0-st (Affymetrix GeneChip Exon 1.0 ST microarrays) [251] were only the cancer data included in my analysis. The GSE41408 data include 48 Prostate cancer samples from radical prostatectomies [252]. The GSE6605 data contains metastatic prostate tumour samples obtained from 4 patients which were profiled on three different platforms (HG_U95B, HG_U95C and HG_U95Av2) [253], and were used as 3 independent sets. The GSE6606 was similar to the previous one but the sample size was larger. The GSE16560 contains 281 prostate cancer cases from Sweden and the data is widely known as the 'Swedish-cohort' [254]. This set was profiled using Human 6k Transcriptionally Informative Gene Panel for DASL. The GSE21034 set [255] contained 370 prostate samples from which only the 131 cancer cases were

included in the analysis. These samples were also profiled using Affymetrix Human Exon 1.0 ST.

4.2.4. Ovarian Cancer Data.

Table 4.4. Ovarian cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE17260	41000	110	Agilent-014850 - G4112F
GSE19829-GPL8300	12558	42	HG_U95Av2
GSE9891	54621	285	HG-U133_Plus_2
GSE29450	54675	10	HG-U133_Plus_2
GSE30161-1	54675	38	HG-U133_Plus_2
GSE30161_ 2	54675	11	HG-U133_Plus_2
GSE19829-GPL570	54675	28	HG-U133_Plus_2
GSE23554	22283	28	HG-U133A
GSE3149	22283	153	HG-U133A
GSE29175_1	22277	25	HG-U133A_2
GSE29175_ 2	22277	13	HG-U133A_2
GSE19161	658	61	CustOva1a520455F
GSE16708_ 1	48803	17	Illumina HumanWG-6 v3.0
GSE16708_2	48803	7	Illumina HumanWG-6 v3.0

Table 4.4 gives a summary of ovarian cancer data sets used later in the chapter for bimodality investigation. The GSE16708 consist of 7 serous cell lines,17 serous tumour and 9 normal [256]. All samples have been performed on Illumina HumanHT-12 V3.0 expression beadchip and we used the first 2 groups as separate sets. The GSE19161 set obtained expression from 61 patients with advanced ovarian cancer and profiled using CustOva1a520455F platform. The GSE23554 set contained 28 advanced stage serous epithelial ovarian cancers and they were profiled using HG-U133A [257]. The GSE29175 contained 38 ovarian cancer cell lines conducted using HG-U133A_2 [258]. The dataset has been divided into 13 samples as ovarian clear cell carcinoma (OCCC) and 25 samples as non-OCCC. The GSE29450 set includes 10 clear cell ovarian cancer and 10 normal samples using the Affymetrix-human_U133Plus2.0Arrays (HG-U133_Plus_2)[259]. The GSE30161 contains expression data from 58 ovarian cancer patients stage III-IV profiled using the Affymetrix-

human_U133Plus2.0Arrays (HG-U133_Plus_2). In our analysis we only included 38 samples of stage III-C as one set and 11 samples of stage III-B [260]. The GSE3149 contains expression of 182 ovarian tumours which were profiled using HG-U133A/ Affymetrix Human Genome U133A Array [261]. The GSE9891 consisted of 285 samples from the AOCS (Australian Ovarian Cancer Study) and these were profiled on the affymetrix U133_plus2 platform [262]. The GSE19829 consisted of gene expression 28/42. The GSE17260 contains 110 samples from patients with advanced-stage serous ovarian cancer and profiled on Agilent-014850 Whole Human Genome Microarray 4x44K G4112F [263].

4.2.5. Leukaemia Cancer Data.

Table 4.5. Leukaemia cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE10255	22283	161	HG-U133A
GSE10358	54675	279	HG-U133_Plus_2
GSE12995	22283	175	HG-U133A
GSE15434	54675	251	HG-U133_Plus_2
GSE17855	54675	237	HG-U133_Plus_2
GSE34861	36846	191	NimbleGen Human Expression Array
GSE38611	33297	136	HuGene-1_0-st
GSE39381	33297	21	HuGene-1_0-st
GSE39671	54675	130	HG-U133_Plus_2
GSE44247	27068	20	HuGene-1_0-st

Table 4.5 summarises leukaemia cancer data sets used later in this chapter for bimodality investigation. The GSE34861 consisted of 191 samples of adult B-lineage acute lymphoblastic leukaemia (ALL) and 3 normal; these were done using NimbleGen Human Expression Array [264]. The GSE38611 contains gene expression of B-cell chronic lymphocytic leukaemia (B-CLL) obtained from 136 patients and analysed on Affymetrix Human Gene 1.0 ST Array. The GSE39381 contains gene expression of 21 primary Plasma Cell Leukaemia (pPCL) that were performed using Affymetrix Human Gene 1.0 ST Array. The

GSE39671 contained 130 samples from patients with chronic lymphocytic leukaemia (CLL) that were conducted on Affymetrix Human Genome U133 Plus 2.0 [265]. The *GSE44247* contained 20 cancer samples performed on Affymetrix Human Gene 1.0 ST Array. The *GSE10255* contains 161 samples from primary acute lymphoblastic leukaemia patients which was profiled using Affymetrix Human Genome U133A Array [266].

The GSE10358 contained 279 acute myeloid leukaemia (AML) patient samples and was performed using Affymetrix Human Genome U133 Plus 2.0 [267]. The GSE12995 set included 175 paediatric B-progenitor ALL samples that were studied using Affymetrix U133A [268]. The GSE15434 contains expression of 251 samples of AML with normal karyotype (AML-NK) which were profiled on Affymetrix Human Genome U133 Plus 2.0 Array [269]. The GSE17855 contained gene expression microarray data of 237 patients with AML that were profiled using Affymetrix Human Genome U133 Plus 2.0 [270].

4.2.6. Lung Cancer Data

Table 4.6 Lung cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE18842	54675	46	HG-U133_Plus_2
GSE20189	22277	81	HG-U133A_2
GSE29016	48803	72	Illumina HumanWG-6 v3.0 expression beadchip
GSE31267 _1	48803	12	Illumina HumanWG-6 v3.0 expression beadchip
GSE31267 _2	48803	12	Illumina HumanWG-6 v3.0 expression beadchip
GSE32175	54675	10	HG-U133_Plus_2
GSE32863	48803	58	Illumina HumanWG-6 v3.0 expression beadchip
GSE32989	48701	68	Illumina HumanWG-6 v2.0 expression beadchip
GSE33072	33297	131	HuGene-1_0-st
GSE37745	54675	196	HG-U133_Plus_2
GSE42127	48803	176	Illumina HumanWG-6 v3.0 expression beadchip

Table 4.6 gives a summary of lung cancer data sets used later in this chapter for bimodality investigation. The GSE29016 set contained the expression profiling of 72 lung carcinomas using Illumina HumanHT-12 V3.0 expression

beadchip. The GSE31267 set contained 12 samples of stage I lung patients and 12 stage II, and was conducted using Illumina HumanHT-12 V3.0 expression beadchip. In this project stage I samples were used as one set, and stage II samples is another set. The GSE32175 contained 10 lung cancer samples and 10 matched normal samples and were profiled using Affymetrix Human Genome U133 Plus 2.0 Array. The GSE32863 contained 58 lung adenocarcinoma and 58 adjacent non-tumour lung tissues. These were analysed using the Illumina HumanWG-6 v3.0 expression beadchip. The GSE 32989 set contained 58 non-small cell lung cancer (NSCLC) and 2 HBEC-KT cell lines and were profiled using Illumina HumanWG-6 v2.0 expression beadchip.

The GSE33072 contained the expression data of 131 NSCLC and was profiled using Affymetrix Human Gene 1.0 ST Array. The GSE37745 contained 169 samples from NSCLC patients profiled on Affymetrix Human Genome U133 Plus 2.0 Array. The GSE42127 included 176 samples of NSCLC patients who were profiled using Illumina HumanWG-6 v3.0 expression beadchip. (This could also be used as 4 sets 2/2). The GSE18842 contained 43 tumours and 45 controls, and was profiled using Affymetrix Human Genome U133 Plus 2.0 Array. The GSE20189 contained 81 lung cancer samples which were profiled using Affymetrix Human Genome U133A 2.0 Array.

4.2.7. Breast Cancer Data.

Table 4.7. Breast cancer data sets used in this study

Dataset	*Probe ID*	*Sample*	*Microarrays platform*
GSE11121	22283	200	HG-U133A
GSE1456-GPL96	22283	39	HG-U133A
GSE15852	22283	43	HG-U133A
GSE2034_1	22283	179	HG-U133A
GSE2034_2	22283	107	HG-U133A
GSE22820	41000	176	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
GSE24450	41000	183	Illumina HumanHT-12 V3.0 expression beadchip
GSE38959	45015	30	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
GSE6532-GPL570	54675	87	HG-U133_Plus_2
GSE7390	22283	198	HG-U133A

Table 4.7 gives a summary of lung cancer data sets used later in this chapter for bimodality investigation. The GSE2034 contains 180/179 lymph node-negative relapse-free patients and 106/107 lymph-node negate patients that developed to metastasis. These were profiled using Affymetrix Human Genome U133A Array and were used as two sets in the analysis. The data set was downloaded from GEO (Gene Expression Omnibus). The data was derived from analysis by Wang et al., to predict patients' risk based on gene expression profiles [232]. The GSE6532 contained 87 samples from primary breast tumours performed using Affymetrix Human Genome U133 Plus 2.0, while the rest of the datasets were performed on a different platform. The GSE7390 contained 198 samples from primary breast tumours performed on Affymetrix Human Genome U133A Array. The GSE11121 dataset was downloaded from GEO (Gene Expression Omnibus) and contains 200 lymph node-negative breast cancer patients who were not treated by systemic therapy after surgery. The data was derived from a study that aimed to find prognostic motifs [227]. Gene expression profiling of patients was done using the Affymetrix HG-U133A microarray platform compromising 22283 probe sets.

The GSE38959 data contains 30 triple negative breast cancer (TNBC) and 13 normal mammary ductal cells with a laser microbeam microdissection system that was profiled using Agilent-014850 Whole Human Genome Microarray 4x44K G4112F. The GSE1456 set contained tissue material which was collected from all breast cancer patients receiving surgery at Karolinska Hospital (Sweden) between 1994 and 1996. However, my analysis included only 30 samples of luminal A type that were analysed on MMM platform. The GSE15852 contained samples from 43 tumours and 43 normal tissues that were profiled using Affymetrix Human Genome U133A Array. Also, as before, only the cancer cases were used in the investigation. The GSE22820 set contained 176 primary breast cancer patients and 10 normal breast samples. Only the cancer cases were included in the analysis and these were profiled using Agilent-014850 Whole Human Genome Microarray 4x44K G4112F. The GSE24450 contained 183 breast tumours obtained from the Helsinki University hospital and profiled using Illumina HumanHT-12 V3.0 expression beadchip.

4.3. Pre-processing

I focused on bimodal genes detection among cancer patients, acknowledging that my proposed method is sensitive to outliers. For example, if a gene with 100 samples, 2 of which samples are extremely higher or lower than the others, my method will assign a large bimodal index for this gene. There is currently no way to distinguish between bimodality and outliers in their biological context. Therefore, in order to control the outliers, I follow the method proposed earlier [216] by ignoring the subgroup if they are 5% of the total sample. Furthermore, the log base 2 transform has been applied to the dataset that is not normalised.

4.4. P-value estimation using gamma distribution.

Due to the large-scale data set used in this chapter and with the intensive computing time of the sequential Monte Carlo approach [226], I found it more convenient to find another method to compute the *p* value for significance analysis. After careful physical examination of the distribution of the indices, I noticed that all indices calculated by *h*BI method follow a gamma distribution, as illustrated in **Figure 4.1**. As a result, I have replaced the sequential Monte Carlo approach to deliver the significance analysis presented in *Chapter 3* by a fitting gamma distribution using the MASS package implemented in the R – fitdistr [271]. The results from both methods are almost similar in term of significance. This has been investigated in some data and has showed large correlation between the two measurements (*p* values from 2 methods) where the correlation coefficient between them is 0.96 and 0.99 for simulated data and real data respectively.

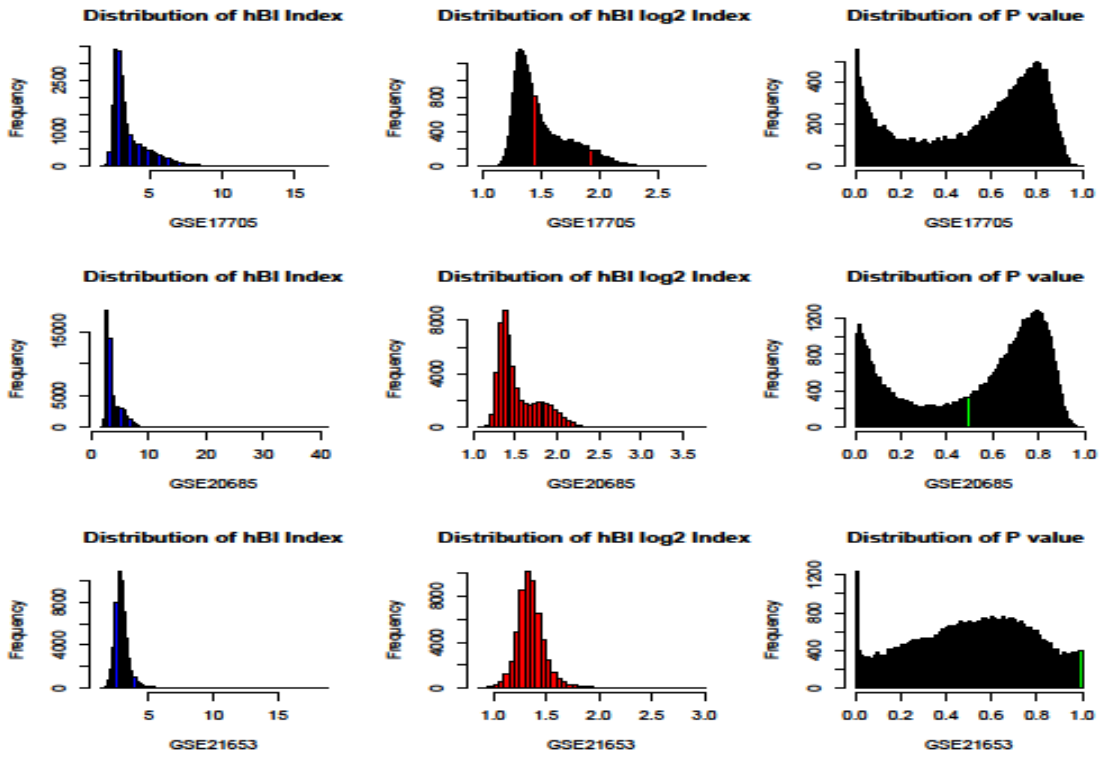


Figure 4.1. The figure shows the gamma distribution of index obtained by *h*BI for three different datasets. The left/first column is index distribution of *h*BI, the middle/second column is the log2 normalised index, while the right/third column shows the *p* value distribution using the gamma fitting function.

In order to demonstrate the power of using Gamma over other distributions, an evaluation analysis being conducted for goodness fit. Here I consider Kolmogorov Smirnov test (KS) for verifying that *h*BI comes from a population with gamma distribution. In this investigation, six distributions were fitted, which are Normal, Weibull, Gamma, Exponential, Poisson and Geometric distributions. **Figure 4.2** shows clearly that Gamma is the best fit among others for such index data.

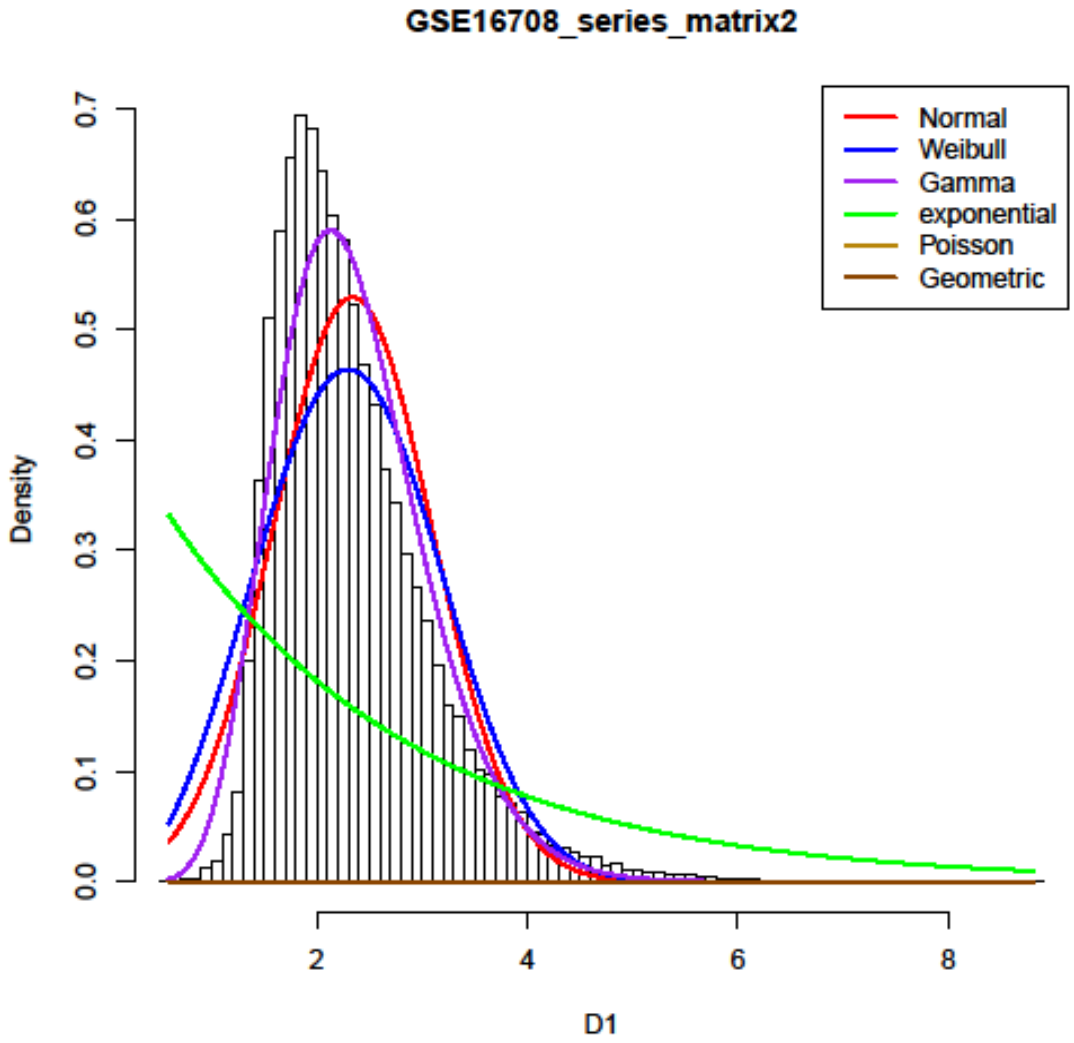


Figure 4.2. Illustrates different distribution fitting for GSE16708 data set.

To confirm the distribution of the data, whether it follows a Gamma distribution or not, I have used test statistic obtained from Kolmogorov Smirnov test, which is the maximum distance between the two cdfs. Applying this to all mentioned above distributions, I have found the minimum distance always linked to Gamma among the 75 datasets. Therefore, it can be concluded based on results of all data in section 4.2, that hBI indices follow a gamma distribution. Figure 4.3 shows boxplot of the D statistics that shows Gamma is the best fit among other distributions as it has the smallest Distance.

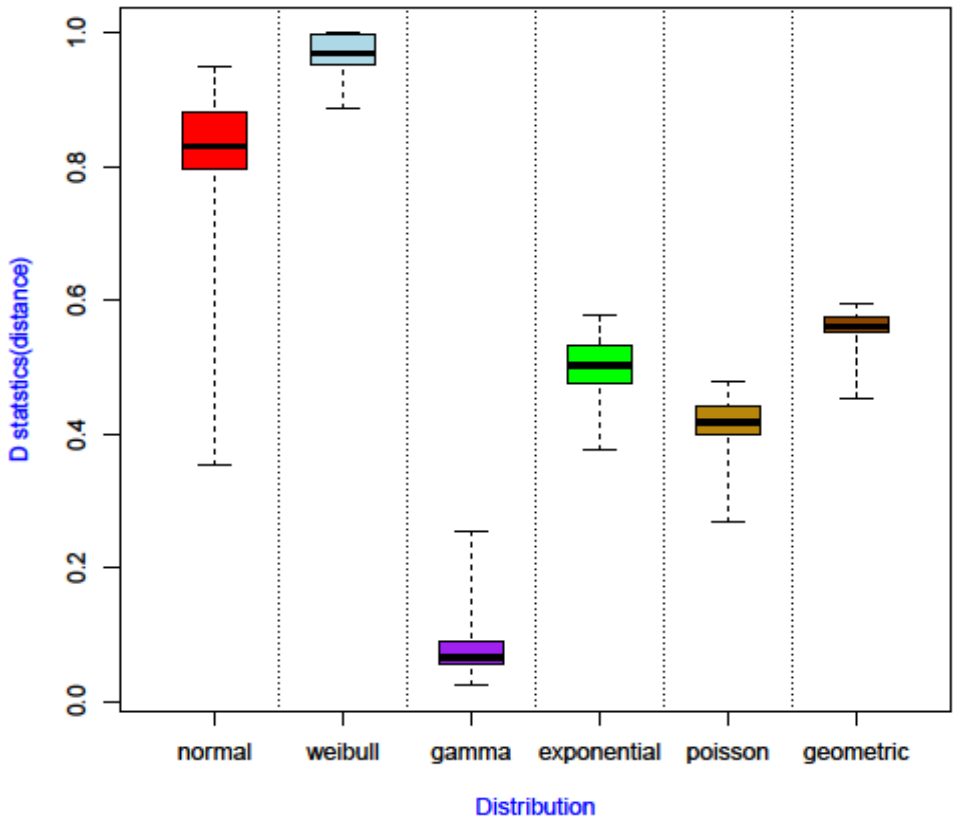


Figure 4.3: shows the maximum distance between the two *cdfs* obtained from (actual distribution and the fitted distribution) on all datasets used in this chapter

4.5 Results and Discussion

In order to verify my hypothesis, I undertook a comprehensive analysis on different cancer types that performed using different platforms from independent laboratories or research groups. Applying the *h*BI algorithm to the data described in Section 4.2 above, I found that the bimodality phenomena were common across cancer. I identified an average of 2% and 6% of bimodal genes at the *p* value thresholds 0.01 and 0.05 respectively, among all cancer types (**Figure 4.4**).

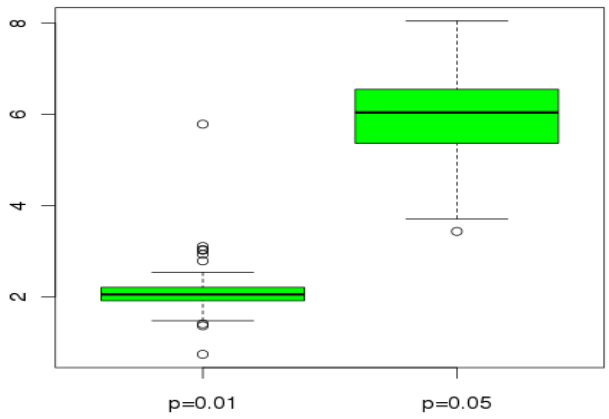


Figure 4.4. Boxplot of bimodal genes percentage among cancer types. The x-axis represented the two *p* value threshold, while the y-axis represented the percentage of bimodal genes

I have also investigated cancer types in relationship with bimodality, and have identified the probability of bimodality in each type. In another words the percentage of bimodality could be higher in some cancer types than in others. In the colon cancer data the average percentage was 2.01% at p = 0.01, and 6.06% at p = 0.05, while these figures were slightly higher in the liver cancer data at 2.5% and 6.14% respectively. In addition, the lower percentages were linked to the prostate data and the leukaemia data with 1.8% at p=0.01, while leukaemia and lung cancer were associated with lower percentages at p= 0.05, with 5.35% and 4.89%. The ovarian cancer data were the second highest at p=0.01 while the breast cancer data ranked third in this respect (**Figure 4.5**).

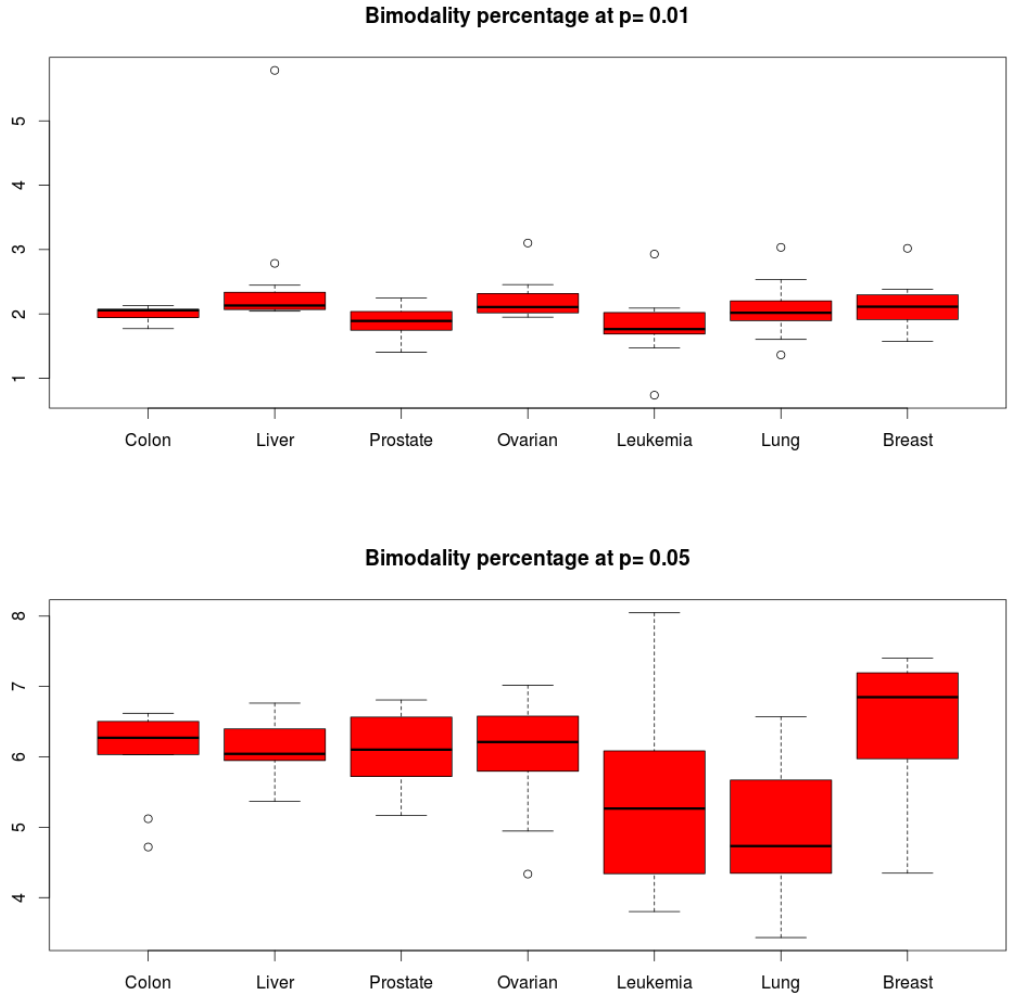


Figure 4.5. Boxplot of percentage of bimodal genes in different cancer types. The upper box represented the bimodality percentage at *p* =0.01, and lower panel at *p*= 0.05. The x-axis represented the different cancer types, and the y-axis represented the percentage of bimodal genes.

Considering the effect of using different microarray platforms, the bimodality percentage was calculated based on the same platforms aiming to find any difference. **Figure 4.6** shows the main platforms in which each one might have more than one version. Clearly there is no dependable platform in terms of bimodality, so that further examination is required in which each version could be analysed separately.

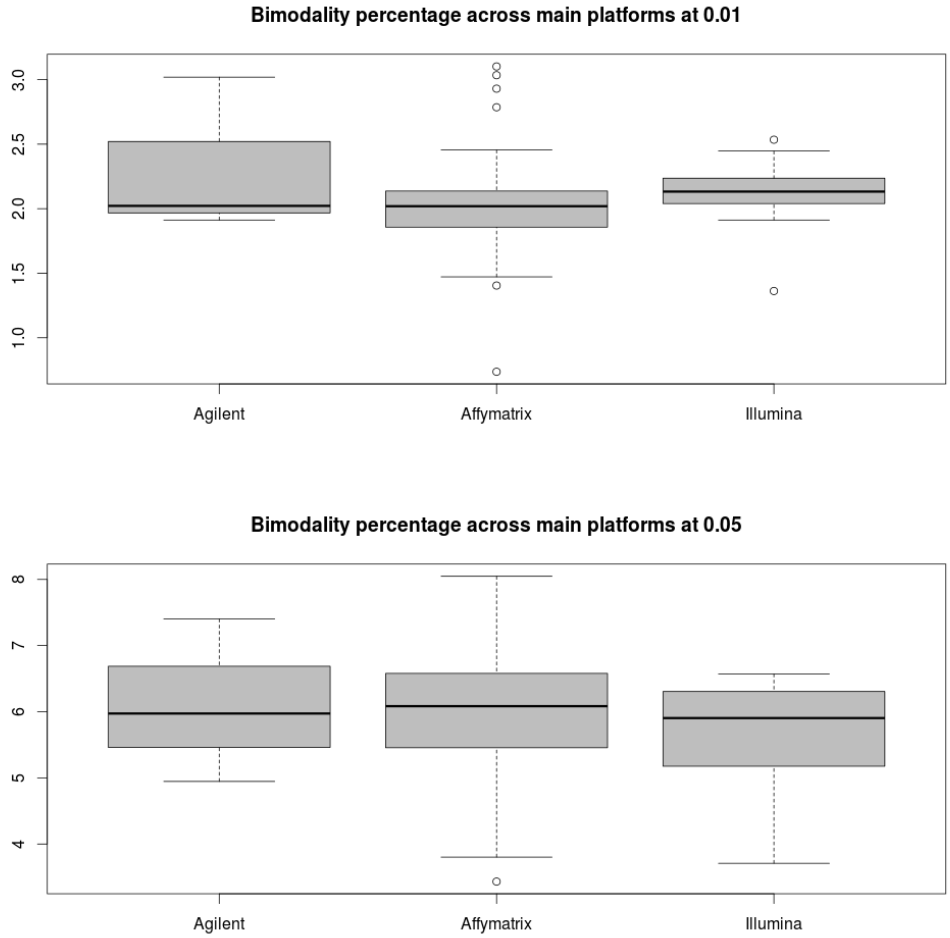


Figure 4.6. Boxplot of bimodal genes percentage in the main platforms. The upper box represents the bimodality percentage at p = 0.01, and the lower box at p = 0.05. The x-axis represents the different platforms, and the y-axis the percentage of bimodal genes.

To do this the same platform versions with more than five sets to be included had to be identified (**Figure 4.7**), which showed that there was no evidence of bimodality-platform association.

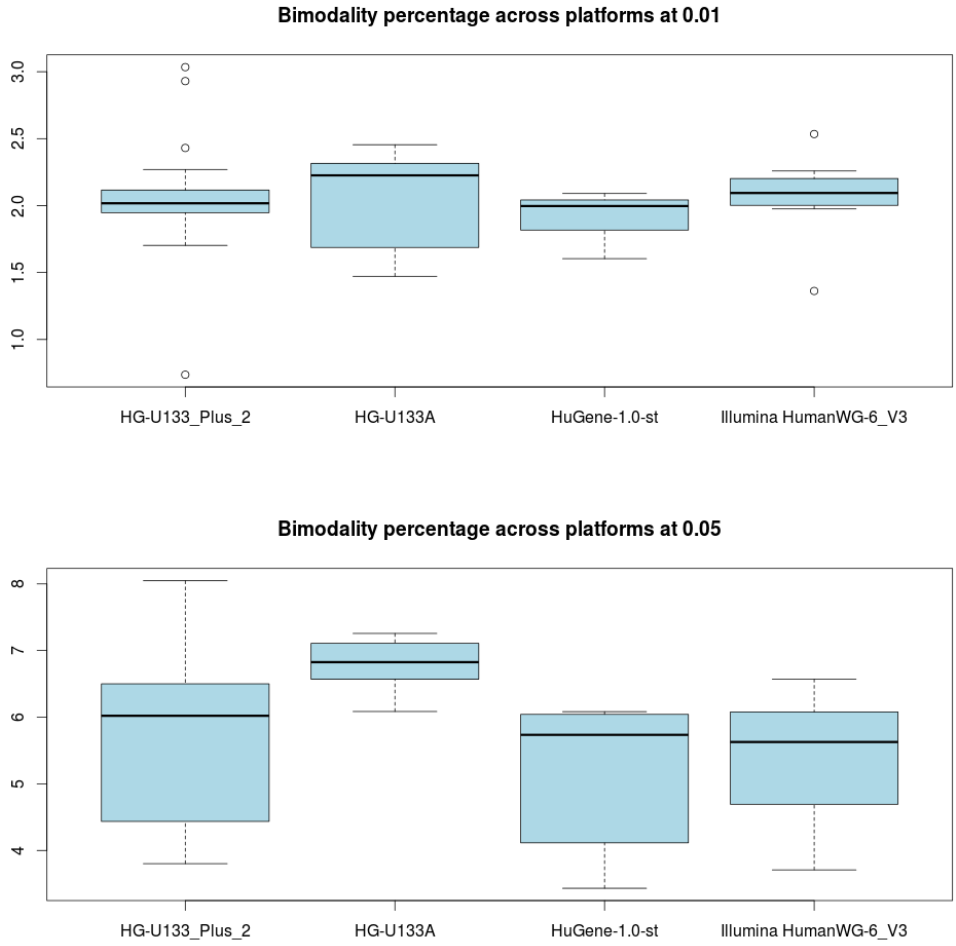


Figure 4.7. Boxplot of bimodal genes percentage in different versions/generations of one platform. The upper box represented the bimodality percentage at p = 0.01, and the lower box at p = 0.05. The x-axis represented the different platforms and the y-axis the percentage of bimodal genes.

The bimodality phenomenon in seven cancer types was comprehensively studied here. The analysis was enlarged to cover different studies using different arrays of platforms, and the results have proved the presence of bimodal genes in all cancer types, platforms and laboratories. As a result, bimodality was found to be a common phenomenon of gene expression that was particularly related to the heterogeneity of cancer, the properties of the disease, specific genetic reactions/characteristics among patients, and/or environmental effects but was not related to the platforms or laboratories. These findings also support the results obtained by Bessarabova and collaborators from his small scale investigation of bimodality in breast cancer [216].

In addition, the results have drawn attention to the genetic contribution that may diversify the patients expressions [272]. This would suggest identifying a subgroup of patients who share a common factor for further analysis.

4.6. Conclusion

The variability in gene expression across patients due to genetic variation or environmental factors is common and this is important in case of cancer. The heterogeneous distribution of gene expression, which indicates that a subset of patients show similar expressions to non-cancer patients. It also has been noticed that even within a specific subgroups of cancer patients, the same genes are not necessarily expressed in the same way among all samples. Thus, its important task to identify those genes with bimodal behaviour for its importance in cancer as seen in some studies. In relation to the detection of bimodal genes it is important to discover a set of genes that are tightly regulated around two conditions at the transcript level. Genes with bimodal behaviour in multiple cancer datasets are strongly evidenced to be ideal biomarker candidates for the treatment of the cancer. In this chapter, I investigated bimodal behaviour/nature among various cancer types, platforms and laboratories. The results showed that bimodality is common in cancer data sets and unrelated to platforms or experiment protocols. This suggests that, on the basis of many factors including, but not limited to, genetic information and geographical or environmental aspects, gene expression from different patients may also be different. Ignoring such inter-individual differences will increase false discoveries and negatively affect the accuracy of the bimodality detection algorithm. In my view, grouping patients into well-defined group with respect to

these factors will help in identifying more reliable biomarkers. Thus, understanding the molecular mechanism of cancer at the systems level may open up new horizons for understanding the bimodality links to cancer.

Chapter 5

The Hierarchical Integrative Model

Abstract

A comprehensive survey of integration methods has motivated my proposed Hierarchical Integrative Model (HIM) for revealing unbiased truth from multiple experimental data sets. The algorithm is based on the assumption that, if an experiment is well designed and carried out, replicate data are random samples of a underlying linear density that is *a priori* unknown. I refer to it as a 'mean pattern model' and it is the basis of HIM. I show here how HIM is constructed, based on this mean pattern model, and evaluate HIM using simulated data. HIM's application to real data is discussed in Chapter 6.

5.1. Introduction

DNA microarrays technology has been extensively utilised in the biomedical field, and it has become standard practice to identify targeted gene expression signatures. Those genes are useful for predicting the stages of the disease and providing an insight to its diagnosis and treatment. Cancer research has benefited from this technology, which has been used in many studies to identify the most important genes that contribute to cancer [73-80]. However, various issues have been linked to this technology, such as noise resulting from technical considerations and/or sample variations, while the consequence of a limited budget is reduced samples in most experiments, which render accurate and precise inference from data invisible. Although independent analysis of a single data is a common practice, inconsistent findings make integrative method for analysis urgently required [276].

It has been noted that biomarkers identified by different studies have a small proportion of overlap because of molecular heterogeneity [81-83]. This can be associated with the use of different procedures/protocols, normalisation, platforms, or experimental samples; therefore, relying on one dataset will not always provide reliable results [91, 273-275]. This has made the identification of reliable biomarkers – and hence the diagnosis/prognosis of cancer – quite complicated. For example, in a cancer disease it is important to explore universal cancer signatures [84-86]. Although cancer is a highly heterogeneous disease, it is believed that there are common gene expression patterns across different cancers [87]. This confers the ability to understand gene regulation

which, owing to the weak signal from that gene in the specific data, may not be understood using a single data set [89].

The better way to carry out such an investigation is to integrate multiple data sets for analysis rather than relying on a single set, since this will help to identify the most important signatures among data sets. Moreover, compared with a single data analysis, the integration approach increases the accuracy and the consistency of the results because it eliminates the possible error in independent study [276]. In addition, it will allow researchers to benefit from the accumulating amount of data placed in repositories such as GEO as illustrated in **Figure 1.1** [13, 14], NASC [277], Array Express [15].

Carrying out a reliable analysis can be satisfactorily incorporated using an integrative study, and will give insight to the hidden relations/patterns in multiple heterogeneous or homogeneous biological data sets. Heterogeneous integration uses different types of data, e.g., gene expression, DNA methylation, and protein expression, while the homogeneous focuses on one data type. Generally, integrating datasets from multiple experiments to target one study will deliver predictions that are less biased because individual laboratory/sample-specific biases can have less impact for truth finding. Indeed, a study investigated the importance of the sample size and found that large sample size improve data analysis accuracy [88].

The integrative analysis, in which results or data from different sources are combined, has the merit of strengthening the overall image through projecting molecules onto different aspects to examine. However, integrating different datasets still remains a challenge, due to the difficulty and the complexity of the

task, as well as the lack of proper statistical tools. The next section is a summarised survey of the published methods.

5.2. Gene expression Data Integration Methods

The integrative study can be divided into two categories based on gene expression data: partial pattern discovery (PPD) and full pattern discovery (FPD). As illustrated in Figure 5.1, the first category is the widely used where common genes among datasets are targeted for analysis. The second category examines all genes from all data sets for analysis.

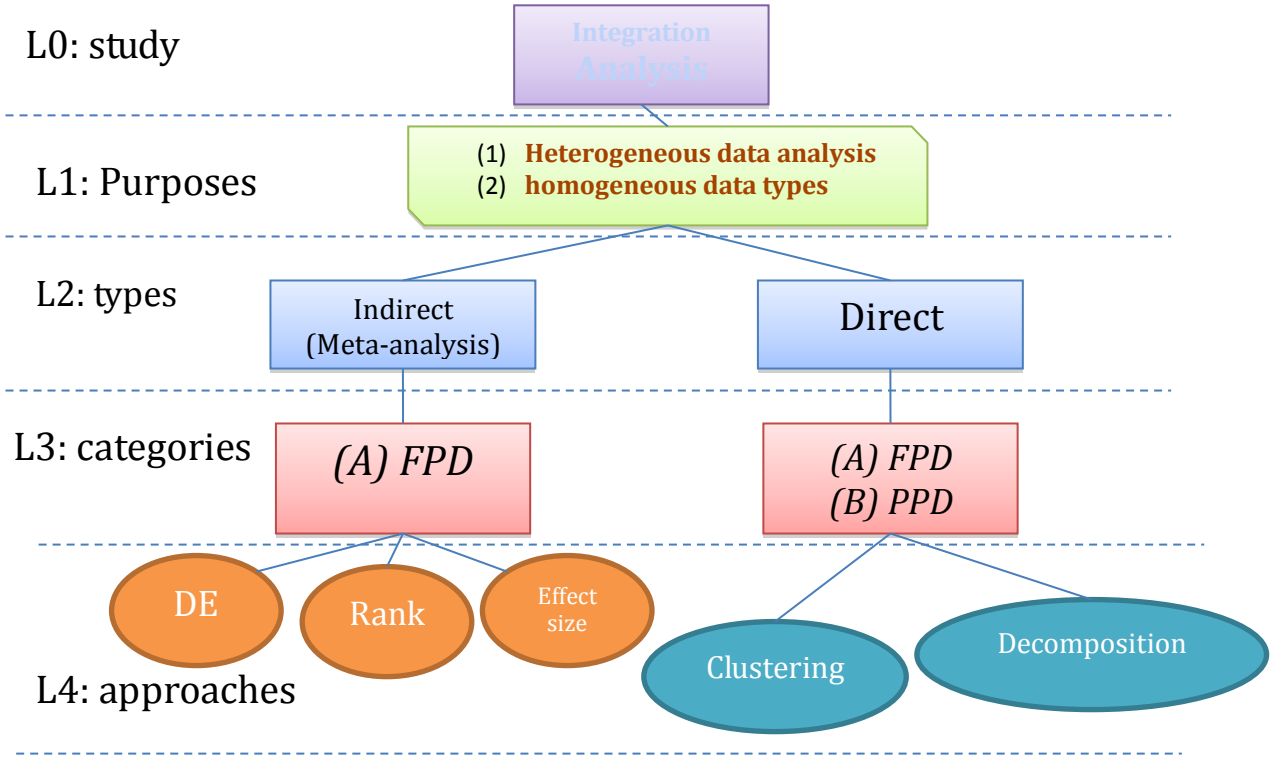


Figure 5.1. Flowchart illustrating integration studies classifications used in the study; FPD: full pattern discovery, PPD: partial pattern analysis, and DE: differentially expressed genes

A major limitation of PPD is that the data must be organised into a single matrix which means that only common genes across data/species are considered for analysis. The uncommon genes are missed from analysis and important information in data will therefore be lost. In contrast, FPD has the benefit of using all data for modelling/analysis. A good example of this is the direct

merging of gene expression, using one of the methods that will be explained in the sections that follow.

There are also two general approaches for integrative study of gene expression data sets. The direct approach re-analyses primary/raw experimental data type, such as laboratories, platforms, species or cancers. The indirect approach combines the published results. Although the latter approach is widely used, it depends on long lists of genes from microarray studies that use different gene annotations, identifiers and/or /nomenclatures; furthermore, most differentially expressed genes lists are presented in tables or graphics formats that render them unreachable by search engines [273, 278]. Additionally, due to differences in protocols, small sample sizes of individual studies will affect the analysis negatively [279-281]

In the sub-section below, I discuss briefly the methods that are applied to integrate results for further analysis. Meta-analysis is most commonly used in this type/category and has been very successful in many applications. Here, the customary approach is to compare the differentially expressed genes in two or more species/platforms/types in order to find the agreements or disagreements. Some researchers have looked into whether some genes that are co-expressed in one species are found co-expressed in another species [66]. This co-expression analysis has been widely used for cross-species analysis. Stuart *et al* developed a computational method for studying different species that included yeast, flies, worms, and human microarrays, in order to find the conserved biological functions [61]. Another algorithm is presented to map an annotated gene module in a species to its homologous genes in another species: this has been used to investigate the regulatory programs in six different organisms [282].

5.2.1. Indirect integration

As mentioned in a Tseng's review [280], indirect integration or meta-analysis can generally be divided into three categories related to the differentially expressed gene information. I agree with the author's classification about the first three categories being indirect, whereas the other two belong more to the next section (5.2.2).

First, combining *p* values from multiple studies has been widely used due to its simplicity and its ability to work with many forms and outcomes. This is because this type of analysis relies on the significance/summary statistics (univariate summary statistics) of each datasets; not on the complex underlying data structures. The most common method is the Fisher's combined probability test (Fisher, 1932). The Fisher method combines the log-transformed *P* values from *K* studies (i.e. microarrays) to Chi-square scores with 2*K* degrees of freedom. One of the initial efforts of employing the Fisher's test in the context of microarrays was made by Rhodes *et al* (2002), who proposed a statistical model that combined *p* values to integrate multiple microarrays from two different platforms. Using the simple Fisher's method, they computed the *p* values for each gene from the individual studies to estimate an overall *p* value [90]. The same group also developed a comprehensive meta analysis framework in which they applied the Student's test [91]. One issue arising from the use of Fisher's method is that it is biased towards the small *p* values. For example, if one gene has very small *p* value in one study but is of no significance to the others, Fisher will classify it as significant (large Fisher's score). Different variations of Fisher's test have been proposed to enhance the power of the traditional Fisher's method. One extended the Fisher's statistics to involve unequal positive weights for each k determined by an expert or prior

information[283]. This method called weighted Fisher's statistics that has shown an improvement on the power of Fisher's method. However, it has some limitations where it led to miscalculations in the case of equal weights or if any of the weights is zero. Another is trimmed Fisher's method that proposed to overcome the effect of extreme values or outliers [284]. The choice of weight is largely depends on expert opinion which somewhat subjective. Therefore Li and Tseng proposed the adaptively weighted Fisher's statistics [92]. They used the data to estimate the weights which increase the statistical power to improve biological interpretation.

Another method of combining P values is the minimum P value statistic [285]. This mainly takes for a gene the minimum p value among the different studies (microarrays) and thus its clearly sensitive to outlier [286]. An enhancement to this was proposed that uses rth smallest p value [287]. Both methods were used for combining microarrays data [286, 288]. More recently, a new method proposed called *r*th ordered *p* value[288]. It aims to test the alternative hypothesis that a gene is expressed in percentage of studies.

Although there are many powerful statistical tools for the same purpose from summed log-transformation of *p* values (Fisher's method) to the adaptively-weighted Fisher's method [90-92, 288], researchers prefer to use Venn diagrams to compare the *p* values of different studies. For example, Venn Mapper offers a comparison of the overlap of differentially expressed genes among different cancer types and calculates the statistical significance of the overlaps [289]. The advantage of this method is that it is easy to use and does not require additional analysis. However, by working with this type of tool, it is not possible to estimate the average magnitude of differential expression. In

contrast, Fisher's method does introduce statistics for integration significance, while a Venn diagram simply gives a summary.

Secondly, a considerable number of studies have focused on more effective ways of meta-analysis by combining effect sizes in their multiple data set analysis [93-95]. The effect size is the standardised mean between two phenotypes (i.e., cancer/normal) in each set, which are then combined as an overall mean. This method has improved the modelling capability for the inter-study variation. There are two models commonly used for this approach: namely the fixed effect model and the random effect model (FEM and REM). Choi *et al* (2003) employed these models for differential expression studies of cancer where multiple datasets were compared. They used the homogenous test *Q* statistics to determine which model was appropriate/better/suitable, since the large *q* values indicate the need for REM while the small *Q* values indicate the need for FEM [93].

The Bayesian approach has also been used to estimate effect size [94-96]. As both of the previous two methods are sensitive to outliers, this could be a serious problem as I am aware of the large noise in microarrays. Hu and his group introduced an extended efficient method to identify the differentially expressed genes of lung cancer in two platforms of microarrays (Hu 2005), by modelling the effect size and measuring the gene chip quality of the probe sets of microarrays.

Thirdly, the average rank combining method has encouraged some researchers to overcome the outlier sensitivity by the *p* value and effect size methods [97-99]. In the rank-based approach that was used by Xu and co-workers for cancer gene expression from multiple experiments [74], the rankings

aggregation represented the results of statistical hypotheses for differential expression across studies for interesting genes, and then ordering the genes for further analysis [99]. The advantage of this kind of analyses is that they provide a good estimation for noisy data and/or small sample size data.

All these previous methods are categorised as indirect integrative study, while the next section focuses on direct analysis.

5.2.2. *Direct integration*

Moving away from independent analysis (indirect integration) to raw (direct) expression integration, it can be seen that this type has attracted a lot of researchers for its usefulness in many applications[74]. However, data from different expression studies poses a number of difficulties which arise from the fact that each data exercise has been conducted using a distinct gene-expression platform and has been carried out in a different laboratory using a different protocol. As a result, and due to these variations, direct combining cannot be useful because the two data sets are not comparable. This kind of analysis may suffer from study-specific concerns, as well from the heterogeneity seen across studies. It really depends on the quality of the data set.

5.2.2.1. Integration by Normalisation

Researchers have used normalisation procedures to avoid variations such as robust multi-array analysis (RMA) [290], median rank scores and quintile [291], cross-platform normalisation (XPN) [292], and ratio-adjusted gene-wise normalization (rGN) [293]. The main aim was to eliminate the data-specific effect and make the data comparable. These methods select genes that are common to different microarray platforms/studies for an integration analysis.

A brief discussion of these types of method is given here. One group combined rescaled expressions of the selected genes among datasets into one matrix for analysis [294]. Their linear scaling normalisation was based on multiplying each column in each data set by the inverse slope of a least-squares linear fit of the sample versus a first sample. They included five different cancer gene expression data sets to find the biomarkers that differentiated between the primary and metastasised cancers. Bayesian mixture model was also used to integrate re-scaled data to one matrix for analysis [295, 296]. This normalisation aimed to estimate the probability of expression (*poe*) for each set, and then to combine the results into one matrix. However, model-based transformation outweighs simple linear normalisation by its ability to involve biological information into estimating the posterior probabilities of expression as well as to reduce the effect of outliers [295, 296].

Jiang and his group used a distribution transformation process to transform different data sets into a comparable distribution before integration [297]. Others used the median rank scores and quintile discretisation to transfer all expression to a comparable level/expression/scale, after which they applied support vector machine (SVM) to predict the sample type as cancerous or normal; they found the prediction accuracy was enhanced with integration [291]. They also proved that using such an integrative method helped to increase the accuracy of classification. Shabalin and colleagues proposed a modal-based method for normalisation, which they called cross-platform normalisation (XPN). This is based on linking gene/sample clustering of the given datasets [292]. Additionally, normalised linear transformation methods have been used that utilise the linear mapping of the two different platforms in order to obtain an identical scale for each gene between the two sets [298].

The disadvantage of the methods explained above is that including the common genes across multiple arrays may led to failure to notice important genes which through oversight are missed by one array. Another issue related to some types in this category is the lack of ability to remove the systematic error produced when sample processed in multiple batches which is usually referred to batch effect [299, 300].

5.2.2.2. Decomposition, correlation and clustering

Other methods were proposed to tackle the issues of using normalisation for integration mentioned in the previous section. One of these focused on dimensionality reduction such as Principal Component Analysis (PCA) and Non-negative Matrix Factorisation (NMF). A second approach benefited from correlation analysis [100], while a third approach focused on modelling or gene clustering across datasets [101]. The following paragraphs briefly discuss the methods proposed for this category.

One of the early examples of this type of integrative study was carried out by Culhane *et al* (2003) who proposed an attractive method called Co-inertia analysis (CIA). This uses the coupling technique to identify the correlation of gene expression patterns among different cancer cell lines from two different platforms [301]. CIA is a multivariate analysis scheme that discovers relationships across different sets through maximising the covariance between them [302]. One advantage of using this method is that it does not require a pre-filtering of data, especially genes, and as shown by Fagan, can also be used to integrate different data sources [303]. Another team benefited from using the reduction technique to analyse the cell cycle expression data from humans and yeasts [304]. Singular value decomposition (SVD) was used to

identify robustness in response patterns. Engelmann *et al.* (2008) employed a kernel Principal Component Analysis (kPCA) to select an appropriate kernel for each dataset and used these for further analysis and in this particular case for clustering [89]. kPCA is the non-linear form of PCA where it projects the data into a high-dimensional feature space in order to detect the non-linear relations.

Another successful tool is NMF [57, 58], both for dimension reduction and visualization [102]. It decomposes an expression matrix into two matrices. One matrix contains meta-gene expression while the other is a coefficient matrix that captures the relationship between genes and meta-genes. NMF showed merit in the elucidation of cancer subtypes when it was applied to leukaemia and brain cancer microarray data [62]., The same group also used consensus clustering to enhance model selection and to make use of the stochastic nature of NMF [62, 305]. The sparse NMF was used to improve molecular cancer class discovery by enforcing an additional sparseness constraint on the coefficient matrix [306].

NMF has also been used as a tool to cluster genes and to predict functional relations in the microarray data of yeast [307]. Tamayo and his group proposed a method called meta-gene projection to perform similar analyses [60]. They have applied this approach on leukaemia and lung cancer datasets and have proved that it reduces noise and can work with many different platforms by meta-gene projection that benefit from reduction and from keeping the same biological features. The main disadvantage of NMF is that it fails to deal with complex pattern across multiple datasets since it is a linear algorithm, and also because all entries must be positive.

iCluster was then proposed by Shen for clustering samples of different genomic data using the joint latent variable/factor analysis of the decomposed combined matrix [103]. The main idea is that the latent variable model tries to find a linear relationship between the observed data and the latent variables. Each latent variable is treated as a cluster label and in integration context this can explain the co-clusters among sets. However, this method is designed to cluster samples, which means that the sample size across sets must be the same. Despite the successes of these methods they are still sensitive to large noise whilst giving global structure.

The Bayesian approach has been extensively used for data integration. It employs *prior* distributions for integrative study [105]. One of its first application was by Troyanskaya *et al* [106], who introduced a general approach to integrating multi-source data to estimate gene function. The method was based mainly on a Bayesian network that studied the relation of different data in terms of biological function [106], although this supervised approach benefited from human knowledge of integrated predictions. Other studies used a Bayesian model-based clustering approach to detect differentially-expressed genes in an integrative context. They used a hierarchical model for integration and mixture prior to distinguishing between differentially-expressed genes and the alternatives [101]. They proposed a normal mixture prior-based on the mean difference of the two sets where the Bayesian estimation of the mixture distribution was done using the Markov Chain Monte Carlo (MCMC). An additional study also introduced a Bayesian hierarchical model to pool different microarrays from the same platform to identify differentially expressed genes (DEGs) [107]. Similarly, Scharp *et al*. (2009) applied a hierarchical Bayesian model to different microarrays with the aim of finding such genes [108].

More recently, Wang *et al*. (2013) have proposed an integrative Bayesian analysis of genomics data (iBAG) to identify important biomarkers that are linked to clinical outcomes [109]. They integrated different types of cancer data (gene expression, methylation and patient survival rate); however, this is beyond the scope of this thesis as it considers the clinical data in the integration. Savage *et al*. (2010) implemented a mixture model approach and used a hierarchical Dirichlet process to integrate two datasets, specifically gene expression and transcription factor binding [308]. The same group has also developed a new integrative method named multiple dataset integration (MDI). This method has the ability to integrate many different types simultaneously, including time series data, and also can integrate more than 2 datasets. Each set is modelled separately, using a Dirichlet-multinomial allocation mixture model, but allows the pairwise dependencies between clusters among models [309]. Lock and Dunson (2013) have proposed an integrative clustering method that is generally similar to the MDI: however, the pairwise dependency has been replaced by an overall consensus clustering [310].

Another group has developed a statistical tool for cross-species clustering that will identify the co-expression patterns [311]. The advantage of this model is its ability to cluster each data independently while allowing each to borrow strength from others.

5.3. The Proposed Method

5.3.1. Motivation

The importance of multiple data integration has been explained earlier in this chapter. Due to lacking a proper method for integrative study of data sets of different dimentions, I developed a simple method to overcome the limitation..

5.3.2. Mean Pattern Model

A common expression matrix contains rows that represent genes/proteins/DNA copy numbers and columns in which represents the samples/replicates. Here, a mean pattern model was based on the assumption that with modern high-resolution microarray technology, if an experiment is well designed, the variance among replicates of each feature (i.e. genes) is small. Let's denote a matrix by $X = \{ \mathbf{x}_n \}_{n=1}^{N} \in \Re^d$ where N is number of features (genes) and d is number of samples. The expression values of each feature (i.e. gene) vector follow one density, such as a Gaussian distribution $\mathbf{x}_n \sim G(\mu_m, \sigma_m^2)$, where $\mathbf{x}_n$ stands for the expression (or later on for differential expressions) of the n^{th} feature (gene) and $G(\mu_m, \sigma_m^2)$ stands for the m^{th} Gaussian with a sample mean μ_m and sample variance σ_m^2. A whole matrix (dataset) is then a random sample of a mixture of several Gaussian densities. In a complex situation each feature will have distinguish mean than others where in such case, the number of mixtures is equal to the number of features (m=n). Often, however this is not a valid case in real biological data in which some correlation between features are preserved[312]. A univariate feature space was obtained from the whole

multivariate feature space. The density function in this univariate feature space characterized the hidden structure from which the multivariate feature space is a random sample.

5.3.3. General method

The general idea behind the proposed method is illustrated in **Figure 5.2**. All observations from both datasets were converted to vector and then used it to construct $\bigcup \mu_k$ **.** This was done by applying clustering on this vector to estimates all possible Gaussian distributions or more specifically means. For easy implementation, I used the R packages Mclust or Kmeans for this purpose. This step identified the mean underlying differential expressions (MUDE), top layer in the figure. After this step, I also generated random samples from each centre identified to be the Mean Observed Differential Expression (MODE) for each set. Each observed differential expression (ODE) is a random sample of its own MODE. Finally, I maximised the likelihood function to estimate the parameters for each layer.

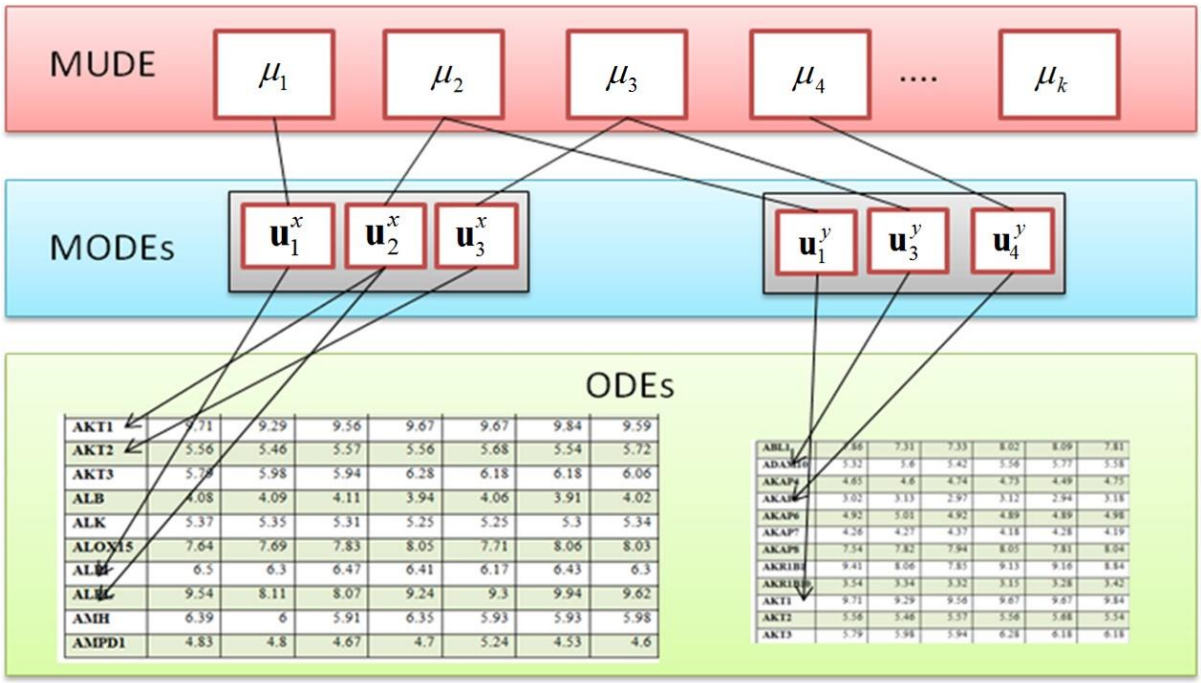


Figure 5.2 Flowchart illustrating the proposed method. MUDU is the mean underlying differential expression, MODE is the mean observed deferential expression for each data set and ODE is the observed differential expression.

5.3.4. Algorithm

One observed differential expression (ODE) data set derived from a microarray experiment is denoted by $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_X}) \subset \Re^{N_x \times d_x}$ and the other is denoted by $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N_y}) \subset \Re^{N_y \times d_y}$, where N is the number of feature/rows (i.e. probe id) and *d* is the length of each feature (i.e samples). I assume that $\mathbf{U}_x = (\mathbf{u}_1^x, \mathbf{u}_2^x, \cdots, \mathbf{u}_{K_x}^x) \subset \Re^{K_x \times d_x}$ are the centres of **X** and that $\mathbf{U}_y = (\mathbf{u}_1^y, \mathbf{u}_2^y, \cdots, \mathbf{u}_{K_y}^y) \subset \Re^{K_y \times d_y}$ are the centres of **Y**. I refer to $\mathbf{U}_x$ and $\mathbf{U}_y$ as the mean observed differential expressions (MODEs). I assume that both $\mathbf{u}_k^x$ and $\mathbf{u}_k^y$ are random samples drawn from a mean underlying differential expressions (MUDE) model, which is a linear and *a priori* an unknown space, i.e.

$$\mathbf{u}_k^z \sim \sum_{m=1}^{K_0} w_m N(\mu_m, \sigma_m^2)\mathbf{i} \tag{5.1}$$

where $z = x/y$: i.e., *z* was the **X** space index or the **Y** space index. In other words, each MODE is a set of random samples of MUDE. The likelihood of MODEs can be defined as

$$L_{\mathbf{U}_z|\boldsymbol{\mu}} = \prod_{k=1}^{K_z} \sum_{m=1}^{K_0} w_m \, p(\mathbf{u}_k^z \mid \mu_m) \tag{5.2}$$

Each ODE is a set of random samples of its own MODE. The likelihood function for one ODE is defined as

$$L_{Z|\mathbf{U}_z} = \prod_{n=1}^{N_z} \sum_{k=1}^{K_z} w_k^z \, p(\mathbf{z}_n \mid \mathbf{u}_k^z) \tag{5.3}$$

The whole likelihood is the multiplication of all afore-mentioned likelihood functions. The log-likelihood function with Langrage coefficients is thus defined as

$$\begin{aligned} O = & \sum_{n=1}^{N_x} \log \sum_{k=1}^{K_x} w_k^x \, p(\mathbf{x}_n \mid \mathbf{u}_k^x) + \sum_{n=1}^{N_y} \log \sum_{k=1}^{K_y} w_k^y \, p(\mathbf{y}_n \mid \mathbf{u}_k^y) \\ & + \sum_{k=1}^{K_x} \log \sum_{m=1}^{K_0} w_m \, p(\mathbf{u}_k^x \mid \mathbf{i}\mu_m) + \sum_{k=1}^{K_y} \log \sum_{m=1}^{K_0} w_m \, p(\mathbf{u}_k^y \mid \mathbf{i}\mu_m) \\ & - \lambda_x \left(\sum_{k=1}^{K_x} w_k^x - 1 \right) - \lambda_y \left(\sum_{k=1}^{K_y} w_k^y - 1 \right) - \lambda_0 \left(\sum_{m=1}^{K_0} w_m - 1 \right) \end{aligned} \tag{5.4}$$

where the probability function is defined as a normal function with $\beta_z = \sigma_z^{-2}$, $\beta_m = \sigma_m^{-2}$ and $z = x/y$

$$p(\mathbf{z}_n \mid \mathbf{u}_k^z) = \left(\frac{\beta_z}{2\pi} \right)^{d_z/2} \exp\left(-\frac{\beta_z (\mathbf{z}_n - \mathbf{u}_k^z)^2}{2} \right) \tag{5.6}$$

and

$$p(\mathbf{u}_k^z \mid \mu_m) = \left(\frac{\beta_m}{2\pi}\right)^{d_z/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^z - \mathbf{i}\mu_m)^2}{2}\right) \tag{5.7}$$

Parameters thus include $\mathbf{u}_k^z$, β_z, μ_m, β_m, w_k^z, and w_m.

The derivative of the objective function with respect to $\mathbf{u}_k^z$ is

$$\begin{aligned}
&\nabla O(\mathbf{u}_k^z) = \sum_{n=1}^{N_z} \frac{w_k^z}{p(\mathbf{z}_n)} \left(\frac{\beta_z}{2\pi}\right)^{d_z/2} \exp\left(-\frac{\beta_z(\mathbf{z}_n - \mathbf{u}_k^z)^2}{2}\right)\beta_z(\mathbf{z}_n - \mathbf{u}_k^z) - \\
&\frac{1}{p(\mathbf{u}_k^z)} \sum_{m=1}^{K_0} w_m \left(\frac{\beta_m}{2\pi}\right)^{d_z/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^z - \mathbf{i}\mu_m)^2}{2}\right)\beta_m(\mathbf{u}_k^z - \mathbf{i}\mu_m) \\
&= \beta_z \sum_{n=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)(\mathbf{z}_n - \mathbf{u}_k^z) - \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m(\mathbf{u}_k^z - \mathbf{i}\mu_m) \\
&= \beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)\mathbf{z}_n - \beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)\mathbf{u}_k^z - \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m\mathbf{u}_k^z + \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m\mathbf{i}\mu_m \\
&= \beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)\mathbf{z}_n + \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m\mathbf{i}\mu_m - \left(\beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n) + \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m\right)\mathbf{u}_k^z
\end{aligned}$$

The update rule for $\mathbf{u}_k^z$ is:

$$\mathbf{u}_k^z = \frac{\beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)\mathbf{z}_n + \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m\mathbf{i}\mu_m}{\beta_z \sum_{i=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n) + \sum_{m=1}^{K_0} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^z)\beta_m} \tag{5.8}$$

The derivative of the objective function with respect to β_z is

$$\begin{aligned}
&\nabla O(\beta_z) = \sum_{n=1}^{N_z} \frac{1}{p(\mathbf{z}_n)} \sum_{k=1}^{K_z} w_k^z \left(\frac{1}{2\pi}\right)^{d_z/2} \exp\left(-\frac{\beta_z(\mathbf{z}_n - \mathbf{u}_k^z)^2}{2}\right)\left\{\frac{d_z}{2}\beta_z^{\,d/2-1} - \beta_z^{\,d_z/2}\frac{(\mathbf{z}_n - \mathbf{u}_k^z)^2}{2}\right\} \\
&= \frac{1}{2}\sum_{n=1}^{N_z}\sum_{k=1}^{K_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)\left\{d_z\beta_z^{-1} - (\mathbf{z}_n - \mathbf{u}_k^z)^2\right\} \\
&= \frac{1}{2}d_z\beta_z^{-1}\sum_{n=1}^{N_z}\sum_{k=1}^{K_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n) - \frac{1}{2}\sum_{i=1}^{N_z}\sum_{k=1}^{K_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)(\mathbf{z}_n - \mathbf{u}_k^z)^2
\end{aligned}$$

The update rule for β_z^{-1} is:

$$\beta_z^{-1} = \frac{\sum_{n=1}^{N_z}\sum_{k=1}^{K_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)(\mathbf{z}_n - \mathbf{u}_k^z)^2}{d_z N_z} \tag{5.9}$$

The derivative of the objective function with respect to μ_m is

$$\nabla O(\mu_m) = \sum_{k=1}^{K_x} \frac{w_m}{p(\mathbf{u}_k^x)} \left(\frac{\beta_m}{2\pi}\right)^{d_x/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^x - \mathbf{i}\mu_m)^2}{2}\right)\beta_m \mathbf{i}^t(\mathbf{u}_k^x - \mathbf{i}\mu_m) +$$
$$\sum_{k=1}^{K_y} \frac{w_m}{p(\mathbf{u}_k^y)} \left(\frac{\beta_m}{2\pi}\right)^{d_y/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^y - \mathbf{i}\mu_m)^2}{2}\right)\beta_m \mathbf{i}^t(\mathbf{u}_k^y - \mathbf{i}\mu_m)$$
$$= \sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x)\beta_m \mathbf{i}^t(\mathbf{u}_k^x - \mathbf{i}\mu_m) + \sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)\beta_m \mathbf{i}^t(\mathbf{u}_k^y - \mathbf{i}\mu_m)$$
$$= \sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x)\beta_m \mathbf{i}^t\mathbf{u}_k^x - \mathbf{i}^t\mathbf{i}\mu_m\beta_m\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x) + \sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)\beta_m \mathbf{i}^t\mathbf{u}_k^y - \mathbf{i}^t\mathbf{i}\mu_m\beta_m\sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)$$

The update rule for μ_m is:

$$\mu_m = \frac{\mathbf{i}^t\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x)\mathbf{u}_k^x + \mathbf{i}^t\sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)\mathbf{u}_k^y}{d_x\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x) + d_y\sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)} \tag{5.10}$$

The derivative of the objective function with respect to β_m is

$$\nabla O(\beta_m) = \sum_{k=1}^{K_x} \frac{w_m}{p(\mathbf{u}_k^x)} \left(\frac{1}{2\pi}\right)^{d_x/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^x - \mathbf{i}\mu_m)^2}{2}\right)\left\{\frac{d_x}{2}\beta_m^{d/2-1} - \frac{1}{2}\beta_m^{d_x/2}(\mathbf{u}_k^x - \mathbf{i}\mu_m)^2\right\}$$
$$+ \sum_{k=1}^{K_y} \frac{w_m}{p(\mathbf{u}_k^y)} \left(\frac{1}{2\pi}\right)^{d_y/2} \exp\left(-\frac{\beta_m(\mathbf{u}_k^y - \mathbf{i}\mu_m)^2}{2}\right)\left\{\frac{d_y}{2}\beta_m^{d/2-1} - \frac{1}{2}\beta_m^{d_y/2}(\mathbf{u}_k^y - \mathbf{i}\mu_m)^2\right\}$$
$$= \frac{1}{2}\sum_{k=1}^{K_x} p(\mu_m \mid \mathbf{u}_k^x)\left\{d_x\beta_m^{-1} - (\mathbf{u}_k^x - \mathbf{i}\mu_m)^2\right\} + \frac{1}{2}\sum_{k=1}^{K_y} p(\mu_m \mid \mathbf{u}_k^y)\left\{d_y\beta_m^{-1} - (\mathbf{u}_k^y - \mathbf{i}\mu_m)^2\right\}$$

The update rule for β_m^{-1} is:

$$\beta_m^{-1} = \frac{\sum_{k=1}^{K_x} p(\mu_m \mid \mathbf{u}_k^x)(\mathbf{u}_k^x - \mathbf{i}\mu_m)^2 + \sum_{k=1}^{K_y} p(\mu_m \mid \mathbf{u}_k^y)(\mathbf{u}_k^y - \mathbf{i}\mu_m)^2}{d_x\sum_{k=1}^{K_x} p(\mu_m \mid \mathbf{u}_k^x) + d_y\sum_{k=1}^{K_y} p(\mu_m \mid \mathbf{u}_k^y)} \tag{5.11}$$

The derivative of the objective function with respect to w_k^z is

$$\nabla O(w_k^z) = \frac{1}{w_k^z}\sum_{n=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n) - \lambda_z$$

$$\rightarrow w_k^z = \frac{1}{\lambda_z}\sum_{n=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)$$

$$\sum_{k=1}^{K_z} w_k^z = \frac{1}{\lambda_z}\sum_{k=1}^{K_z}\sum_{n=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n)$$

$$\rightarrow 1 = \frac{N_z}{\lambda_z} \rightarrow \lambda_z = \frac{1}{N_z}$$

The update rule for w_k^z is:

$$w_k^z = \frac{1}{N_z}\sum_{n=1}^{N_z} p(\mathbf{u}_k^z \mid \mathbf{z}_n) \tag{5.12}$$

The derivative of the objective function with respect to w_m is

$$\nabla O(w_m) = \frac{1}{w_m}\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x) + \frac{1}{w_m}\sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y) - \lambda_0$$

$$\rightarrow \sum_{m=1}^{K_0}\left\{\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x) + \sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)\right\} = \lambda_0 \sum_{m=1}^{K_0} w_m$$

$$\rightarrow K_x + K_y = \lambda_0$$

The update rule for w_m is:

$$w_m = \frac{1}{K_x + K_y}\left\{\sum_{k=1}^{K_x} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^x) + \sum_{k=1}^{K_y} p(\mathbf{i}\mu_m \mid \mathbf{u}_k^y)\right\} \tag{5.13}$$

5.4. Experimental design/ Simulated data

Simulated data were designed to evaluate the method in identifying such pair information/common patterns across different datasets. As one of the main purposes of integrative study is to identify how genes are similarly expressed or differentially expressed across many datasets. In cancer researches, a common signatures is important tasks, where it believed that common gene signatures

may exists [84, 87]. Only genes that show significant differential expression worth to be included for integration investigation. This is because other genes are not biologically meaningful/interesting, at least in this context. Also, including large data for analysis will cost a huge computing time. Thus, in the following simulated data I have stuck to this and sampled 400 ~ 1000 genes. However, I have expanded the most complex scenario to involve 20,000 genes aiming to demonstrate algorithms ability in genome wide analysis. For a fair comparison, different scenarios were designed to integrate two datasets and three datasets. Each one includes variables parameter such as noise levels, clusters number and pairing structures, as exemplified in the following sections.

Many methods were developed to generate/synthesis datasets. One of the most used in literature is that assume gene expression are drawn from Gaussian distributions [313-316]. Similarly, I have sampled genes from Gaussian mixtures with variables centres that represent different clusters.

5.4.1. Two datasets integration

For *scenario one:* The mean underlying differential expressions (MUDE) model was a Gaussian mixture with mean values of DE; $\mu = \{-2,-1,1,2\}$. Two data sets were generated $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_X}) \subset \Re^{N_x \times d_x}$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{Ny}) \subset \Re^{N_y \times d_y}$ where space dimensions d_x and d_y are 5 and 3 respectively. Four clusters were chosen for both datasets and 100 vectors (genes) were randomly drawn from each centre, $D \sim N(\mu_m, \sigma_m^2)$ where $\mathbb{D}$ is either X or Y, and $\sigma = 0.1$. This scenario was also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and

$\sigma = 0.5$, and was designed to investigate performance at different noise levels. (see **Table 5.1: S1** panel)

Table 5.1. Overall design of simulated data (S means scenarios 1-3)

Cluster	Data1	Data2
K1	100	100
K2	100	100
K3	100	100
K4	100	100

S1

Cluster	Data1	Data2
K1	100	×
K2	100	×
K3	100	100
K4	100	100
K5	100	100
K6	100	100
K7	100	100
K8	100	100
K9	100	100
K10	100	100

S2

Cluster	Data1	Data2
K1	100	×
K2	100	×
K3	100	100
K4	100	100
K5	100	100
K6	100	100
K7	100	100
K8	100	100
K9	×	100
K10	×	100

S3

Scenario 2, MUDE model was composed of ten differential expression means $\mu = \{-5,-4,-3-2,-1,1,2,3,4,5\}$. Two data sets were generated, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_X}) \subset \Re^{N_x \times d_x}$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N_X}) \subset \Re^{N_y \times d_y}$, where the space dimensions (d_x and d_y) were 5 and 3 respectively. Ten clusters were designed for dataset one and eight clusters were designed for dataset two. Two datasets had eight clusters in common and data set one has 2 clusters unique. 100 vectors (genes) were randomly drawn from each centre for data one, while the last eight centres were used for data set two $D \sim N(\mu_m, \sigma_m^2)$,where D is either $\mathbf{X}$ or $\mathbf{Y}$, and $\sigma = 0.1$. This scenario was also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and $\sigma = 0.5$, and was designed to investigate the performance at different cluster structures and test the ability of the algorithm to identify the unique cluster in one data. (**Table 5.1: S2**).

Scenario 3, I design a mean underlying differential expressions (MUDE) model with ten differential expression means; $\mu = \{-5,-4,-3-2,-1,1,2,3,4,5\}$. Two data

sets were generated $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_X}) \subset \Re^{N_x \times d_x}$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N_y}) \subset \Re^{N_y \times d_y}$ where the space dimensions (d_x and d_y) are 5 and 3 respectively. 100 vectors are randomly drawn from the first eight centres for data one and from the last 8 centres for data set two. This means there are only 6 clusters common across the two sets and 2 unique clusters for each data set one. Each vector is drawn from the hidden model, $D \sim N(\mu_m, \sigma_m^2)$,where D is either X or Y, and $\sigma = 0.1$. This scenario is also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and $\sigma = 0.5$. This scenario was designed to investigate the performance at different cluster structures and to test the algorithm's ability to identify the unique cluster in one data. (**Table 5.1: S3**)

5.4.2. Three data sets

In this section I have repeated the previous scenarios but this extended to three data sets. In *Scenario 4,* a mean underlying differential expressions (MUDE) model was designed with four differential expression means; $\mu = \{-2, -1, 1, 2\}$. Three data sets are generated X, Y and Z where the space dimensions are 5, 4 and 3 respectively. 100 vectors are randomly drawn from each centre for each dataset. Each vector is drawn from the hidden model, $D \sim N(\mu_m, \sigma_m^2)$, where D is X, Y or Z, and $\sigma = 0.1$. This scenario is also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and $\sigma = 0.5$. This scenario was designed to investigate performance at different noise level on more than 2 datasets. (**Table 5.2: S4**).

Table 5.2. Overall design of simulated data for 3 integration sets (S represents scenarios 4-6)

K	D1	D2	D3
K1	100	100	100
K2	100	100	100
K3	100	100	100
K4	100	100	100

S4

K	D1	D2	D3
K1	100	×	×
K2	100	×	×
K3	100	100	×
K4	100	100	×
K5	100	100	100
K6	100	100	100
K7	100	100	100
K8	100	100	100
K9	100	100	100
K10	100	100	100

S5

K	D1	D2	D3
K1	100	×	×
K2	100	×	100
K3	100	100	100
K4	100	100	100
K5	100	100	100
K6	100	100	100
K7	100	100	100
K8	100	100	100
K9	×	100	100
K10	×	100	×

S6

Scenario 5: a mean underlying differential expressions (MUDE) model was designed with ten differential expression means; $\mu = \{-5,-4,-3-2,-1,1,2,3,4,5\}$. Three data sets were generated, $_{X,Y}$ and $_{Z}$ where the space dimensions were 5, 4 and 3 respectively. One hundred vectors were randomly drawn from each centre for data one while only 8 centres were used for data set two and 6 centres used for data 3.This means there are only 6 clusters common across the three sets and 2 unique clusters belong to data set one: also there are 2 clusters common between data 1 and data 2 but not data 3. Each vector is drawn from the hidden model, $D \sim N(\mu_m, \sigma_m^2)$,where D is $_{X}$, $_{Y}$ or $_{Z}$, and $\sigma = 0.1$. Also, this scenario is repeated/extended using different noise levels ,specifically $\sigma = 0.3$ and $\sigma = 0.5$. This scenario was designed to investigate the performance at different cluster size with more than 2 datasets. (**Table 5.2: S5**).

For *Scenario 6:* MUDE model was designed with ten differential expression means; $\mu = \{-5,-4,-3-2,-1,1,2,3,4,5\}$. Three data sets are generated $_{X,Y}$ and $_{Z}$ where the space dimensions are 5, 4 and 3 respectively. 100 vectors are randomly drawn from the first eight centres for data one and from the last 8 centres for data set two and the middle 8 centres for data 3 (**Table 5.2: S6**).

This means there are only 6 clusters common across the three sets. Each vector is drawn from the hidden model, $D \sim N(\mu_m, \sigma_m^2)$,where D is X, Y or Z, and $\sigma = 0.1$. This scenario is also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and $\sigma = 0.5$. This scenario was designed to investigate the performance of integrate more than 2 datasets at different complex cluster structures and to test the ability of the algorithm to identify the unique cluster in one data or in two sets only.

5.4.3. Large scale data integration (full genome data integration)

Scenario 7: This scenario was designed for genome wide integration. This scenario is similar to scenario 5 but with a very large mean vector. In each data set 21,000 genes were simulated. MUDE model was designed with fifty means; $\mu = \{-25:-1\&1:25\}$. Three data sets were generated X, Y and Z where the space dimensions were 5, 4 and 3 respectively. Here 500 vectors were randomly drawn from the 42 centres for data one and from the last 42 centres for data set two and the middle 42 centres for data set three. This means there are only 34 clusters common across the three sets. Each vector was drawn from the hidden model, $D \sim N(\mu_m, \sigma_m^2)$,where D was X, Y or Z, and $\sigma = 0.1$. This scenario was also repeated/extended using different noise levels, specifically $\sigma = 0.3$ and $\sigma = 0.5$. This scenario was designed to evaluate the performance of integration algorithms for large datasets (>20,000 genes/set).

5.5. Evaluation measurements

I introduced two measurements to the evaluations. The first was to evaluate whether a cluster size was correctly estimated. The second was to evaluate

whether gene-gene independence/dependence was correctly maintained after clustering the data sets. In the design, a gene either belonged to an independent cluster or belonged to a dependent cluster. An independent gene was unique to one data set, such as genes in some clusters of Scenarios Two, Three, Five and Six. A dependent gene demonstrated similar differential expression across datasets. For each classification (a gene was classified into a cluster), I calculated an error as:

$$E_{mi} = 1 - \frac{2 * [match(\hat{\Re}_{mi}, \Re_{mi}) - 1]}{[length(\hat{\Re}_{mi}) + length(\Re_{mi})] - 2} \quad (5.14)$$

where E_{mi} lies between 0 and 1

$\Re_{mi}$ is a list of designed genes that belong to the same label of i^{th} gene in m^{th} data set.

$\hat{\Re}_{mi}$ is a list of predicted genes that belong to the same label of i^{th} gene in m^{th} data set.

$match(\hat{\Re}_{mi}, \Re_{mi})$ is the total number of common genes in both designed and predicted.

The total error ($totE$) is calculated below:

$$totE = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N_m} E_{mi}}{\sum_{m=1}^{M} N_m} \quad (5.15)$$

where M is the number of data sets and N_m is the total number of genes in the m^{th} data set

5.6. Benchmark algorithms

I used two benchmark algorithms for the comparison. They were BCC [310] and MDI [309]. The Multiple Dataset Integration (MDI) method [309] clusters each data set separately while concurrently modelling dependence between the clusters. It assumes a finite mixture model for each set while the dependence between data sets is exploited using the Dirchlet mixture coefficients describing their cluster membership. Similarly to MDI, Bayesian Consensus Clustering (BCC) [310] proposed different dependence modelling. MDI models the dependence pair wisely between data sources while BCC focuses on adherence to the overall clustering. Those methods outweigh the separate modelling in which each dataset clustered independently and then followed by manual integration [317, 318], since the joint modelling allow borrowing strength through/across many data sets.

5.7. Results and Discussions

Evaluation of the three scenarios on two data sets using three variance values is shown in **Table 5.3**. The first measurement is beyond the scope of this thesis but is examined here to make a full comparison. In this scenario, the cluster size prediction shows that the proposed algorithm performed well in this term under all variance values. The others have lower percentage at the large s.d of $\sigma = 0.5$. However, in terms of dependency/independency relations I compute the error percentage as proposed in the evaluation criteria section. They all performed well with s.d=0.1 with no errors at all. This means that the relationship is preserved after clustering. This is also the case at variance =0.3 with the 0.3% maximum error belonging to my proposed method and 0.26% and 0.03% for BCC and MDI respectively. The error figures increased with the large

s.d at 0.5 with a maximum 8.4% obtained, while the minimum error was observed in the HIM model.

Table 5.3. The average results of 50 simulations, including all scenarios in two datasets. Figures in **bold** mean the best performance and the underlined means the worst performance. k%: percentage of correctly cluster estimation, Error: is average error percentage. S indicates for Scenarios

S	Noise level	BCC		MDI		HIM	
		K%	Error%	K%	Error%	K%	Error%
1	0.1	**100**	**0**	**100**	**0**	**100**	**0**
	0.3	96.6	0.26	**100**	**0.03**	**100**	0.31
	0.5	16.6	8.4	60	7.8	**100**	**5.6**
2	0.1	NA		NA		**100**	**0**
	0.3					**60**	**1.5**
	0.5					**0**	**34.6**
3	0.1	36.6	53.1	**100**	72.8	86.6	**0.56**
	0.3	73.3	51.07	43.3	75	**80**	**2.7**
	0.5	**66.6**	52.8	3.3	72.8	0	**49**

In scenario 2 only, my proposed method is able to integrate such data with different sizes. The maximum error rate is 34.6% for simulated data at the large variance. Scenario 3 shows the merit of the proposed method in terms of the error rate. Despite the fact that MDI correctly gained 100% cluster numbers estimation for 0.1 noise, it has the maximum error rate with 72.8% since it barely distinguishes the different cluster designs due to the so-called correlation clustering it applies.

Next I wanted to see how far each method estimated the clusters as well as the variation of the error obtained. Therefore I used the boxplots in **Figure 5.3** and **Figure 5.4** to evaluate the stability/variations of each method. **Figure 5.3** shows the cluster estimated by each method for each data set among 50 simulations. It can be seen that BCC was the worst for scenario one at large noise. However BCC was the best for scenario 3 at the large noise level while both MDI and HIM always estimated the clusters below the correct clusters.

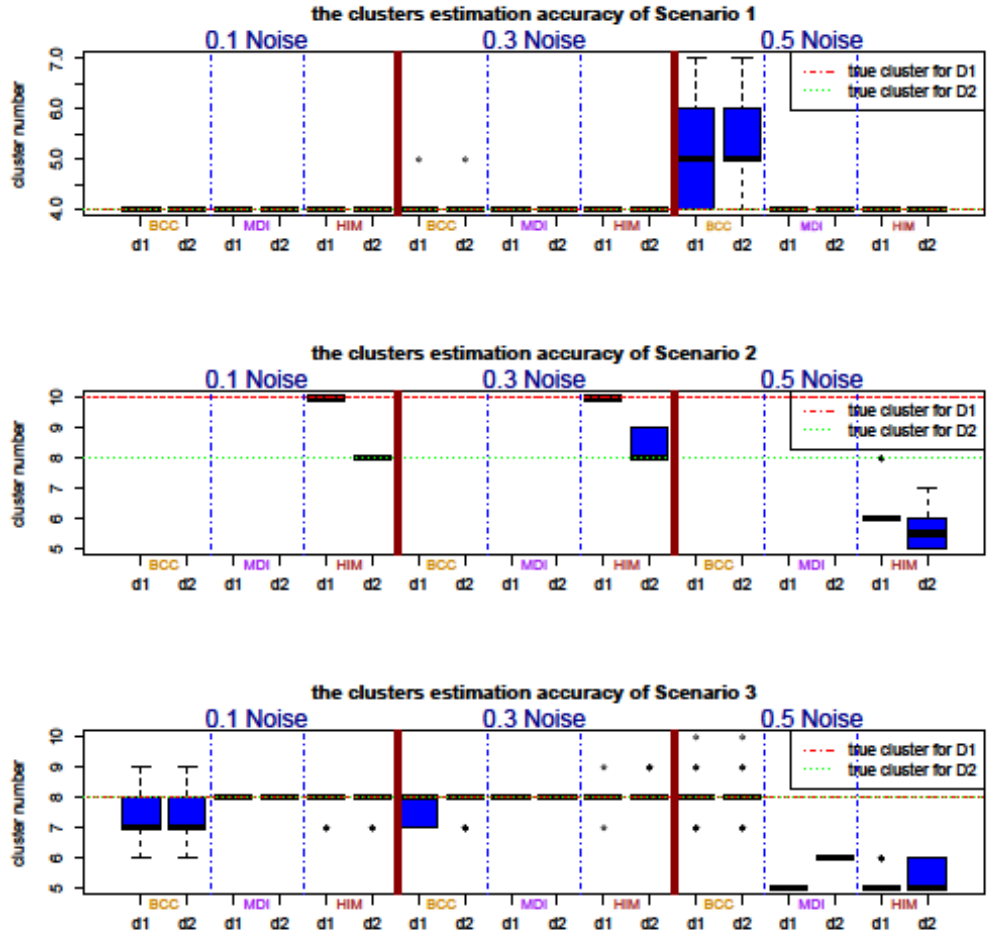


Figure 5.3. Boxplot to illustrate the cluster estimation accuracy for each dataset on different algorithms and scenarios on 2 sets integration. D1, D2 are Datasets 1 and 2.

According to **Figure 5.4,** HIM always has the least error rate in all simulations. In scenario 3 both MDI and BCC gain a relatively larger error rate compared to HIM. Although HIM fails to estimate the number of clusters correctly in scenario 3 noise 0.5, it is still able to maintain the designed structure. In contrast to HIM, BCC gained the highest percentage for cluster estimation but fails to capture the relation of the designed data.

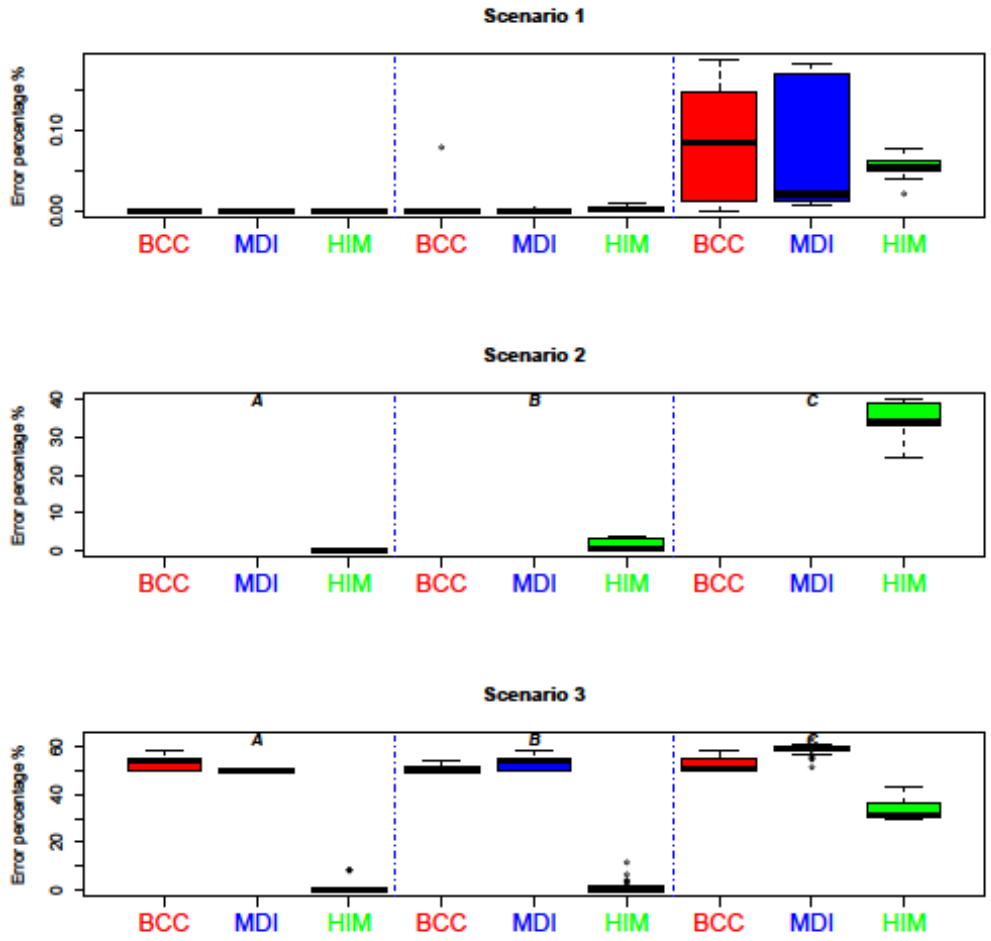


Figure 5.4. Boxplot of the error by all methods among 50 runs; A, B and C signify 0.1, 0.3 and 0.5 noise levels respectively on 2 datasets integration.

All methods included in this study/evaluation/comparison are able to work with more than two datasets. Here I present the result of those algorithms on extended scenarios. **Table 5.4** indicates that BCC performed the worst in terms of both clusters' size/estimation and the error rate. In the first scenario(S4) the cluster estimation showed that the proposed algorithm performed well for all variance values, with 100% correctly-estimated accuracy. In scenario 2(S5), where only HIM was able to integrate data with such a design, the error rate was high, with a maximum 58.2% at the large variance of 0.5. Although this scenario worked well in the two datasets integration it seems to have been less

accurate within 3 sets. This drop could be related to the low estimation of cluster numbers (**Figure 5.5**).

Similar to the 2 datasets application (Scenario 3), BCC and MDI failed to capture the designed pairing information with less than 45% of the correctly-maintained relationship being preserved after clustering. Interestingly, the number of clusters in BCC increased with the noise level increasing, while the others decreased as can be seen in **Figure 5.5***.* The reason for this is that BCC always assumes the large number of clusters; then after the clustering equips only the relevant cluster. Thus, the number of used clusters will definitely be larger with the large noise. In contrast to that, MDI and HIM predict fewer clusters with the large noise as they merge the overlapped clusters together, or, in other words, shift the centres to estimate a new centre in between. This is a hot topic in the machine learning world but is beyond the scope of this thesis.

Table 5.4 Average results of 50 simulations including all scenarios for three datasets integration. Figures in **bold** mean the best performance**.** S = scenario, underline figures are the worst, K: percentage of correctly cluster estimation, Error is percentage of average error.

S	Noise level	BCC		MDI		HIM	
		K%	Error%	K%	Error%	K%	Error%
4	0.1	15	24.12	**100**	**0**	**100**	**0**
	0.3	0	49.52	**100**	**0**	**100**	0.09
	0.5	0	56.37	61.1	6.6	**100**	**4.33**
5	0.1	NA		NA		**100**	**38.8**
	0.3					**55**	**39.57**
	0.5					**0**	**58.23**
6	0.1	5	71.63	**100**	66.66	**100**	**41.66**
	0.3	0	73.37	33.33	66.02	**45**	**41.71**
	0.5	0	76.64	0	65.22	0	**50.24**

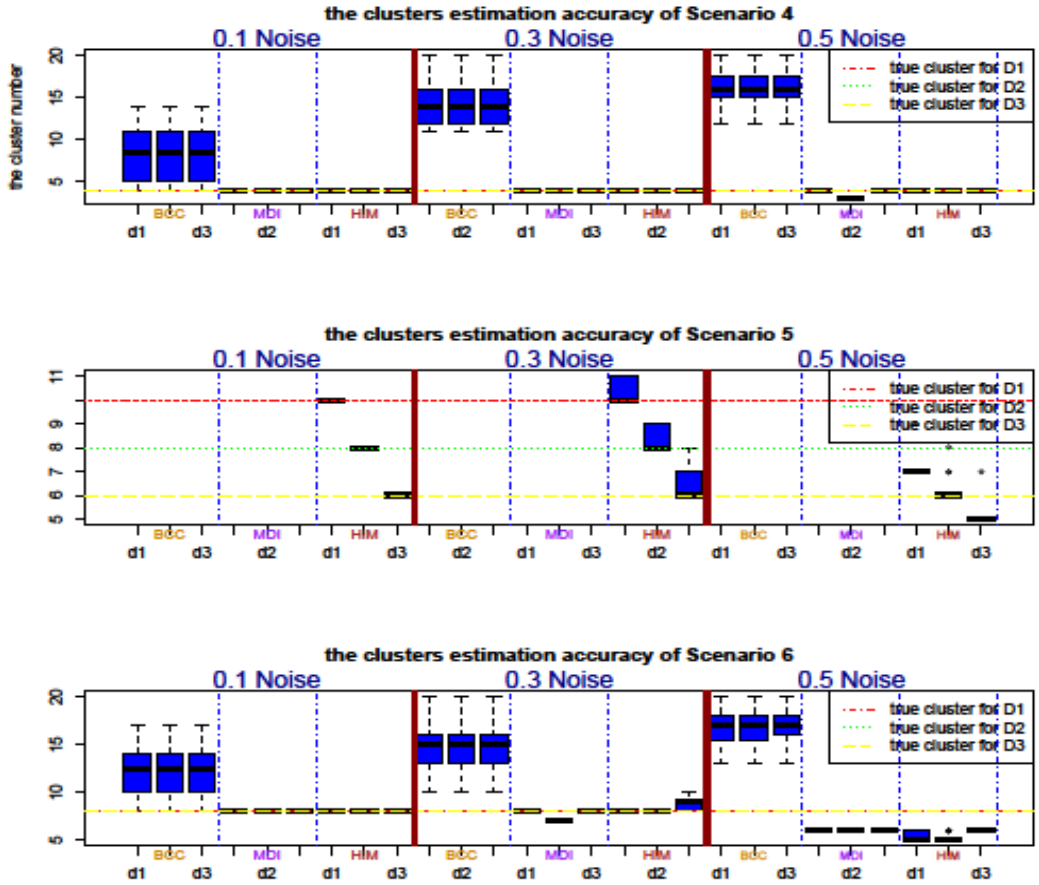


Figure 5.5. Boxplot to illustrate the cluster estimation accuracy for each data set on different algorithms and scenarios on 3 integration sets. D1, D2, D3=Datasets 1,2 and 3.

Figure 5.6 showed that HIM always had the least error rate in all simulations. In scenario 6 both MDI and BCC gained a large error rate compare to HIM. Although HIM failed to estimate the number for clusters correctly in scenario 3(S6) noise 0.5, it is still able to maintain almost 50% of designed structure.

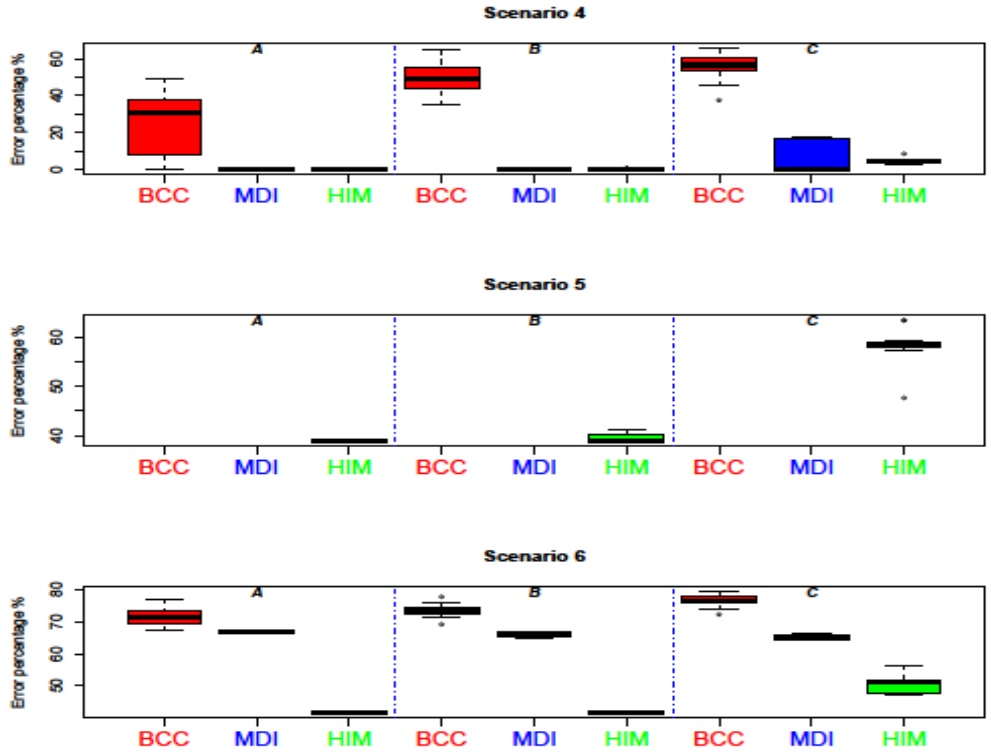


Figure 5.6 Boxplot of error by all methods among 50 runs. A, B and C mean 0.1, 0.3 and 0.5 noise level respectively on 3 datasets integration.

Finally, It worth to see the ability of those algorithms on a large scale data. One issue arisen of such analysis is huge computing cost . For example in this case three months were needed to finished this scenario and hence that I repeat them only 10 times. MDI is the slowest where it takes 6 weeks to finish one sub-scenario (i.e. noise level=0.1). BCC also takes long time (~ 2 weeks) for a sub-scenario. The number of sampling for both MDI and BCC used in this study were set to the default (10,000).

Table 5.4. The average results of ten simulations for scenario 7 (large scale integration). Figures in **bold** mean the best performance and the underlined means the worst performance. k%: percentage of correctly cluster estimation, Error: is average error percentage. S indicates for Scenarios

S	Noise level	BCC		MDI		HIM	
		K%	Error%	K%	Error%	K%	Error%
7	0.1	0	73.45	0	80.74	0	**1.64**
	0.3	0	73.80	0	80.65	0	**6.77**
	0.5	0	78.43	0	82.42	0	**45.43**

Surprisingly, all algorithms failed to estimate cluster numbers correctly (**Table 5.4**). This is due the complexity of large number of data to be analysed. It can be seen in **Figure 5.7** that HIM's cluster estimation was the closest to the correct number of clusters especially for the small and medium noise levels. BCC has the same issue as in small data integration where it always estimates large cluster number. However, this can be explained by the way BCC cluster estimation proposed. They suggest to put a large cluster number then map the data into the clusters. In contrast MDI tends to have a very small numbers of clusters, this also encountered for the previous scenarios. One way to explain this is that in such large data it is recommended to increase the number of mixtures to be used. Although there is no clear guide on this in their paper, I have found that they used in their data number of mixtures equal to the total number of genes over two. Different numbers of mixture were used to see if this can improve the estimation or not. Specifically 100,200,1000 and 2000 were chosen as the number of mixtures with no improvement noticed.

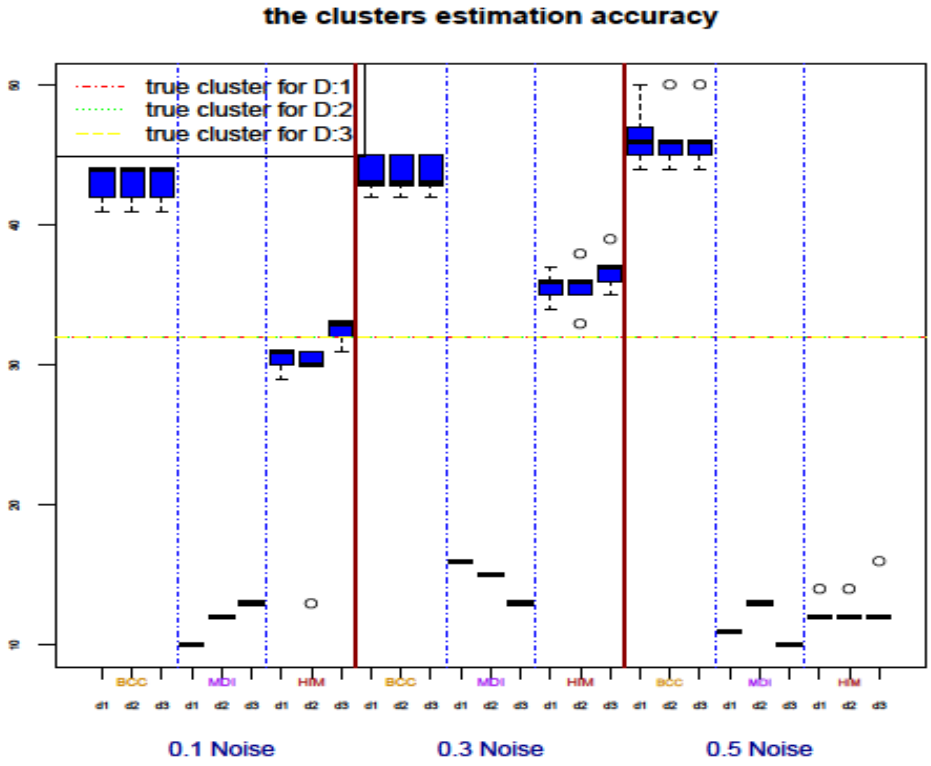


Figure 5.7. Boxplot to illustrate the cluster estimation accuracy for each data set based on different algorithms for scenarios7. D1, D2 and D3 are Datasets 1, 2 and 3.

The poor cluster number estimation had led to large errors(Table 5.4) for BCC and MDI where they only maintain the dependency/independency pairing for less than 26%, while HIM has very small error percentages (1.65% and 6.77%) in the small and medium noise. In the large noise HIM also gained large error and this can be linked to the small number of cluster estimated (Figure 5.8).

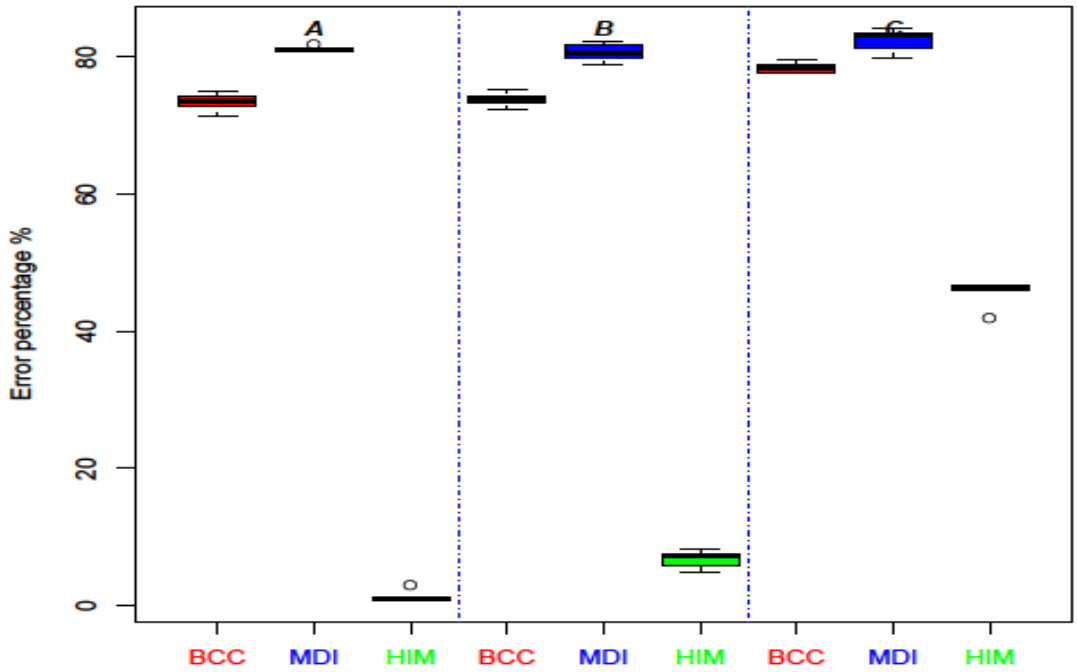


Figure 5.8 Boxplot of error by all methods among 10 runs. A, B and C mean 0.1, 0.3 and 0.5 noise level respectively for large scale integration(Scenario 7).

5.8. Conclusion

In this chapter, a comprehensive survey of integrative methods was given. The limitations of those method also discussed and considered in the design of the proposed method. A mean pattern model for integrative study was introduced based on the assumption that an experiment is well designed. Therefore, I assume that all the replicates (observations) in an experiment are random samples of a mixture of one-dimensional Gaussians. This simplification makes modelling of differential expressions of integrative study much easier. I have used simulated data to show that the mean pattern model outperforms the existing algorithms used for integrative study. The main issue is the poor cluster estimation; especially when large data are included for analysis. One way to overcome the common issue in clustering is to filter the data by including only the differentially expressed genes in each data set. In the next chapter I will

apply the algorithms to the real data to show how the mean pattern model explores biological relations/meaning inside biological data.

Chapter 6

Application of the Integrative Study

Abstract

Integrative study has been an important approach for revealing unbiased truth from multiple biological experiments for one investigation. The challenge in integrative study is large-scale data and the complexity of different data types. Based on the proposed method for integration, I show in this chapter that the mean pattern model implemented in a hierarchical structure works well compared with benchmark algorithms in real data.

6.1. Introduction

In this chapter, I have applied BCC, MDI and HIM to different real data sets to compare their performance. The first data, which is hereafter named Application One, is cross-species data, which aimed to integrate data from different species (Mouse and Rat). The second data, which is after this named Application Two, is cross-cancer data, which integrates four different cancer datasets from two cancer type.

6.2. Application 1.

6.2.1. Data

The data which was downloaded from gene expression omnibus (GEO) with the accession number GSE42081, includes expression from Mouse and Rat at three different cell development stages (figure 6.1), each containing two replicates The stages are blastocyst (B), inner cell mass (ICM) and Morula (M) [23].

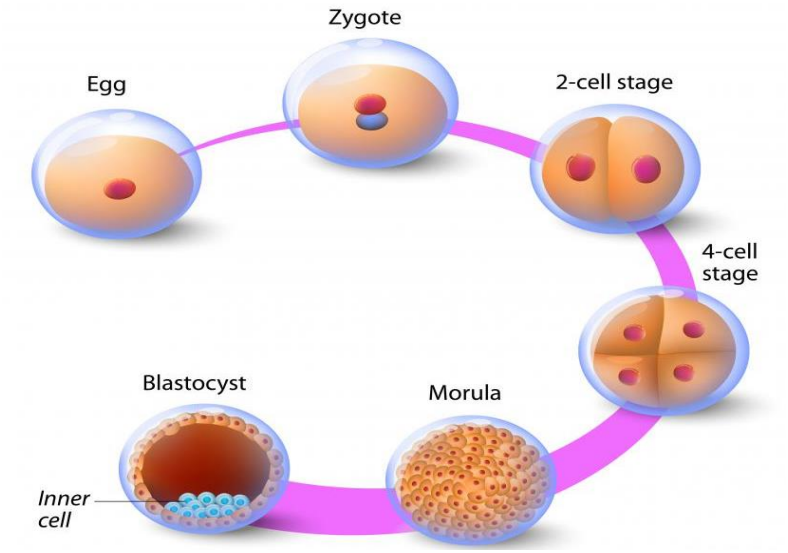


Figure 6.1 illustration of the development stages of embryo from fertilization of the egg to the implantation of the blastocyst. (downloaded form wisegeek; http://www.wisegeek.com/what-is-mesenchyme.htm)

It's an important study of the regulatory networks in the preimplantation embryo that determine cell fate decisions. More specifically, the transition between morula to blastocysts is attractive, as during this time the pluripotent cells are created [23]. In Casanova study, they performed a microarray analysis to study the different regulation patterns of mouse and rat. Three populations of cells were collected from the two organisms. Morula and blastocysts stages embryos and inner cell mass cells isolated from the blastocysts are the three different populations used in this study.

6.2.2. Data pre-processing

To use all data, I took the difference expression of; ***i***) B and ICM as mod 1; ***ii***) ICM and M as mod 2; and ***iii***) B and M as mod 3 for each species (Mouse and Rat) (**Figure 6.2**), as done in the original research. It is an important task to remove bimodal genes. Including such genes in a study may affect the result of modelling where the proposed method in chapter 5 designed to assume that all replicates/samples in a gene are random samples from the same Gaussian. Thus, removing such genes with bimodal behaviour before using the HIM algorithm is required.

In order to remove the bimodal/outlier genes, MSG algorithm proposed in Chapter 2 was used. The reason for using MSG instead of hBI is that hBI requires more than two replicates to be used as explained in chapter 3. In this case MSG can be used for bimodality detection for two replicates, as the differential expressed gene identified is a bimodal. After modelling the data using MSG, I removed the top 10% of genes that were associated with the large posterior probability obtained by MSG.

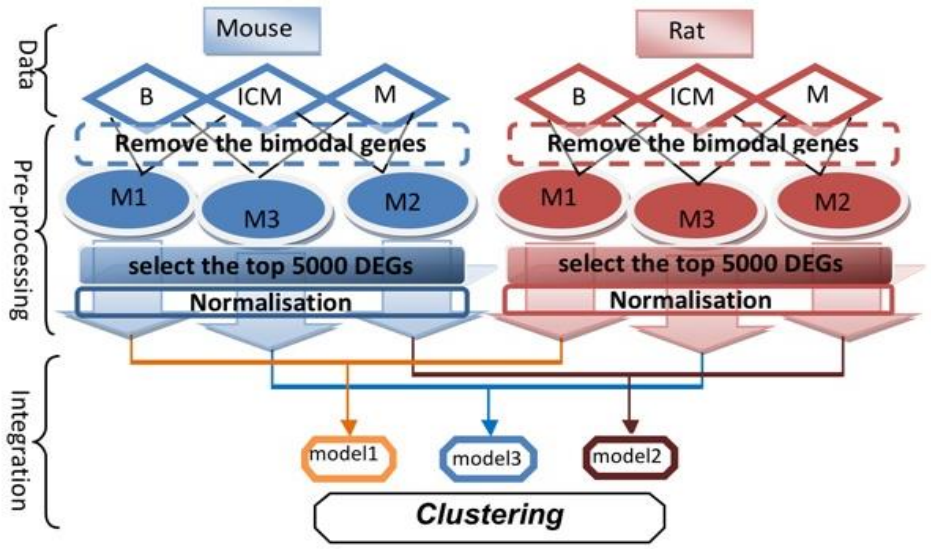


Figure 6.2. Diagram of the integration process of two species that include 3 stages

Having removed such bimodal genes I used MSG again to identify the differentially expressed genes (DEGs). I then selected the top 5000 genes differentially expressed from each mode (M1:M3) to be integrated across species. In order to make the expression level comparable, normalisation was applied before integration.

6.2.3. Integration Results

It was noticed that allowing each algorithm to estimate the cluster number gives very different results. Thus, it is difficult to compare the results obtained by each algorithm where the number of clusters between them is different. For example, using *Mclust* packages for large noisy data in HIM always returns very small clusters (i.e., K=3), whereas this is always the other way around for BCC as shown in the simulated data in the previous chapter. As a corrective action, I used the maximum cluster number estimated by MDI to be the number of cluster for both BCC and HIM. I have rely on MDI to estimate the cluster for the reason that I cannot enforce MDI to be cluster into a predefine cluster number while this possible for HIM and BCC. More enhancement, I replaced the *Mclust*

package used for clustering in HIM with *Kmeans* package as there is no need to estimate the cluster number.

Using BCC, MDI and HIM to integrate model 1, model 2 and model 3, as in **Table 6.1**, it was found that 5.6% of genes had the same cluster label among species for model 1 by BCC while the percentage increased to 8.16% and 8.28% respectively for MDI and HIM. However, MDI linked to the lower percentage for model 2 (2.9%) and to the higher percentage in model 3 (14.6%).

Table 6.1. Percentage of common genes assigned to the same cluster across species

Method	Model 1	Model 2	Model 3
BCC	5.64%	7.5%	6.5%
MDI	8.16%	2.94%	14.6%
HIM	8.43%	8.28%	6.78%

Table 6.2 shows that BCC always has common clusters while MDI only has unique clusters at model 3. As expected, HIM has unique clusters in all models, which is further supportive evidence for the merit of the HIM algorithm in such data as seen in the simulated data.

Table 6.2. shows the cluster numbers for each species in all three models. Bold numbers signify input cluster number for BCC and HIM. Underline italic numbers signify some unique clusters.

species	Model 1		Model 2		Model 3	
	Mouse	**Rat**	**Mouse**	**Rat**	**Mouse**	**Rat**
BCC	27	27	20	20	28	28
MDI	**27**	**27**	**28**	**28**	**26**	***28***
HIM	18	*23*	20	*25*	23	*26*

Further evaluation was conducted using the R package ClusterProfiler [319] to map each cluster for GO enrichment molecular function (MF), biological process (BP), and cellular component (CC). **Figure 6.3** shows the MF mapping of clustered genes for model 1 of the mouse species using BCC, MDI and HIM. It can be seen that some clusters are conserved to certain molecular functions. This is clear in clusters 6, 8 and 16 from HIM. **Figures S6.1:S6.5** show the molecular function mapping of clustered genes from models 2 and 3 of the mouse species, and all models for rat species using the three algorithms.

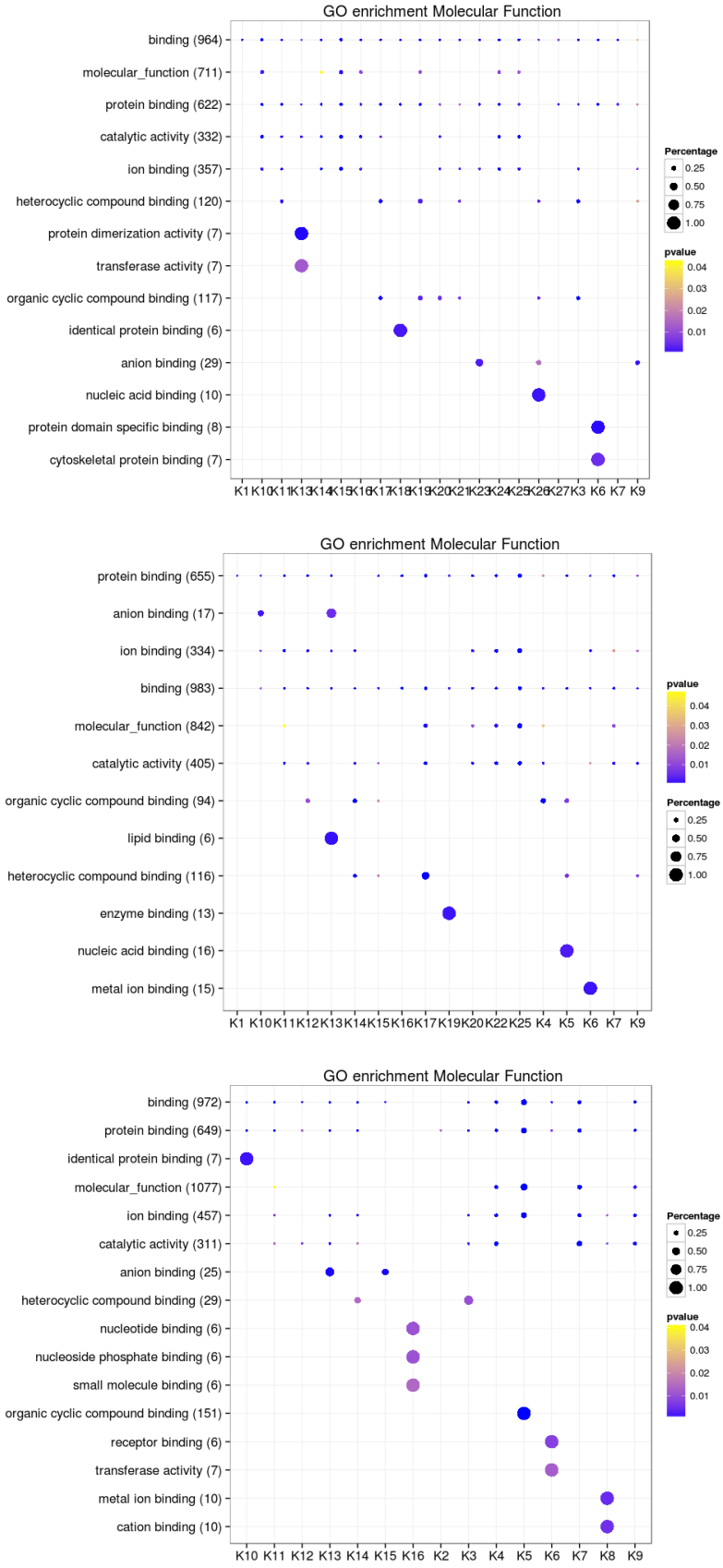


Figure 6.3 Molecular function mapping for clustered genes for model 1 of Mouse species, identified by BCC (left), MDI (centre) and HIM (right); the horizontal axis is the cluster label and the vertical line is for molecular function.

Figure 6.4 shows the cellular component enriched for each cluster in model 2 of the rat species. Also this has shown some homogenous clusters that conserved to certain CC. **Figures S6.6:S6.10** shows the cellular component mapping of clustered genes from models 1 and 3 of the rat species and all models for mouse species using the three algorithms.

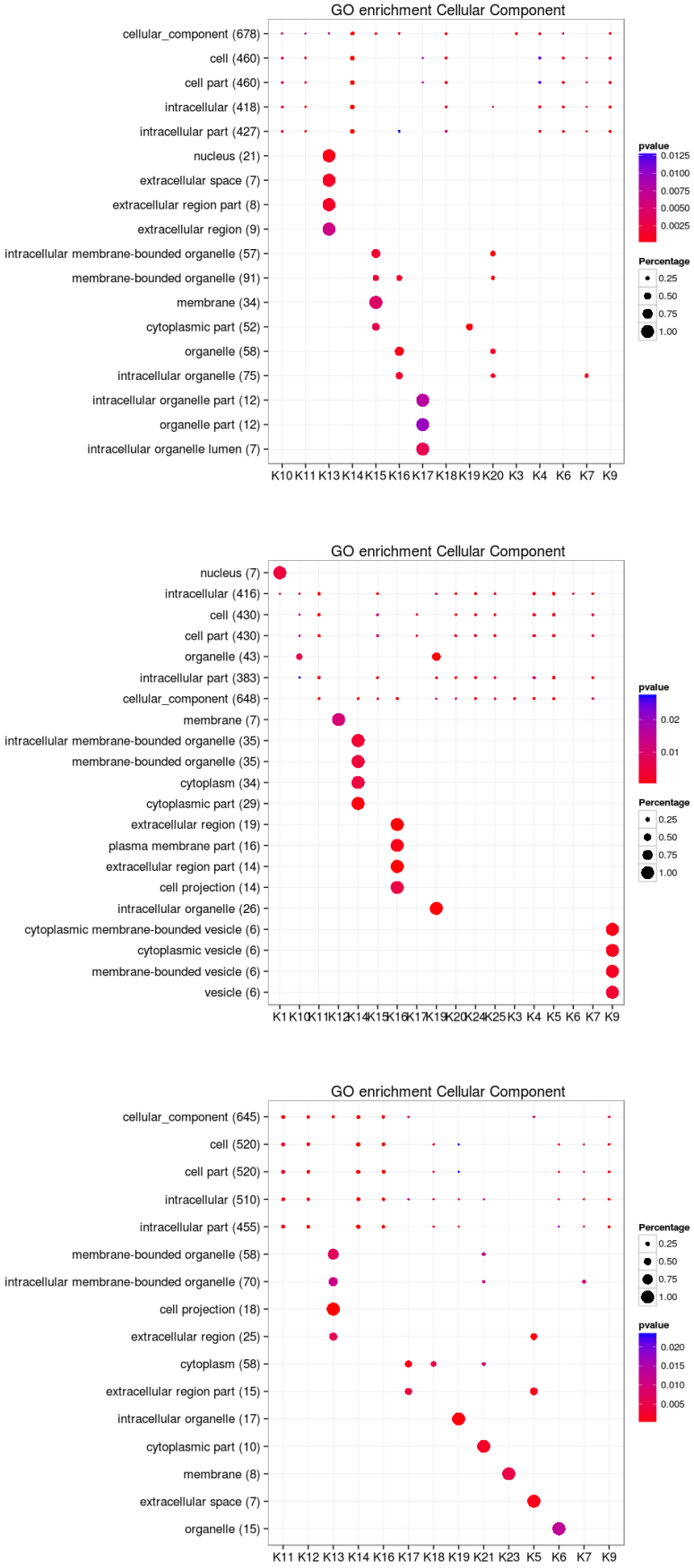


Figure 6.4. Cellular component mapping for clustered genes for model 2 of Rat species identified by BCC (left), MDI (centre) and HIM (right). The horizontal axis is the cluster label and the vertical line is for cellular component.

Figure 6.5 shows the biological process enriched for each cluster in model 3 of the mouse species. This showed that some clusters are conserved to certain BP. **Figures S6.11:S6.15** shows the biological process mapping of clustered genes from model 1and 2 of the mouse species and all models for rat species using the three algorithms.

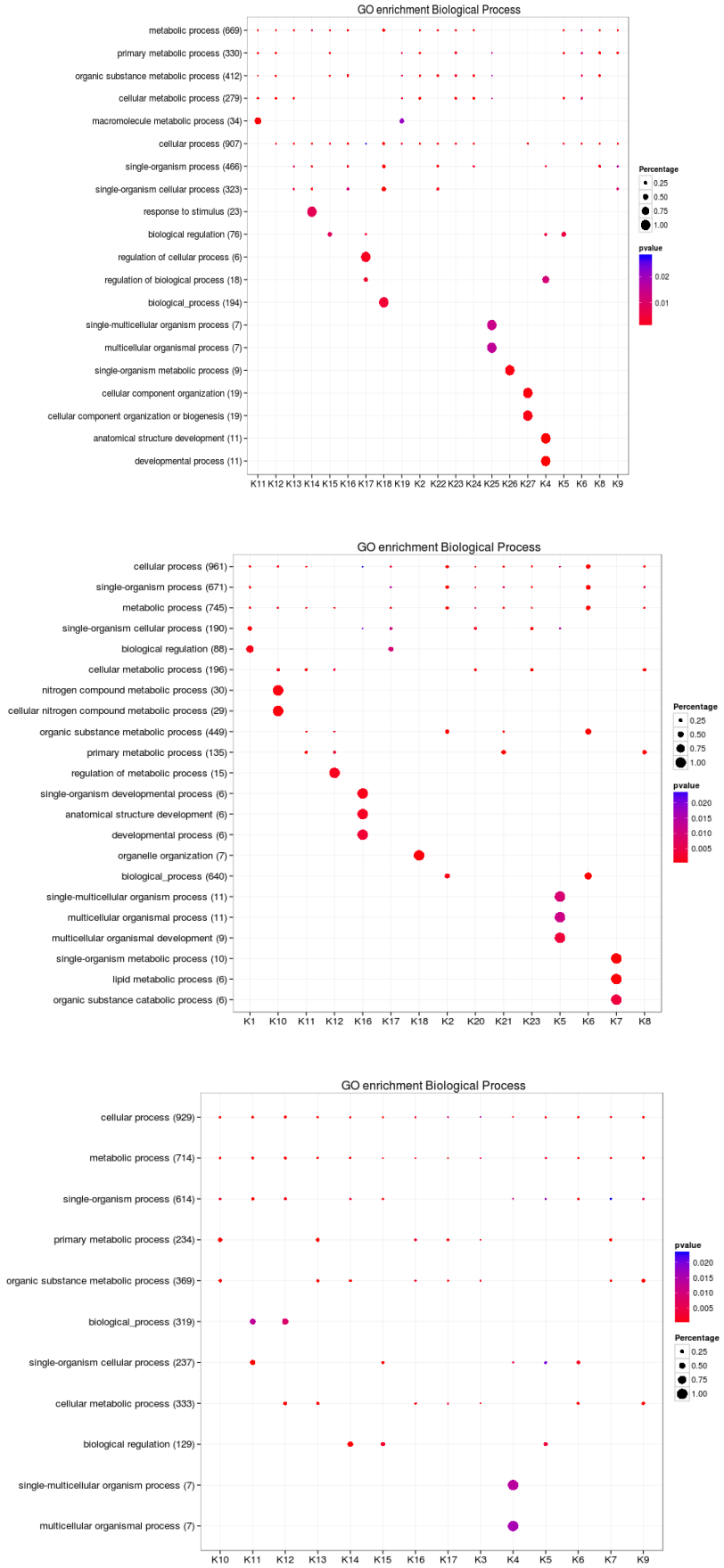


Figure 6.5. Biological process mapping for clustered genes for model 3 of mouse species identified by BCC (left), MDI (centre) and HIM (right). The horizontal axis is the cluster label and the vertical line is for biological process.

Table 6.3 shows genes that cluster in the same cluster among the two species. Although the top 5000 DEGs were selected from each species in the three models, it was found that out of 5000 genes, 1470, 1292 and 1519 were differentially expressed in both species in model 1, model 2 and model 3 respectively. However only 150 genes were common among all models.

Table 6.3. Genes identified as being in the same cluster across species. ALL means having same cluster by all methods. B-H in same cluster by BCC and HIM, B-M in same cluster by BCC and MDI, and M-H in same cluster by MDI and HIM

ALL	B_H	B_M				M-H
Bloc1s5	Acsl4	Apmap	Epha1	Sbds	Slc38a5	Asz1
Lgals1	Myh11	Capn13	Meis3	Slc22a18	Snx27	Cd79b
Lsm3	Wdr4	Car7	Nckap1l	Fgl1	Ssrp1	Cyfip1
Nsmce2		Ccdc102a	Nr1i3	Foxi1	Stk25	Ift43
Pfkfb3		Ccnb1	Pdhx	Gpr84	Supt20	Mfge8
Tmbim6		Cdc23	Pou2f1	Ict1	Tmed9	Nck2
		Cops3	Prss42	Krt26	Trio	Ptbp3
		Cox7a2l	Psmb4	Ly6h	Vps45	Rtcb
		Dapl1	Ring1	Med8	Wdr43	Wipf1
		Ddc	S1pr2	Meis1		Zfp292

To present a fair comparison I selected some genes for further comparison against the literature. One important work is the study that produced the data [23]. I found that the gene Stat3 cluster was in the same cluster for model 1 as well as model 3, but was not differentially expressed from model 2. This totally agreed with the finding by Casanova *et al* since they identified this gene to be up-regulated in model 1 and 3 but not differentially expressed in model 2 [23].

The same study also revealed that the gene Lifr had the same regulated process as the previous gene and interestingly this was also clustered into the neighbouring cluster of the previous one. This suggests that the gene had the same direction as the other but might have a different magnitude. The gene Aph1a was identified as up-regulated in the mouse and slightly changed in the

rat [23]; this was also extracted by HIM and classified into the high positive cluster 24 for the mouse and in the less positive cluster 17 for the rat at model 1, and in 25 and 18 for model one as illustrated in **Figure 6.6.**

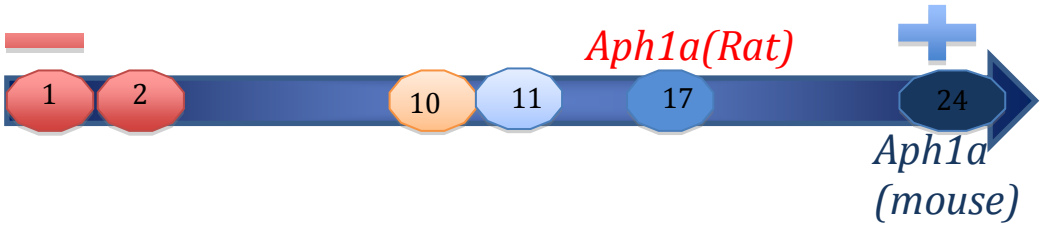


Figure 6.6 HIM clusters label distribution; illustration for the Aph1a cluster in both species by HIM

Another important gene is Skp2 which was identified in two different regulation directions by HIM, specifically in cluster 4 for the mouse and cluster 15 in the rat. This also agreed with findings by Casanova *et al* (2012) as the gene was down-regulated for the mouse while up-regulated in the rat. They noticed a different regulation for the gene Myl9 in all models, while my approach identified the same but only for model 2 while the gene was not differentially expressed in models 1 and 3 in both species. The gene Myc was found by Casanova *et al* to be down-regulated for the rat and not changed in the mouse; however, I identified this gene to be down-regulated in both species at model 1 with different magnitude; in cluster 5 for the mouse and 2 for the rat.

HIM successfully clustered the gene Gsk3β into the next neighbouring clusters, 16 and 17 which reflects the same finding by Casanova *et al* (2012). Gene Bmp4 was identified with a similar expression pattern among species and the same was gained by HIM where it was classified in a very close up cluster. Finally, the gene Wnt6 was found to show a different regulation between mouse and rat [23]. HIM therefore classified them into two far clusters making them consistent with the findings of Casanova *et al*.

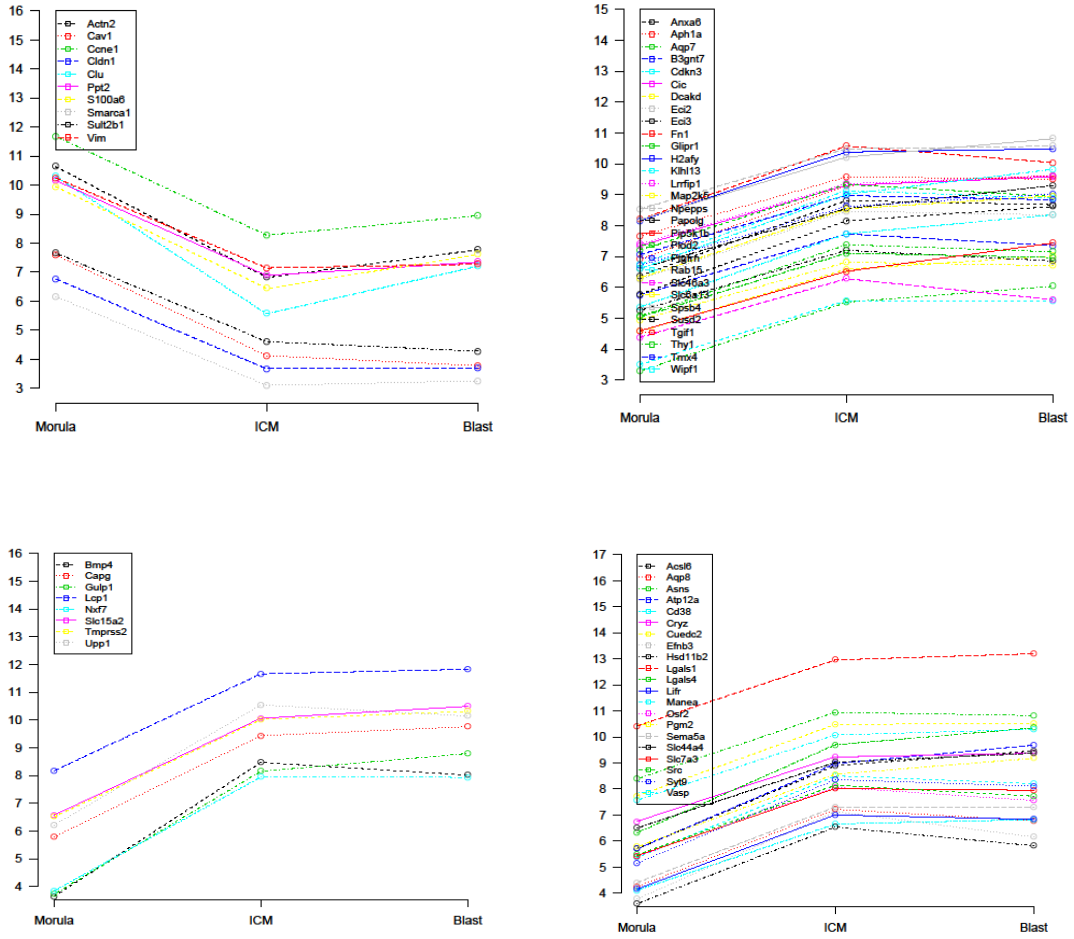


Figure 6.7. Pattern analysis of the raw expression at all 3 stages for each cluster

Figure 6.7 shows the pattern of the raw expression for four clusters, which also supports the fact that each cluster has a unique pattern. It can be seen that the majority of genes have different regulatory direction.

6.3. Application 2.

6.3.1. Data sets

Four cancer datasets were downloaded from Gene Expression Omnibus (GEO) (**Table 6.4**). The first data was a prostate cancer type that contained 6 normal samples and 6 cancer samples. The second was also prostate data that included 6 pairs of samples; the third set was breast cancer data that contained

6 normal and 6 cancer samples; while the fourth was also a breast cancer set that included 7 normal and 7 cancer samples.

Table 6.4. The data sets used for Application 2

Accession#	Type	Sample	Log 2
GDS1439	Prostate	6N / 6C	×
GDS4114	Prostate	6N / 6C	√
GDS4114	Breast	6N / 6C	√
GDS2250	Breast	7N / 7C	√

Note: The data sets used for application 2, normal (N) and cancer (C), log 2 if the data is already logged (√) or not

Figure 6.9 explains the process for this application. In order to remove the bimodal/outlier genes the *h*BI algorithm proposed in Chapter 3 was applied. Using *h*BI the top 10% of genes associated with the smallest *p* values were removed, after which the data was modelled using MSG to find the top 1000 genes differentially expressed between the two stages (normal and cancer) independently for each set. As previously noted it was not possible to use a critical posterior cut-off point for selection as the other algorithms were not able to integrate different gene numbers among the data.

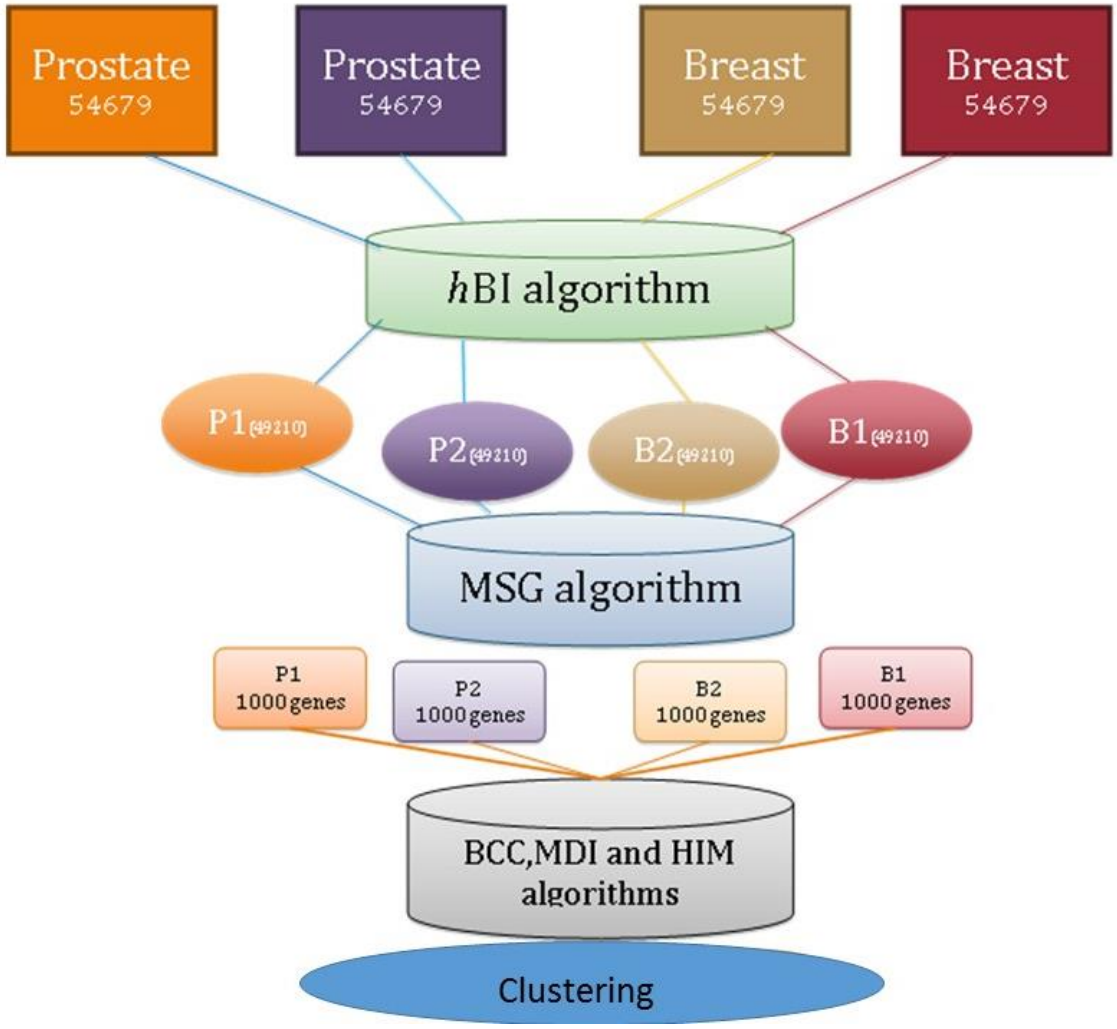


Figure 6.8. Flow chart of the second application. The numbers mean gene number after each process. P = Prostate; B = breast

6.3.2. Results and Discussions

In order to assess the integration clustering by BCC, MDI and HIM, the Biological Homogeneity Index (BHI) [320] was applied, using the R package clValid [321]. Clusters that have many genes and share GO annotations will have a high BHI score, with the value being between 0 and 1 where 1 is a completely homogenous cluster. Here four different BHI scores were included, namely Biological Process (BP), Cellular Component (CC), Molecular Function (MF), and ALL which included all three categories. **Table 6.5** shows the four different BHI scores by each method for each data set, and **Table 6.6** shows the BHI scores for the combined cluster of all four data sets. In most cases HIM

is either the best (as indicated by bold figures in the tables) or the second best (indicated by underlined italic figures). Interestingly, BCC was the best among all categories for data 1 and this could be linked to the even distribution of cluster size (**Table 6.7**) as the HIM and MDI have also only 5 and 4 clusters respectively, compared to 6 identified by BCC. However, among these, at least two clusters contained fewer than 20 genes. The reason for not using BHI for application one is because there is no available annotation packages for mouse or rat. Although I have downloaded the available packages for mouse and rat, but both of them returns zeros or NA.

Table 6.5. Comparison of BHI scores for each data cluster obtained using BCC, MDI and HIM. Figures in bold are the best performance and the underlined figures are the second best performance.

Data	Method	BP	MF	CC	ALL
1	BCC	**0.183**	**0.235**	**0.223**	**0.172**
	MDI	*0.097*	*0.142*	*0.135*	*0.131*
	HIM	0.093	0.128	0.114	0.115
2	BCC	0.121	**0.155**	*0.164*	0.189
	MDI	*0.123*	*0.149*	**0.173**	*0.192*
	HIM	**0.124**	0.138	0.152	**0.196**
3	BCC	**0.151**	**0.157**	**0.167**	0.202
	MDI	*0.142*	0.142	0.123	**0.222**
	HIM	0.121	*0.143*	*0.152*	*0.209*
4	BCC	*0.122*	**0.139**	0.127	0.227
	MDI	0.119	0.117	*0.136*	**0.243**
	HIM	**0.125**	*0.132*	**0.146**	*0.236*

Table 6.6. Comparison of BHI scores for combined clusters obtained using BCC, MDI and HIM for all data. Figures in bold are the best performance while the italic is the second best performance.

method	*BP*	*MF*	*CC*	*All*
BCC	0.132	0.171	*0.161*	**0.191**
MDI	**0.196**	**0.219**	**0.209**	0.157
HIM	*0.134*	*0.175*	0.153	*0.186*

Table 6.7: Cluster size for each cluster in each data by all methods.

Method	label	data1	data2	data3	data4
HIM	**1**	0	35	40	59
	2	9	77	203	574
	3	337	432	332	172
	4	638	352	294	0
	5	14	65	115	187
	6	2	39	16	8
BCC	**1**	89	124	80	118
	2	137	100	65	74
	3	145	125	147	149
	4	131	156	89	124
	5	152	163	194	194
	6	346	332	425	341
MDI	**5**	337	328	304	596
	9	0	0	0	3
	22	0	0	6	0
	31	0	0	0	58
	44	642	377	261	206
	45	9	167	158	45
	49	12	128	271	92

The clusterProfiler package was also used to explore whether a cluster conserved to MF, BP or CC. **Figures 6.9-6.11** shows the mapping of clustered genes for data 1 (Prostate1) to MF, BP or CC, using BCC, MDI and HIM. Interestingly, it can be seen that the results of this are consistent with the previous findings for BHI. However, this may reveal part of the issue discovered earlier where BCC showed a high index/score. It can also be seen that MDI and HIM have some more conserved clusters but fewer – in this case 3 clusters each (since this package ignores clusters with a few data points) – whereas some clusters in BCC have the same enriched process. In other words they are overlapped by other clusters. This supports the notion that relying on BHI gives only part of the story because BHI assesses each cluster independently and does not look at the overlap between clusters.

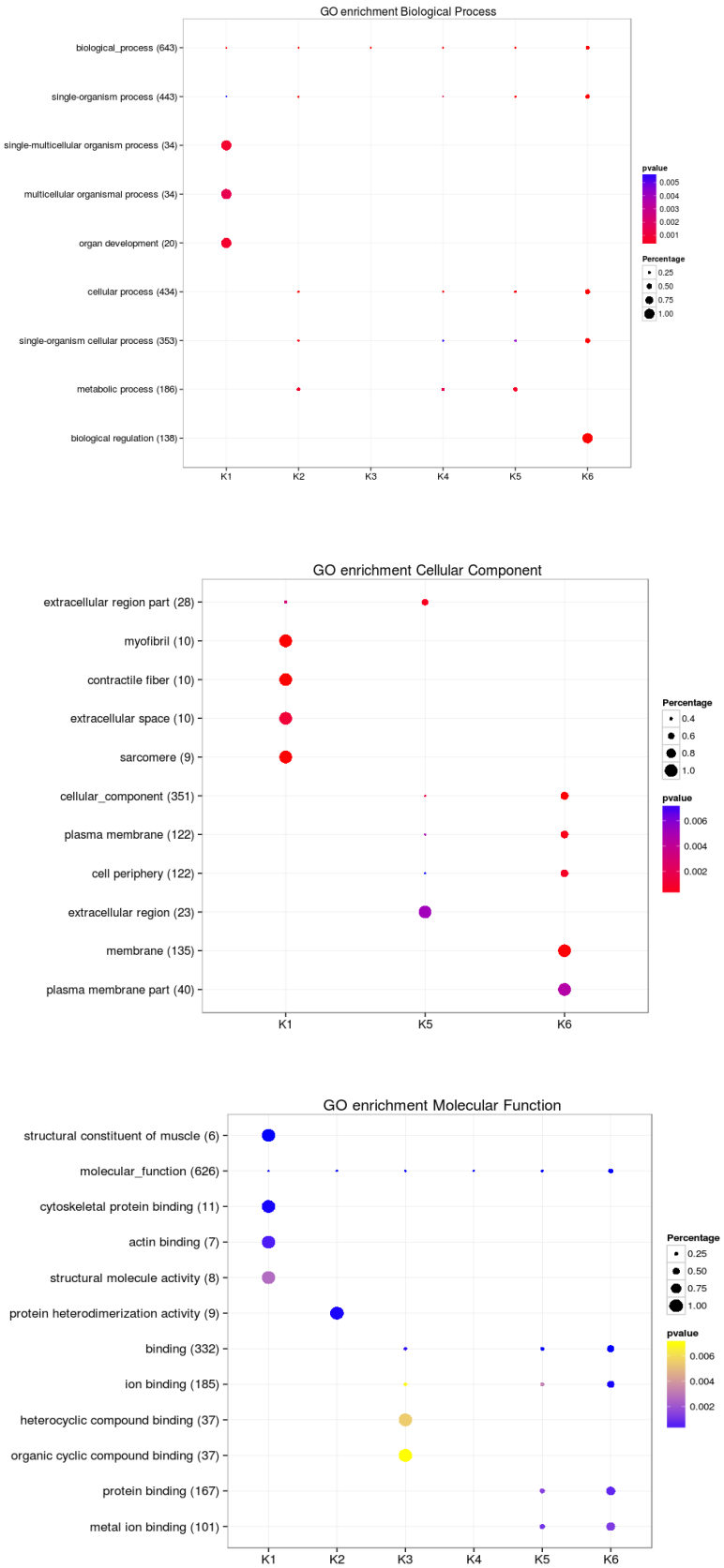


Figure 6.9. BP (top), CC (middle) and MF(bottom) mapping for clustered genes for data 1 that were identified by BCC.

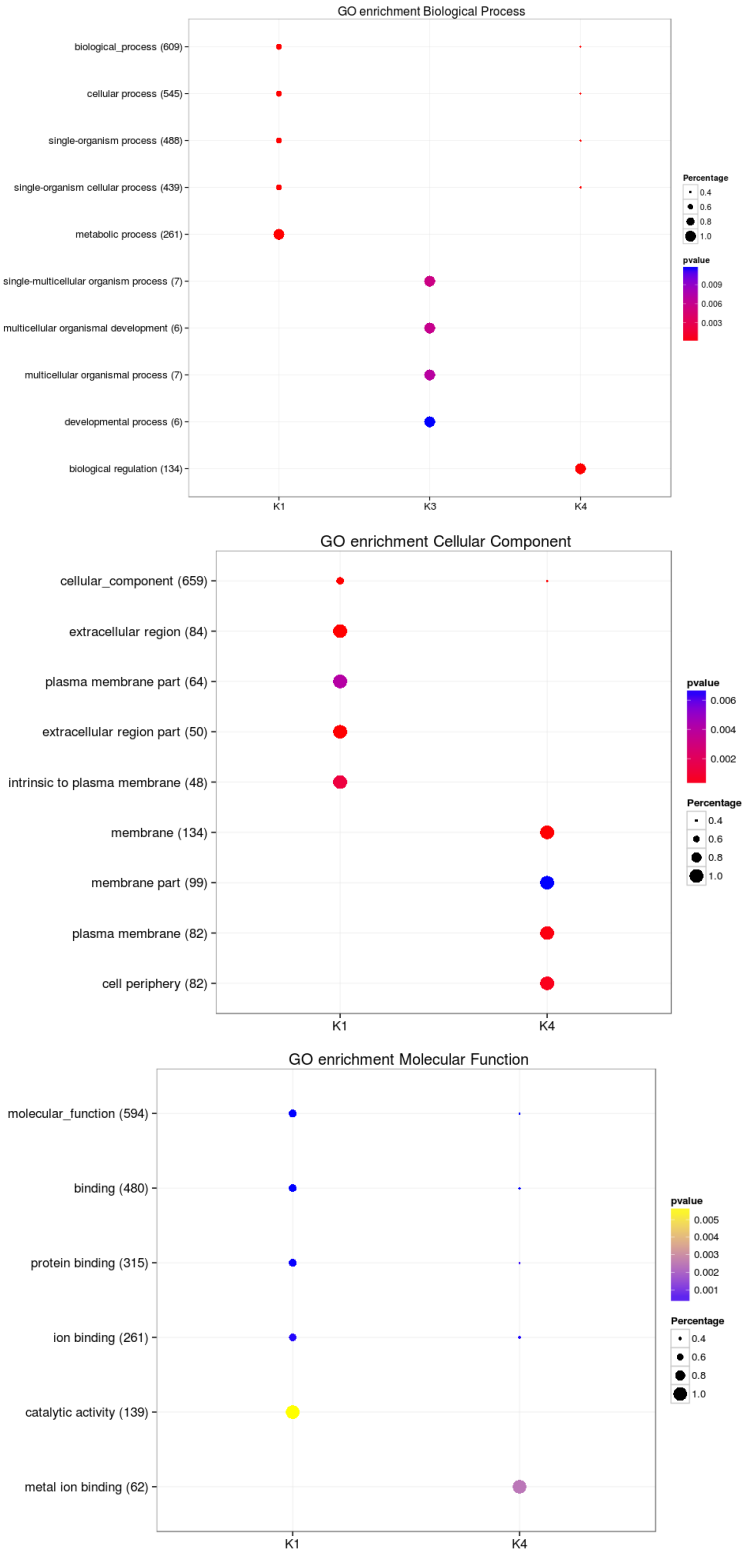


Figure 6.10. BP (top), CC (middle) and MF(bottom) mapping for clustered genes for data 1 that were identified by MDI.

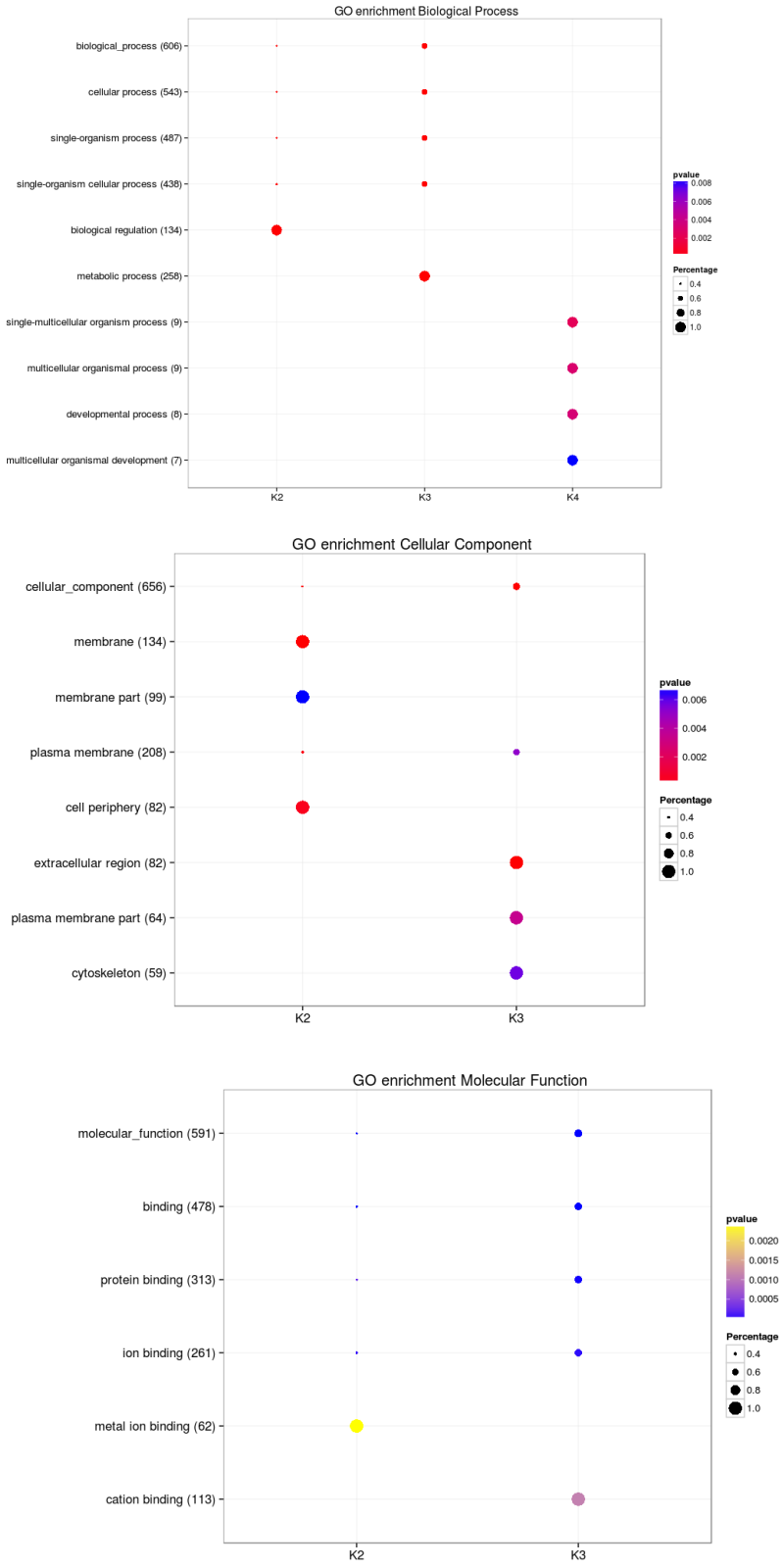


Figure 6.11. BP (top), CC (middle) and MF(bottom) mapping for clustered genes for data 1 that were identified by HIM.

Figures S6.16:S6.21 show the molecular function, biological process and cellular components mapping of clustered genes from data 2-4 using the three

algorithms. This has a similar pattern to the one illustrated above and has also shown some conserved clusters.

6.3.2.1. Common genes

I identified 134 genes common across all four cancer sets from the top 10,000 differentially-expressed genes. Table 6.8 shows the percentage of genes that have similar cluster label among all data and within cancer specific type. The results are with expectation where the percentage of common genes sharing the same label is doubled to the specific type.

Table 6.8 Percentage of genes that share the same cluster label

Methods	All data	prostate	breast
BCC	24.6%	55.9%	55.97%
MDI	34.32%	57.46%	70.14%
HIM	25.37%	63.43%	57.46%

6.4. Conclusion

This chapter has presented an evaluation of the proposed integration method against others using some evaluation methods. In this chapter all proposed methods in previous chapters have been used for two real applications. I showed that the mean pattern model implemented in a hierarchical structure was comparable to, and sometimes outperforms benchmark algorithms for real data. One issue remains unsolved is the cluster number estimation, as there is no common cluster number were found for the real data. This would need further investigation.

Chapter 7

The Extensions of MSG for Cross-species Studies

Abstract

Cross-species studies using microarray gene expressions have helped in discovering genetic diversity among different species, with results that are fundamental to comparative genomics. Various approaches have been used for cross-species studies, such as homogeneity tests and cluster analysis. A homogeneity test provides a homogeneity significance ranking for each gene pair, whilst cluster analysis looks at the discovery of co-expressed meta-genes. I propose a unified method to extract both homogeneous and heterogeneous expression patterns across species. The homogeneous genes stand for genes which have the same differential expression pattern across two species. For instance, a homogeneous gene may show similar differential expression magnitude and the same differential expression direction in both species. The heterogeneous genes stand for genes which show different differential expression pattern across two species. For instance, a heterogeneous gene may show up regulation in one species but down regulation in the other species. Or one heterogeneous gene may show differential expression in one species but not the other species. The number of homogeneous genes can be used to interpret the similarity between two species. For instance, in drug design, homogeneous genes may be used for designing a drug which can be used for two similar diseases, such as cancer in relation with hormone malfunction. But the heterogeneous genes are used to interpret why two species respond to a stress differently. Such a gene is also useful for developing personalised drug for fighting a disease. The basic idea being to model the sum and difference of expressions across species using multi-scale

Gaussians which reveal information about homogeneous and heterogeneous expression patterns respectively. However, the difficulty of gene expression dynamics across stages being unable currently to be identified effectively motivated me to consider a further extension of MSG to develop a new method, whose basic idea is to model cross-species differential expressions of all stages using a multi-scale Gaussian mixture. This multi-scale Gaussian mixture is easily able to explore dynamic differential expression pattern across all stages.

7.1. Introduction

A cross-species study compares multiple data sets from biological/medical experiments to reveal how genes are conserved among distantly related species [44]. For instance, it is assumed that the primary structure of an orthologous gene will be conserved when species evolve, even if their evolutionary distance is large [322]. Biological hypotheses in one species on conserved genes can therefore be tested in another species if the two species are related. For instance, clinical trials for disease intervention are conducted on mice prior to being tested on humans, and there exist many such human-mice species comparison studies [323-327]. However, due to factors such as sample variation and technique resolution that pose challenges in cross-species studies, gene expression levels vary significantly across species [44, 322, 328-330].

Two types of quantitative analytical approach are typically used for cross-species studies. The first approach tests homogeneity of gene expressions across species. For example, a homogeneity test based on correlation was used to determine the genetic components of alcohol consumption between human and rats [331]. Those genes with large correlation coefficients across the two species were classified as homogeneous genes. As an extension to correlation analysis, which captures only second order and linear information, some studies looked at higher order statistics such as mutual information to detect early stress responses in rodent models of lung injury across species [332]. Evolutionary conservation is a complex implementation of correlation analysis for cross-species studies [333-335]. Both correlation analysis and

mutual information methods rely on large sample sizes. When examining homogeneity across more than two species, the Fisher combined probability test [336] can be used when the species number is sufficiently large. It combines *p* values derived from significance analysis carried out separately in each species to deliver a combined *p* value of the dependence across species [337-339]. The Fisher combined probability test is unreliable when some *p* values are extremely small [340].

The second approach uses multivariate analysis such as unsupervised and supervised learning algorithms for cross-species studies. Supervised learning algorithms are used to examine whether a pattern conserved in one species can be a predictive factor for the other species. The algorithms used for supervised cross-species studies include artificial neural network [341], linear discriminant analysis [145] and k-nearest neighbours [145]. For instance, the artificial neural network algorithm was used for identifying conserved and divergent transcriptional modules across species [342], linear discriminant analysis was used for lung injury biomarker detection across multiple species [343], while the k-nearest neighbour algorithm was used for probe sequence identification [344].

As phenotypic data are not always available, or are difficult to acquire, unsupervised learning algorithms are more often used in real applications. For example, non-negative matrix factorisation (NMF) [57, 58] has been used to analyse common meta-genes across two species [60-62]. NMF decomposes a gene expression matrix into a meta-gene expression matrix and a coefficient matrix of individual gene contributions to the meta-genes. The expression level of a gene is then a linear combination of the expression levels of all meta-genes. Based on the magnitude of coefficients, one can quantify the

relationship of each gene to a meta-gene, giving a partition or cluster of the data. Common meta-genes can then be identified across species by inspecting the separate NMF models. Cluster analysis is another set of important unsupervised learning algorithms which partition data explicitly where each cluster corresponds to a meta-gene.

Clustering algorithms including the k-means algorithm, mixture models, and self-organizing maps [43] have also been widely used for cross-species studies [64, 66, 69, 345]. During cluster analysis, each data point is assigned to a cluster and a cluster represents a meta-gene. When using cluster analysis for cross-species, one can model a gene expression matrix for each species separately [68, 346, 347], or model a combined gene expression matrix from multiple species [322, 345, 348]. Separate clustering suits any data size as each species is modelled individually to generate a cluster model through an unsupervised learning process. Co-expressed patterns across species are then extracted after individual cluster models have been generated for each species. Cross-species clustering can also be carried out using a combined expression matrix where gene expression matrices are normalized separately and then merged into one matrix. Clusters of genes from multiple species directly show the co-expression of genes across species. However, this approach requires all expression matrices to contain the same number of samples.

A major distinction between the two quantitative approaches is the focus on co-expressed genes versus the focus on co-expressed meta-genes. A homogeneity test can detect exactly which subset of genes is co-expressed across species and measure the significance of co-expressions for gene ranking. Cluster analysis does not produce statistics of significance but summarises information based on meta-genes. One critical issue with cluster

analysis is that cluster structures can often be badly estimated when data noise is large, leading to false identifications of co-expressed genes.

7.2. MSG for cross-species study of regulation patterns

7.2.1. Motivation

A major common drawback with the methods discussed is that they cannot be used to detect heterogeneous patterns, such as when a gene shows up regulation in one species but down regulation in the other species. Clearly, a homogeneity test will attribute such a gene with a low significance for homogeneity; similarly, cluster analysis only identifies genes with similar expressions. Therefore, researchers are motivated to consider a method which detects heterogeneous as well as homogeneous differential expression patterns (DEPs) across species. This is done by modelling the sum and difference of DEs across two species using multi-scale Gaussians (MSG). Simulated and real data are used to illustrate the detection of both homogeneous and heterogeneous DEPs using the proposed method.

7.2.2. Proposed method

7.2.2.1. MSG model

This is similar to the main model discussed in Chapter 2, section 2.3

7.2.2.2. Using MSG

To use multi-scale Gaussian to model differential expressions (DEs) for a microarray expression data, I denoted a DE matrix by $X = \{ \mathbf{x}_n \}_{n=1}^{N} \in \Re^d$. It had

N rows of genes or probe sets. I used the biological significance [349, 350] of the expression data and denoted it by $Z = \mathrm{E}(X) \in \Re$.

7.2.3 Cross-species MSG (CSMSG) for differential expression pattern discovery

I am interested in discovering two kinds of cross-species differential expression patterns (DEPs). The first refers to the subset of genes with similar DE direction as well as magnitude across species - homogeneous DEP (DEP0). The second refers to a subset of genes with opposite DE directions across species - heterogeneous DEP (DEP1). DEP0s and DEP1s inform how a gene demonstrates similar or different responses to stress across species.

I denoted two DE matrices for two species by $X = \{ \mathbf{x}_n \}_{n=1}^{N} \in \Re^{d_x}$ and $Y = \{ \mathbf{y}_n \}_{n=1}^{N} \in \Re^{d_y}$. Both matrices had N rows of genes or probe sets. I then derived a co-differential expression vector $Z = \mathrm{E}(X) \otimes \mathrm{E}(Y)$, where $\otimes = (+, -)$. The sum and difference of DEs were then used for revealing DEP0 and DEP1 respectively across species. The MSG posterior probability of a homogeneous (heterogeneous) DEG would then lie in the tail regions of the density function of the sum (difference) of DEs.

7.2.4. Experimental design

Simulated data were designed to evaluate the method of identifying DEP0s and DEP1s. A thousand (1000) genes were simulated across two species, with ten control and ten test samples for each species. The control and test samples of non-DEGs were random samples of a normal distribution of mean ten and a

varying standard deviation. The standard deviation values were 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3 and 1.4, and the number of samples was five and ten. The test samples of DEGs were random samples of a normal distribution with mean ten and plus (minus) one unit for up (down) regulated genes.

Table 7.1 gives the designated number of genes for each combination of null/up-regulated/down-regulated behaviour across the two species. The first 800 were non-DEGs and the next 200 were DEGs (homogeneous or heterogeneous). Therefore a design vector (actually a label vector) was composed of 800 zeroes and 200 ones. The design vector was denoted by ***t***. For testing homogeneous DEGs, $t_{1:900} = 0$, $t_{951:1000} = 0$ and $t_{901:950} = 1$. For testing heterogeneous DEGs, $t_{1:950} = 0$ and $t_{951:1000} = 1$.

Table 7.1. Experimental design for the simulated data with combinations of non-DEGs (Null), up-regulated DEGs (Up) and down-regulated DEGs (Down) across the two species.

	800	25	25	25	25	25	25	25	25
Species 1	Null	Null	Null	Up	Down	Up	Down	Up	Down
Species 2	Null	Up	Down	Null	Null	Up	Down	Down	Up

To evaluate how accurate an algorithm would be in identifying hhDEGs, AUR was used. AUR – which stands for 'area under ROC' curve (where ROC stands for 'receiver operating characteristic') [143, 144] – is typically used as a robustness measure in two-class classification analysis tasks. A ROC curve describes how the sensitivity (also called the true positive rate) varies, along with the false positive rate (also called the false alarm rate). Varying the cutting point for classification between non-DEGs and DEGs gives the multiple pairs of false positive rates and sensitivities on a ROC curve. A robust classifier is characterized by a ROC curve which is close to the top-left corner, or, equivalently, a large AUR. In order to demonstrate the power of MSG, three

modified ***t***-test algorithms were used for comparison: these were eBayes [54], SAMr [351], and Cyber-T [35].

A data set was also downloaded from the Gene Expression Omnibus (GEO), accession number GSE44337. The data set originated from a study on conserved gene differentiation in aggressive B lymphomas across human species (human diffuse large B cell) and mouse species (B6 iMyc). The human species expression data was generated using the GPL570 platform with 54,675 probe sets, and was composed of three samples of the wide type and nine tumour samples. The mouse species was generated using the GPL1261 platform with 45,101 probe sets, and was composed of three wide type samples and seven tumour samples. As both data sets had unmatched sample numbers between wide type and tumour, we used a one-to-one sample pairing approach, i.e., each wide type sample was paired with a tumour sample to generate a DE. This generated a 27-dimension DE vector for each gene for the human species and a 21-dimension DE vector for each gene for the mouse species.

Because different data sets in a cross-species study may use different platforms, we matched probe sets to gene symbols. For gene symbols with more than one probe set, we selected the probe set with the maximum variance. For this real data, we also applied modified ***t***-test algorithms for comparison.

7.2.5 Results of *CSMSG*

7.2.5.1 Simulated data

7.2.5.1.1. Detecting DEP0

The AUC measures for the DEP0 detection simulations for the data set with ten replicates and variable noise in data were illustrated in **Figure 7.1**. It can be seen that MSG achieved the highest AUC measure among four algorithms, and also that the performance of all four algorithms depended on noise in data. Importantly, when noise level in data was increased, the difference of AUC measure among four algorithms was enlarged. For instance when σ was 1.4, the median AUC of MSG was greater than 0.95, while that of eBayes and SAMr was lower than 0.9. The difference of median AUC was roughly larger than 0.05.

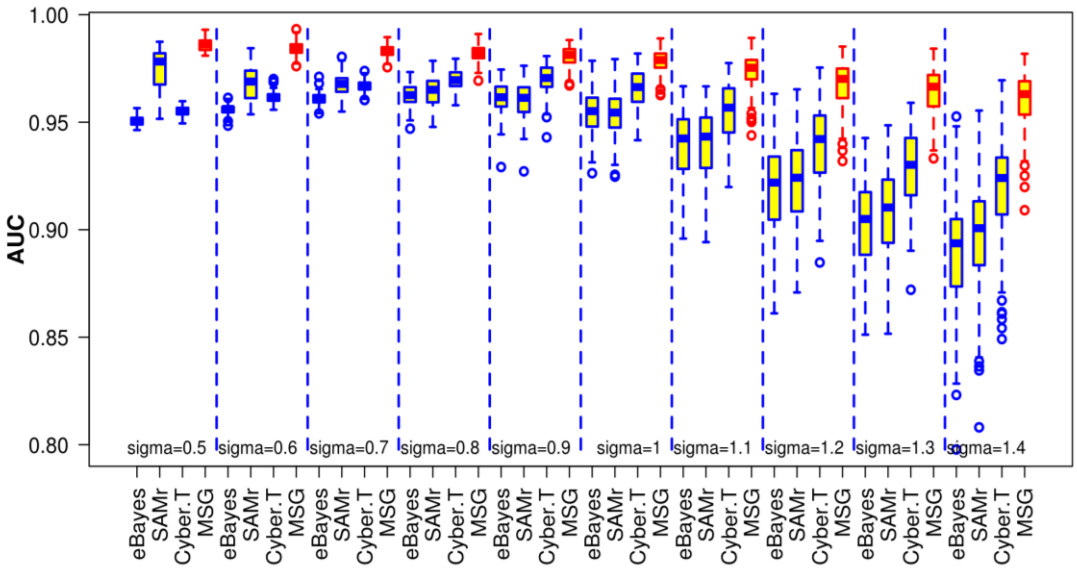


Figure 7.1. AUC measures for detecting DEP0 for sample size ten.

7.2.5.1.2. Detecting DEP1

The AUC measures for detecting DEP1s when sample size was ten and noise level varied, are shown in Figure **7.2**, where the same pattern can be seen.

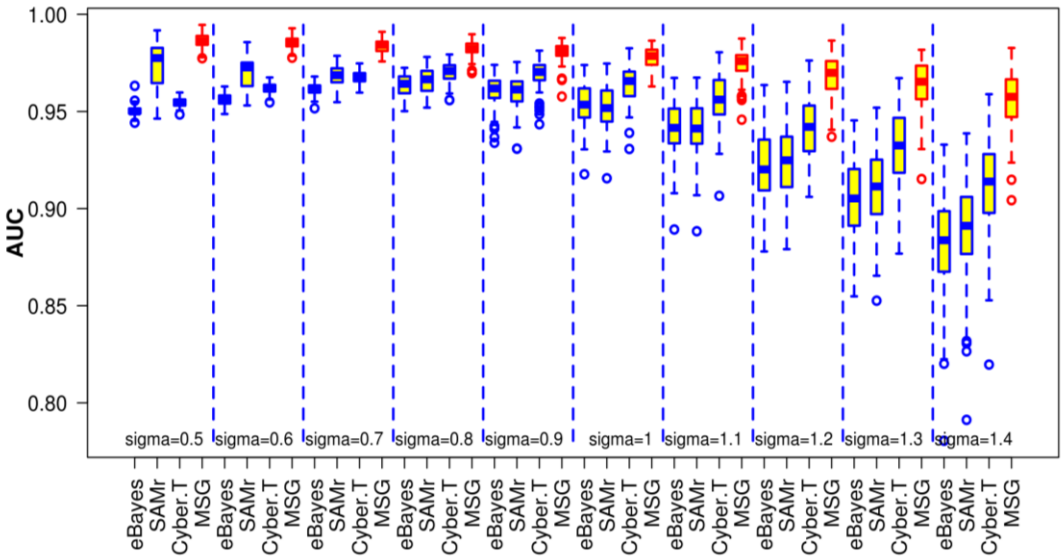


Figure 7.2. AUC measures for detecting DEP1 for sample size ten.

Figure 7.3 shows the AUC measures for detecting DEP0 for sample size five, while the AUC measures for detecting DEP1 for sample size five are shown in **Figure 7.4**. It can be seen that the difference of median AUC between MSG and three modified ***t***-test algorithms was further enlarged compared with that shown in Figures 7.1 and 7.2.

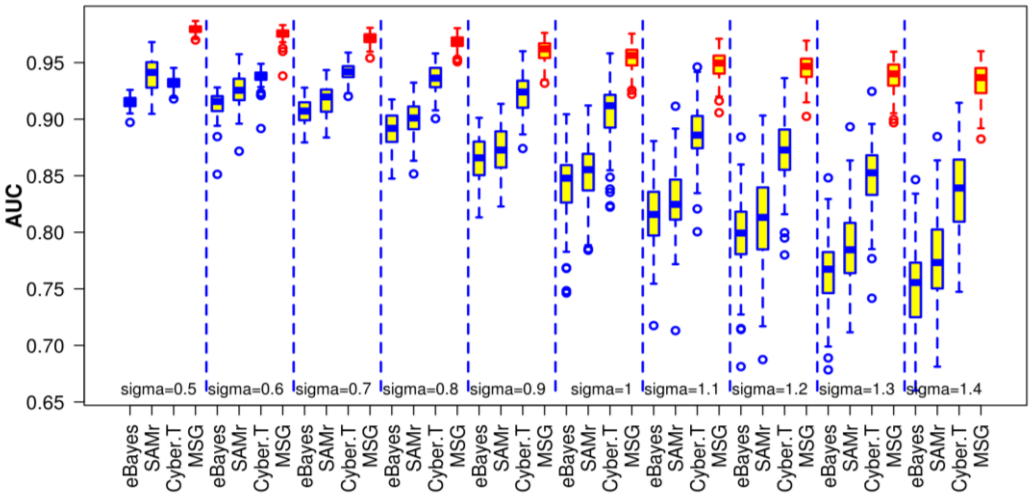


Figure 7.3. **AUC measures for detecting DEP0 for sample size five.**

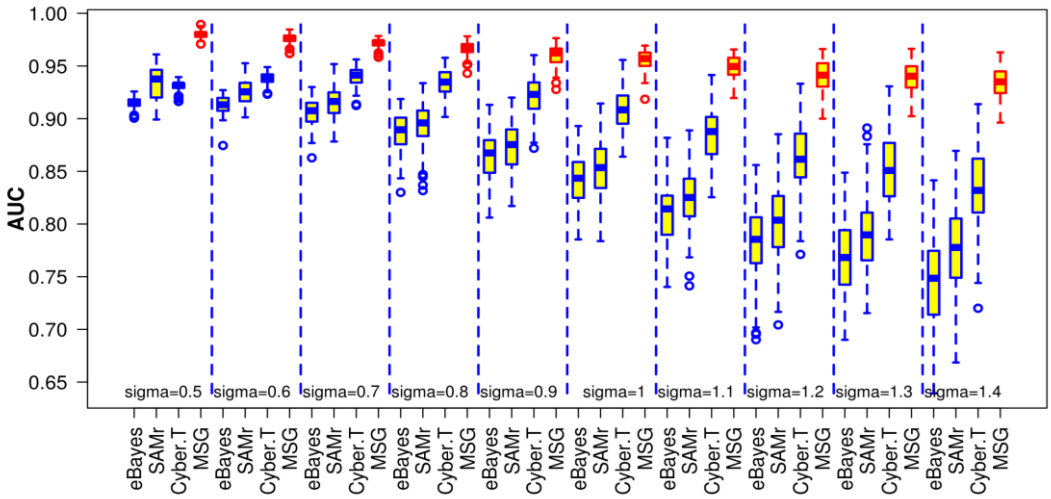


Figure 7.4. AUC measures for detecting DEP1 for sample size five

7.2.5.2. Real data

We extracted 27 DEs for each gene in the human data set and 21 DEs for the mouse data set. From a probe set - gene symbol mapping, 13,992 probe sets were selected for the cross-species study. Using 0.95 as the critical posterior probability, MSG identified 546 homogeneous DEGs and 309 heterogeneous DEGs. Using critical *p* value 0.05, eBayes, SAMr and Cyber-T identified far more DEGs compared with MSG, so that altogether 272 homogeneous DEGs and 218 heterogeneous DEGs were agreed by all algorithms. **Figure 7.5** is a Venn diagram showing these four algorithms for two types of DEGs between two species.

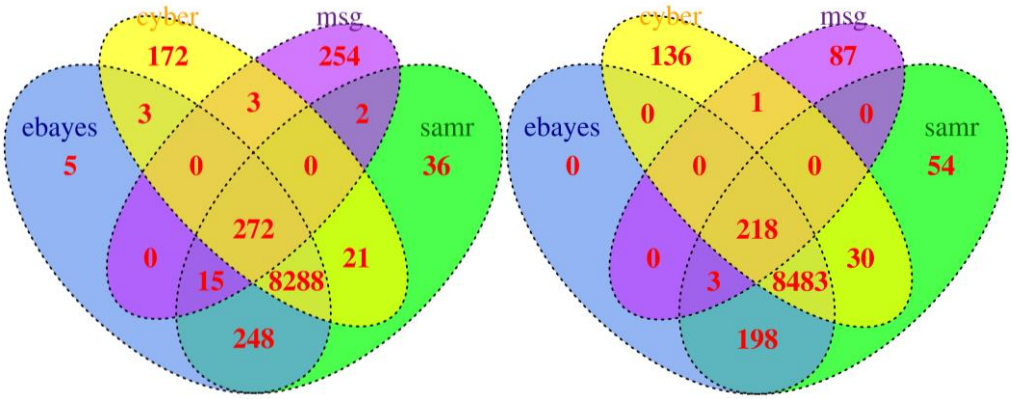


Figure 7.5 Venn diagram using four algorithms for detecting two types of DEGs. The left panel shows the identifying of homogeneous DEGs between two species, and the right panel shows the identifying of heterogeneous DEGs**.**

Figure 7.6 shows the top ten homogeneous DEGs identified by MSG, where six genes showed homogeneous up-regulation in both species and the rest showed homogeneous down-regulation in both species. These top ten genes were then mapped to the Gene Ontology Biological Processes. It was found that 74 were identical among 87 biological processes for the human species and 74 biological processes for the mice species. This shows that the biological processes are well conserved for these homogeneous DEGs. **Figures S7.1 – S7.3** show the top ten homogeneous DEGs identified by eBayes, SAMr and Cyber-T respectively. It can be seen that there was some confusion with regard to homogeneity. Some of the homogeneous DEGs predicted by eBayes were weak. Some of the homogeneous DEGs predicted by SAMr were close to zero DE, while some of the homogeneous DEGs predicted by Cyber-T demonstrated larger variance.

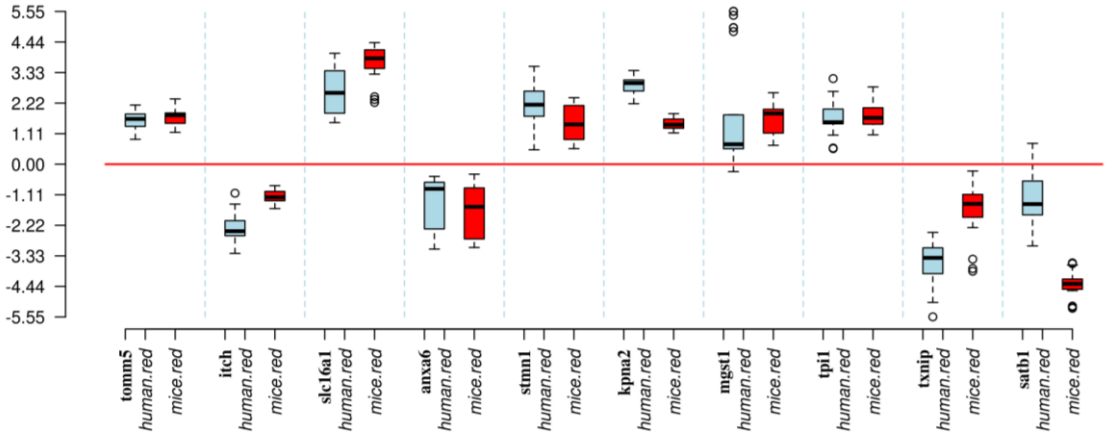


Figure 7.6. Expressions of the top ten homogeneous genes identified by MSG across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DEs. The horizontal line in red represents zero DE.

Figure 7.7 shows the top ten heterogeneous DEGs identified by MSG. When mapping these top ten genes to Gene ontology biological processes, it was found that 68 were identical among 73 biological processes for the human species and 77 biological processes for the mice species. **Figures S7.4 – S7.6** show the predicted heterogeneous DEGs using the modified ***t***-test algorithms. Though most of them were fine, some demonstrated unreasonable patterns; for instance, some heterogeneous DEGs predicted by eBayes showed median DE close to zero DE.

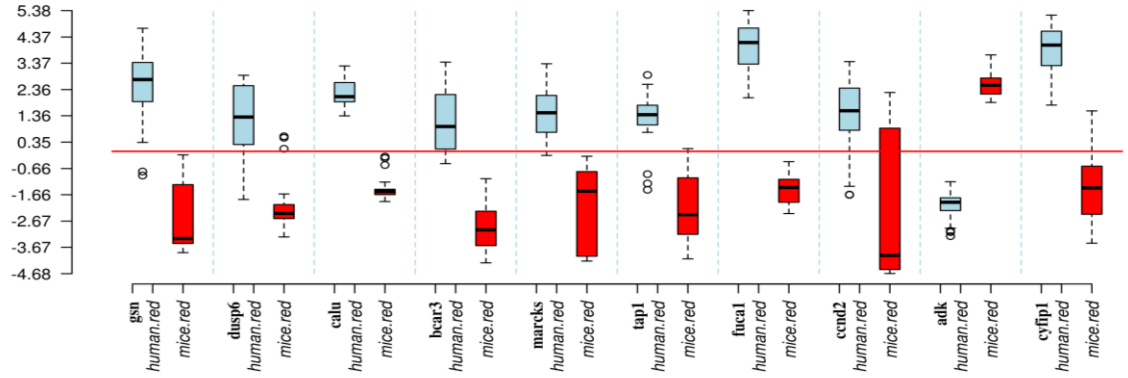


Figure 7.7. Expressions of the top ten heterogeneous genes identified by MSG across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DE. The horizontal line in red represents zero DE.

Table 7.2 summarizes the top ten homogeneous and heterogeneous DEGs across the human and mice species. Among them, several have been tested both in human and mice. For instance, TOMM5 has been tested in human and mice tissue for examining the histopathology value and mice used as the baseline assay for high-throughput phenotyping [352]. CALU has been examined in human tissue for using IRF5 as a tumour suppressor in splenic marginal-zone lymphoma [353]. BCAR3 has been studied in relation to B lymphoma in both mice and human tissue [354]. ANXA6 was studied in human tissue for lymphoma [355], while STMN1 has been studied in human tissue for its anti-cancer activity [356].

Table 7.2. Top ten homogeneous and heterogeneous differentially co-expressed genes.

Homogeneous			Heterogeneous		
Symbol	FC_{human}	FC_{mice}	Symbol	FC_{human}	FC_{mice}
tomm5	1.61125	1.721	sn	2.489049	-2.4431
Itch	-2.35208	-1.169	dusp6	1.269211	-2.0094
slc16a1	2.602503	3.691961	calu	2.225852	-1.41457
anxa6	-1.40117	-1.71861	bcar3	1.141033	-2.92931
stmn1	2.127447	1.50151	marcks	1.48766	-1.94119
kpna2	2.883727	1.477786	tap1	1.262543	-2.22466
mgst1	1.695297	1.64594	fuca1	3.98821	-1.40672
tpi1	1.719474	1.76643	ccnd2	1.412139	-2.45862
Txnip	-3.59066	-1.64936	adk	-2.09514	2.562235
satb1	-1.25254	-4.37684	cyfip1	3.805123	-1.39123

Note: FC = Fold Change.

Finally I used the R package clusterProfiler to map clustered homogeneous DEGs and heterogeneous DEGs identified by MSG to GO molecular functions. Before the mapping, mixture models implemented in Mclust (an R package) were used to cluster DEGs.

The mapping of clustered homogeneous DEGs for the human species is shown in **Figure 7.8.** It can be seen that most clusters were conserved with certain molecular functions. **Figures S7.7 – S7.9** show the mappings of clustered homogeneous DEGs for the mice species, clustered heterogeneous DEGs for the human species, and clustered heterogeneous DEGs for the mice species that show very conserved molecular functions for clustered DEGs. Two species also show a large degree of diversity of mapping pattern; i.e., DEGs from different species tend to be clustered in different ways.

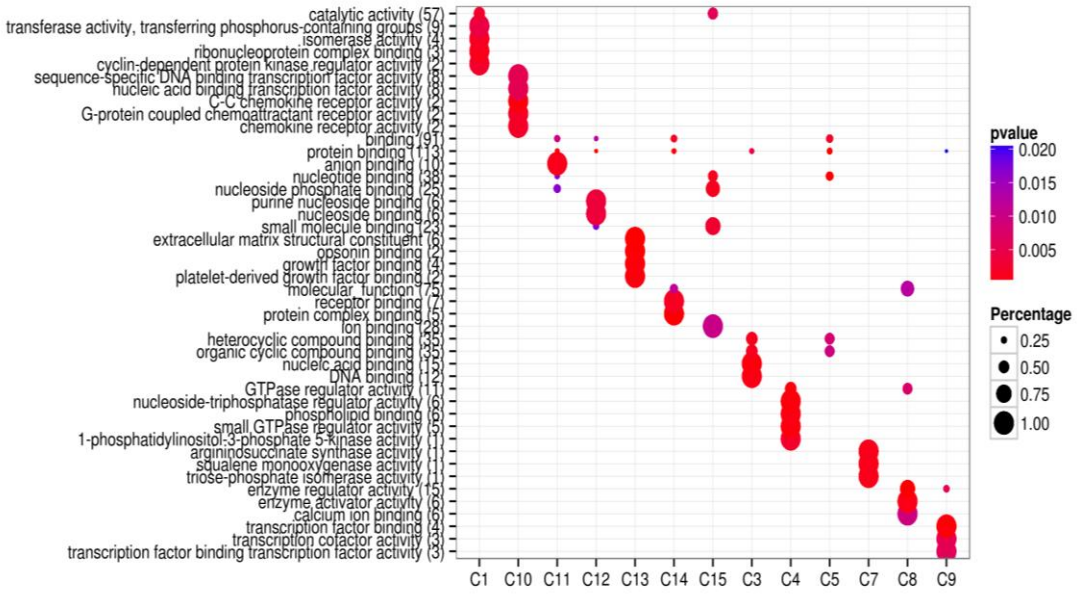


Figure 7.8. Molecular function mapping for clustered homogeneous DEGs identified by MSG for the human species. C1 – C15 represent clusters identified by Mclust.

7.3. MSG for gene expression dynamic in different stages across species

7.3.1. Motivation

Most cluster models search for non-overlapping centres. However when examining how genes show differential expression (DE) across species, the null hypothesis assumes no difference between the means from the data of different

experimental conditions; i.e., $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu = 0$, where μ_1 is the mean of the data of experimental condition 1, μ_2 is the mean of the data of experimental condition 2, and μ is the mean of the data of one experiment. A very small DE implies the impossibility of rejecting the null hypothesis. A large absolute value of DE implies the likelihood of rejecting the null hypothesis. Because of this, if DE data is clustered, it is of more interest to place cluster centres at the origin of a Euclidean space. In other words, the cluster centre of differentially expressed genes (DEGs) and the cluster centre of non-DEGs (NEGs) should overlap at the origin.

Therefore a multivariate cross species method is proposed, the basic idea of which is the employment of a multi-scale Gaussian mixture (MSG). MSG was used to partition a DE matrix derived from multi-stages gene expression data from two species into two clusters. One contained genes with small DE values across all stages and the other contained genes with large DE values at least at one stage. A gene which shows small DE values across all stages is certainly a gene which plays a key role for the similarity between species. A set of genes which show large DE values across all stages is not sufficient to include genes which play a temporal role of species diversity. Thus it is of interest to search for genes which contribute to temporal species diversity.

7.3.2. The property of DE

DE is commonly used for testing the null hypothesis that a gene shows no difference of expression between two experimental conditions. Studies on the robustness of biological systems have indicated that a biological system normally dispatches a small number of genes to respond to a stress [129, 130]. This means that a majority of DEs distribute densely around the origin (zero)

and their null hypotheses cannot be rejected. The other subset of DEs distributes sparsely away from the origin and their null hypotheses are rejected. In a one dimensional case, a sharp peak is seen around zero (a Gaussian distribution with a very small variance) and two long tails (a Gaussian distribution with a very large variance or a uniform distribution).

In terms of hypothesis test, we have three categories of hypotheses. First, those DEs which are around zero are statistically non-differential. Second, those DEs which are positive and away from the origin are statistically up-regulated. Third, those DEs which are negative and away from the origin are statistically down-regulated. It can be seen that all DEs are tested against the origin for the null hypothesis, $H_0 : \mu = 0$. This makes it a natural consideration to use DSG as described below.

7.3.3. Multivariate Multi-Scale Gaussian (MVMSG)

A DE matrix is denoted by $\mathcal{D} = \{ \mathbf{x}_n \}_{n=1}^{N}$, where $\mathbf{x}_n \in \Re^M$, N is the number of data points, and M is the number of samples (data dimension). A multi-scale Gaussian mixture [357, 358] is a special case of mixture models [59, 359] where all Gaussians in a mixture have similar mean vectors. A mixture model of K Gaussians is expressed by

$$f(\mathbf{x} \mid \omega) = \sum_{k=1}^{K} w_k \mathcal{G}(\mathbf{x} \mid \mu_k, \Gamma_k) \qquad (7.1)$$

where $\mu_k \in \Re^M$, $\Gamma_k \in \Re^{M \times M}$ and w_k are the centre, the covariance matrix and mixing coefficient (weight) of the k[th] Gaussian. $\mathcal{G}(\mathbf{x} \mid \mu_k, \Gamma_k)$ is the Gaussian distribution centred at μ_k with a covariance matrix Γ_k, which is defined as

$$\mathcal{G}(\mathbf{x} \mid \mu_k, \Gamma_k) = (2\pi)^{-\frac{M}{2}} \left|\Gamma_k\right|^{-\frac{M}{2}} \exp(-\Delta_k) \tag{7.2}$$

where $\Delta_k = \frac{1}{2}(\mathbf{x} - \mu_k)^t \Gamma_k^{-1}(\mathbf{x} - \mu_k)$. The parameter set of the mixture model is expressed by $\omega = (w_k, \mu_k, \Gamma_k)_{k=1}^K$. When

$$\mu_1 \approx \mu_2 \approx \cdots \approx C \tag{7.3}$$

it is then an MSG. Note that C is a constant. **Figure 7.9** illustrates an example of multiple Gaussians with zero mean and 20 variable variances (from one to 20). When we draw samples from them to form a new distribution, this is a MSG.

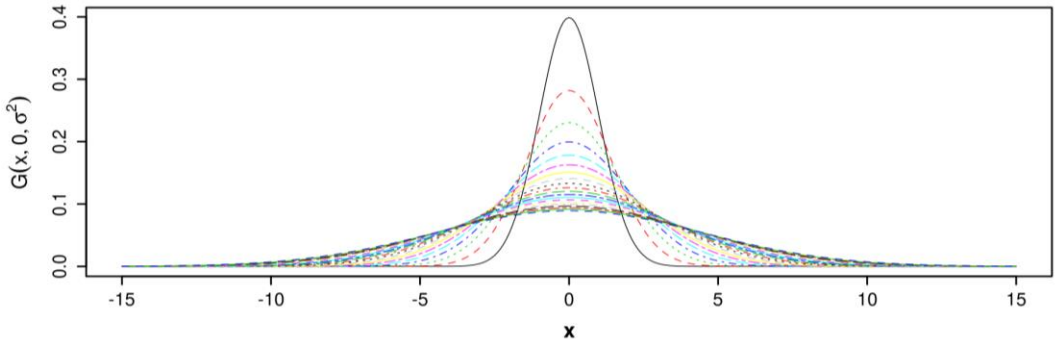


Figure 7.9. Gaussians with 20 variances

If K is two, MSG becomes a special-case dual-scale Gaussian mixture (DSG) defined below

$$f(\mathbf{x} \mid \omega) = w_0 \mathcal{G}(\mathbf{x} \mid \mu_0, \Gamma_0) + w_1 \mathcal{G}(\mathbf{x} \mid \mu_1, \Gamma_1) \tag{7.4}$$

where $\omega = \{\mu_0, \mu_1, \Gamma_0, \Gamma_1, w_0, w_1\}$ is the model parameter set and $w_0 + w_1 = 1$.

7.3.3.1. Homogeneous DSG

Homogeneous DSG (hoDSG) was introduced for independent and homogeneous variance in all dimensions

$$\Gamma_k = \mathrm{diag}(\sigma_k^2) \tag{7.5}$$

where $k \in \{1,2\}$. In this model, all replicates were assumed to have identical and independent distribution. Equation (7.2) was rewritten as

$$\mathcal{G}(\mathbf{x} \mid \mu_k, \Gamma_k) = (2\pi)^{-\frac{M}{2}} (\beta_k)^{\frac{M}{2}} \exp(-\tfrac{1}{2}\beta_k(\mathbf{x}-\mu_k)^2) \tag{7.6}$$

where $\beta_k = \sigma_k^{-2}$. The likelihood that N data points (genes) were considered as random samples of parameterised model was defined as

$$P(\mathcal{D} \mid \omega) = \prod_{n=1}^{N} f(\mathbf{x}_n \mid \omega) \tag{7.7}$$

Inverse gamma distributed priors were placed (conjugated) on the variances:

$$IG(\sigma_k^2 \mid a_k, b_k) = \frac{b_k^{a_k} \sigma_k^{-2(a_k+1)}}{\Gamma(a_k)} \exp(-b_k \sigma_k^{-2}) \tag{7.8}$$

The mixing coefficients were modelled using non-informative priors and μ_0 and μ_1 were each assigned a Gaussian prior with zero mean and small variances τ_m as shown below

$$\mu_k \sim \mathcal{G}(0, \tau_k^2 \mathbf{I}) \tag{7.9}$$

In this study, the prior means were set to zero on the basis of the modelling requirement of DE values as previously mentioned (section 7.5.2). This constraint was made to force all the hypothesis testing to have the same benchmark – i.e., zero for the null hypothesis. In the study both τ_k is 0.01. The posterior was then defined below and was approximated by a product of a Likelihood model and a prior model:

$$P(\omega|\mathcal{D},\ \alpha)=\frac{P(\mathcal{D}|\omega)P(\omega|\ \alpha)}{P(\mathcal{D}|\alpha)}\propto P(\mathcal{D}|\omega)P(\omega|\ \alpha) \tag{7.10}$$

where $\alpha=\{a_0,a_1,b_0,b_1,\tau_0,\tau_1\}$ was the hyper-parameter set. The posterior was defined below with the constants removed

$$\begin{aligned}
&P(\mathcal{D}|\omega)P(\omega|\ \alpha)\propto \\
&\prod_{n=1}^{N}\sum_{k=0}^{1}w_k(\beta_k)^{\frac{M}{2}}\exp(-\frac{1}{2}\beta_k(\mathbf{x}_n-\mu_k)^2) && \text{LK}\\
&\prod_{k=0}^{1}\beta_k^{(a_k+1)}\exp(-b_k\beta_k) && \text{VP}\\
&\coprod_{k=0}^{1}\coprod_{m=1}^{M}\exp\left(-\frac{\upsilon_k\mu_{km}^2}{2}\right) && \text{MP}
\end{aligned} \tag{7.11}$$

where LK was the likelihood, VP was the variance prior, MP was the mean vector prior and $\upsilon_k=\tau_k^{-2}$. The log-posterior was defined as

$$\begin{aligned}
&O=\log P(\mathcal{D}|\omega)P(\omega|\ \alpha)\\
&\propto\sum_{n-0}^{N}\log\sum_{k=0}^{1}w_k(\beta_k)^{\frac{M}{2}}\exp(-\frac{1}{2}\beta_k(\mathbf{x}_n-\mu_k)^2)\\
&+\sum_{k=0}^{1}\left[(a_k+1)\log\beta_k-b_k\beta_k\right]-\frac{1}{2}\sum_{k=0}^{1}\sum_{m=1}^{M}\upsilon_k\mu_{km}^2
\end{aligned} \tag{7.12}$$

The model parameters were estimated by maximising the log-posterior, giving the iterative variance update rule as

$$\sigma_k^2=\beta_k^{-1}=\frac{\sum_{n=1}^{N}\vartheta_{nk}(\mathbf{x}_n-\mu_k)^2+2b_k}{M\sum_{n=1}^{N}\vartheta_{nk}+2(a_k+1)} \tag{7.13}$$

where ϑ_{nk} is defined as:

$$\vartheta_{nk}=\frac{w_k\mathcal{G}(\mathbf{x}_n|\mu_k,\Gamma_k)}{f(\mathbf{x}_n;\omega)} \tag{7.14}$$

The iterative update rule for the mixing coefficients was defined as

$$w_k = \frac{1}{N}\sum_{n=1}^{N}\vartheta_{nk} \tag{7.15}$$

and the iterative update rule for the centres (mean vectors) was defined as

$$\mu_{km} = \frac{\beta_k \sum_{n=1}^{N}\vartheta_{nk}x_{nm}}{\upsilon_k + \beta_k \sum_{n=1}^{N}\vartheta_{nk}} = \frac{\sum_{n=1}^{N}\vartheta_{nk}x_{nm}}{\frac{\upsilon_k}{\beta_k} + \sum_{n=1}^{N}\vartheta_{nk}} = \frac{\sum_{n=1}^{N}\vartheta_{nk}x_{nm}}{\frac{\sigma_k^2}{\tau_k^2} + \sum_{n=1}^{N}\vartheta_{nk}} \tag{7.16}$$

where $m \in [1, M]$.

7.3.3.2. Heterogeneous DSG

For a heterogeneous DSG (*he*DSG), it was assumed that replicated samples displayed different variances. In this model, replicates were presumed to be independent, but with different variances. Hence, the covariance matrix was defined as

$$\Gamma_k = \begin{pmatrix} \sigma_{k1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{k2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{kM}^2 \end{pmatrix} \tag{7.17}$$

The Gaussian density for *he*DSG was defined as

$$\mathcal{G}(\mathbf{x} \mid \mu_k, \Gamma_k) = (2\pi)^{-\frac{M}{2}} \prod_{m=1}^{M}\sqrt{\beta_{km}}\,\exp\left(-\tfrac{1}{2}\beta_{km}(x_m - \mu_{km})^2\right) \tag{7.18}$$

where $k \in \{1,2\}$. The log-posterior of this model was defined as

$$\begin{aligned} O &= \log P(\mathcal{D} \mid \omega)P(\omega \mid \alpha) \\ &\propto \sum_{n=1}^{N}\log\sum_{k=0}^{1} w_k \prod_{m=1}^{M}\sqrt{\beta_{km}}\,\exp\left(-\tfrac{1}{2}\beta_{km}(x_{nk} - \mu_{km})^2\right) \\ &+ \sum_{k=0}^{1}\sum_{m=1}^{M}[(a_k + 1)\log\beta_{km} - b_k\beta_{km}] - \frac{1}{2}\sum_{k=0}^{1}\sum_{m=1}^{M}\upsilon_k\mu_{km}^2 \end{aligned} \tag{7.19}$$

The model parameters were also estimated by maximising the log-posterior. The variance update rule was defined as

$$\sigma_{km}^2 = \beta_{km}^{-1} = \frac{\sum_{n=1}^{N} \vartheta_{nk}(x_{nk} - \mu_{km})^2 + 2b_k}{\sum_{n=1}^{N} \vartheta_{nk} + 2(a_k + 1)} \quad (7.20)$$

Others were similar to *ho*DSG.

7.3.4Training DSG

The training process of DSG was initiated by assigning 0.5 to w_k, 1 to σ_{km}^2, and 0 to μ_k. The hyper-parameters were determined based on the empirical DE distribution. For both Gaussians, the variance hyper-parameters a_0 and a_1 were set to one, and b_0 was set to be the standard deviation of 90% of the smallest absolute DEs since it was assumed that the majority of DEs corresponded to those of null genes. b_1 was the 90th percentile of all DEs.

7.3.5 Prediction

After the training convergence was approached, the null density (with a small variance) and the alternative density (with a large variance) were estimated for each gene. Equation (7.1) defined the full density. The Bayes rule was then used to predict whether a gene was differentially expressed across species or not. The null probability of a gene, whose DE vector was denoted by $\mathbf{x}$, was defined as

$$\Pr_0(\mathbf{x}) = \frac{w_0 \mathcal{G}(\mathbf{x} \mid \mu_0, \Gamma_0)}{f(\mathbf{x} \mid \omega)} \quad (7.21)$$

The alternative probability was defined as

$$\Pr_1(\mathbf{x}) = \frac{w_1 \mathcal{G}(\mathbf{x} \mid \mu_1, \Gamma_1)}{f(\mathbf{x} \mid \omega)} \tag{7.22}$$

A gene was predicted as a DEG if

$$\Pr_1(\mathbf{x}) > \Pr_0(\mathbf{x}) \tag{7.23}$$

Otherwise a gene was predicted as a NEG.

7.3.6. Data

7.3.6.1. Background

Pre-implantation embryonic development (PED) goes through a number of developing stages where cells grow and divide from fertilization till the embryo forms into a blastocyst which is implanted into the tissue of the uterus. A complete PED involves two major processes; zygote genome activation and embryonic cell compaction. Zygote genome activation is a genetic transition process in which the embryonic cells degrade maternal genetic material fertilized up to this stage. Along with the degradation of maternal material, zygote DNA is activated and maternal and paternal genetic material are combined [360]. Embryonic cell compaction is a process for compacting loosely-connected embryonic cells to form a tight structure called morula, which eventually forms the blastocyst [360, 361].

PED is critical for human and animal health because defects in this early stage result in the malfunction of embryonic development, leading to future problems such as hypertension and intrauterine growth retardation [362-365]. Intensive studies on embryonic development processes have been carried out involving different species, such as Drosophila [366, 367], human and mouse [365], mouse and rat [368], and bovine [369], to name but a few. It was found that

roughly 15700 genes are expressed in mice during PED [370]. Other mammalian species expect similar genes expressed due to the relatively conserved nature of PED [360]. This promotes the studies of PED across species using biological data such as oxygen consumption [371], glucose transporters [372], amino acid transport [373], and gene expression [374].

7.3.6.2. GSE18290 data set

The raw expression data used in this section was downloaded from the Gene Expression Omnibus (GEO); its accession number was GSE18290. The data was generated for a gene expression-based cross-species PED study, using three species; human, bovine and mouse. **Table 7.3** summarizes the data for these three species in this data with six stages. The stages were one-cell, two-cell, four-cell, eight-cell, morula and blastocyst.

Table 7.3 Details of the data downloaded from GEO for the three species used in the analysis.

	Human	**Bovine**	**Mouse**
Platform	GPL570	GPL2112	GPL339
Probe sets	54675	24128	22690
Replicates	3	2	3

7.3.6.3. Common gene extraction

These three species were experimented at different platforms and the mapping from probe set IDs to gene symbols was therefore different. The platforms used by these three species were compared first to find common gene symbols. Among multiple probe sets which were mapped to one gene symbol, the one probe set that was selected had the largest variance in expressions. This ended up with 1180 genes which were common to all three species.

7.3.6.4. Data organization

These three species were organized as three raw expression matrices, each with 1180 rows for 1180 genes and a number of columns for replicates at six stages. The expressions of six stages were sequentially ordered from one-cell to blastocyst in columns. The main focus was to study how the human species was similar or different from the bovine and mouse species. For this purpose, we built two models: the human species versus the bovine species (referred to as the bovine model) and the human species versus the mouse species (referred to as the mouse model). What we wanted to explore was a subset of genes which were similar between three species and a subset of genes which showed significantly differential expression between the human species and the bovine/mouse species. Two DE matrices were formed from three raw expression matrices. One matrix was for the bovine model and the other was for the mouse model. Each matrix was formed using the base two logarithm of fold change ratio between two species, i.e. between the human species and the bovine/mouse species. The DSG was used to model each of these two matrices.

7.3.7. Results of MVMSG

7.3.1.1. Comparison between *ho*DSG and *he*DSG

The first thing was to define what a DEG was in this cross-species PED study. As noted, there were two models – the bovine model and the mouse model – both of which were used to examine how the human species was different from the other two species. My wish was to examine which gene showed conserved expression for PED across three species and which gene showed significant DE for PED between human and bovine/mouse species. A gene that showed

conserved expression across three species was called a cross-species NEG or simply NEG. A gene that showed DE between the human species and other two species was called a cross-species DEG or simply DEG.

Figure 7.10 shows the Venn diagrams of detected NEGs and DEGs between the human species and other two species using *ho*DSG and *he*DSG. NEGs were identified when the posterior probability was less than 0.05. DEGs were identified when the posterior probability was greater than 0.95. Both NEGs and DEGs were identified from the bovine as well as the mouse model. More predictions were made by *he*DSG and all the predictions made by *ho*DSG were covered by the predictions made by *he*DSG.

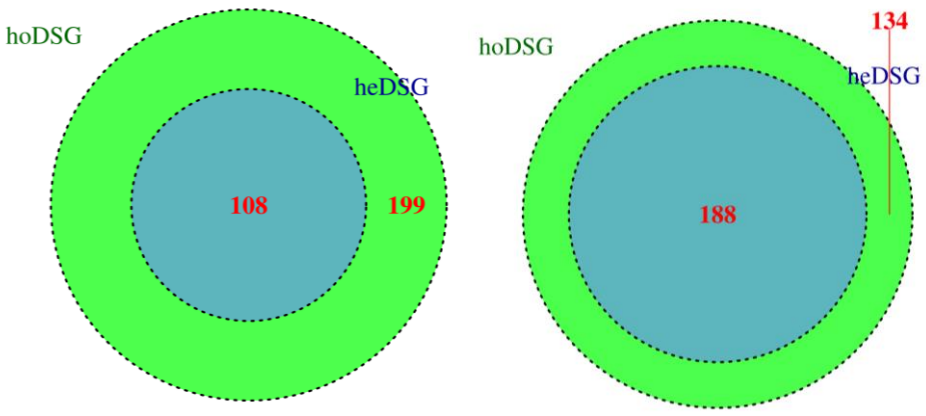


Figure 7.10. Venn diagrams for comparing *ho*DSG and *he*DSG for DEGs (right panel) and non-DEGs (left panel), between human and other two species.

7. 3.1.2. Comparison between *ho*DSG and modified *t*-test

Three modified *t*-test algorithms were used for this cross-species PED study. These algorithms were eBayes [54, 116], SAMr [351], and Cyber-T [35]. **Figure 7.11** shows the Venn diagrams for detecting NEG (left panel) and DEG (right panel) across three species using four algorithms which agreed for 48 predictions of NEGs. They showed little difference in expressions between the human species and the bovine species and between the human species and

the mouse species. They therefore played a key role for these three species to share similar PED process.

The thresholds for selection were 0.05 for the posterior probabilities generated by the *ho*DSG model and 0.05 for the *p* values generated by three modified *t*-test algorithms. A gene was predicted as NEG expressed only when it showed very insignificant difference in expression between the human and bovine species as well as between the human and mouse species. The four algorithms also agreed for 114 predictions of DEGs across three species. The identification of these DEGs also followed the same pattern when identifying NEGs. This means that these 114 DEGs showed significant DE between the human and bovine species as well as between the human and mouse species.

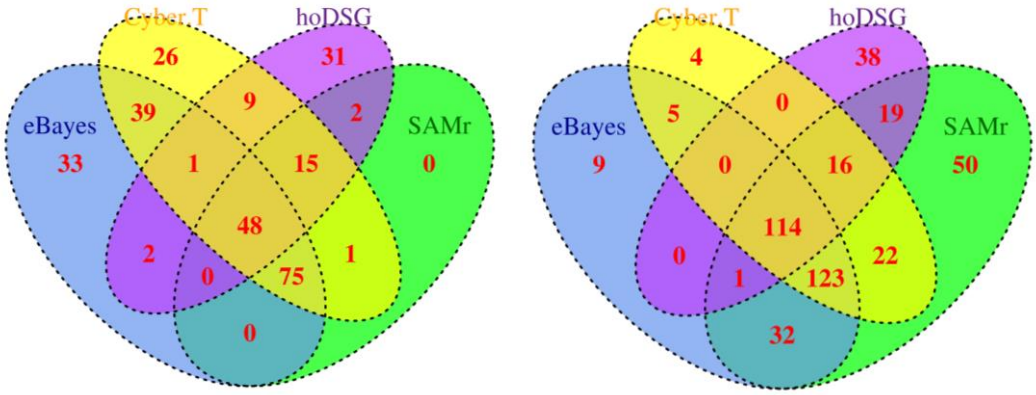


Figure 7.11. Venn diagrams for comparing four algorithms for identifying NEG (left panel) and DEG (right panel) across three species.

DSG and the modified *t*-test algorithms were developed based on different assumptions. The modified *t*-test algorithms followed the principle of *t*-test in that a gene is predicted as a differentially expressed one unless the mean difference of expressions from two conditions is sufficiently large and the pooled variance is sufficiently small. For this six-stage cross-species PED study, the aim was to ascertain the underlying cross-species gene expression dynamics.

Unfortunately, however, this kind of pattern may not be easily explored using the modified *t*-test algorithms.

It worth to mentioned that the overlaps among these method are small. The reason behind that is each method was designed to tackle an issue as explained in the previous chapters. In this thesis, I have compared them on both simulated and real data and showed that the proposed method outweighs others.

A principal component analysis was carried out using the first two principal components to draw two-dimensional volcano plots to visualize the relationship between PED stages and prediction values (posterior probabilities from *ho*DSG and *p* values from the modified *t*-test algorithms). The left panel of **Figure 7.12** shows the PCA-volcano plot of the Cyber-T model built for the bovine model. This displayed a flying bird pattern in which two wings represented two groups of DEGs – down-regulated and up-regulated ones. The *ho*DSG model did not work this way, but instead explored DEGs for all directions of PED stages. The right panel of Figure 7.12 clearly shows a cylindrical pattern of the PCA-volcano plot for the bovine model. On the top, all DEGs were sparsely distributed away from the centre of the cylinder, with different directions indicating DEs at different PED stages.

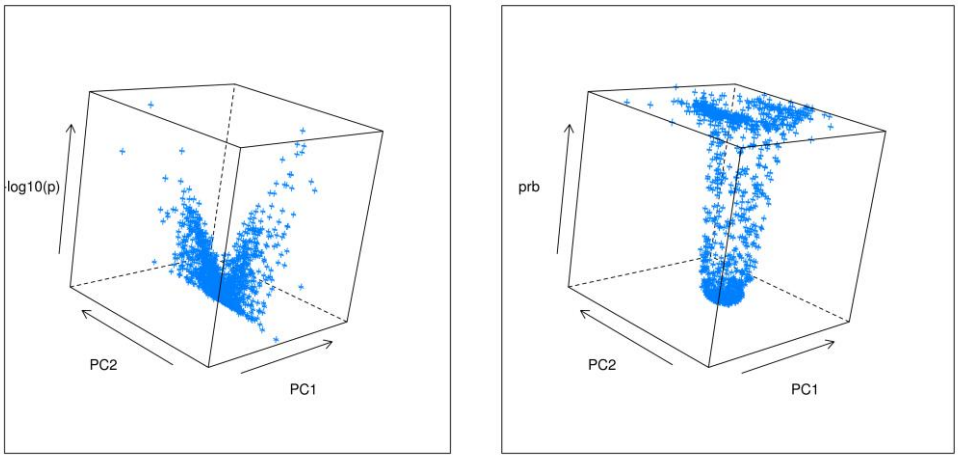


Figure 7.12. 2D volcano visualization for the Cyber-T model (left panel) and the *ho*DSG model (right panel) built for human-bovine cross-species study. The Z-axis of the Cyber-T volcano plot uses negative base ten-logarithm of *p* values, but the Z-axis of the *ho*DSG volcano plot uses posterior probabilities. PC1 and PC2 are the first two principal components derived from principal component analysis.

Figure 7.13 shows the top ten cross-species DEGs. Nine of them showed down-regulation in the human species compared with the other two species and only one showed up-regulation in the human species. **Figure 7.14** shows the cross-species NEGs. In both cases, all showed very similar expressions.

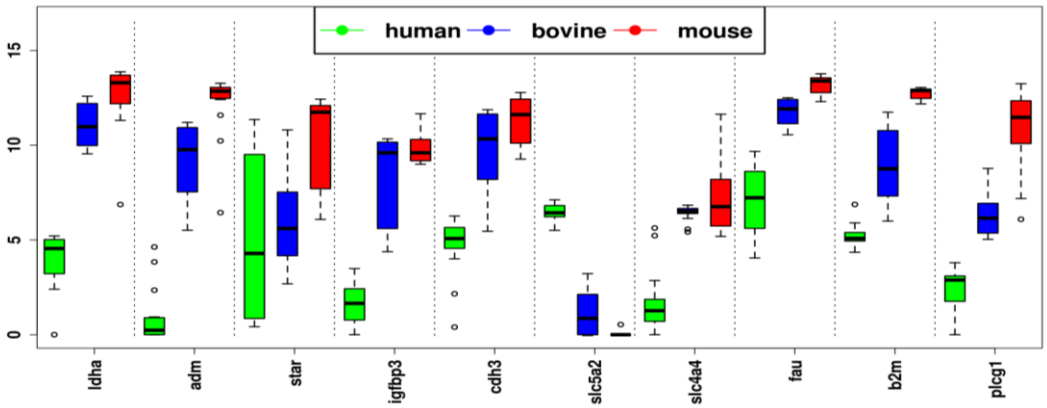


Figure 7.13. Top ten cross-species DEGs detected by *ho*DSG.

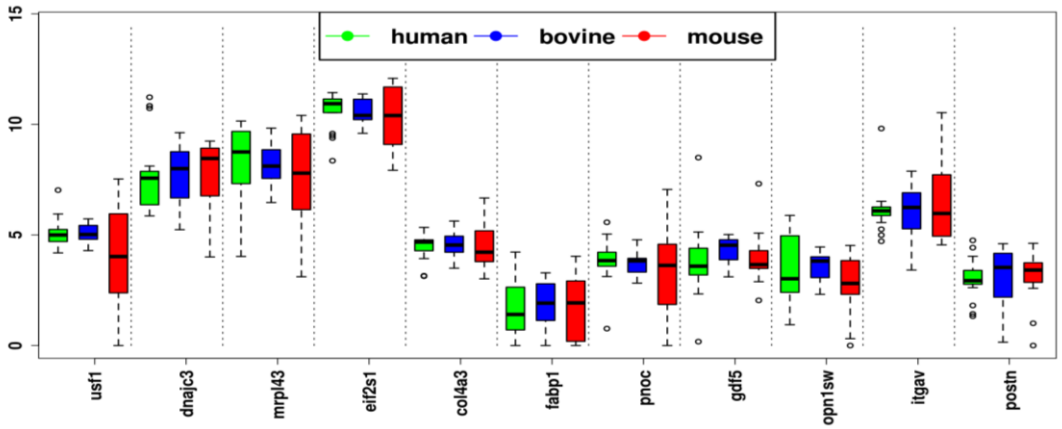


Figure 7.14. Top ten cross-species NEGs detected by *ho*DSG

The *ho*DSG model was also used to explore cross-species DE dynamics across PED stages. **Figure 7.15** shows the DEGs detected by *ho*DSG, demonstrating highest level of change in the one-cell stage of PED for both the bovine and the mouse model. This pattern can also be explored for the other five stages of PED.

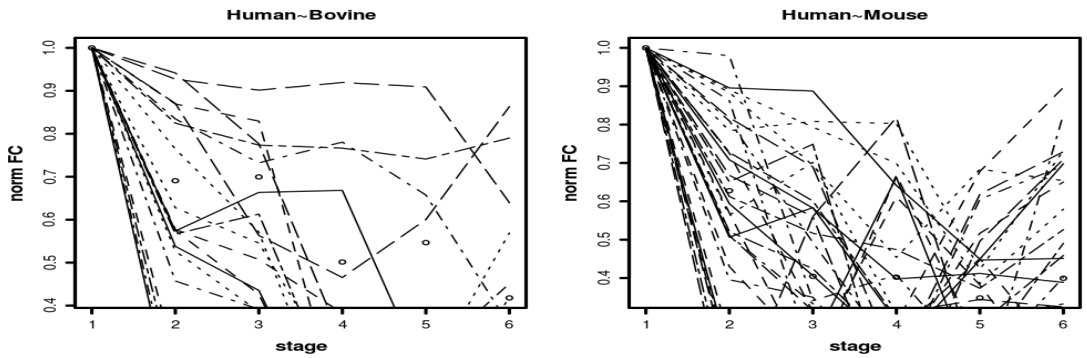


Figure 7.15. Cross-species DEGs at one cell stage of PED detected by *ho*DSG, where the vertical axes stand for normalized fold change.

An arbitrary fold change threshold (6) was used to select genes which showed cross-species DE at a specified PED stage, and was examined to see how *ho*DSG and three modified *t*-test algorithms detected them. **Table 7.4** shows the fold change values for six genes in six stages, where the detected genes all showed large absolute fold change values at the blastocyst PED stage, but smaller absolute fold change values at the other five PED stages.

Table 7.4. Six genes were selected with absolute fold change values larger than 6 in the stage of blastocyst and absolute fold change values less than 6 in other five stages. The figures are absolute fold change values for six genes across six PED stages for the human-bovine model. M stands for morula stage and B for the blastocyst PED stage

gene	1 cell	2 cell	4 cell	8 cell	M	B
cpe	2.6	0.9	0.3	1.6	5.1	6.5
slc7a5	0.4	1.0	0.6	0.3	3.7	6.2
fdps	4.2	2.6	4.0	3.4	1.1	6.7
vapb	3.5	3.7	5.6	1.3	4.4	7.1
plcg1	1.8	3.0	2.7	4.5	4.0	7.4
fgfr3	1.7	2.6	2.1	1.0	5.7	6.8

Table 7.5 shows the posterior probability derived from the *ho*DSG model and *p* values derived from three modified *t*-test models. The posterior probabilities of the *ho*DSG model showed that these genes were differentially expressed. However all three modified *t*-test models failed to recognize stage-specific cross-species DEGs. The analyses for other species at other stages showed similar patterns.

Table 7.5. The posterior probabilities of *ho*DSG and *p* values of three modified *t*-test algorithms for six genes, shown in Table 7.4.

Gene	DSG	eBayes	SAMr	Cyber-T
Cpe	1	0.97	0.91	1
slc7a5	1	0.95	0.91	1
Fdps	1	0.95	0.91	1
Vapb	1	0.38	0.84	0.73
plcg1	1	0.27	0.61	0.39
fgfr3	1	0.08	0.23	0.10

7.4. Conclusion

Here I have presented a new method for exploring homogeneous and heterogeneous DEPs across species using multi-scale Gaussian mixtures. The algorithm is motivated by the limitations in existing homogeneity tests and cluster analysis techniques for cross-species studies. Here an MSG model was used to characterise the sum and difference of DEs across species for homogeneous and heterogeneous DEP discovery respectively. I used

simulated as well as real data to illustrate the efficacy of this method in revealing both homogeneous and heterogeneous DEPs. I also presented a new method for cross-species PED study, by identifying genes which are conserved between species (non-differentially expressed across species) and which are diverse between species (differentially expressed across species) during PED.

The new model is called multivariate dual-scale Gaussian mixture, and the gene expression dynamics in PED across species can be satisfactorily explored by the algorithm. In applying this new algorithm to the data, I show that it is superior to the modified *t*-test algorithms for this purpose. The modified *t*-test algorithms are simple and have been used for cross-species, but they are used for detecting homogeneous cross-species DEGs only, and not stage-specific cross-species DEGs.

Chapter 8
Conclusion and future projects

8.1. Reflections on the thesis

This chapter reflects on the thesis as a whole and links its achievements to its objectives, which included a complete system for integration. It highlights some key improvements to the current work for future investigation.

In attempting to address some issues toward integration this thesis has successfully achieved the three main aims that were noted in Section 1.2, as follows:

- Chapter 2 emphasised the limitations of current algorithms for identifying differentially expressed genes in details with illustrative examples. It also introduced the MSG method for differential gene identification since it predicts more robustly accurately, and this method was widely investigated on both simulated and real data. However, this is the pre-processing method for the whole system: I had selected the top differential gene for integration as one of the main applications had only 2 replicates and none of the modified *t*-test methods worked well in such data.
- The *h*BI introduced in Chapter 3 allowed the use of different variance for the 2 modes in order to identify bimodal genes. This proved that using such a method enhanced bimodality prediction. However, this method is sensitive to the outliers and I used ad hoc methods to consider subgroups with only 5 % or less as outliers. This chapter met the designated objective 2 in section 1.2. It also identified the bimodal genes which are important in cancer research.

This method outweigh others as it is able to identify different types of bimodal distributions. It can be see that a clear bimodal genes was missed by all other methods and identified by hBI. It was used in my integration system as a pre-filtering method where my proposed method was based on a mean pattern model.

- A comprehensive analysis of bimodality among large cancer data was successfully applied in Chapter 4, and revealed that bimodality is common in cancer. I found that 2% of genes were bimodal at the critical *p* value *p*=0.01, and this could reach up to 8% at the significance level of *p*=0.05. As a result, I removed the top 10% as bimodal and outlier in the integrative analysis.
- HIM was presented in Chapter 5 and was compared to current integration methods, showing comparable performance. Chapter 6 included various applications using the proposed integration system.
- The thesis concluded with developments in MSG for cross species studies in Chapter 7. One such development was cross-species differential expression pattern discovery (CSMSG). The aim was to explore homogeneous and heterogeneous DEPs across species using multi-scale Gaussian mixtures. The second was called multivariate multi-scale Gaussian mixture (MVMSG). This aimed to identify genes that are conserved/diverged between species in many stages.

The proposed integration system has been shown to be a highly useful technique for integrating biological/medical data. The system contains three main components which can also be used separately for a specific purpose. The first component is used to identify the differentially expressed genes, the second to detect the bimodal genes and the third to apply the mean pattern model for integration.

8.2 The Limitations

Like most clustering methods, the proposed integration model suffers from estimating the number of clusters, especially when there is a large vector through pooling all data into it to estimate the number of clusters. However, this can be enhanced by vectorising the mean value for each gene in each set, instead of using all sample expressions. Another issue is that using normalising techniques does not always make the data comparable, particularly when different data types are included. Consequently, I have stuck to my proposed method to normalise the data as MDI did. Another issue is that the estimation of parameters is not always good enough if the data is noisy.

It has been demonstrated that hBI required large samples to be included in the analysis. This limits the use of this algorithm in small samples. In addition, this model is sensitive to outliers. For example, if one sample is higher than other samples it will gain small p value. One way to solve this issue was used in chapter 4 is to remove the subgroup of samples if they are less than 5% of the total samples. However this is arbitrary set in order to make a fair comparison with other methods.

8.3 Future studies and the use of models

8.3.1. The HIM model

The most obvious direction for future research is to employ a Bayesian learning framework to enhance the parameter estimation. I used Likelihood maximisation in the designing of my algorithm, as it is faster as well as good for debugging and development procedures. Also an important future study will be to

investigate my model in the context of using different genomics data. This study aims to integrate methylation data, DNA copy number and gene expression data. It is has been known that epigenetic and genetic variables interplay. However, their complex association is still far from being explored. Modelling their relationship is even more difficult. My future work is to examine how epigenetic variables associate with genetic variables, specifically, how possibly genes are regulated by epigenetic variables.

8.3.2. The MSG model

As has been seen, MSG is simple to implement and can be used in different contexts. It can also be extended to deal with more general problem settings in my future work.

*8.3.3. The h*BI model

I admit that I have introduced a hyper-parameter, i.e., a trade-off parameter. In order to remove this hyper-parameter, my future research will employ the Bayesian learning framework to overcome this difficulty. Nevertheless, the simulations that I have documented all show that my new algorithm is better than the benchmark algorithms in simulated data sets. In the application to real data sets, I show that my new algorithm is partially consistent with benchmark algorithms and that it also provides some new insights to the analysis of bimodal genes. Importantly, most of the bimodal genes predicted by my new algorithm do show typical bimodality, although a small percentage of my unique predictions was unfortunately not favoured by benchmark algorithms. I therefore look forward to some even more advanced approach, such as meta-analysis of prediction, to deliver even more robust predictions of bimodal genes.

It has been reported that some genes always showing bimodal distribution in cancer such as HER2 and they become a good candidate for subgroups identifications or so. Thus, I have downloaded more than 200 cancer datasets and going to identify bimodal genes in each dataset. Further analysis of those genes with the clinical data information provided with each dataset are required for more understanding of this phenomenon.

Bibliography

1. Crick, F.H., *On protein synthesis.* Symp of the Soc for Exp Biol, 1958. **12**: p. 138-163.
2. Alwine, J.C., Kemp, D.J.,Stark, G.R., *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proceedings of the National Academy of Sciences, 1977. **74**(12): p. 5350-5354.
3. Liang, P., Pardee, AB., *Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction.* Science, 1992. **257**(5072): p. 967-971.
4. Lockhart, D.J., Dong, Helin., Byrne, Michael.C., Follettie, Maximillian.T., Gallo, Michael.V., Chee, Mark.S., Mittmann, Michael., Wang, Chunwei., Kobayashi, Michiko., Norton, Heidi., Brown, Eugene.L.,, *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotech, 1996. **14**(13): p. 1675--1680.
5. Schena, M., Shalon, Dari.,Davis, Ronald.W., Brown, Patrick.O., *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.* Science, 1995. **270**(5235): p. 467-470.
6. Velculescu, V.E., Zhang, Lin.,Vogelstein, Bert., Kinzler, Kenneth.W., *Serial Analysis of Gene Expression.* Science, 1995. **270**(5235): p. 484-487.
7. Baldi, P., Frontmatter, G.Wesley.Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. 2002, United Kingdom: © Cambridge University Press.
8. Schulze, A., Downward, Julian., *Navigating gene expression using microarrays -- a technology review.* Nat Cell Biol, 2001. **3**(8): p. E190-E195.
9. 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., et al. , *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**: p. 531-7.
10. Hedenfalk, I., Duggan, David., Chen, Yidong., Radmacher, Michael., Bittner, Michael., Simon, Richard., Meltzer, Paul., Gusterson, Barry., Esteller, Manel., Raffeld, Mark.,Yakhini, Zohar., Ben-Dor, Amir., Dougherty, Edward., Kononen, Juha., Bubendorf, Lukas., Fehrle, Wilfrid., Pittaluga, Stefania., Gruvberger, Sofia., Loman, Niklas., Johannsson, Oskar., Olsson, Håkan., Wilfond, Benjamin., Sauter, Guido., Kallioniemi, Olli-P., Borg, Åke., Trent, Jeffrey., *Gene-Expression Profiles in Hereditary Breast Cancer.* New England Journal of Medicine, 2001. **344**(8): p. 539-548.
11. Ramaswamy, S., Ross, Ken.N., Lander, Eric.S., Golub, Todd.R., *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**(1): p. 49-54.
12. Penkett, C.J., Bähler, Jürg. , *Navigating Public Microarray Databases.* Comparative and Functional Genomics, 2004. **5**(6-7): p. 471-479.
13. Barrett, T., Suzek, T.O., Dennis B.Troup., Wilhite, Stephen.E., Ngau, Wing-Chi., Ledoux, P., Rudnev, R., Lash, Alex.E., Fujibuchi, W., Edgar,

R.,, *NCBI GEO: mining millions of expression profiles—database and tools.* Nucl. Acids Res, 2005. **33**: p. 562-566.
14. Barrett, T., Dennis B.Troup., Wilhite, Stephen.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M.,, *NCBI GEO: mining tens of millions of expression profiles—database and tools update* Nucl. Acids Res, 2007. **35**: p. 760-765.
15. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Ele, Holloway., Misha, Kapushesky., Patrick, Kemmeren., Gonzalo, Garcia., Lara, Ahmet., Oezcimen., Philippe, Rocca-Serra., Susanna-Assunta, Sansone.,, *ArrayExpress-a public repository for microarray gene expression data at the EBI.* nucl. Acids Res, 2003. **31**(1): p. 68-71.
16. Gregory, P.-S., Pablo Tamayo., *Microarray data mining: facing the challenges.* SIGKDD Explor. Newsl., 2003. **5**(2): p. 1-5.
17. Sui, Y., Zhao, Xiaoyue., Speed, Terence.P., Wu, Zhijin., *Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments.* Journal of Computational Biology, 2009. **16**(11): p. 1501-1515.
18. Okoniewski, M., Miller, Crispin., *Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.* BMC Bioinformatics, 2006. **7**(1): p. 276.
19. Wang, Z., Gerstein, Mark., Snyder, Michael., *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
20. Zhao, S., Fung-Leung, Wai-Ping., Bittner, Anton., Ngo, Karen., Liu, Xuejun., *Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells.* PLoS ONE, 2014. **9**(1): p. e78644.
21. Xu, X., Zhang, Yuanhao., Williams, Jennie., Antoniou, Eric., McCombie, W., Wu, Song., Zhu, Wei., Davidson, Nicholas., Denoya, Paula., Li, Ellen., *Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets.* BMC Bioinformatics, 2013. **14**(Suppl 9): p. S1.
22. Fu, X., Fu, Ning., Guo, Song., Yan, Zheng., Xu, Ying., Hu, Hao., Menzel, Corinna., Chen, Wei., Li, Yixue., Zeng, Rong., Khaitovich, Philipp., *Estimating accuracy of RNA-Seq and microarrays with proteomics.* BMC Genomics, 2009. **10**(1): p. 161.
23. Casanova, E.A., Okoniewski, Michal.J., Cinelli, Paolo., *Cross-Species Genome Wide Expression Analysis during Pluripotent Cell Determination in Mouse and Rat Preimplantation Embryos.* PLos ONE, 2012. **7**(10): p. e47107.
24. Wilcoxon, F., *Individual Comparisons by Ranking Methods.* Biometrics Bulletin`, 1945. **6**: p. 80-3.
25. Tusher, V.G., Tibshirani, R., Chu, G., *Significance analysis of microarrays applied to the ionizing radiation response.* PNAS, 2001. **98**: p. 5116-21.
26. Dondrup, M., Hüser, A.T., Mertens, D., Goesmann, A., *An evaluation framework for statistical tests on microarray data.* J Biotechnol, 2009. **140**: p. 18-26.
27. Vardhanabhuti, S., Blakemore, S.J., Clark, S.M., Ghosh, S., Stephens, R.J., Rajagopalan, D., *A comparison of statistical tests for detecting differential expression using Affymetrix oligonucleotide microarrays.* OMICS, 2006. **10**: p. 555-66.

28. Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J., Charnock-Jones, D.S., Print, C.G., Smith, S.K., *Independent component analysis of microarray data in the study of endometrial cancer.* Oncogene, 2004. **23**: p. 6677-83.
29. Cole, S.W., Galic, Z., Zack, J.A., *Controlling false-negative errors in microarray differential expression analysis: a PRIM approach.* Bioinformatics, 2003. **19**: p. 1808-16.
30. Kang, C.H., Anraku, M., Cypel, M., Sato, M., Yeung, J., Gharib, S.A., Pierre, A.F., de Perrot, M., Waddell, T.K., Liu, M., Keshavjee, S., *Transcriptional signatures in donor lungs from donation after cardiac death vs after brain death: a functional pathway analysis.* J Heart Lung Transplant, 2011. **30**: p. 289-98.
31. Zhang, Y., Zan, L., Wang, H., *Screening candidate genes related to tenderness trait in Qinchuan cattle by genome array.* Mol Biol Rep, 2011. **38**: p. 2007-14.
32. Korkor, M.T., Meng, F.B., Xing, S.Y., Zhang, M.C., Guo, J.R., Zhu, X.X., Yang, P., *Microarray analysis of differential gene expression profile in peripheral blood cells of patients with human essential hypertension.* Int J Med Sci, 2011. **8**: p. 168-79.
33. Lin, A.Y., Chua, M.S., Choi, Y.L., Yeh, W., Kim, Y.H., Azzi, R., Adams, G.A., Sainani, K., van de Rijn, M., So, S.K., Pollack, J.R., *Comparative profiling of primary colorectal carcinomas and liver metastases identifies LEF1 as a prognostic biomarker.* PLoS One, 2011. **6**: p. e16636.
34. Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., Smyth, G.K., *A comparison of background correction methods for two-colour microarrays.* Bioinformatics, 2007. **23**: p. 2700-7.
35. Baldi, P., Long, A.D., *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.* Bioinformatics, 2001. **17**: p. 509-19.
36. Wright, G.W., Simon, R.M., *A random variance model for detection of differential gene expression in small microarray experiments.* Bioinformatics, 2003. **19**: p. 2448-55.
37. Smyth, G., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Statistical applications in genetics and molecular biology, 2004. **3**: p. 1.
38. Delmar, P., Robin, S., Daudin, J.J., *VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data.* Bioinformatics, 2005. **21**: p. 502-8.
39. Jaffrezic, F., Marot, G., Degrelle, S., Hue, I., Foulley, J.L., *A structural mixed model for variances in differential gene expression studies.* Genet Res, 2007. **89**: p. 19-25.
40. Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.L., Visvader, J.E., Smyth, G.K., *ROAST: rotation gene set tests for complex microarray experiments.* Bioinformatics, 2010. **26**: p. 2176-82.
41. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci USA, 1998. **95**: p. 14863-8.
42. MacQueen, J. *Some Methods for classification and Analysis of Multivariate Observations*. in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*. 1967. University of California Press.
43. Kohonen, T., *Self-Organizing Maps*. 2001, Berlin: Springer.

44. Kristiansson, E., Osterlund, T, Gunnarsson, L, Arne, G, Joakim, Larsson, DG, Nerman, O, *A novel method for cross-species gene expression analysis.* BMC Bioinformatics, 2013. **14**: p. 70.
45. Hor, C., Yang, CB, Yang, ZJ, Tseng, CT, *Prediction of protein essentiality by the support vector machine with statistical tests.* Evol Bioinform Online, 2013. **9**: p. 387-416.
46. Blades, N., Browman, K, *Estimating the number of essential genes in a genome by random transposon mutagenesis*. 2002: Johns Hopkins University.
47. Cheng, J., Wu, W, Zhang, Y, Li, X, Jiang, X, Wei, G, Tao, S, *A new computational strategy for predicting essential genes.* BMC Genomics, 2013. **14**: p. 910.
48. Zhong, J., Wang, J, Peng, W, Zhang, Z, Pan, Y, *Prediction of essential proteins based on gene expression programming.* BMC Genomics, 2013. **s4**(s7).
49. Plaimas, K., Eils, R, König, R, *Identifying essential genes in bacterial metabolic networks with machine learning methods.* BMC Syst Biol, 2010. **4**: p. 56.
50. Bijlsma, J., Burghout, P, Kloosterman, TG, Bootsma, HJ, De, JA, *Development of genomic array footprinting for identification of conditionally essential genes in Streptococcus pneumoniae.* Appl Environ Microbiol, 2007. **73**: p. 1514-24.
51. Gerdes, S., Edwards, R, Kubal, M, Fonstein, M, Stevens, R, Osterman, A, *Essential genes on metabolic maps.* Current Opinion in Biotechnology, 2006. **17**: p. 448-56.
52. Hensel, M., Shea, JE, Gleeson, C, Jones, MD, Dalton, E, et al., *Simultaneous identification of bacterial virulence genes by negative selection.* Science, 1995. **269**: p. 400-3.
53. Sassetti, C., Boyd, DH, Rubin, EJ, *Comprehensive identification of conditionally essential genes in mycobacteria.* PNAS, 2001. **98**: p. 12712-7.
54. Efron, B., Tibshirani, R, Storey, JD, Tusher, V, *Empirical Bayes analysis of a microarray experiment.* Journal of American Statistical Association, 2001. **96**: p. 1151-60.
55. Oshlack, A., Chabot, AE, Smyth, GK, Gilad, Y, *Using DNA microarrays to study gene expression in closely related species.* Bioinformatics, 2007. **23**: p. 1235-40.
56. Adjaye, J., Herwig, R, Herrmann, D, Wruck, W, Benkahla, A, Brink, TC, Nowak, M, Carnwath, JW, Hultschig, C, Niemann, H, Lehrach, H, *Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays.* BMC Genomics, 2004. **5**: p. 8.
57. Lee, D.D., Seung, H.S., *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**: p. 788-91.
58. Lee, D.D., Seung, H.S., *Algorithms for non-negative matrix factorization.* Adv Neural Info Proc Syst, 2001. **13**: p. 556–62.
59. Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006: Springer.
60. Tamayo, P., Scanfeld, D., Ebert, B.L., Gillette, M.A., Roberts, C.W., Mesirov, J.P., *Metagene projection for cross-platform, cross-species characterization of global transcriptional states.* PNAS, 2007. **104**: p. 5959-64.

61. Stuart, J.M.e.a., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**: p. 249-255.
62. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P., *Metagenes and molecular pattern discovery using matrix factorization.* PNAS, 2004. **101**: p. 4164-9.
63. Burghout, P., Bootsma, HJ, Kloosterman, TG, Bijlsma, JJ, de Jongh, CE, et al., *Search for genes essential for pneumococcal transformation: the RADA DNA repair protein plays a role in genomic recombination of donor DNA.* J Bacteriol, 2007. **189**: p. 6540-50.
64. Mi, Z., Shen, K., Song, N., Cheng, C., Song, C., Kaminski, N., Tseng, G.C., *Module-based prediction approach for robust inter-study predictions in microarray data.* Bioinformatics, 2010. **26**: p. 2586-93.
65. Kleckner, N., Roth, J, Botstein, D, *Genetic engineering in vivo using translocatable drugresistance elements. New methods in bacterial genetics.* J Mol Biol, 1977. **116**: p. 125-59.
66. Lu, Y., Huggins, P., Bar-Joseph, Z., *Cross species analysis of microarray expression data.* Bioinformatics, 2009. **25**: p. 1476-83.
67. Fels, S., Zane, GM, Blake, SM, Wall, JD, *Rapid transposon liquid enrichment sequencing (TnLE-seq) for gene fitness evaluation in underdeveloped bacterial systems.* Appl Environ Microbiol, 2013. **79**: p. 7510-7.
68. Lu, Y., Yi, Y, Liu, P, Wen, W, James, M, Wang, D, You, M, *Common human cancer genes discovered by integrated gene-expression analysis.* PLoS One, 2007. **2**: p. e1149.
69. Lu, Y., He, X, Zhong, S, *Cross-species microarray analysis with the OSCAR system suggests an INSR->Pax6->NQO1 neuro-protective pathway in aging and Alzheimer's disease.* Nucleic Acids Res, 2007. **35**: p. W105-14.
70. Lamichhane, G., Zignol, M, Blades, NJ, Geiman, DE, Dougherty, A, Grosset, J, Broman, KW, Bishai, WR, *A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis.* PNAS, 2003. **100**: p. 7213-8.
71. Zomer, A., Burghout, P, Bootsma, HJ, Hermans, PW, van Hijum, SA, *ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data.* PLoS One, 2012. **7**: p. e43012.
72. Gallagher, L., Ramage, E, Jacobs, MA, Kaul, R, Brittnacher, M, Manoil, C, *A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate.* PNAS, 2007. **104**: p. 1009-14.
73. Dhanasekaran, S., . Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM., *Delineation of prognostic biomarkers in prostate cancer.* Nature, 2001. **412**(6849): p. 822-826.
74. Xu, L., Tan, Aik Choon., Naiman, Daniel Q., Geman, Donald., Winslow, Raimond L., *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data.* Bioinformatics, 2005. **21**(20): p. 3905-3911.
75. Higgins, J.P.T., Shinghal, Rajesh., Gill, Harcharan., Reese, Jeffrey H., Terris, Martha., Cohen, Ronald J., Fero, Michael., Pollack, Jonathan R., van de Rijn, Matt., Brooks, James D., *Gene Expression Patterns in Renal Cell Carcinoma Assessed by Complementary DNA Microarray.* The American Journal of Pathology, 2003. **162**(3): p. 925-932.

76. Yagi, T., Morimoto, Akira., Eguchi, Mariko., Hibi, Shigeyoshi., Sako, Masahiro., Ishii, Eiichi., Mizutani, Shuki., Imashuku, Shinsaku., Ohki, Misao., Ichikawa, Hitoshi., *Identification of a gene expression signature associated with pediatric AML prognosis.* Blood, 2003. **102**(5): p. 1849-1856.
77. Iacobuzio-Donahue, C.A., Maitra, Anirban., Olsen, Mari., Lowe, Anson W., Van Heek, N. Tjarda., Rosty, Christophe., Walter, Kim., Sato, Norihiro., Parker, Antony., Ashfaq, Raheela., Jaffee, Elizabeth., Ryu, Byungwoo., Jones, Jessa., Eshleman, James R., Yeo, Charles J., Cameron, John L., Kern, Scott E., Hruban, Ralph H., Brown, Patrick O., Goggins, Michael., *Exploration of Global Gene Expression Patterns in Pancreatic Adenocarcinoma Using cDNA Microarrays.* The American Journal of Pathology, 2003. **162**(4): p. 1151-1162.
78. Bullinger, L., Döhner, Konstanze., Bair, Eric., Fröhling, Stefan., Schlenk, Richard F., Tibshirani, Robert., Döhner, Hartmut., Pollack, Jonathan R., *Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia.* New England Journal of Medicine, 2004. **350**(16): p. 1605-1616.
79. Valk, P.J.M., Verhaak, Roel G.W., Beijen, M. Antoinette., Erpelinck, Claudia A.J., van Doorn-Khosrovani, Sahar Barjesteh van Waalwijk., Boer, Judith M., Beverloo, H. Berna., Moorhouse, Michael J., van der Spek, Peter J., Löwenberg, Bob., Delwel, Ruud., *Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia.* New England Journal of Medicine, 2004. **350**(16): p. 1617-1628.
80. Wang, Y., Chen, Jiajia., Li, Qinghui., Wang, Haiyun., Liu, Ganqiang., Jing, Qing., Shen, Bairong., *Identifying novel prostate cancer associated pathways based on integrative microarray data analysis.* Computational Biology and Chemistry, 2011. **35**(3): p. 151-158.
81. Sorlie, T., Perou, Charles M., Tibshirani, Robert., Aas, Turid., Geisler, Stephanie., Johnsen, Hilde., Hastie, Trevor., Eisen, Michael B., van de Rijn, Matt., Jeffrey, Stefanie S., Thorsen, Thor., Quist, Hanne., Matese, John C., Brown, Patrick.O., Botstein, David., Lonning, Per Eystein., Borresen-Dale, Anne-Lise., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proceedings of the National Academy of Sciences, 2001. **98**(19): p. 10869-10874.
82. Alizadeh, A.A., Eisen, Michael B., Davis, R.Eric., Ma, Chi., Lossos, Izidore.S., Rosenwald, Andreas., Boldrick, Jennifer.C. Sabet, Hajeer., Tran, Truc., Yu, Xin., Powell, John.I., Yang, Liming., Marti, Gerald.E., Moore, Troy., Hudson, James., Lu, Lisheng., Lewis, David.B., Tibshirani, Robert., Sherlock, Gavin., Chan, Wing.C., Greiner, Timothy.C., Weisenburger, Dennis.D., Armitage, James.O., Warnke, Roger., Levy, Ronald., Wilson, Wyndham., Grever, Michael.R., Byrd, John.C., Botstein, David., Brown, Patrick.O., Staudt, Louis.M.,, *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-511.
83. Ein-Dor, L., Kela, Itai., Getz, Gad., Givol, David., Domany, Eytan., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2): p. 171-178.
84. Yao, J., Zhao, Qi., Yuan, Ying., Zhang, Li., Liu, Xiaoming., Yung, W.K.Alfred., Weinstein, John.N., *Identification of Common Prognostic*

Gene Expression Signatures with Biological Meanings from Microarray Gene Expression Datasets. PLos ONE, 2012. **7**(9): p. e45894.
85. Buffa, F.M., Harris, A.L., West, C.M., Miller, C.J., *Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene.* Br J Cancer, 2010. **102**(2): p. 428-435.
86. Daves, M., Hilsenbeck, Susan., Lau, Ching., Man, Tsz-Kwong., *Meta-analysis of multiple microarray datasets reveals a common gene signature of metastasis in solid tumors.* BMC Medical Genomics, 2011. **4**(1): p. 56.
87. Markert, E.K., Levine, A.J., Vazquez, A., *Proliferation and tissue remodeling in cancer: the hallmarks revisited.* Cell Death Dis, 2012. **3**: p. e397.
88. Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, TR., Mesirov, JP., *Estimating dataset size requirements for classifying DNA microarray data. .* J Comput Biol 2003. **10**(2): p. 119-142.
89. Engelmann, J.C., Schwarz, Roland., Blenk, Steffen., Friedrich, Torben., Seibel, Philipp.N., Dandekar, Thomas., Müller, Tobias., *Unsupervised Meta-Analysis on Diverse Gene Expression Datasets Allows Insight into Gene Function and Regulation.* Bioinformatics and Biology Insights, 2008. **2**: p. 265.
90. Rhodes, D., Barrette, TR., Rubin, MA., Ghosh, D., Chinnaiyan, AM., *Meta- analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer.* Cancer Research, 2002. **62**: p. 4427 - 4433.
91. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A.M., *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* PNAS, 2004. **101**: p. 9309-14.
92. Li, J., Tseng, GC., *An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies.* Ann. Appl. Stat., 2011. **5**(2A): p. 994-1019.
93. Choi, J., Yu, U ., Kim, S., Yoo, OJ., *Combining multiple microarray studies and modeling inter-study variation.* Bioinformatics, 2003(Suppl 19): p. i84 - i90.
94. Marot, G., Foulley, Jean-Louis., Mayer, Claus-Dieter., Jaffrézic, Florence., *Moderated effect size and P-value combinations for microarray meta-analyses.* Bioinformatics, 2009. **25**(20): p. 2692-2699.
95. Hu, P., Greenwood, CMT., Beyene, J., *Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.* BMC Bioinformatics, 2005. **6**: p. 128.
96. Garrett-Mayer, E., Parmigiani, Giovanni., Zhong, Xiaogang., Cope, Leslie., Gabrielson, Edward.,, *Cross-study validation and combined analysis of gene expression microarray data.* Biostatistics, 2008. **9**(2): p. 333-354.
97. Zintzaras, E., Ioannidis, John.PA., *Meta-analysis for ranked discovery datasets: Theoretical framework and empirical demonstration for microarrays.* Computational Biology and Chemistry, 2008. **32**(1): p. 39-47.

98. Hong, F., Breitling, Rainer., *A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.* Bioinformatics, 2008. **24**(3): p. 374-382.
99. DeConde, R.P., Hawley,S., Falcon,S., Clegg,N., Knudsen,B., Etzioni,R., *Combining results of microarray experiments: a rank aggregation approach.* Stat Appl Genet Mol Biol, 2006. **5**(1): p. Article 15.
100. Parmigiani, G., Garrett-Mayer, ES., Anbazhagan, R., Gabrielson, E., *A cross-study comparison of gene expression studies for the molecular classification of lung cancer.* Clinical Cancer Research, 2004. **10**: p. 2922 - 2927.
101. Jung, Y., . Oh, M., Shin, D., Kang, S., Oh, H., *Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering. .* Biometrical Journal, 2006. **48**(3): p. 435-450.
102. Devarajan, K., *Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology.* PLoS Comput Biol, 2008. **4**(7): p. e1000029.
103. Shen, R., Olshen, Adam.B., Ladanyi, Marc., *Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.* Bioinformatics, 2009. **25**(22): p. 2906-2912.
104. Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B., *JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES.* The annals of applied statistics, 2013. **7**(1): p. 523-542.
105. Kristensen, V.N., Lingjaerde, Ole.Christian., Russnes, Hege.G., Vollan, Hans.Kristian.M., Frigessi, Arnoldo., Borresen-Dale, Anne-Lise.,, *Principles and methods of integrative genomic analyses in cancer.* Nat Rev Cancer, 2014. **14**(5): p. 299-313.
106. Troyanskaya, O.G., Dolinski , Kara.,Owen, Art.B., Altman, Russ.B.,Botstein , David., *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8348-8353.
107. Conlon, E., Song, Joon., Liu, Jun., *Bayesian models for pooling microarray studies with multiple sources of replications.* BMC Bioinformatics, 2006. **7**(1): p. 247.
108. Scharpf, R.B., Tjelmeland, Håkon., Parmigiani, Giovanni., Nobel, Andrew.B., *A Bayesian Model for Cross-Study Differential Gene Expression.* Journal of the American Statistical Association, 2009. **104**(488): p. 1295-1310.
109. Wang, W., Baladandayuthapani, Veerabhadran., Morris, Jeffrey.S., Broom, Bradley.M., Manyam, Ganiraju., Do, Kim-Anh., *iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data.* Bioinformatics, 2013. **29**(2): p. 149-159.
110. Wang, J., Wen, S., Symmans, W.F., Pusztai, L., Coombes, K.R., *The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data.* Cancer Inform, 2009. **7**: p. 199-216.
111. Yang, Y., Tashman, Adam., Lee, Jung., Yoon, Seungtai., Mao, Wenyang., Ahn, Kwangmi., Kim, Wonkuk., Mendell, Nancy., Gordon, Derek., Finch, Stephen., *Mixture modeling of microarray gene expression data.* BMC Proceedings, 2007. **1**(Suppl 1): p. S50.

112. Ertel, A., Tozeren, A., *Switch-like genes populate cell communication pathways and are enriched for extracellular proteins.* BMC Bioinformatics, 2008. **9**: p. 3.
113. Teschendorff, A.E., Naderi, A., Barbosa-Morais, N.L., Caldas, C., *PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer.* Bioinformatics, 2006. **22**: p. 2269-75.
114. Hirakawa, A., Sato, Y., Hamada, C., Yoshimura, I., *A new test statistic based on shrunken sample variance for identifying differentially expressed genes in small microarray experiments.* Bioinform Biol Insights, 2008. **2**: p. 145-56.
115. Kim, M., Cho, S.B., Kim, J.H., *Mixture-model based estimation of gene expression variance from public database improves identification of differentially expressed genes in small sized microarray data.* Bioinformatics, 2010. **26**: p. 486-92.
116. Smyth, G., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. 3.
117. Reid, R.W., Fodor, A.A., *Determining gene expression on a single pair of microarrays.* BMC Bioinformatics, 2008. **9**: p. 489.
118. Kendziorski, C.M., Newton, M.A., Lan, H., Gould, M.N., *On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.* Stat Med, 2003. **22**: p. 3899-914.
119. Hein, A.M., Richardson, S., *A powerful method for detecting differentially expressed genes from GeneChip arrays that does not require replicates.* BMC Bioinformatics, 2006. **7**: p. 353.
120. Turro, E., Bochkina, N., Hein, A.M., Richardson, S., *BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips.* BMC Bioinformatics, 2007. **8**: p. 439.
121. Lonnstedt, I., Speed, T.P., *Replicated microarray data.* Statistica Sinica, 2002. **12**: p. 31-46.
122. Tripathi, A., King, C., de la Morenas, A., Perry, V.K., Burke, B., Antoine, G.A., Hirsch, E.F., Kavanah, M., Mendez, J., Stone, M., Gerry, N.P., Lenburg, M.E., Rosenberg, C.L., *Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients.* Int J Cancer, 2008. **122**: p. 557-66.
123. Long, A.D., Mangalam, H.J., Chan, B.Y., Tolleri, L., Hatfield, G.W., Baldi, P., *Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12.* J Biol Chem, 2001. **276**: p. 19937-44.
124. Jemtland, R., Holden, M., Reppe, S., Olstad, O.K., Reinholt, F.P., Gautvik, V.T., Refvem, H., Frigessi, A., Houston, B., Gautvik, K.M., *Molecular disease map of bone characterizing the postmenopausal osteoporosis phenotype.* J Bone Miner Res, 2011. **26**: p. 1793-901.
125. Jungke, P., Ostrow, G., Li, J.L., Norton, S., Nieber, K., Kelber, O., Butterweck, V., *Profiling of hypothalamic and hippocampal gene expression in chronically stressed rats treated with St. John's wort extract (STW 3-VI) and fluoxetine.* Psychopharmacology, 2011. **213**: p. 757-72.
126. López-Romero, P., *Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library.* BMC Genomics, 2011. **12**: p. 64.

127. Kim, R.D., Park, P.J., *Improving identification of differentially expressed genes in microarray studies using information from public databases.* Genome Biol, 2004. **5**: p. R70.
128. Miller, W.R., Larionov, A.A., Renshaw, .L, Anderson, T.J. et al., *Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole.* Pharmacogenet Genomics, 2007. **17**: p. 813-26.
129. Cai, J., Xie, D, Fan, Z, Chipperfield, H, Marden, J, Wong, WH, Zhong, S, *Modeling Co-Expression across Species for Complex Traits: Insights to the Difference of Human and Mouse Embryonic Stem Cells.* PLOS Computational Biology, 2010. **6**: p. e1000707.
130. Casanova, E., Okoniewski, MJ, Cinelli, P, *Cross-Species Genome Wide Expression Analysis during Pluripotent Cell Determination in Mouse and Rat Preimplantation Embryos.* PLoS ONE, 2012. **7**: p. e47107.
131. Walder, C., Kim, K., Scholkopf, B., *Sparse Multiscale Gaussian Process Regression*, in *25 th International Conference on Machine Learning*. 2008: Helsinki, Finland.
132. Sarkar, S., Ghosh, K., Bhaumik, K., *A weighted sum of multi-scale Gaussians generates new near-ideal interpolation functions*, in *Conf Proc IEEE Eng Med Biol Soc*. 2005. p. 6387-90.
133. Chang, T.R., Chen, E.L., Poon, P.W., Chung, P.C., Chiu, T.W., *Responses of central auditory neurons modeled with finite impulse response (FIR) neural networks.* Comput Methods Programs Biomed, 2004. **74**: p. 151-65.
134. Liang, D., Deng, W., Wang, X., Zhang, Y., *Multivariate Image Analysis in Gaussian Multi-Scale Space for Defect Detection.* Journal of Bionic Engineering, 2009. **6**: p. 298-305.
135. Jin, C., Kim, H., *Pixel-level singular point detection from multi-scale Gaussian filtered orientation field.* Pattern Recognition, 2010. **43**: p. 3879-90.
136. Sakai, T., Narita, M., Komazaki, T., Nishiguchi, H., Imiya, A., *Image Hierarchy in Gaussian Scale Space* Advances in Imaging and Electron Physics, 2011. **165**: p. 175-263.
137. Webb, A., *Statistical Pattern Recognition*. 2002, Chichester: John Wiley & Sons Ltd.
138. Dembélé, D., *A flexible microarray data simulation model.* Microarrays, 2013. **2**(2): p. 115-130.
139. Käll, L., Storey, J.D., Noble, W.S., *QVALITY: non-parametric estimation of q-values and posterior error probabilities.* Bioinformatics, 2009. **25**: p. 964-6.
140. Käll, L., Storey, J.D., Noble, W.S., *Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry.* Bioinformatics, 2008. **24**: p. i42-8.
141. Storey, J.D., *A direct approach to false discovery rates.* J. R. Stat. Soc., 2002. **64**: p. 479-98.
142. Käll, L., Storey, J.D., MacCoss, M.J., Noble, W.S., *Posterior error probabilities and false discovery rates: two sides of the same coin.* J Proteome Res, 2008. **7**: p. 40-4.
143. Krzanowski, W.J., Hand, D.J., *ROC curves for continuous data*. 2009: CRC Press.
144. Metz, C.E., *Basic principles of ROC analysis. .* Seminars in Nuclear Medicine, 1978. **8**: p. 283-288.

145. Duda, R.O., Hart, P.E., Stork, D.G., *Pattern Classification*. 2nd ed. 2000: Wiley-Interscience.
146. Vapnik, V., *The Nature of Statistical Learning Theory.* 1995, New York: Springer-Verlag.
147. Tipping, M.E., *Sparse Bayesian learning and the relevance vector machine.* J Mach Learn Res, 2001. **1**: p. 211-44.
148. Tavazoie, S.F., Alarcón, .C, Oskarsson, T., Padua, D. et. al., *Endogenous human microRNAs that suppress breast cancer metastasis.* Nature 2008. **451**: p. 147-52.
149. Wong, S.Y., Haack, H., Kissil, J.L., Barry, M. et. al., *Protein 4.1B suppresses prostate cancer progression and metastasis.* PNAS, 2007. **104**: p. 12784-9.
150. Chen, H., Boutros, P., *VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.* BMC Bioinformatics, 2011. **12**: p. 35.
151. Palmieri, D., Fitzgerald, D., Shreeve, S.M., Hua, E., Bronder, J.L., Weil, R.J., Davis, S, Stark, A.M., Merino, M.J., Kurek, R., Mehdorn, H.M., Davis, G., Steinberg, S.M., Meltzer, P.S., Aldape, K., Steeg, P.S., *Analyses of resected human brain metastases of breast cancer reveal the association between up-regulation of hexokinase 2 and poor prognosis.* Mol Cancer Res, 2009. **7**: p. 1438-45.
152. Liao, L.Y., Sun, Y.M. Chau, G.Y., Chau, Y.P., Lai, T.C., Wang, J.L., Horng, J.T., Hsiao, M., Tsou, A.P., *Identification of SOX4 target genes using phylogenetic footprinting-based prediction from expression microarrays suggests that overexpression of SOX4 potentiates metastasis in hepatocellular carcinomaOverexpression of SOX4 potentiates metastasis in HCC.* Oncogene, 2008. **27**: p. 5578-89.
153. Lai, Y.H., Cheng, J., Cheng, D., Feasel, M.E., Beste, K.D., Peng, J., Nusrat, A., Moreno, C.S., *SOX4 interacts with plakoglobin in a Wnt3adependent manner in prostate cancer cells.* BMC Cell Biology, 2011. **12**: p. 50.
154. Tsuji, K., Kawauchi, S., Saito, S., Furuya, T., Ikemoto, K., Nakao, M., Yamamoto, S., Oka, M., Hirano, T., Sasaki, K., *Breast cancer cell lines carry cell line-specific genomic alterations that are distinct from aberrations in breast cancer tissues: comparison of the CGH profiles between cancer cell lines and primary cancer tissues.* BMC Cancer, 2010. **10**: p. 15.
155. Venere, M., Snyder, A., Zgheib, O., Halazonetis, T.D., *Phosphorylation of ATR-interacting protein on Ser239 mediates an interaction with breast-ovarian cancer susceptibility 1 and checkpoint function.* Cancer Res, 2007. **67**: p. 6100-5.
156. Ewton, D.Z., Hu, J., Vilenchik, M., Deng, X., Luk, K.C., Polonskaia, A., Hoffman, A.F., Zipf, K., Boylan, J.F., Friedman, E.A., *Inactivation of mirk/dyrk1b kinase targets quiescent pancreatic cancer cells.* Mol Cancer Ther, 2011. **10**: p. 2104-14.
157. Luyimbazi, D., Akcakanat, A., McAuliffe, P.F., Zhang, L., Singh, G., Gonzalez-Angulo, A.M., Chen, H., Do, K.A., Zheng, Y., Hung, M.C., Mills, G.B., Meric-Bernstam, F., *Rapamycin regulates stearoyl CoA desaturase 1 expression in breast cancer.* Mol Cancer Ther, 2010. **9**: p. 2770-84.
158. Muzikar, K.A., Nickols, N.G., Dervan, P.B., *Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression.* PNAS, 2009. **106**: p. 16598-603.

159. Wu, W., Chaudhuri, S., Brickley, D.R., Pang, D., Karrison, T., Conzen, S.D., *Microarray analysis reveals glucocorticoid-regulated survival genes that are associated with inhibition of apoptosis in breast epithelial cells.* Cancer Res, 2004. **64**: p. 1757-64.
160. De, S., Chen, J., Narizhneva, N.V., Heston, W., Brainard, J., Sage, E.H., Byzova, T.V., *Molecular pathway for cancer metastasis to bone.* J Biol Chem, 2003. **278**: p. 39044-50.
161. Derosa, C.A., Furusato, B., Shaheduzzaman, S., Srikantan, V., Wang, Z., Chen, Y., Siefert, M., Ravindranath, L., Young, D., Nau, M., Dobi, A., Werner, T., McLeod, D.G., Vahey, M.T., Sesterhenn, I.A., Srivastava, S., Petrovics, G., *Elevated osteonectin/SPARC expression in primary prostate cancer predicts metastatic progression.* Prostate Cancer Prostatic Dis, 2012. **in press**.
162. Thomas, R., True, L.D., Bassuk, J.A., Lange, P.H., Vessella, R.L., *Differential expression of osteonectin/SPARC during human prostate cancer progression.* Clin Cancer Res, 2000. **6**: p. 1140-9.
163. Blumenthal, R.D., Leon, E., Hansen, H.J., Goldenberg, D.M., *Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers.* BMC Cancer, 2007. **7**: p. 2.
164. Bajaj, J., Maliekal, T.T., Vivien, E., Pattabiraman, C., Srivastava, S., Krishnamurthy, H., Giri, V., Subramanyam, D., Krishna, S., *Notch signaling in CD66+ cells drives the progression of human cervical cancers.* Cancer Res, 2011. **71**: p. 4888-97.
165. Kolla, V., Gonzales, L.W., Bailey, N.A., Wang, P., Angampalli, S., Godinez, M.H., Madesh, M., Ballard, P.L., *Carcinoembryonic cell adhesion molecule 6 in human lung: regulated expression of a multifunctional type II cell protein.* Am J Physiol Lung Cell Mol Physiol, 2009. **206**: p. L1019-30.
166. Gu, Z., Rubin, M.A., Yang, Y., Deprimo, S.E., Zhao, H., Horvath, S., Brooks, J.D., Loda, M., Reiter, R.E., *Reg IV: a promising marker of hormone refractory metastatic prostate cancer.* Clin Cancer Res, 2005. **11**: p. 2237-43.
167. Taylor, B.S., Pal, M., Yu, J., Laxman, B., Kalyana-Sundaram, S., Zhao, R., Menon, A., Wei, J.T., Nesvizhskii, A.I., Ghosh, D., Omenn, G.S., Lubman, D.M., Chinnaiyan, A.M., Sreekumar, A., *Humoral response profiling reveals pathways to prostate cancer progression.* Mol Cell Proteomics, 2008. **7**: p. 600-11.
168. You, Z., Dong, Y., Kong, X., Beckett, L.A., Gandour-Edwards, R., Melamed, J., *Midkine is a NF-kappaB-inducible gene that supports prostate cancer cell survival.* BMC Med Genomics, 2008. **1**: p. 6.
169. Trojan, L., Schaaf, A., Steidler, A., Haak, M., Thalmann, G., Knoll, T., Gretz, N., Alken, P., Michel, M.S., *Identification of metastasis-associated genes in prostate cancer by genetic profiling of human prostate cancer cell lines.* Anticancer Res, 2005. **25**: p. 183-91.
170. Hengstler, J.G., Lange, Jost., Kett, Alexandra., Dornhöfer, Nadja., Meinert, Rolf., Arand, Michael., Knapstein, Paul G., Becker, Roger., Oesch, Franz., Tanner, Berno., *Contribution of c-erbB-2 and Topoisomerase IIα to Chemoresistance in Ovarian Cancer.* Cancer Research, 1999. **59**(13): p. 3206-3214.
171. Kristensen, V.N., Edvardsen, Hege., Tsalenko, Anya., Nordgard, Silje H., Sorlie, Therese., Sharan, Roded., Vailaya, Aditya., Ben-Dor, Amir., Lonning, Per Eystein., Lien, Sigbjorn., Omholt, Stig., Syvanen, Ann-

Christine., Yakhini, Zohar., Borresen-Dale, Anne-Lise., *Genetic variation in putative regulatory loci controlling gene expression in breast cancer.* Proceedings of the National Academy of Sciences, 2006. **103**(20): p. 7735-7740.
172. Bertucci, F.B., Daniel., *Reasons for breast cancer heterogeneity.* Journal of Biology, 2008. **7**(2): p. 6.
173. Hu, J., *Cancer outlier detection based on likelihood ratio test.* Bioinformatics, 2008. **24**: p. 2193-9.
174. 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., Chinnaiyan, A.M., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**: p. 644-8.
175. Slamon, D., Clark, GM., Wong, SG., Levin, WJ., Ullrich, A., McGuire, WL., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science, 1987. **235**(4785): p. 177-182.
176. Mason, C., Hanson, Robert., Ossowski, Vicky., Bian, Li., Baier, Leslie., Krakoff, Jonathan., Bogardus, Clifton., *Bimodal distribution of RNA expression levels in human skeletal muscle tissue.* BMC Genomics, 2011. **12**(1): p. 98.
177. Lim, T.-O.B., Rugayah. Morad, Zaki. Hamid, Maimunah A., *Bimodality in Blood Glucose Distribution: is it universal?* Diabetes Care, 2002. **25**(12): p. 2212-2217.
178. Fan, J.M., Susanne.J. Zhou, Yue. Barrett-Connor, Elizabeth., *Bimodality of 2-h Plasma Glucose Distributions in Whites.* Diabetes Care, 2005. **28**(6): p. 1451-1456.
179. Dozmorov, I.K., Nicholas. Tang, Yuhong. Shields, Alan. Pathipvanich, Parima. Jarvis, James,N. Centola, Michael., *Hypervariable genes—experimental error or hidden dynamics.* Nucleic Acids Research, 2004. **32**(19): p. e147.
180. Blomquist, T., Crawford, E.L., Yoon, Y., Hernandez, D.A., Khuder, S., Ruppel, P.L., Peters, E., Oldfield, D.J., Austermiller, B., Anders, J.C. and Willey, J.C., *Pattern of antioxidant and DNA repair gene expression in normal airway epithelium associated with lung cancer diagnosis.* Cancer research, 2009. **69**(22): p. 8629-8635.
181. Willey, J.C., *expression a La Bimode. .* Cancer informatics, , 2010. **9**: p. 37-38.
182. Khalil, I.G., Hill, C., *Systems biology for cancer.* Current Opinion in Oncology, 2005. **17**(1): p. 44-48.
183. Hellwig, B., Hengstler, J.G., Schmidt, M., Gehrmann, M.C., Schormann, W., Rahnenführer, J., *Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes.* BMC Bioinformatics, 2010. **11**: p. 276.
184. Blenkiron, C., Goldstein, Leonard., Thorne, Natalie., Spiteri, Inmaculada., Chin, Suet-Feung., Dunning, Mark., Barbosa-Morais, Nuno., Teschendorff, Andrew., Green, Andrew., Ellis, Ian., Tavare, Simon., Caldas, Carlos., Miska, Eric., *MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.* Genome Biology, 2007. **8**(10): p. R214.
185. Chin, S., Teschendorff, Andrew., Marioni, John.,Wang, Yanzhong., Barbosa-Morais, Nuno., Thorne, Natalie., Costa, Jose., Pinder, Sarah.,

van de Wiel, Mark., Green, Andrew., Ellis, Ian., Porter, Peggy., Tavare, Simon., Brenton, James., Ylstra, Bauke., Caldas, Carlos., *High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.* Genome Biology, 2007. **8**(10): p. R215.

186. Gort, M., Broekhuis, Manda., Otter, Renée., Klazinga, Niek., *Improvement of best practice in early breast cancer: actionable surgeon and hospital factors.* Breast Cancer Research and Treatment, 2007. **102**(2): p. 219-226.
187. Ellsworth, R.E., Hooke,Jeffrey.A., Shriver,Craig.D., Ellsworth,Darrell.L. , *Genomic Heterogeneity of Breast Tumor Pathogenesis.* Clinical Medicine Insights: Oncology 2009. **3**: p. 77-85.
188. Bradford, L.D., *CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants.* Pharmacogenomics, 2002. **3**(2): p. 229-243.
189. Ellsworth, R.E.D., David.J.; Shriver, Craig.D.; Ellsworth, Darrell.L. , *Breast Cancer in the Personal Genomics Era.* Current Genomics, 2010. **11**(3): p. 146-161.
190. Agus, D.B.A., Robert W.; Fox, William D.; Lewis, Gail D.; Higgins, Brian.; Pisacane, Paul I.; Lofgren, Julie A.; Tindell, Charles; Evans, Douglas P.; Maiese, Krista; Scher, Howard I.; Sliwkowski, Mark X., *Targeting ligand-activated ErbB2 signaling inhibits breast and prostate tumor growth.* Cancer Cell, 2002. **2**(2): p. 127-137.
191. Harris, M., *Monoclonal antibodies as therapeutic agents for cancer.* The Lancet Oncology, 2004. **5**(5): p. 292-302.
192. Nahta, R.E., Francisco., *HER2 therapy: Molecular mechanisms of trastuzumab resistance.* Breast Cancer Research, 2006A. **8**(6): p. 215.
193. Nahta, R.E., Francisco.J., *Herceptin: mechanisms of action and resistance.* Cancer Letters, 2006B. **232**(2): p. 123-138.
194. Chabner, B.A.R., Thomas G., *Chemotherapy and the war on cancer.* Nat Rev Cancer, 2005. **5**(1): p. 65-72.
195. Muhsin, M.G., Joanne.; Kirkpatrick, Peter., *Gefitinib.* Nat Rev Cancer, 2003. **3**(8): p. 556-557.
196. Jänne, P.A.E., Jeffrey A.; Johnson, Bruce E., *Epidermal Growth Factor Receptor Mutations in Non–Small-Cell Lung Cancer: Implications for Treatment and Tumor Biology.* Journal of Clinical Oncology, 2005. **23**(14): p. 3227-3234.
197. Kris, M.G., Natale, Ronald B.,Herbst, Roy S., Lynch, Thomas J., Prager, Diane., Belani, Chandra P., Schiller, Joan H., Kelly, Karen., Spiridonidis, Harris., Sandler, Alan., Albain, Kathy S., Cella, David., Wolf, Michael.K., Averbuch, Steven.D., Ochs, Judith.J., Kay, Andrea.C., *Efficacy of Gefitinib, an Inhibitor of the Epidermal Growth Factor Receptor Tyrosine Kinase, in Symptomatic Patients With Non–Small Cell Lung Cancer.* JAMA: The Journal of the American Medical Association, 2003. **290**(16): p. 2149-2158.
198. Wu, H.-C.C., De-Kuan.; Huang, Chia-Ting., *Targeted Therapy for Cancer.* J. Cancer Mol., 2006. **2**(2): p. 57-66.
199. Giaccone, G.H., Roy S.; Manegold, Christian.; Scagliotti, Giorgio.; Rosell, Rafael.; Miller, Vincent.; Natale, Ronald B.;Schiller, Joan H.;von Pawel, Joachim.; Pluzanska, Anna.; Gatzemeier, Ulrich.; Grous, John.; Ochs, Judith S.; Averbuch, Steven D.; Wolf, Michael K.; Rennie, Pamela.; Fandi, Abderrahim.; Johnson, David H., *Gefitinib in Combination With Gemcitabine and Cisplatin in Advanced Non–Small-Cell Lung Cancer: A*

Phase III Trial—INTACT 1. Journal of Clinical Oncology, 2004. **22**(5): p. 777-784.
200. Paez, J.G.J., Pasi A. Lee, Jeffrey C.; Tracy, Sean.; Greulich, Heidi.; Gabriel, Stacey.; Herman, Paula.; Kaye, Frederic J.; Lindeman, Neal.; Boggon, Titus J.; Naoki, Katsuhiko.; Sasaki, Hidefumi.; Fujii, Yoshitaka.; Eck, Michael J.; Sellers, William R.; Johnson, Bruce E.; Meyerson, Matthew., *EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy.* Science, 2004. **304**(5676): p. 1497-1500.
201. Perou, C.M., Sorlie, Therese., Eisen, Michael.B., van de Rijn, MattJeffrey, Stefanie S. Rees, Christian A. Pollack, Jonathan R. Ross, Douglas T. Johnsen, Hilde Akslen, Lars A. Fluge, Oystein Pergamenschikov, Alexander Williams, Cheryl Zhu, Shirley., X. Lonning, Per E. Borresen-Dale, Anne-Lise Brown, Patrick O. Botstein, David, *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-752.
202. DeRisi, J.L., Iyer, Vishwanath.R., Brown, Patrick O., *Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale.* Science, 1997. **278**(5338): p. 680-686.
203. Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., *Empirical bayes analysis of a microarray experiment.* Journal of the American Statistical Association, 2001. **96**: p. 1151-60.
204. Karn, T., Pusztai, Lajos., RuckhÃ¤berle, Eugen., Liedtke, Cornelia., MÃ¼ller, Volkmar., Schmidt, Marcus., Metzler, Dirk., Wang, Jing., Coombes, Kevin R., GÃ¤tje, Regine., Hanker, Lars., Solbach, Christine., Ahr, Andre., Holtrich, Uwe., Rody, Achim., Kaufmann, Manfred., *Melanoma antigen family A identified by the bimodality index defines a subset of triple negative breast cancers as candidates for immune response augmentation.* European Journal of Cancer, 2012. **48**(0): p. 12-23.
205. Kernagis, D.N., Hall, Allison H. S., Datto, Michael B., *Genes with Bimodal Expression Are Robust Diagnostic Targets that Define Distinct Subtypes of Epithelial Ovarian Cancer with Different Overall Survival.* The Journal of Molecular Diagnostics, 2012. **14**(3): p. 214-222.
206. Wappett, M.a.D., Austin and Yang, Zheng Rong and Al-Watban, Abdullatif and Bradford, James R. and Dry, Jonathan R., *Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs.* BMC Genomics, 2016. **17**(65): p. 1-15.
207. Bottomly, D., Ryabinin, Peter.A., Tyner, Jeffrey.W., Chang, Bill.H., Loriaux, Marc.M., Druker, Brian.J., McWeeney, Shannon.K., Wilmot, Beth., *Comparison of methods to identify aberrant expression patterns in individual patients: augmenting our toolkit for precision medicine.* Genome Medicine 2013. **5**(103).
208. Paliwal, S.I., Pablo.A. Campbell, Kyle.Hilioti, Zoe. Groisman, Alex. Levchenko, Andre., *MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast.* Nature, 2007. **446**(7131): p. 46-51.
209. Ertel, A., *Bimodal Gene Expression and Biomarker Discovery.* Cancer Informatics, 2010. **9**: p. 11-14.
210. Kanehisa, M.G., Susumu. Hattori, Masahiro. Aoki-Kinoshita, Kiyoko.F. Itoh, Masumi. Kawashima, Shuichi. Katayama, Toshiaki. Araki, Michihiro. Hirakawa, Mika., *From genomics to chemical genomics: new*

developments in KEGG. Nucleic Acids Research, 2006. **34**(suppl 1): p. D354-D357.
211. Ashburner, M., . Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Isseltarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G., *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. **25**(1): p. 25-29.
212. Sandberg, R.E., Ingemar, *The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles.* Genome Biology, 2005. **6**(8): p. R65.
213. Ertel, A.V., Arun. Byers, Stephen. Ochs, Michael. Tozeren, Aydin, *Pathway-specific differences between tumor cell lines and normal and tumor tissue cells.* Molecular Cancer, 2006. **5**(1): p. 55.
214. Ertel, A., Tozeren, A, *Human and mouse switch-like genes share common transcriptional regulatory mechanisms for bimodality.* BMC Genomics, 2008. **9**(1): p. 628.
215. Zhao, H.-Y.Y., Patrick Y. K. Fang, Kai-Tai, *Identification of Differentially Expressed Genes with Multivariate Outlier Analysis.* Journal of Biopharmaceutical Statistics, 2004. **14**(3): p. 629-646.
216. Bessarabova, M., Kirillov, E., Shi, W., Bugrim, A., Nikolsky, Y., Nikolskaya, T., *Bimodal gene expression patterns in breast cancer.* BMC Genomics, 2010. **11**: p. S8.
217. Tibshirani, R., Hastie, T., *Outlier sums for differential gene expression analysis.* Biostatistics, 2007. **8**: p. 2-8.
218. Wu, B., *Cancer outlier differential gene expression detection.* Biostatistics, 2007. **8**: p. 566-75.
219. Smith, T.D.A., Makov, U., *Statistical Analysis of Finite Mixture Distributions*. 1985: John Wiley & Sons.
220. Everitt, B.S., Hand D.J., *Finite mixture distributions*. 1981: Chapman & Hall.
221. Lian, H., *MOST: detecting cancer differential gene expression.* Biostatistics, 2008. **9**: p. 411-8.
222. Wang, Y., Rekaya, R., *LSOSS: Detection of Cancer Outlier Differential Gene Expression.* Biomark Insights, 2010. **5**: p. 69-78.
223. Chen, L.-A.C., Dung-Tsa.; Chan, Wenyaw., *The distribution-based p-value for the outlier sum in differential gene expression analysis.* Biometrika, 2010. **97**(1): p. 246-253.
224. Teschendorff, A., . Miremadi, Ahmad. Pinder, Sarah. Ellis, Ian. Caldas, Carlos., *An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer.* Genome Biology, 2007. **8**(8): p. R157.
225. Gormley, M., Tozeren, A., *Expression profiles of switch-like genes accurately classify tissue and infectious disease phenotypes in model-based classification.* BMC Bioinformatics, 2008. **9**: p. 486.
226. Besag, J., Clifford, Peter, *Sequential Monte Carlo p-values.* Biometrika, 1991. **78**(2): p. 301-304.
227. Schmidt, M., Böhm, Daniel., von Törne, Christian., Steiner, Eric., Puhl, Alexander., Pilch, Henryk., Lehr, Hans-Anton., Hengstler, Jan G.., Kölbl, Heinz.,Gehrmann, Mathias.,, *The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer.* Cancer Research, 2008. **68**(13): p. 5405-5413.

228. Chen, H., Boutros, Paul, *VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R.* BMC Bioinformatics, 2011. **12**(1): p. 35.
229. Jovov, B., Araujo-Perez, Felix., Sigel, Carlie S., Stratford, Jeran K., McCoy, Amber N., Yeh, Jen Jen., Keku, Temitope., *Differential Gene Expression between African American and European American Colorectal Cancer Patients.* PLos ONE, 2012. **7**(1): p. e30168.
230. Tan, E.-H., Ramlau, R., Pluzanska, A., Kuo, H.-P., Reck, M., Milanowski, J., Au, J. S.-K., Felip, E., Yang, P.-C., Damyanov, D., Orlov, S., Akimov, M., Delmar, P.,Essioux, L., Hillenbach, C., Klughammer, B., McLoughlin, P. Baselga, J., *A multicentre phase II gene expression profiling study of putative relationships between tumour biomarkers and clinical response with erlotinib in non-small-cell lung cancer.* Annals of Oncology, 2010. **21**(2): p. 217-222.
231. Shields, D.J., Niessen, Sherry., Murphy, Eric A., Mielgo, Ainhoa., Desgrosellier, Jay S., Lau, Steven K. M., Barnes, Leo A., Lesperance, Jacqueline., Bouvet, Michael., Tarin, David., Cravatt, Benjamin F., Cheresh, David A.,, *RBBP9: A tumor-associated serine hydrolase activity required for pancreatic neoplasia.* Proceedings of the National Academy of Sciences, 2010. **107**(5): p. 2189-2194.
232. Wang, Y., Klijn, Jan G. M., Zhang, Yi, Sieuwerts, Anieta M., Look, Maxime P., Yang, Fei, Talantov, Dmitri, Timmermans, Mieke, Meijer-van Gelder, Marion E., Yu, Jack, Jatkoe, Tim, Berns, Els M. J. J., Atkins, David., Foekens, John A., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* The Lancet, 2005. **365**(9460): p. 671-679.
233. Boimel, P., Smirnova, Tatiana., Zhou, Zhen Ni., Wyckoff, Jeffrey., Park, Hae In., Coniglio, Salvatore., Patel, Purvi., Qian, Bin-Zhi.,Stanley, E., Bresnick, Anne., Cox, Dianne., Pollard, Jeffrey., Muller, William., Condeelis, John., Segall, Jeffrey., *Contribution of CXCL12 secretion to invasion of breast cancer cells.* Breast Cancer Research, 2012. **14**(1): p. R23.
234. Slamon, D., Godolphin, W., Jones, LA., Holt, JA., Wong, SG., Keith, DE., Levin, WJ., Stuart, SG., Udove, J., Ullrich, A., et, al., *Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer.* Science, 1989. **244**(4905): p. 707-712.
235. Tomlinson, D.C., Knowles, M. A., Speirs, V., *Mechanisms of FGFR3 actions in endocrine resistant breast cancer.* International Journal of Cancer, 2011: p. n/a-n/a.
236. Hudlebusch, H.R., Skotte, Julie., Santoni-Rugiu, Eric., Zimling, Zarah Glad., Lees, Michael James., Simon, Ronald., Sauter, Guido., Rota, Rossella., De Ioris, Maria Antonietta., Quarto, Micaela., Johansen, Jens Vilstrup., Jørgensen, Mette., Rechnitzer, Catherine., Maroun, Lisa Leth., Schrøder, Henrik., Petersen, Bodil Laub., Helin, Kristian.,, *MMSET Is Highly Expressed and Associated with Aggressiveness in Neuroblastoma.* Cancer Research. **71**(12): p. 4226-4235.
237. Pawitan, Y., Bjohle, Judith., Amler, Lukas., Borg, Anna-Lena., Egyhazi, Suzanne., Hall, Per., Han, Xia,. Holmberg, Lars., Huang, Fei., Klaar, Sigrid., Liu, Edison.,Miller, Lance., Nordgren, Hans., Ploner, Alexander., Sandelin, Kerstin., Shaw, Peter.,Smeds, Johanna., Skoog, Lambert., Wedren, Sara., Bergh, Jonas., *Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in*

two population-based cohorts. Breast Cancer Research, 2005. **7**(6): p. R953 - R964.
238. Yang, P.-S., Yin, Pen-Hui., Tseng, Ling-Ming., Yang, Chin-Hua., Hsu, Chih-Yi., Lee, Ming-Yuan., Horng, Cheng-Fang., Chi, Chin-Wen., *Rab5A is associated with axillary lymph node metastasis in breast cancer patients.* Cancer Science, 2011. **102**(12): p. 2172-2178.
239. Shalek, A.K., Satija, Rahul., Adiconis, Xian., Gertner, Rona.S., Gaublomme, Jellert.T., Raychowdhury, Raktima., Schwartz, Schraga., Yosef, Nir., Malboeuf, Christine., Lu, Diana., Trombetta, John.J., Gennert, Dave., Gnirke, Andreas., Goren, Alon., Hacohen, Nir., Levin, Joshua.Z., Park, Hongkun., Regev, Aviv., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-240.
240. Gröne, J., Lenze, Dido.,Jurinovic, Vindi., Hummel, Manuela., Seidel, Henrik., Leder, Gabriele.,Beckmann, Georg.,Sommer, Anette.,Grützmann, Robert., Pilarsky, Christian., Mansmann, Ulrich.,Buhr, Heinz-Johannes., Stein, Harald., Hummel, Michael., *Molecular profiles and clinical outcome of stage UICC II colon cancer patients.* International Journal of Colorectal Disease, 2011. **26**(7): p. 847-858.
241. Hong, Y., Ho, Kok Sun., Eu, Kong Weng., Cheah, Peh Yean., *A Susceptibility Gene Set for Early Onset Colorectal Cancer That Integrates Diverse Signaling Pathways: Implication for Tumorigenesis.* Clinical Cancer Research, 2007. **13**(4): p. 1107-1114.
242. Gyorffy, B., Molnar, Bela., Lage, Hermann., Szallasi, Zoltan., Eklund, Aron C., *Evaluation of Microarray Preprocessing Algorithms Based on Concordance with RT-PCR in Clinical Samples.* PLos ONE, 2009. **4**(5): p. e5645.
243. Galamb, O., Spisak, S., Sipos, F., Toth, K., Solymosi, N., Wichmann, B., Krenacs, T., Valcz, G., Tulassay, Z., Molnar, B., *Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor.* Br J Cancer, 2010. **102**(4): p. 765-773.
244. Sabates-Bellver, J., Van der Flier, Laurens G., de Palo, Mariagrazia., Cattaneo, Elisa., Maake, Caroline., Rehrauer, Hubert., Laczko, Endre., Kurowski, Michal A., Bujnicki, Janusz M., Menigatti, Mirco., Luz, Judith., Ranalli, Teresa V., Gomes, Vito., Pastorelli, Alfredo., Faggiani, Roberto., Anti, Marcello., Jiricny, Josef., Clevers, Hans., Marra, Giancarlo., *Transcriptome Profile of Human Colorectal Adenomas.* Molecular Cancer Research, 2007. **5**(12): p. 1263-1275.
245. Jiang, X., Tan, Jing., Li, Jingsong., Kivimäe, Saul., Yang, Xiaojing., Zhuang, Li., Lee, Puay Leng., Chan, Mark T. W., Stanton, Lawrence W., Liu, Edison T., Cheyette, Benjamin N. R., Yu, Qiang., *DACT3 Is an Epigenetic Regulator of Wnt/β-Catenin Signaling in Colorectal Cancer and Is a Therapeutic Target of Histone Modifications.* Cancer Cell, 2008. **13**(6): p. 529-541.
246. Stefanska, B., Huang, Jian., Bhattacharyya, Bishnu., Suderman, Matthew., Hallett, Michael., Han, Ze-Guang., Szyf, Moshe., *Definition of the Landscape of Promoter DNA Hypomethylation in Liver Cancer.* Cancer Research, 2011. **71**(17): p. 5891-5903.
247. Seok, J.Y., Na, Deuk Chae., Woo, Hyun Goo., Roncalli, Massimo., Kwon, So Mee., Yoo, Jeong Eun., Ahn, Ei Yong., Kim, Gwang Il., Choi, Jin-Sub., Kim, Young Bae., Park, Young Nyun., *A fibrous stromal*

component in hepatocellular carcinoma reveals a cholangiocarcinoma-like gene expression trait and epithelial-mesenchymal transition. Hepatology, 2012. **55**(6): p. 1776-1786.

248. Sia, D., Hoshida, Yujin., Villanueva, Augusto., Roayaie, Sasan., Ferrer, Joana., Tabak, Barbara., Peix, Judit., Sole, Manel., Tovar, Victoria., Alsinet, Clara., Cornella, Helena., Klotzle, Brandy., Fan, Jian–Bing., Cotsoglou, Christian., Thung, Swan N., Fuster, Josep., Waxman, Samuel.,Garcia–Valdecasas, Juan Carlos., Bruix, Jordi., Schwartz, Myron E., Beroukhim, Rameen., Mazzaferro, Vincenzo., Llovet, Josep M., *Integrative Molecular Analysis of Intrahepatic Cholangiocarcinoma Reveals 2 Classes That Have Different Outcomes.* Gastroenterology, 2013. **144**(4): p. 829-840.
249. Coulouarn, C., Cavard,Catherine., Rubbia-Brandt, Laura., Audebourg,Anne., Dumont,Florent.,Jacques,Sébastien., Just,Pierre-Alexandre., Clément,Bruno., Gilgenkrantz,Hélène., Perret,Christine., Terris,Benoît., *Combined hepatocellular-cholangiocarcinomas exhibit progenitor features and activation of Wnt and TGFβ signaling pathways.* Carcinogenesis, 2012. **33**(9): p. 1791-1796.
250. Nissim, O., Melis,Marta., Diaz,Giacomo., Kleiner,David.E., Tice,Ashley., Fantola,Giovanni., Zamboni,Fausto., Mishra,Lopa., Farci,Patrizia., *Liver Regeneration Signature in Hepatitis B Virus (HBV)-Associated Acute Liver Failure Identified by Gene Expression Profiling.* PLos ONE, 2012. **7**(11): p. e49611.
251. Brase, J., Johannes,Marc., Mannsperger,Heiko., Falth,Maria., Metzger,Jennifer., Kacprzyk,Lukasz., Andrasiuk,Tatjana., Gade,Stephan., Meister,Michael., Sirma,Huseyin., Sauter,Guido., Simon,Ronald., Schlomm,Thorsten., BeiSZbarth,Tim., Korf,Ulrike., Kuner,Ruprecht., Sultmann,Holger., *TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-beta signaling.* BMC Cancer, 2011. **11**(1): p. 507.
252. Boormans, J.L., Korsten,Hanneke., Ziel-van der.Made, Angelique.J.C., van Leenders,Geert.J.L.H., de Vos,Carola.V., Jenster,Guido., Trapman,Jan.,, *Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer.* International Journal of Cancer, 2013: p. n/a-n/a.
253. Chandran, U., Ma,Changqing., Dhir,Rajiv., Bisceglia,Michelle., Lyons-Weiler,Maureen., Liang,Wenjing., Michalopoulos,George., Becich,Michael., Monzon,Federico.,, *Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process.* BMC Cancer, 2007. **7**(1): p. 64.
254. Sboner, A., Demichelis,Francesca., Calza,Stefano., Pawitan,Yudi., Setlur,Sunita., Hoshida,Yujin., Perner,Sven., Adami,Hans-Olov., Fall,Katja., Mucci,Lorelei., Kantoff,Philip., Stampfer,Meir., Andersson,Swen-Olof., Varenhorst,Eberhard., Johansson,Jan-Erik., Gerstein,Mark., Golub,Todd., Rubin,Mark., Andren,Ove., *Molecular sampling of prostate cancer: a dilemma for predicting disease progression.* BMC Medical Genomics, 2010. **3**(1): p. 8.
255. Taylor, B.S., Schultz,Nikolaus., Hieronymus,Haley., Gopalan,Anuradha., Xiao,Yonghong., Carver,Brett.S., Arora,Vivek.K., Kaushik,Poorvi., Cerami,Ethan., Reva,Boris., Antipin,Yevgeniy., Mitsiades,Nicholas., Landers,Thomas., Dolgalev,Igor., Major,John,E., Wilson,Manda., Socci, Nicholas.D., Lash,Alex.E., Heguy,Adriana., Eastham,James.A.,

Scher,Howard.I., Reuter,Victor.E., Scardino,Peter.T., Sander,Chris., Sawyers,Charles.L., Gerald,William.L., *Integrative Genomic Profiling of Human Prostate Cancer.* Cancer Cell, 2010. **18**(1): p. 11-22.
256. Creighton, C.J., Fountain,Michael.D., Yu,Zhifeng., Nagaraja,Ankur.K., Zhu,Huifeng., Khan,Mahjabeen., Olokpa,Emuejevoke., Zariff,Azam., Gunaratne,Preethi.H., Matzuk,Martin.M., Anderson,Matthew.L.,, *Molecular Profiling Uncovers a p53-Associated Role for MicroRNA-31 in Inhibiting the Proliferation of Serous Ovarian Carcinomas and Other Cancers.* Cancer Research, 2010. **70**(5): p. 1906-1915.
257. Marchion, D.C., Cottrill,Hope.M., Xiong,Yin., Chen,Ning., Bicaku,Elona., Fulp,William.J., Bansal,Nisha., Chon,Hye.Sook., Stickles,Xiaomang.B., Kamath,Siddharth.G., Hakam,Ardeshir., Li,Lihua., Su,Dan., Moreno,Carolina., Judson,Patricia.L., Berchuck,Andrew., Wenham,Robert.M., Apte,Sachin.M., Gonzalez-Bosquet,Jesus., Bloom, Gregory.C. Eschrich,Steven.A., Sebti,Said., Chen,Dung-Tsa., Lancaster,Johnathan.M., *BAD Phosphorylation Determines Ovarian Cancer Chemosensitivity and Patient Survival.* Clinical Cancer Research, 2011. **17**(19): p. 6356-6366.
258. Yamaguchi, K., Mandai,M., Oura,T., Matsumura,N., Hamanishi,J., Baba,T., Matsui,S., Murphy,S.K., Konishi,I., *Identification of an ovarian clear cell carcinoma gene signature that reflects inherent disease biology and the carcinogenic processes.* Oncogene, 2010. **29**(12): p. 1741-1752.
259. Stany, M.P., Vathipadiekal,Vinod., Ozbun,Laurent., Stone,Rebecca.L., Mok,Samuel.C., Xue,Hui., Kagami,Takashi., Wang,Yuwei., McAlpine,Jessica.N., Bowtell,David., Gout,Peter.W., Miller,Dianne.M., Gilks,C.Blake., Huntsman,David.G., Ellard,Susan.L., Wang,Yu-Zhuo., Vivas-Mejia,Pablo., Lopez-Berestein,Gabriel., Sood,Anil.K., Birrer,Michael.J., *Identification of Novel Therapeutic Targets in Microdissected Clear Cell Ovarian Cancers.* PLos ONE, 2011. **6**(7): p. e21121.
260. Ferriss, J.S., Kim,Youngchul., Duska,Linda ., Birrer,Michael., Levine,Douglas.A., Moskaluk,Christopher., Theodorescu,Dan., Lee,JaeK.,, *Multi-Gene Expression Predictors of Single Drug Responses to Adjuvant Chemotherapy in Ovarian Carcinoma: Predicting Platinum Resistance.* PLos ONE, 2012. **7**(2): p. e30550.
261. Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson Jr, J.A., Marks, J.R., Dressman, H.K., West, M., Nevins, J.R., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**: p. 353-7.
262. Tothill, R.W., Tinker,A.V., George,Joshy., Brown,Robert., Fox,S.B., Lade,Stephen., Johnson,D.S., Trivett, M.K., Etemadmoghadam,Dariush., Locandro,Bianca., Traficante,Nadia., Fereday,Sian., Hung,J.A., Chiew,Yoke-Eng., Haviv, Izhak., Australian Ovarian Cancer Study Group., Gertig,Dorota., deFazio,Anna., Bowtell, D.D.L., *Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome.* Clinical Cancer Research, 2008. **14**(16): p. 5198-5208.
263. Yoshihara, K., Tajima, Atsushi., Yahata, Tetsuro., Kodama, Shoji., Fujiwara, Hiroyuki., Suzuki, Mitsuaki., Onishi, Yoshitaka., Hatae, Masayuki., Sueyoshi, Kazunobu., Fujiwara, Hisaya., Kudo, Yoshiki., Kotera, Kohei., Masuzaki, Hideaki., Tashiro, Hironori., Katabuchi, Hidetaka., Inoue, Ituro., Tanaka, Kenichi., *Gene Expression Profile for*

Predicting Survival in Advanced-Stage Serous Ovarian Cancer Across Two Independent Datasets. PLos ONE, 2010. **5**(3): p. e9615.

264. Geng, H., Brennan, Sarah., Milne, T.A., Chen, Wei-Yi., Li, Yushan., Hurtz, Christian., Kweon, Soo-Mi., Zickl, Lynette., Shojaee, Seyedmehdi., Neuberg, Donna., Huang, Chuanxin., Biswas, Debabrata., Xin, Yuan., Racevskis, Janis., Ketterling, R.P., Luger, S.M., Lazarus, Hillard., Tallman, M.S., Rowe, J.M., Litzow, M.R., Guzman, M.L., Allis, C.D., Roeder, R.G., Müschen, Markus., Paietta, Elisabeth., Elemento, O., Melnick, A.M., *Integrative Epigenomic Analysis Identifies Biomarkers and Therapeutic Targets in Adult B-Acute Lymphoblastic Leukemia.* Cancer Discovery, 2012. **2**(11): p. 1004-1023.
265. Chuang, H.-Y., Rassenti, Laura., Salcedo, Michelle., Licon, Kate., Kohlmann, Alexander., Haferlach, Torsten., Foà, Robin., Ideker, Trey., Kipps, T.J., *Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression.* Blood, 2012. **120**(13): p. 2639-2649.
266. Sorich, M.J., Pottier, Nicolas., Pei, Deqing., Yang, Wenjian., Kager, Leo., Stocco, Gabriele., Cheng, Cheng., Panetta, J.C., Pui, Ching-Hon., Relling, M.V., Cheok, M.H., Evans, W.E., *In Vivo Response to Methotrexate Forecasts Outcome of Acute Lymphoblastic Leukemia and Has a Distinct Gene Expression Profile.* PLoS Med, 2008. **5**(4): p. e83.
267. Tomasson, M.H., Xiang, Zhifu., Walgren, Richard., Zhao, Yu., Kasai, Yumi., Miner, Tracie., Ries, R.E., Lubman, Olga., Fremont, D.H., McLellan, M.D., Payton, J.E., Westervelt, Peter., DiPersio, J.F., Link, D.C., Walter, M.J., Graubert, T.A., Watson, Mark., Baty, Jack., Heath, Sharon., Shannon, W.D., Nagarajan, Rakesh., Bloomfield, C.D., Mardis, E.R., Wilson, R.K.. Ley, T.J., *Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia.* Blood, 2008. **111**(9): p. 4797-4808.
268. Mullighan, C.G., Su, Xiaoping., Zhang, Jinghui., Radtke, Ina., Phillips, Letha A.A., Miller, Christopher B., Ma, Jing., Liu, Wei., Cheng, Cheng., Schulman, Brenda A., Harvey, Richard C., Chen, I-Ming., Clifford, Robert J., Carroll, William L., Reaman, Gregory., Bowman, W. Paul., Devidas, Meenakshi., Gerhard, Daniela S., Yang, Wenjian., Relling, Mary V., Shurtleff, Sheila A., Campana, Dario., Borowitz, Michael J., Pui, Ching-Hon., Smith, Malcolm., Hunger, Stephen P., Willman, Cheryl L., Downing, James R., *Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia.* New England Journal of Medicine, 2009. **360**(5): p. 470-480.
269. Klein, H.-U., Ruckert, Christian., Kohlmann, Alexander., Bullinger, Lars., Thiede, Christian., Haferlach, Torsten., Dugas, Martin., *Quantitative comparison of microarray experiments with published leukemia related gene expression signatures.* BMC Bioinformatics, 2009. **10**(1): p. 422.
270. Balgobind, B.V., Van den Heuvel-Eibrink, Marry M., De Menezes, Renee X., Reinhardt, Dirk., Hollink, Iris H.I.M., Arentsen-Peters, Susan T.J.C.M., van Wering, Elisabeth R., Kaspers, Gertjan J.L., Cloos, Jacqueline., de Bont, Evelien S.J.M., Cayuela, Jean-Michel., Baruchel, Andre., Meyer, Claus., Marschalek, Rolf., Trka, Jan., Stary, Jan., Beverloo, H. Berna., Pieters, Rob., Zwaan, C. Michel., den Boer, Monique L., *Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia.* Haematologica, 2011. **96**(2): p. 221-230.

271. Venables, W.N., Ripley, B.D., *Modern Applied Statistics with S*. 4th ed. 2002, NewYork: Springer.
272. Chen, J., Wang, Ying., Shen, Bairong., Zhang, Daqing., *Molecular Signature of Cancer at Gene Level or Pathway Level? Case Studies of Colorectal Cancer and Prostate Cancer Microarray Data.* Computational and Mathematical Methods in Medicine, 2012. **2012**: p. 8.
273. Cahan, P., Rovegno, Felicia., Mooney, Denise., Newman, John C., St. Laurent Iii, Georges., McCaffrey, Timothy A., *Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization.* Gene, 2007. **401**(1–2): p. 12-18.
274. Xu, L., Geman, Donald., Winslow, Raimond., *Large-scale integration of cancer microarray data identifies a robust common cancer signature.* BMC Bioinformatics, 2007. **8**(1): p. 275.
275. Tang, B., Wu, Xuechen., Tan, Ge., Chen, Su-Shing., Jing, Qing., Shen, Bairong., *Computational inference and analysis of genetic regulatory networks via a supervised combinatorial-optimization pattern.* BMC Systems Biology, 2010. **4**(Suppl 2): p. S3.
276. Ren, X., Fu, H., & Jin, Q. , *Integrating heterogeneous genomic data to accurately identify disease subtypes. .* BMC Medical Genomics, 2015. **8**(78).
277. Edgar, R., Domrachev, M., Lash, AE.,, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucl. Acids Res, 2002. **30**(1): p. 207-210.
278. Larsson, O., Sandberg, Rickard., *Lack of correct data format and comparability limits future integrative microarray research.* Nat Biotech, 2006. **24**(11): p. 1322-1323.
279. Hamid, J.S., Hu, Pingzhao., Roslin, Nicole.M., Ling, Vicki., Greenwood, Celia.M.T., Beyene, Joseph., *Data Integration in Genetics and Genomics: Methods and Challenges.* Human Genomics and Proteomics, 2009. **1**(1).
280. Tseng, G.C., Ghosh, Debashis., Feingold, Eleanor., *Comprehensive literature review and statistical considerations for microarray meta-analysis.* Nucleic Acids Research, 2012.
281. Kumar, S., Chintanu., Samarasinghe, Sandhya., *Microarray Data Integration: Frameworks and a List of Underlying Issues.* Current Bioinformatics, 2010. **5**(4): p. 280-289.
282. Bergmann, S., Ihmels, J., Barkai, N., *Similarities and differences in genome-wide expression data of six organisms.* PLoS Biol, 2004. **2**: p. e9.
283. Goods, I.J., *On the Weighted Combination of Significance Tests.* Journal of the Royal Statistical Society; Series B, 1955. **17**(17): p. 264-265.
284. Olkin, I., Saner, Hilary., *Approximations for trimmed Fisher procedures in research synthesis.* Statistical Methods in Medical Research, 2001. **10**(4): p. 267-276.
285. Tippett, L.H.C., *The methods of statistics*. 1931, Williams and Norgate: London.
286. Taminau, J., Lazar, Cosmin., Meganck, Stijn., Nowé, Ann., *Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis.* ISRN Bioinformatics, 2014. **2014**: p. 7.
287. Wilkinson, B., *A statistical consideration in psychologial research.* Psych. Bull, 1951. **48**: p. 156-158.

288. Song, C., Tseng, George.C. , *Hypothesis setting and order statistic for robust genomic meta-analysis.* The Annals of Applied Statistics, 2014. **8**(2): p. 777-800.
289. Smid, M., Dorssers, Lambert C.J., Jenster, Guido., *Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes.* Bioinformatics, 2003. **19**(16): p. 2065-2071.
290. Irizarry, R.A., Bolstad, Benjamin.M., Collin, Francois., Cope, Leslie.M., Hobbs, Bridget., Speed, Terence.P., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Research, 2003. **31**(4): p. e15.
291. Warnat, P., Eils, Roland., Brors, Benedikt,, *Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.* BMC Bioinformatics, 2005. **6**(1): p. 265.
292. Shabalin, A.A., Tjelmeland, Håkon., Fan, Cheng., Perou, Charles.M., Nobel, Andrew.B., *Merging two gene-expression studies via cross-platform normalization.* Bioinformatics, 2008. **24**(9): p. 1154-1160.
293. Cheng, C., Shen, Kui., Song, Chi., Luo, Jianhua., Tseng, George.C., *Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction.* Bioinformatics, 2009. **25**(13): p. 1655-1661.
294. Ramaswamy, S., Ross, K.N., Lander, E.S., Golub, T.R., *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**: p. 49-54.
295. Shen, R., Ghosh, Debashis., Chinnaiyan, Arul., *Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.* BMC Genomics, 2004. **5**(1): p. 94.
296. Choi, H., Shen, Ronglai., Chinnaiyan, Arul., Ghosh, Debashis., *A Latent Variable Approach for Meta-Analysis of Gene Expression Data from Multiple Microarray Experiments.* BMC Bioinformatics, 2007. **8**(1): p. 364.
297. Jiang, H., Deng, Y., Chen, H., Tao, L., Sha, Q., Chen, J., Tsai, C., Zhang, S., *Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.* BMC Bioinformatics, 2004. **5**: p. 81.
298. Xiong, H., Zhang, Ya.,Chen, Xue-Wen., Yu, Jiangsheng., *Cross-platform microarray data integration using the Normalised Linear Transform.* Int. J. Data Min. Bioinformatics, 2010. **4**(2): p. 142-157.
299. Guerra, R., Goldstein, Darlene R.,, *Meta-analysis and combining information in genetics and genomics*. 2009: CRC Press.
300. Chen, C., Grennan, Kay., Badner, Judith., Zhang, Dandan., Gershon, Elliot., Jin, Li., Liu, Chunyu., *Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods.* PLos ONE, 2011. **6**(2): p. e17238.
301. Culhane, A., Perriere, Guy., Higgins, Desmond., *Cross-platform comparison and visualisation of gene expression data using co-inertia analysis.* BMC Bioinformatics, 2003. **4**(1): p. 59.
302. Tomescu, O., Mattanovich, Diethard., Thallinger, Gerhard., *Integrative omics analysis. A study based on Plasmodium falciparum mRNA and protein data.* BMC Systems Biology, 2014. **8**(Suppl 2): p. S4.
303. Fagan, A., Culhane, Aedín.C., Higgins, Desmond.G., *A multivariate analysis approach to the integration of proteomic and gene expression data.* Proteomics, 2007. **7**(13): p. 2162-2171.

304. Alter, O., Brown, P.O., Botstein, D., *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.* PNAS, 2003. **100**: p. 3351-6.
305. Monti, S., Tamayo, Pablo., Mesirov, Jill,. Golub, Todd.,, *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Machine Learning, 2003. **52**(1): p. 91-118.
306. Gao, Y., Church, George.,, *Improving molecular cancer class discovery through sparse non-negative matrix factorization.* Bioinformatics, 2005. **21**(21): p. 3970-3975.
307. Kim, P.M., Tidor, Bruce.,, *Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data.* Genome Research, 2003. **13**(7): p. 1706-1718.
308. Savage, R.S., Ghahramani, Zoubin., Griffin, Jim.E., de la Cruz, Bernard.J., Wild, David.L., *Discovering transcriptional modules by Bayesian data integration.* Bioinformatics, 2010. **26**(12): p. i158-i167.
309. Kirk, P., Griffin, Jim.E. Savage, Richard.S. Ghahramani, Zoubin., Wild, David.L.,, *Bayesian correlated clustering to integrate multiple datasets.* Bioinformatics, 2012. **28**(24): p. 3290-3297.
310. Lock, E.F., Dunson, David.B., *Bayesian consensus clustering.* Bioinformatics, 2013. **29**(20): p. 2610-2616.
311. Cai, J., Xie, Dan., Fan, Zhewen., Chipperfield, Hiram., Marden, John., Wong, Wing H., Zhong, Sheng., *Modeling Co-Expression across Species for Complex Traits: Insights to the Difference of Human and Mouse Embryonic Stem Cells.* PLoS Comput Biol, 2010. **6**(3): p. e1000707.
312. Efron, B., *Correlation and Large-Scale Simultaneous Significance Testing.* Journal of the American Statistical Association, 2007. **102**(477): p. 93-103.
313. Zhang, J., Li, Jian., Deng, Hongwen., *Class-Specific Correlations of Gene Expressions: Identification and Their Effects on Clustering Analyses.* The American Journal of Human Genetics, 2008. **83**(2): p. 269-277.
314. Lonnstedt, I. and T. Speed, *Replicated microarray data.* Statistica Sinica, 2002. **12**: p. 31 - 46.
315. Nykter, M., Aho, Tommi., Ahdesmaki, Miika., Ruusuvuori, Pekka., Lehmussola, Antti., Yli-Harja, Olli, *Simulation of microarray data with realistic characteristics.* BMC Bioinformatics, 2006. **7**(1): p. 349.
316. Jeanmougin, M., deReynies, Aurelien., Marisa, Laetitia., Paccard, Caroline., Nuel, Gregory., Guedj, Mickael., *Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies.* PLos ONE, 2010. **5**(9): p. e12336.
317. Qin, L.-X., *An Integrative Analysis of microRNA and mRNA Expression - A Case Study.* Cancer Informatics, 2008. **6**(CIN-6-Qin-(Li-xuan)): p. 0-0.
318. Lee, H., Kong, Sek.Won., Park, Peter.J., *Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes.* Bioinformatics, 2008. **24**(7): p. 889-896.
319. Guangchuang, Y., Li-Gen, Wang., Yanyan, Han., Qing-Yu, He., *clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.* OMICS: A Journal of Integrative Biology, 2012. **16**(5): p. 284-287.
320. Datta, S., Datta, Somnath., *Evaluation of clustering algorithms for gene expression data.* BMC Bioinformatics, 2006. **7**(Suppl 4): p. S17.

321. Brock, G., Pihur, Vasyl., Datta, Susmita., Datta, Somnath., *clValid: an R package for cluster validation.* Journal of Statistical Software, 2008. **25**(4).
322. Zheng-Bradley, X., Rung, J, Parkinson, H, Brazma, A, *Large scale comparison of global gene expression patterns in human and mouse.* Genome Biol, 2010. **11**: p. r124.
323. Ala, U., Piro, RM, Grassi, E, Damasco, C, Silengo, L, Oti, M, Provero, P, Di Cunto, F, *Prediction of human disease genes by human-mouse conserved coexpression analysis.* PLoS Comput Biol, 2009. **4**: p. e1000043.
324. Sweet-Cordero, A., Mukherjee, S, You, ASH, Roix, JJ, Ladd-Acosta, C, Mesirov, J, Golub, TR, Jacks, T, *An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis.* Nat Genet, 2005. **37**: p. 48-55.
325. Miller, J., Horvath, S, Geschwind, DH, *Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways.* PNAS, 2010. **107**: p. 220-229.
326. Rasche, A., Al-Hasani, H, Herwig, R, *Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 Diabetes mellitus.* BMC Genomics, 2008. **9**: p. 310.
327. Segal, E., Friedman. N, Kaminski, N, Regev, A, Koller, D, *From signatures to models: understanding cancer using microarrays.* Nat Genet, 2005. **37**: p. S38-45.
328. Allison, D., Cui, X, Page, GP, Sabripour, M, *Microarray data analysis: from disarray to consolidation and consensus.* Nat Rev Genet, 2006. **14**: p. 55-6.
329. Consortium, M., *The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **14**: p. 1151-61.
330. Kuo, W., Liu, F, Trimarchi, J, Punzo, C, Lombardi, M, Sarang, J, Whipple, ME, Maysuria, M, Serikawa, K, Lee, SY, McCrann, D, Kang, J, Shearstone, JR, Burke, J, Park, DJ, Wang, X, Rector, TL, Ricciardi-Castagnoli, P, Perrin, S, Choi, S, Bumgarner, R, Kim, JH, III, GFS, Freeman, MW, Seed, B, Jensen, R, Church, GM, Hovig, E, Cepko, CL, Park, P, Ohno-Machado, L, Jenssen, TK, *A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies.* Nat Biotechnol, 2006. **14**: p. 832-40.
331. Tabakoff, B., Saba, L, Printz, M, Flodman, P, Hodgkinson, C, Goldman, D, Koob, G, Richardson, HN, Kechris, K, Bell, RL, Hübner, N, Heinig, M, Pravenec, M, Mangion, J, Legault, L, Dongier, M, Conigrave, KM, Whitfield, JB, Saunders, J, Grant, B, Hoffman, PL, *Genetical genomic determinants of alcohol consumption in rats and humans.* BMC Biol, 2009. **7**: p. 70.
332. Ma, S., Grigoryev, DN, Taylor, AD, Nonas, S, Sammani, S, Ye, SQ, Garcia, JG, *Bioinformatic identification of novel early stress response genes in rodent models of lung injury.* Am J Physiol Lung Cell Mol Physiol, 2005. **289**: p. L468-77.
333. Liao, B., Zhang, JZ, *Evolutionary conservation of expression profiles between human and mouse orthologous genes.* Mol Biol Evol, 2006. **23**: p. 530-40.

334. Dutilh, B., Huynen, MA, Snel, B, *A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation.* BMC Genomics, 2006. **7**: p. 10.
335. Essien, K., Hannenhalli, S, Stoeckert, CJ, *Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to Plasmodium phenotypic diversity.* PLoS One, 2008. **3**: p. e3122.
336. Fisher, R., *Answer to question 14 on combining independent tests of significance.* Amer Statistician, 1948. **2**: p. 30.
337. Hu, P., Greenwood, CMT, Beyene, J, *Statistical methods for meta-analysis of microarray data: a comparative study.* Inf Syst Front, 2006. **8**: p. 9-20.
338. Campain, A., Yang, YH, *Comparison study of microarray meta-analysis methods.* BMC Bioinformatics, 2010. **3**: p. 408.
339. Tseng, G., Ghosh, D, Feingold, E, *Comprehensive literature review and statistical considerations for microarray meta-analysis.* Nucleic Acids Res, 2012. **40**: p. 3785-99.
340. Rice, W., *A consensus combined P-value test and the family-wide signi_cance of component tests.* Biometrics, 1990. **46** p. 303-8.
341. Rumelhart, D.E., McClelland, J.L, *Parallel Dsitributed Processing*. 1986, Cambridge, MA, USA: MIT press.
342. Li, H., Zhan, M, *Identifying Conserved and Divergent Transcriptional Modules by Cross-species Matrix Decomposition on Microarray Data.* J Proteomics Bioinform, 2009. **2**: p. 117.
343. Hu, P., Wang, X, Haitsma, JJ, Furmli, S, Masoom, H, Liu, M, Imai, Y, Slutsky, AS, Beyene, J, Greenwood, CM, dos Santos, C, *Microarray meta-analysis identifies acute lung injury biomarkers in donor lungs that predict development of primary graft failure in recipients.* PLoS One, 2012. **7**: p. e45506.
344. Royce, T., Rozowsky, JS, Gerstein, MB, *Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification.* Nucleic Acids Res, 2007. **35**: p. e99.
345. Cai, J., Xie, D, Fan, Z, Chipperfield, H, Marden, J, Wong, WH, Zhong, S, *Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells.* PLoS Comput Biol, 2010. **6**: p. e1000707.
346. Schlicht, M., Matysiak, B, Brodzeller, T, Wen, X, Liu, H, Zhou, G, Dhir, R, Hessner, MJ, Tonellato, P, Suckow, M, Pollard, M, Datta, MW, *Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium.* BMC Genomics, 2004. **5**: p. 58.
347. Lu, Y., He, X., Zhong, S., *Cross-species microarray analysis with the OSCAR system suggests an INSR->Pax6->NQO1 neuro-protective pathway in aging and Alzheimer's disease.* NAR, 2007. **35**: p. W105–14.
348. Huang, W., Cao, X, Zhong, S, *Network-based comparison of temporal gene expression patterns.* Bioinformatics, 2010. **26**: p. 2944-51.
349. Xiao, Y., Hsiao, T.H., Suresh, U., Chen, H.I., Wu, X., Wolf, S.E., Chen, Y., *A Novel Significance Score for Gene Selection and Ranking.* Bioinformatics, 2012. **in press**.
350. McCarthy, D., Smyth, GK, *Testing significance relative to a fold-change threshold is a TREAT.* Bioinformatics, 2009. **25**: p. 765–71.

351. Tusher, V., Tibshirani, R, Chu, G, *Significance analysis of microarrays applied to the ionizing radiation response.* PNAS, 2001. **98**: p. 5116-21.
352. Vogel, P., Read, RW, Rehg, JE, Hansen, GM, *Cryptogenic organizing pneumonia in Tomm5(-/-) mice.* Vet Pathol, 2013. **50**: p. 65-75.
353. Fresquet, V., Robles, EF, Parker, A, et al, *High-throughput sequencing analysis of the chromosome 7q32 deletion reveals IRF5 as a potential tumour suppressor in splenic marginal-zone lymphoma.* Br J Haematol, 2012. **158**: p. 712-26.
354. Nakayama, Y., Iwamoto, Y, Maher, SE, Tanaka, Y, Bothwell, AL, *Altered gene expression upon BCR cross-linking in Burkitt's lymphoma B cell line.* Biochem Biophys Res Commun, 2000. **277**: p. 124-7.
355. Turtoi, A., Sharan, RN, Srivastava, A, Schneeweiss, FH, *Proteomic and genomic modulations induced by 3B3;-irradiation of human blood lymphocytes.* Int J Radiat Biol, 2010. **86**: p. 888-904.
356. Rana, S., Maples, PB, Senzer, N, Nemunaitis, J, *Stathmin 1: a novel therapeutic target for anticancer activity.* Expert Rev Anticancer Ther, 2008. **8**: p. 1461-70.
357. Akerley, B., Rubin, EJ, Camilli, A, Lampe, DJ, Robertson, HM, Mekalanos, JJ, *Systematic identification of essential genes by in vitro mariner mutagenesis.* PNAS, 1998. **95**: p. 8927-32.
358. Griffin, J., Gawronski, JD, Dejesus, MA, Ioerger, TR, Akerley, BJ, Sassetti, CM, *High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism.* PLoS Pathog, 2011. **7**: p. e1002251.
359. McLachlan, G.J., Peel, D., *Finite Mixture Models*. 2000: Wiley
360. Frei, R., Schultz, GA, Church, RB, *Qualitative and quantitative changes in protein synthesis occur at the 8-16-cell stage of embryogenesis in the cow.* Journal of Reproduction & Fertility, 1989. **86**: p. 637-41.
361. Siddiqui, N., Li, X, Luo, H, Karaiskakis, A, Hou, H, Kislinger, T, Westwood, JT, Morris, Q, Lipshitz, HD, *Genome-wide analysis of the maternal-to-zygotic transition in Drosophila primordial germ cells.* Genome Biology, 2012. **13**: p. R11.
362. Lei, L., Xukun, L, Jurrien, D, *The maternal to zygotic transition in mammals.* Molecular Aspects of Medicine, 2013. **34**: p. 919-38.
363. Bultman, S., Gebuhr, TC, Pan, H, Svoboda, P, Schultz, RM, Magnuson, T, *Maternal BRG1 regulates zygotic genome activation in the mouse.* Genes & Development, 2006. **20**: p. 1744-54.
364. Jarrell, V., Day, BN, Prather, RS, *The Transition from Maternal to Zygotic Control of Development Occurs during the 4-Cell Stage in the Domestic Pig, Sus scrofa: Quantitative and Qualitative Aspects of Protein Synthesis.* Biology of Reproduction, 1991. **44**: p. 62-8.
365. Stanton, J., Macgregor, AB, Green, DP, *Gene expression in the mouse preimplantation embryo.* Reproduction, 2003. **125**: p. 457-68.
366. Thompson, J., Partridge, RJ, Houghton, FD, Cox, CI, Leese, HJ, *Oxygen uptake and carbohydrate metabolism by in vitro derived bovine embryos.* Journal of Reproduction & Fertility, 1996. **106**: p. 299-306.
367. Augustin, R., Pocar, P, Navarrete-Santos, A, Wrenzycki, C, Gandolfi, F, Niemann, H, Fischer, B, *Glucose Transporter Expression is Developmentally Regulated in In Vitro Derived Bovine Preimplantation Embryos.* Molecular Reproduction and Development, 2001. **60**: p. 370-6.

368. Markadieu, N., Delpire, E, *Physiology and pathophysiology of SLC12A1/2 transporters.* European Journal of Physiology, 2014. **466**: p. 91-105.
369. Gardner, D., Leese, HJ, *The role of glucose and pyruvate transport in regulating nutrient utilization by preimplantation mouse embryos.* Development, 1988. **104**: p. 423-9.
370. Ibla, J., Khoury, J, Kong, T, *Transcriptional repression of Na-K-2Cl cotransporter NKCC1 by hypoxia-inducible factor-1.* American Journal of Physiology – Cell Physiology, 2006. **291**: p. C282-9.
371. Wu, G., Bazer, FW, Wallace, JM, Spencer, TE, *Intrauterine growth retardation: Implications for the animal sciences.* J ANIM SCI, 2006. **84**: p. 2316-37.
372. Gao, H., Wu, G, Spencer, TE, Johnson, GA, Li, X, Bazer, FW, *Select Nutrients in the Ovine Uterine Lumen. I. Amino Acids, Glucose, and Ions in Uterine Lumenal Flushings of Cyclic and Pregnant Ewes.* Biology of Reproduction, 2009. **80**: p. 86-93.
373. Belkacemi, L., Nelson, DM, Desai, M, Ross, MG, *Maternal Undernutrition Influences Placental-Fetal Development.* Biology of Reproduction, 2010. **83**: p. 325-31.
374. Gwatkin, R., *Amino acid requirements for attachment and outgrowth of the mouse blastocyst in vitro.* Journal of Cellular Physiololgy, 1966. **68**: p. 335-43.

Supplementary

Table S2.1. Top 20 genes predicted by four algorithms for GDS3138. The *p* values for eBayes cyber and SAM algorithms and the null probabilities of MSG algorithms. The first column is for the probe ID and Gene symbol

probe#symbol	eBayes	cyber	SAM	MSG
1559646_a_at#NCRNA00184	0.000378546	0	1.83E-05	1.70E-29
233626_at#AK024580	0.000504021	6.73E-12	1.83E-05	2.85E-27
202728_s_at#LTBP1	0.000570842	1.10E-09	1.83E-05	2.50E-22
224588_at#XIST	0.000634478	0	1.83E-05	1.72E-39
1561245_at#BC040302	0.000672461	2.95E-14	1.83E-05	2.62E-20
1552579_a_at#ADAM21	0.000680944	2.27E-11	1.83E-05	4.41E-17
208395_s_at#URB1	0.000691311	5.55E-16	1.83E-05	7.87E-21
229275_at#IGFN1	0.000729201	1.44E-15	1.83E-05	9.92E-20
213432_at#MUC5B	0.000810501	2.46E-08	1.83E-05	2.39E-18
209530_at#CACNB3	0.000881455	4.76E-12	2.74E-05	3.58E-21
219643_at#LRP1B	0.00093325	4.52E-09	2.74E-05	1.07E-13
228100_at#C1orf88	0.000956674	7.04E-08	2.74E-05	1.24E-13
242830_at#AI092709	0.000976669	3.34E-08	2.74E-05	4.03E-13
224590_at#XIST	0.000987926	0	1.83E-05	2.72E-35
220784_s_at#UTS2	0.001015663	9.93E-07	4.57E-05	1.77E-13
217439_at#AL122122	0.001046902	1.06E-10	5.49E-05	6.26E-14
204750_s_at#DSC2	0.001050127	3.88E-07	2.74E-05	3.30E-12
1554406_a_at#CLEC7A	0.001070305	3.42E-06	1.83E-05	6.11E-12
219341_at#CLN8	0.001124405	8.29E-13	0.000101	3.74E-15
206534_at#GRIN2A	0.001142977	6.23E-08	5.49E-05	5.76E-13
1553102_a_at#CCDC69	0.009305337	0	0.011294	0.007066
1553575_at#ND6	0.012903396	0	0.016049	0.025642
1554462_a_at#DNAJB9	0.009282069	0	0.010947	0.021451
1554711_at#CALHM3	0.003893065	0	0.00267	0.00015

1555037_a_at#IDH1	0.021629796	0	0.028404	0.138832
1555536_at#ANTXR2	0.039657922	0	0.032968	0.006456
1555673_at#LOC730755	0.013200221	0	0.017055	0.017248
1555764_s_at#TIMM10	0.023415629	0	0.027179	0.168819
200632_s_at#NDRG1	0.009548824	0	0.011559	0.01949
200644_at#MARCKSL1	0.016277437	0	0.019497	0.158759
200696_s_at#GSN	0.008144209	0	0.007791	0.019763
200799_at#HSPA1B	0.051451304	0	0.057202	0.213425
200825_s_at#HYOU1	0.025418387	0	0.026145	0.360469
200831_s_at#SCD	0.003695327	0	0.002734	1.37E-05
200832_s_at#SCD	0.008497816	0	0.009803	0.002794
200875_s_at#NOP56	0.026616607	0	0.027901	0.218476
201015_s_at#JUP	0.006542058	0	0.007124	0.000804
201147_s_at#TIMP3	0.009283089	0	0.007819	0.036568
201148_s_at#TIMP3	0.008835346	0	0.010572	0.011281
1568871_at#BC032557	0.00120076	1.31E-09	2.74E-05	4.52E-12
1561564_at#BC030741	0.001157108	6.79E-08	3.66E-05	1.68E-12
1569719_at#BCL2L14	0.001327253	4.81E-05	5.49E-05	3.38E-11
221900_at#COL8A2	0.005801259	5.55E-16	0.001683	2.91E-22
223749_at#C1QTNF2	0.001333088	1.57E-12	0.000119	1.65E-21
213668_s_at#SOX4	0.00523325	1.75E-12	0.001573	4.31E-20
1556760_a_at#R16907	0.015633875	4.31E-14	0.006392	4.06E-19
232417_x_at#ZDHHC11	0.003449768	4.94E-14	0.001097	1.41E-17
205875_s_at#ATRIP	0.005874117	7.01E-10	0.002341	2.13E-17
217270_s_at#DYRK1B	0.004206346	4.38E-09	0.001372	2.32E-17
207838_x_at#PBXIP1	0.006526518	4.03E-08	0.002588	2.72E-16
227082_at#AI760356	0.002264103	7.67E-11	0.000631	7.56E-16

Table S2.2. Top 20 genes predicted by each of the four algorithms for GDS2865. The *p* values for eBayes cyber and SAM algorithms and the null probabilities for MSG algorithms. The first column is for the probe ID and Gene symbol

probe#symbol	eBayes	cyber	SAM	MSG
201110_s_at#THBS1	1.13E-07	0	4.49E-05	3.22E-117
209839_at#DNM3	3.46E-07	0	4.49E-05	4.37E-182
205523_at#HAPLN1	5.28E-07	0	4.49E-05	1.01E-77
209942_x_at#MAGEA3	7.15E-07	0	4.49E-05	1.79E-97
200665_s_at#SPARC	9.74E-07	0	4.49E-05	1.54E-126
220784_s_at#UTS2	1.02E-06	0	4.49E-05	0
204455_at#DST	1.95E-06	0	4.49E-05	9.03E-49
204932_at#TNFRSF11B	2.40E-06	0	5.61E-05	3.43E-127
214612_x_at#MAGEA2B	2.90E-06	0	6.73E-05	5.88E-101
211675_s_at#MDFIC	3.41E-06	0	4.49E-05	2.35E-54
219936_s_at#GPR87	4.00E-06	0	4.49E-05	1.85E-46
214476_at#TFF2	4.97E-06	0	4.49E-05	3.03E-31
204933_s_at#TNFRSF11B	6.05E-06	0	6.73E-05	1.72E-100
213273_at#ODZ4	6.33E-06	0	6.73E-05	8.48E-40
209488_s_at#RBPMS	7.51E-06	0	6.73E-05	8.81E-51
206710_s_at#EPB41L3	8.01E-06	0	6.73E-05	1.59E-37
201884_at#CEACAM5	8.27E-06	0	7.85E-05	7.92E-97
209487_at#RBPMS	8.56E-06	0	6.73E-05	6.97E-51
205524_s_at#HAPLN1	8.79E-06	2.22E-16	4.49E-05	6.24E-30
212681_at#EPB41L3	8.92E-06	0	6.73E-05	6.61E-37
201042_at#TGM2	0.000585	0	0.001683	4.19E-63
201108_s_at#THBS1	1.80E-05	0	7.85E-05	1.33E-52
201109_s_at#THBS1	3.73E-05	0	0.000146	5.44E-96
201242_s_at#ATP1B1	1.48E-05	0	7.85E-05	5.75E-48
201243_s_at#ATP1B1	0.000413	0	0.001313	9.01E-34

201362_at#IVNS1ABP	0.000566	0	0.001705	2.31E-30
201667_at#GJA1	9.40E-05	0	0.000348	1.73E-54
202625_at#LYN	0.0005	0	0.0012	7.51E-12
202626_s_at#LYN	1.22E-05	0	5.61E-05	2.13E-25
203021_at#SLPI	1.31E-05	0	7.85E-05	1.10E-74
203691_at#PI3	2.58E-05	0	8.98E-05	7.58E-47
203757_s_at#CEACAM6	8.97E-06	0	6.73E-05	1.69E-54
203889_at#SCG5	0.000114	0	0.00037	1.38E-38
203980_at#FABP4	1.03E-05	0	6.73E-05	5.45E-30
204597_x_at#STC1	7.01E-05	0	0.000247	3.58E-31
204620_s_at#VCAN	0.000111	0	0.000213	3.67E-15
200696_s_at#GSN	0.000131	4.20E-08	4.49E-05	2.52E-08
203180_at#ALDH1A3	1.20E-05	2.93E-13	4.49E-05	1.24E-23
205680_at#MMP10	4.56E-05	1.82E-09	4.49E-05	1.25E-13
208078_s_at#SIK1	1.99E-05	3.44E-11	4.49E-05	6.81E-18
214496_x_at#MYST4	3.09E-05	4.70E-10	4.49E-05	1.14E-15
216316_x_at#GK3P	4.72E-05	1.06E-12	4.49E-05	1.04E-13
219304_s_at#PDGFD	2.46E-05	2.81E-11	4.49E-05	3.28E-17
221805_at#NEFL	1.36E-05	0	4.49E-05	2.97E-23
212667_at#SPARC	1.02E-05	0	7.85E-05	2.34E-63
211657_at#CEACAM6	3.40E-05	0	0.000135	3.04E-63
208116_s_at#MAN1A1	5.87E-05	0	0.000236	1.90E-57
209035_at#MDK	1.74E-05	0	7.85E-05	5.64E-56

Table S2.3, The genes which have zero *p* values given by Cyber-T. The table shows the p values and the null probabilities corresponding to the zero p genes identified by cyber for the prostate cancer data set, GDS2865.

	eBayes	cyber	SAM	MSG
200665_s_at#SPARC	9.74E-07	0	4.49E-05	1.54E-126
201042_at#TGM2	0.000585	0	0.001683	4.19E-63
201108_s_at#THBS1	1.80E-05	0	7.85E-05	1.33E-52
201109_s_at#THBS1	3.73E-05	0	0.000146	5.44E-96
201110_s_at#THBS1	1.13E-07	0	4.49E-05	3.22E-117
201242_s_at#ATP1B1	1.48E-05	0	7.85E-05	5.75E-48
201243_s_at#ATP1B1	0.000413	0	0.001313	9.01E-34
201362_at#IVNS1ABP	0.000566	0	0.001705	2.31E-30
201667_at#GJA1	9.40E-05	0	0.000348	1.73E-54
201884_at#CEACAM5	8.27E-06	0	7.85E-05	7.92E-97
202625_at#LYN	0.0005	0	0.0012	7.51E-12
202626_s_at#LYN	1.22E-05	0	5.61E-05	2.13E-25
203021_at#SLPI	1.31E-05	0	7.85E-05	1.10E-74
203691_at#PI3	2.58E-05	0	8.98E-05	7.58E-47
203757_s_at#CEACAM6	8.97E-06	0	6.73E-05	1.69E-54
203889_at#SCG5	0.000114	0	0.00037	1.38E-38
203980_at#FABP4	1.03E-05	0	6.73E-05	5.45E-30
204455_at#DST	1.95E-06	0	4.49E-05	9.03E-49
204597_x_at#STC1	7.01E-05	0	0.000247	3.58E-31
204620_s_at#VCAN	0.000111	0	0.000213	3.67E-15
204749_at#NAP1L3	0.000231	0	0.000673	2.29E-48
204751_x_at#DSC2	6.29E-05	0	0.000247	2.92E-48
204818_at#HSD17B2	1.00E-05	0	7.85E-05	5.52E-52
204932_at#TNFRSF11B	2.40E-06	0	5.61E-05	3.43E-127

204933_s_at#TNFRSF11B	6.05E-06	0	6.73E-05	1.72E-100
204990_s_at#ITGB4	8.91E-05	0	0.000269	5.43E-22
205009_at#TFF1	0.00026	0	0.000572	1.41E-13
205016_at#TGFA	0.000172	0	0.000482	4.59E-29
205376_at#INPP4B	0.000116	0	0.00037	3.22E-34
205523_at#HAPLN1	5.28E-07	0	4.49E-05	1.01E-77
205992_s_at#IL15	0.001422	0	0.00414	6.77E-36
206025_s_at#TNFAIP6	0.000647	0	0.002008	1.29E-27
206224_at#CST1	0.000558	0	0.001683	1.96E-34
206424_at#CYP26A1	0.000183	0	0.00046	4.82E-22
206710_s_at#EPB41L3	8.01E-06	0	6.73E-05	1.59E-37
206884_s_at#SCEL	0.000117	0	8.98E-05	2.17E-10
207387_s_at#GK	3.15E-05	0	7.85E-05	9.94E-26
207781_s_at#ZNF711	0.005569	0	0.015505	8.56E-09
207850_at#CXCL3	0.001893	0	0.005711	5.97E-22
208116_s_at#MAN1A1	5.87E-05	0	0.000236	1.90E-57
209016_s_at#KRT7	2.60E-05	0	8.98E-05	5.49E-36
209035_at#MDK	1.74E-05	0	7.85E-05	5.64E-56
209487_at#RBPMS	8.56E-06	0	6.73E-05	6.97E-51
209488_s_at#RBPMS	7.51E-06	0	6.73E-05	8.81E-51
209631_s_at#GPR37	1.00E-05	0	5.61E-05	1.74E-26
209839_at#DNM3	3.46E-07	0	4.49E-05	4.37E-182
209942_x_at#MAGEA3	7.15E-07	0	4.49E-05	1.79E-97
210095_s_at#IGFBP3	0.000185	0	0.000572	7.45E-53
211549_s_at#HPGD	0.000203	0	0.000595	1.88E-34
211573_x_at#TGM2	0.00084	0	0.002524	1.68E-47
211657_at#CEACAM6	3.40E-05	0	0.000135	3.04E-63
211675_s_at#MDFIC	3.41E-06	0	4.49E-05	2.35E-54

211776_s_at#EPB41L3	8.29E-05	0	0.000303	4.35E-44
211959_at#IGFBP5	0.000176	0	0.000505	2.68E-30
212667_at#SPARC	1.02E-05	0	7.85E-05	2.34E-63
212681_at#EPB41L3	8.92E-06	0	6.73E-05	6.61E-37
213273_at#ODZ4	6.33E-06	0	6.73E-05	8.48E-40
213711_at#KRT81	0.000266	0	0.000684	2.11E-18
214078_at#AF070581	0.000456	0	0.001077	5.45E-12
214476_at#TFF2	4.97E-06	0	4.49E-05	3.03E-31
214612_x_at#MAGEA2B	2.90E-06	0	6.73E-05	5.88E-101
215966_x_at#GK3P	0.00016	0	0.00037	8.05E-17
217028_at#CXCR4	0.003327	0	0.009525	4.94E-27
217771_at#GOLM1	0.000101	0	0.000348	1.04E-38
217787_s_at#GALNT2	5.39E-05	0	0.00018	5.15E-29
219327_s_at#GPRC5C	1.86E-05	0	7.85E-05	1.62E-35
219936_s_at#GPR87	4.00E-06	0	4.49E-05	1.85E-46
220468_at#ARL14	3.51E-05	0	7.85E-05	1.42E-19
220784_s_at#UTS2	1.02E-06	0	4.49E-05	0
221618_s_at#TAF9B	0.008075	0	0.020172	3.70E-34
221731_x_at#VCAN	4.86E-05	0	0.000101	1.63E-22
221805_at#NEFL	1.36E-05	0	4.49E-05	2.97E-23
41469_at#PI3	7.79E-05	0	0.00028	2.00E-39

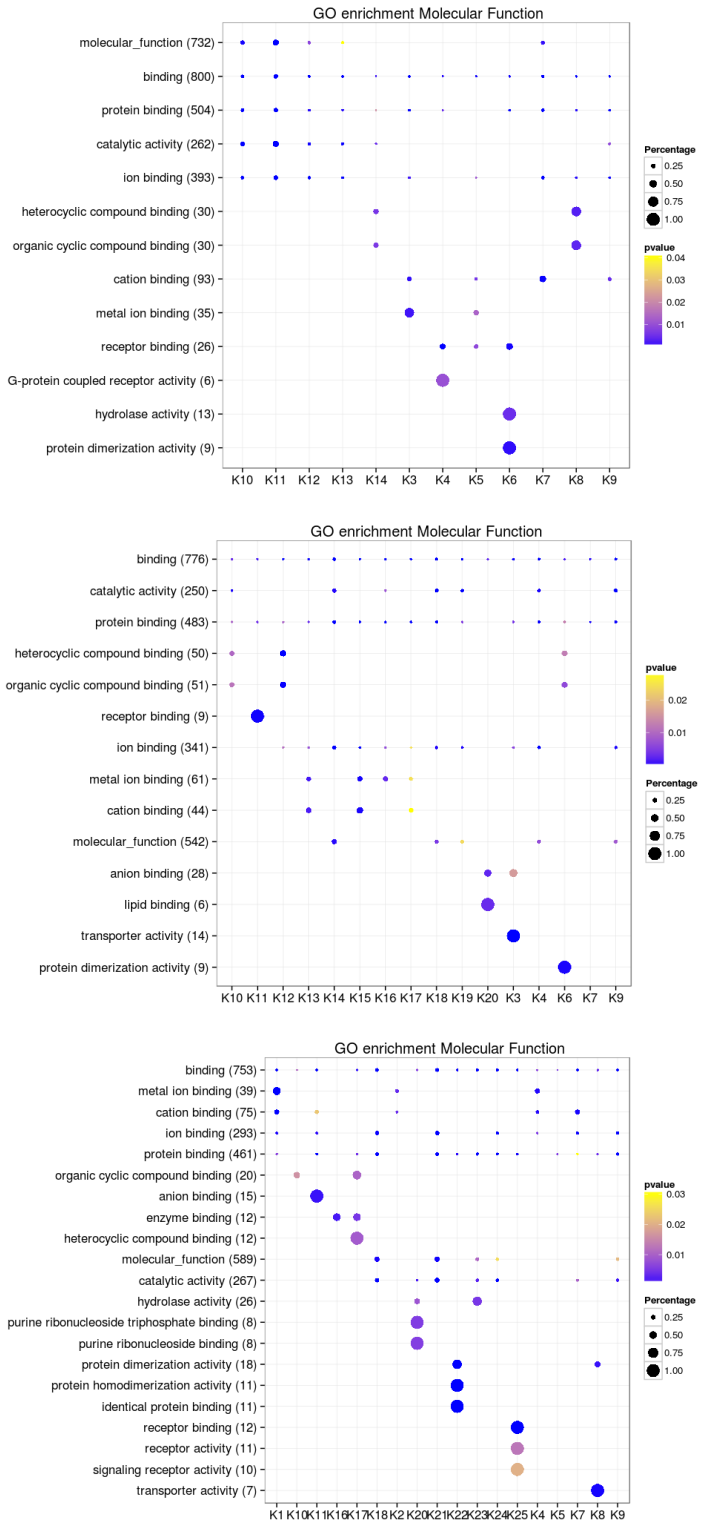


Figure S6.1. Molecular function mapping of Mouse species for clustered genes of model 2, by BCC(top), HIM(middle) and MDI(bottom)

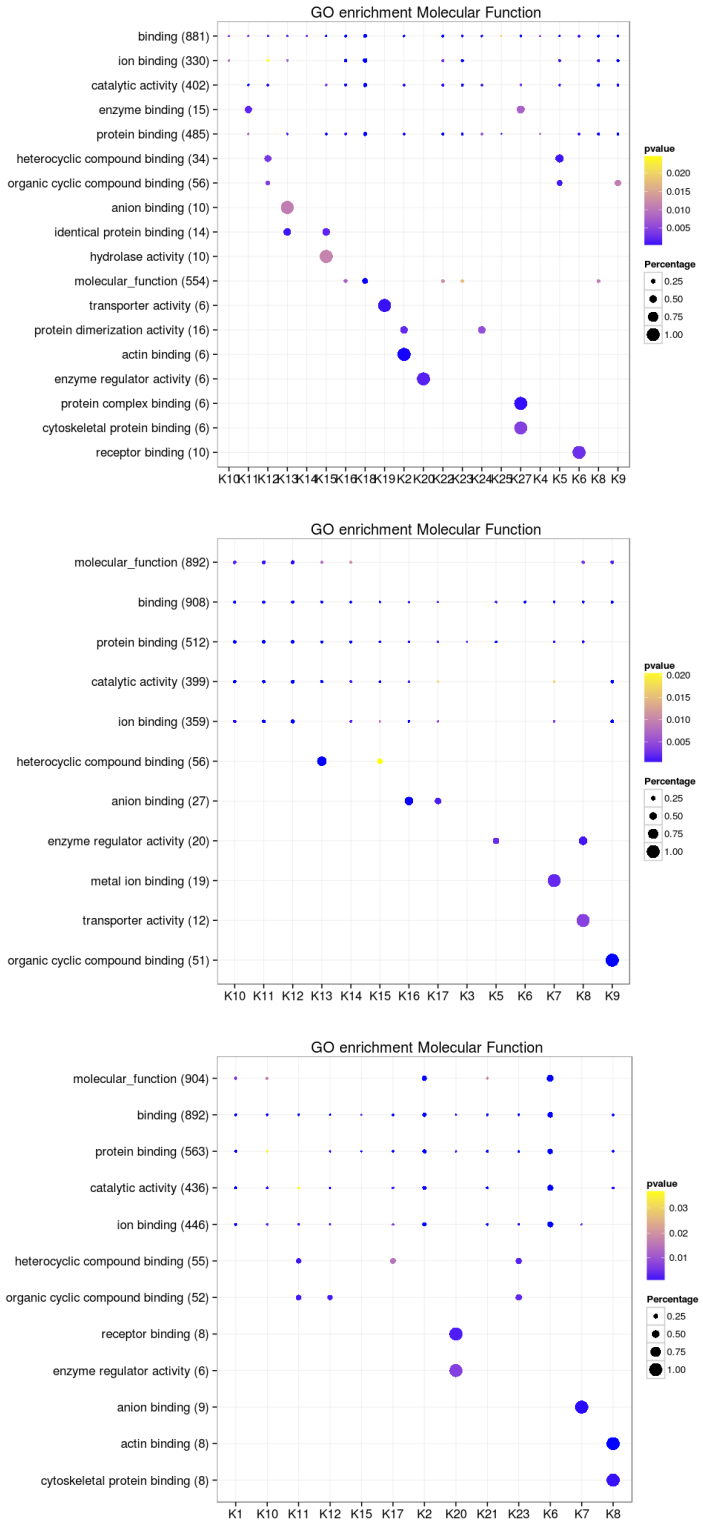


Figure S6.2. Molecular function mapping of Mouse species for clustered genes of model 3 by BCC(top), HIM(middle) and MDI(bottom)

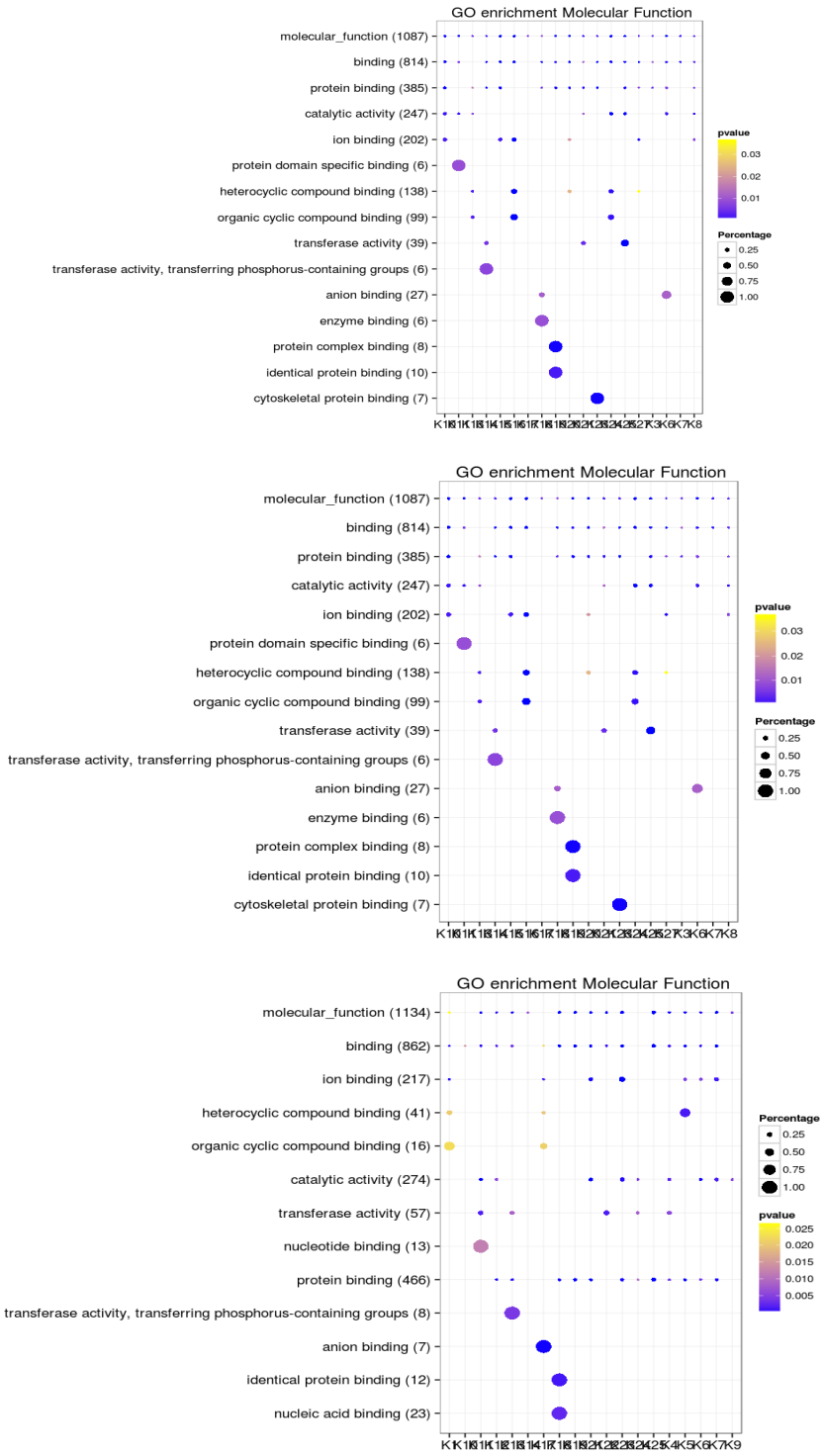


Figure S6.3. Molecular function mapping of Rat species for clustered genes of model 1 identified by BCC(top), HIM(middle) and MDI(bottom).

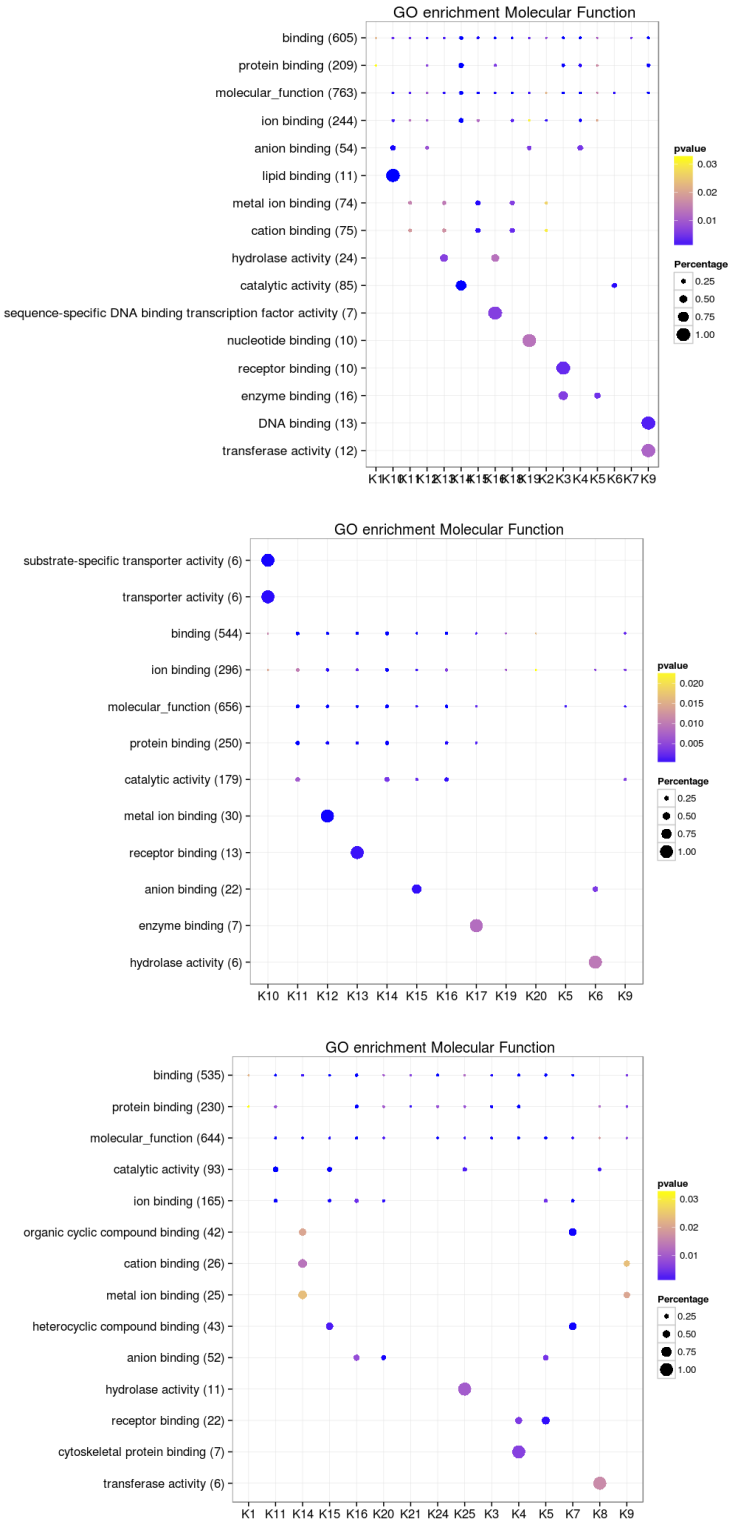


Figure S6.4. Molecular function mapping of Rat species for clustered genes of model2 identified by BCC(top), HIM(middle) and MDI(bottom).

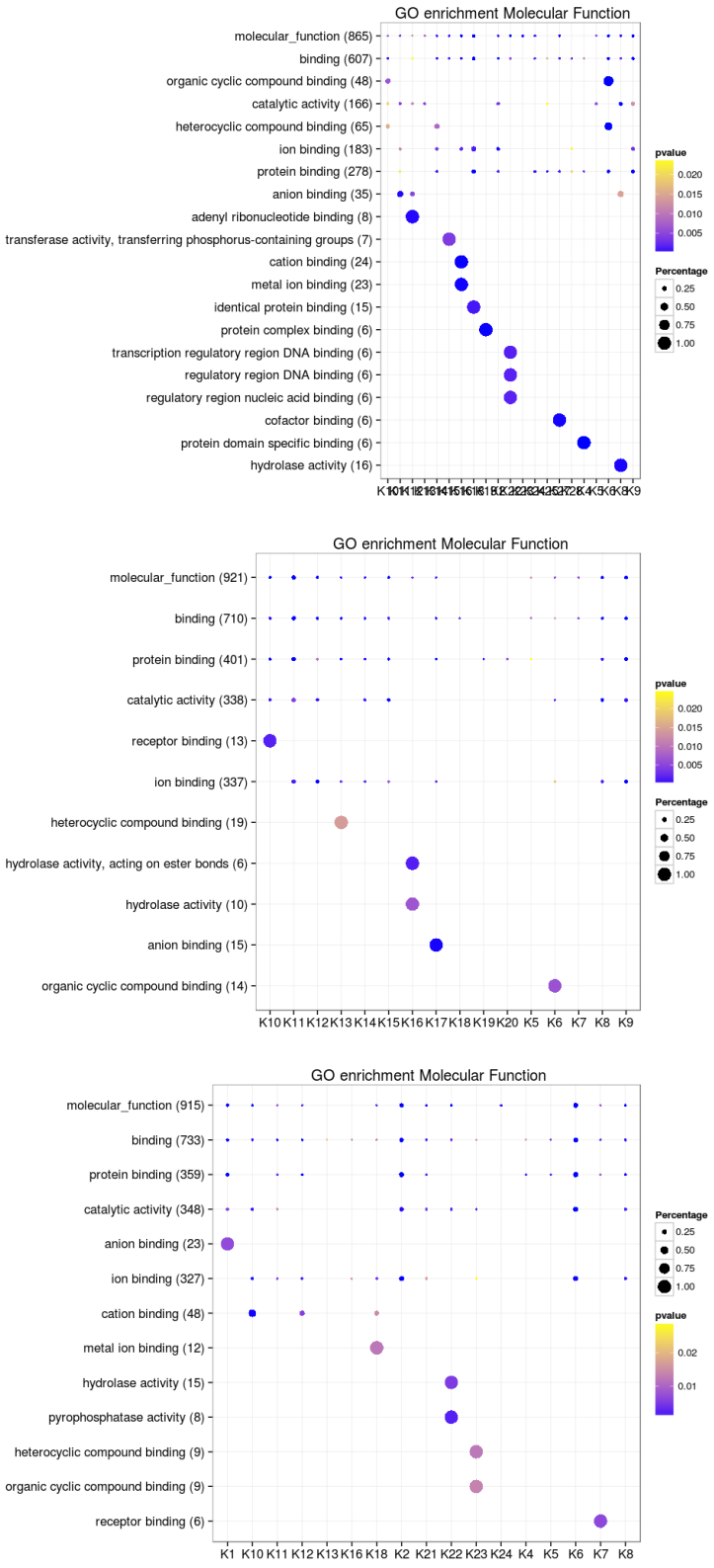


Figure S6.5. Molecular function mapping of Rat species for clustered genes of model3 identified by BCC(top), HIM(middle) and MDI(bottom).

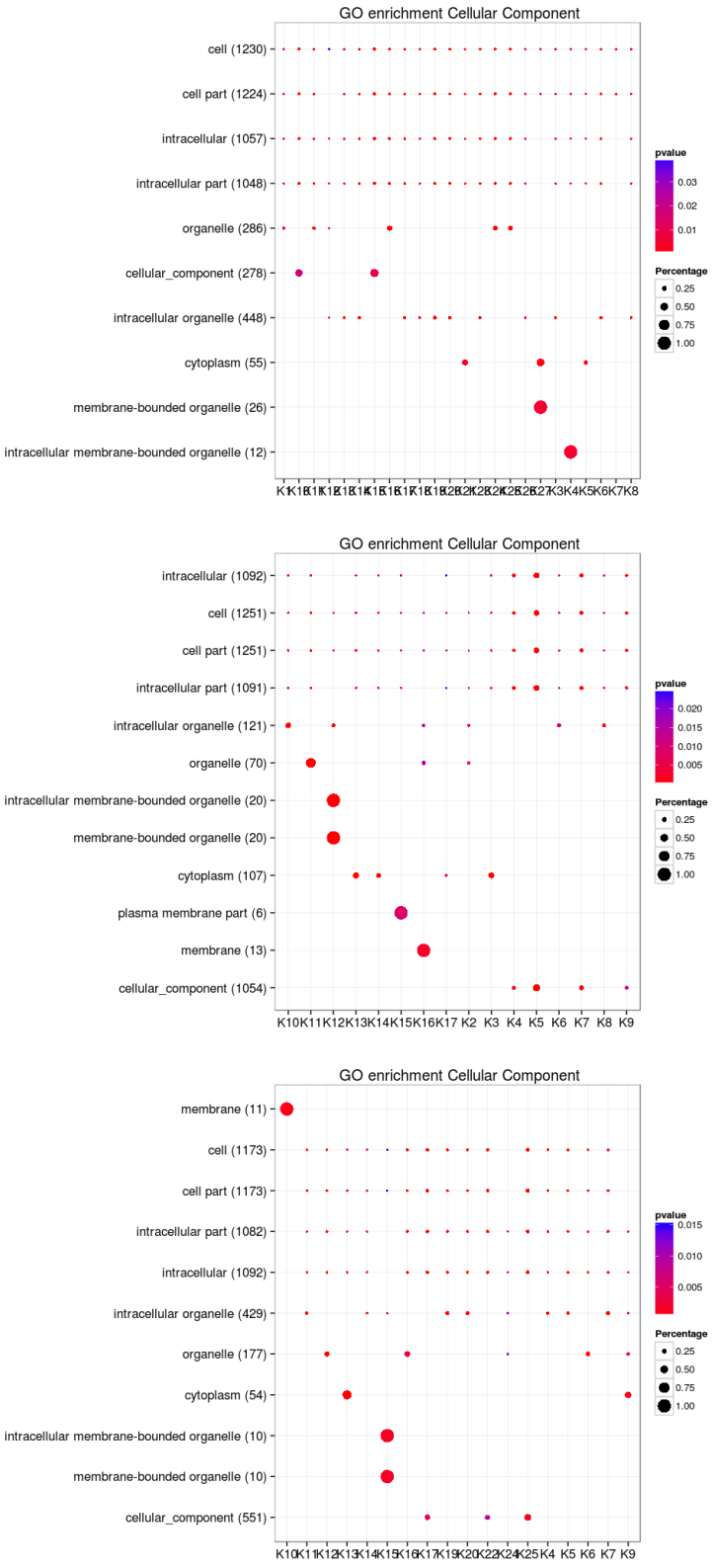


Figure S6.6. Cellular component mapping of Mouse species for clustered genes of model1 identified by BCC(top), HIM(middle) and MDI(bottom).

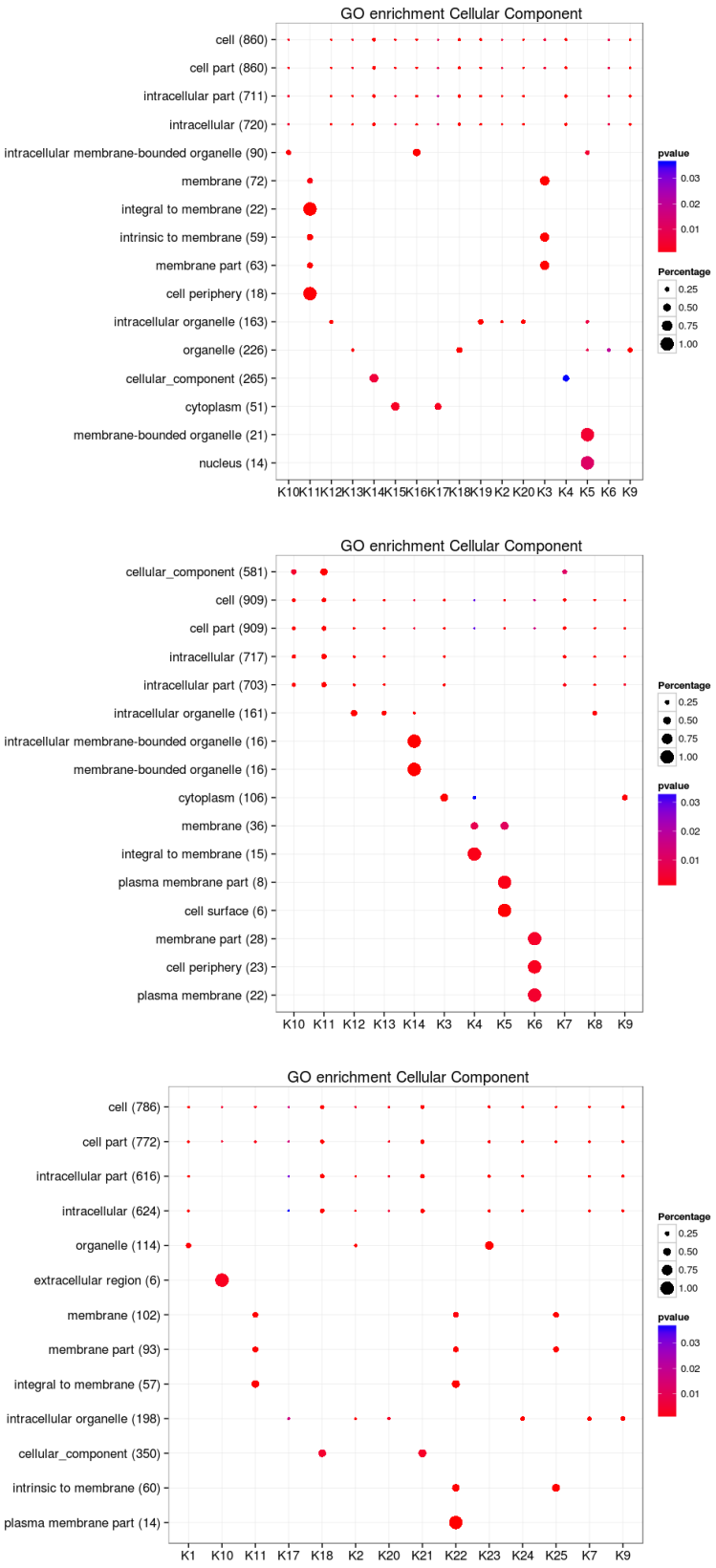


Figure S6.7. Cellular component mapping of Mouse species for clustered genes of model2 identified by BCC(top), HIM(middle) and MDI(bottom).

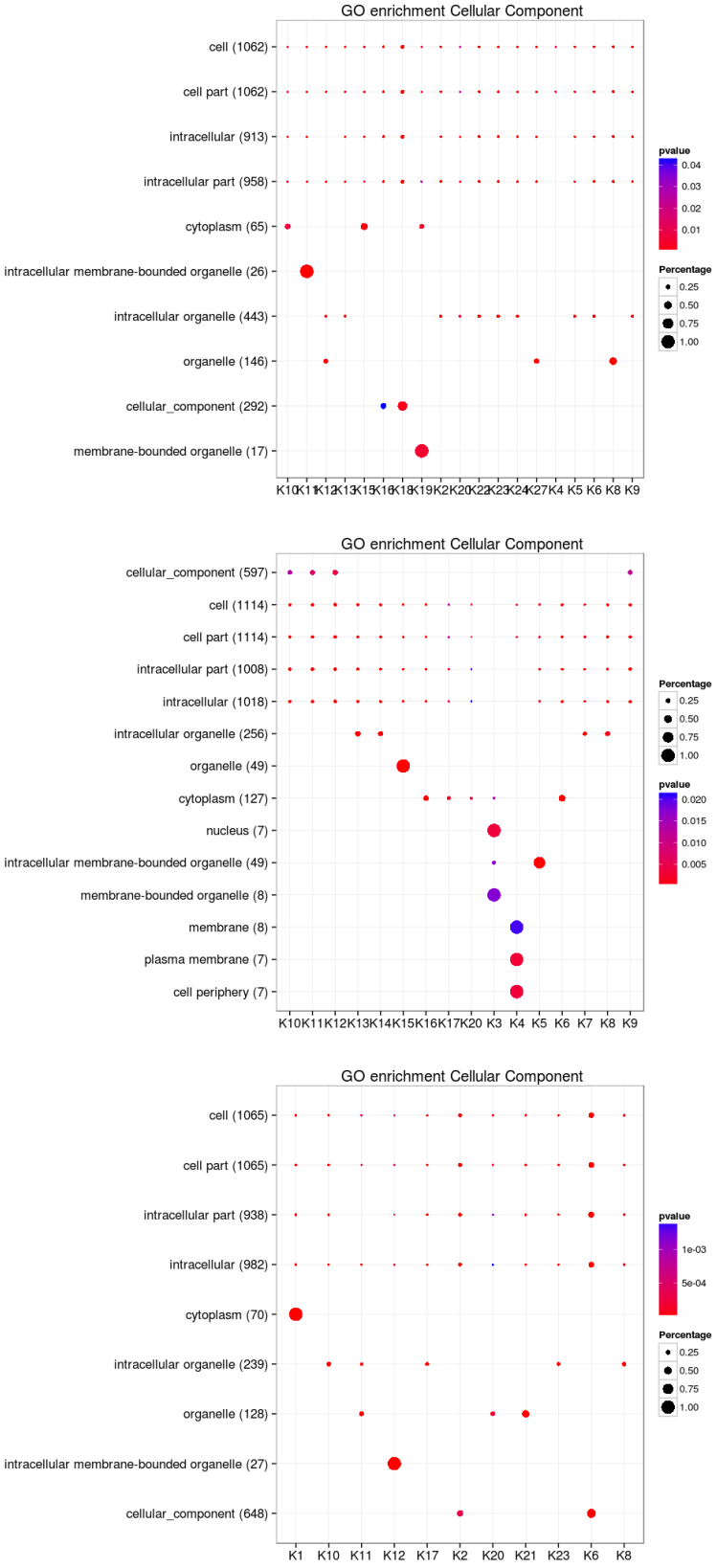


Figure S6.8. Cellular component mapping of Mouse species for clustered genes of model3 identified by BCC(top), HIM(middle) and MDI(bottom).

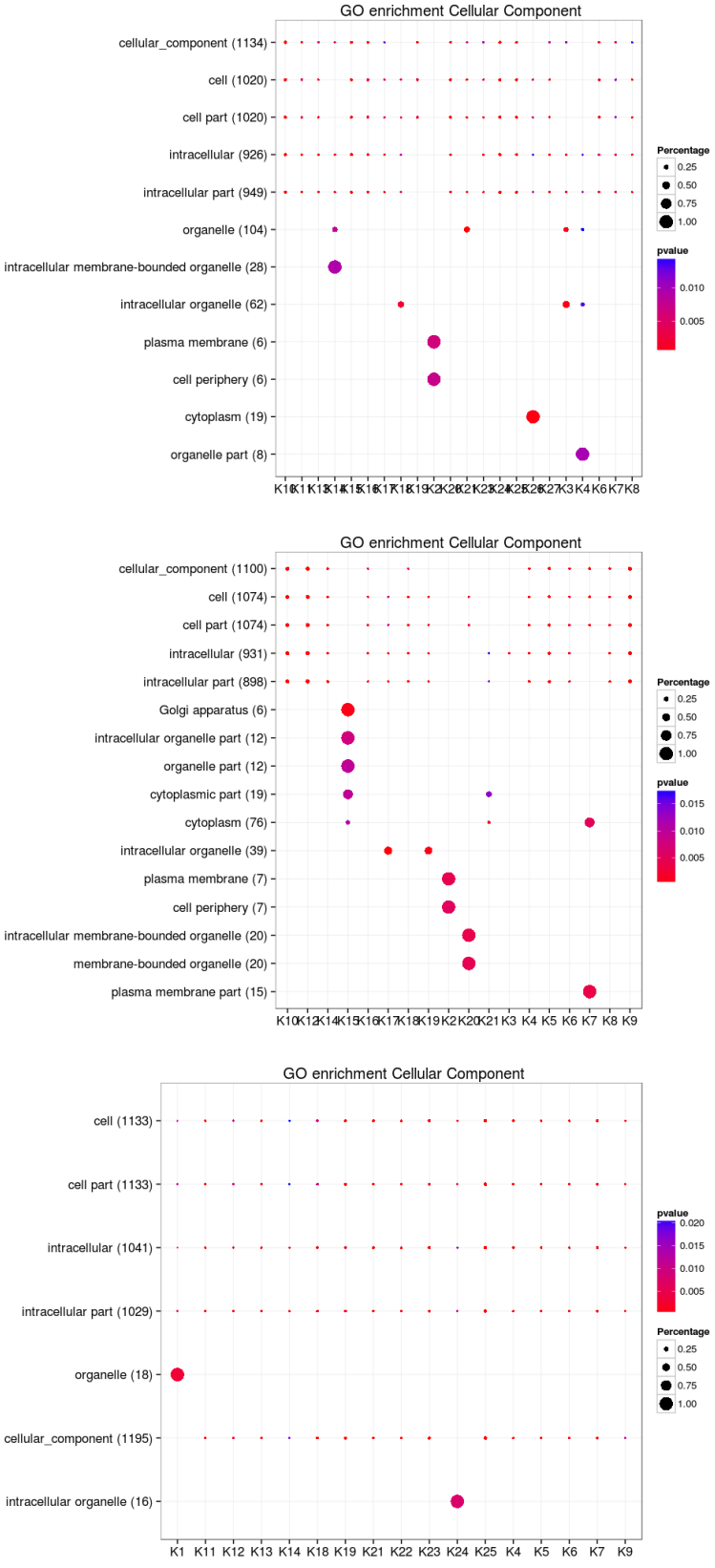


Figure S6.9. Cellular component mapping of Rat species for clustered genes of model1 identified by BCC(top), HIM(middle) and MDI(bottom).

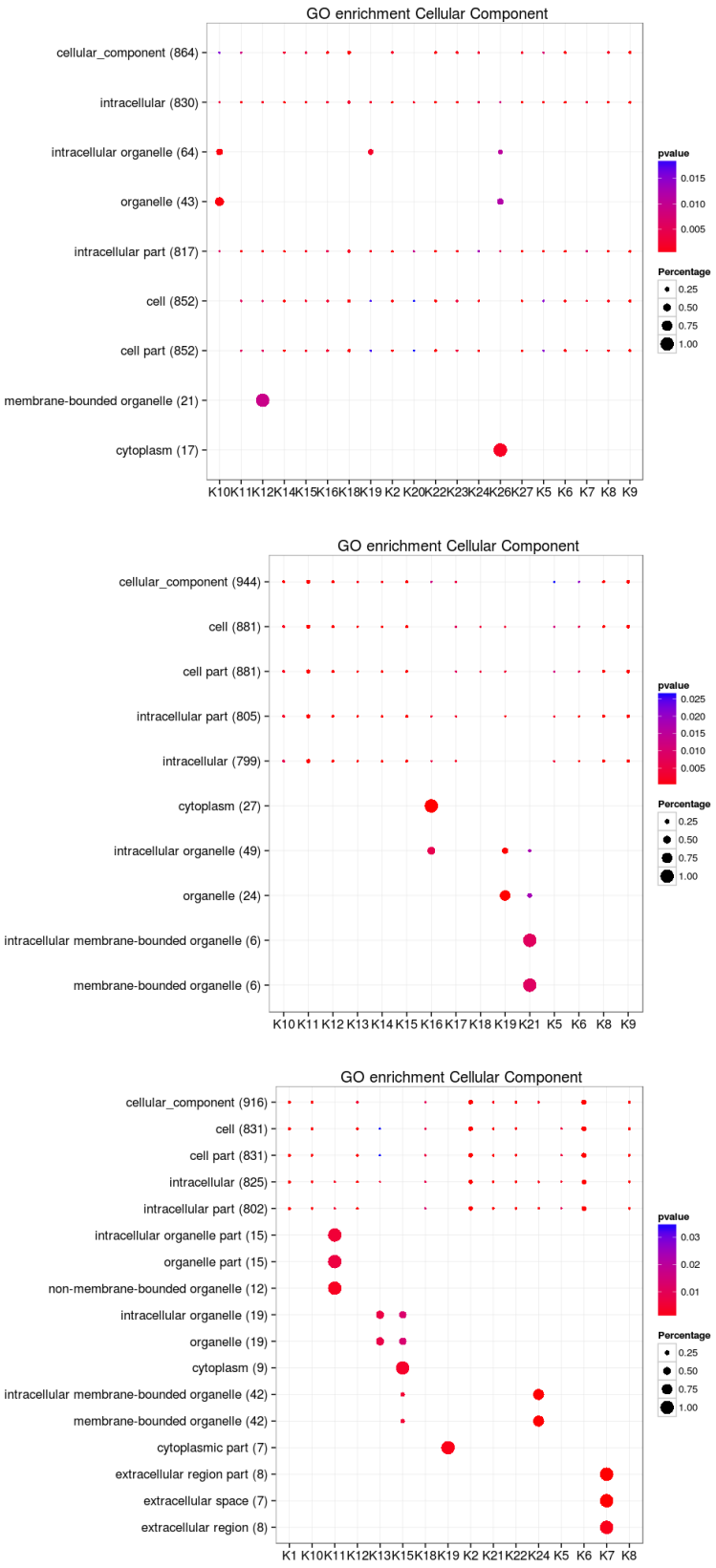


Figure S6.10. Cellular component mapping of Rat species for clustered genes of model3 identified by BCC(top), HIM(middle) and MDI(bottom).

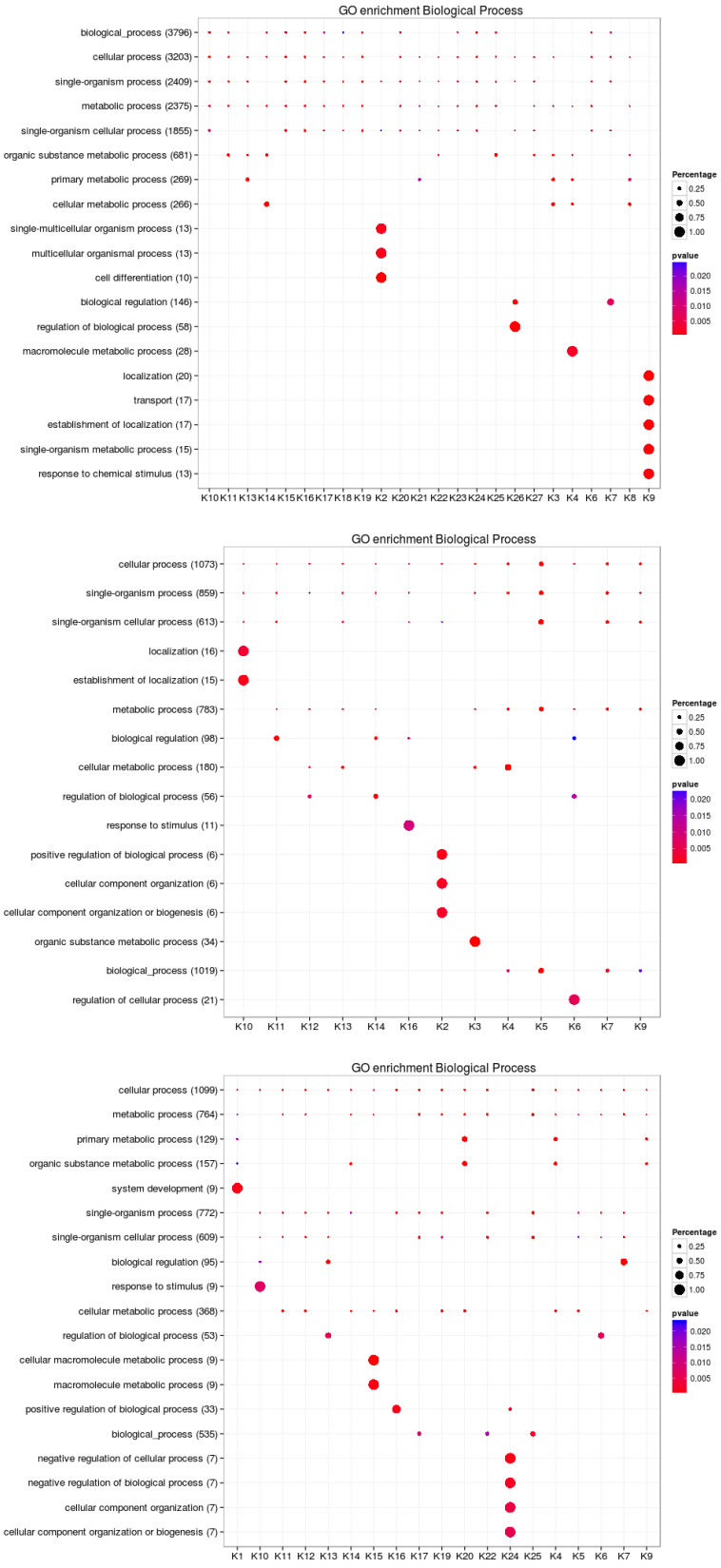


Figure S6.11. Biological process mapping of Mouse species for clustered genes of model1 identified by BCC(top), HIM(middle) and MDI(bottom).

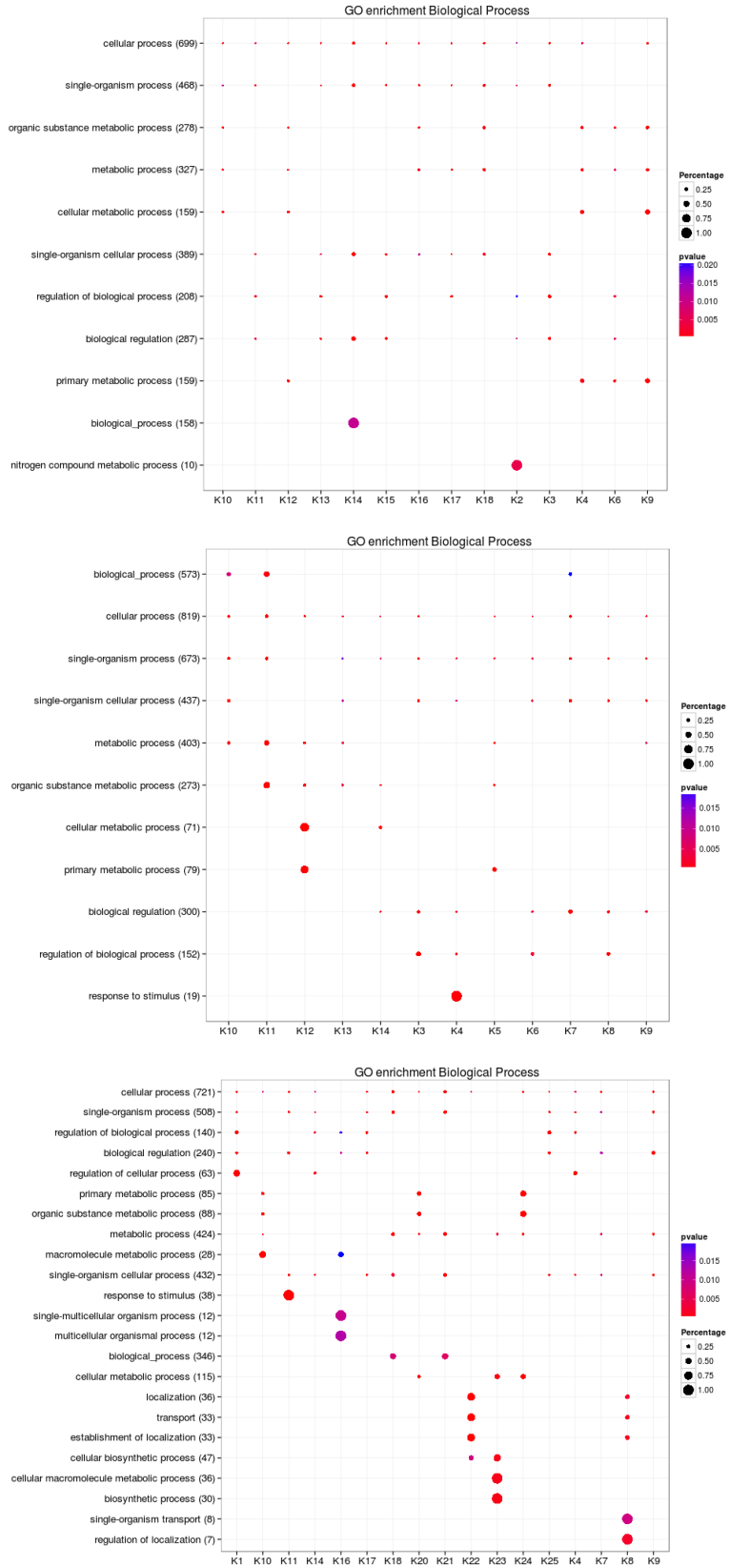


Figure S6.12. Biological process mapping of Mouse species for clustered genes of model2 identified by BCC(top), HIM(middle) and MDI(bottom).

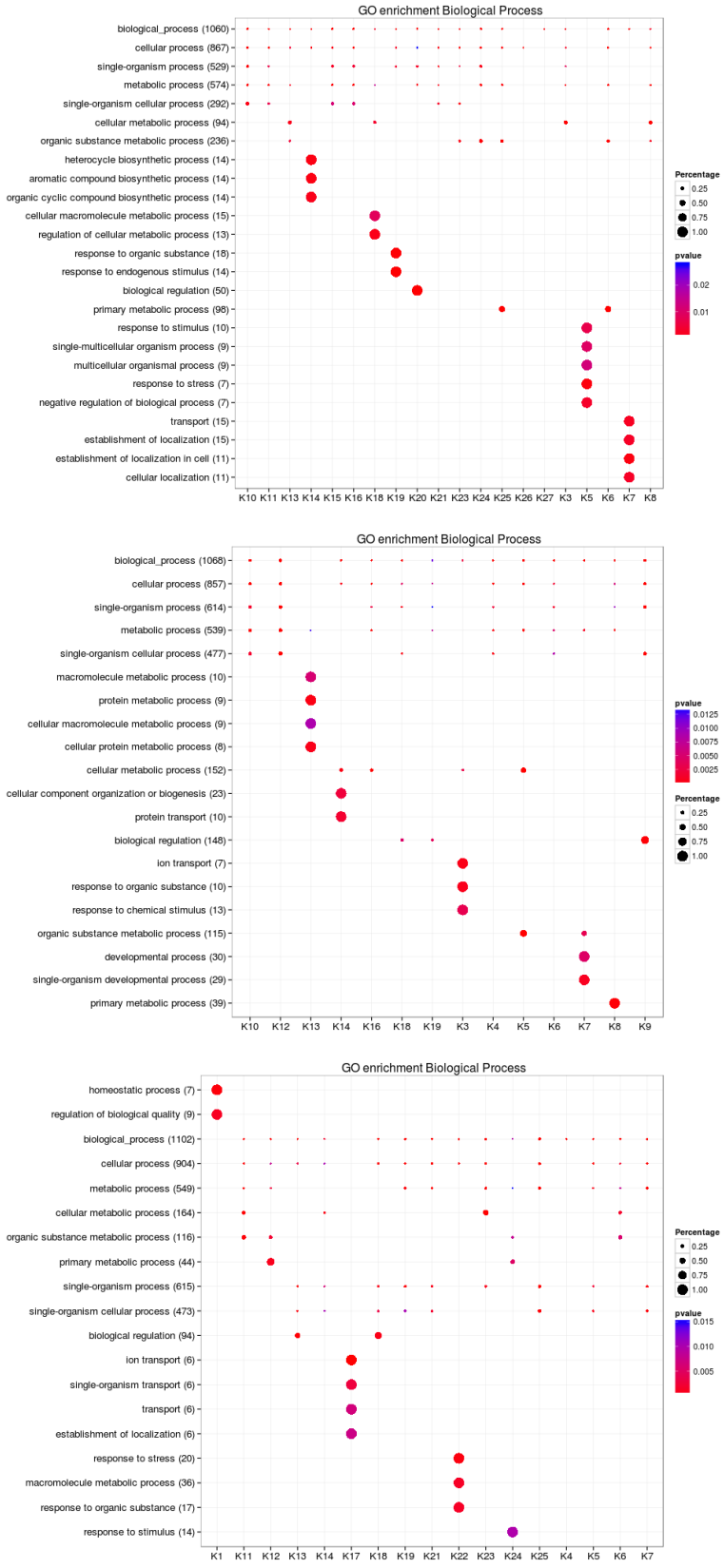


Figure S6.13. Biological process mapping of Rat species for clustered genes of model1 identified by BCC(top), HIM(middle) and MDI(bottom).

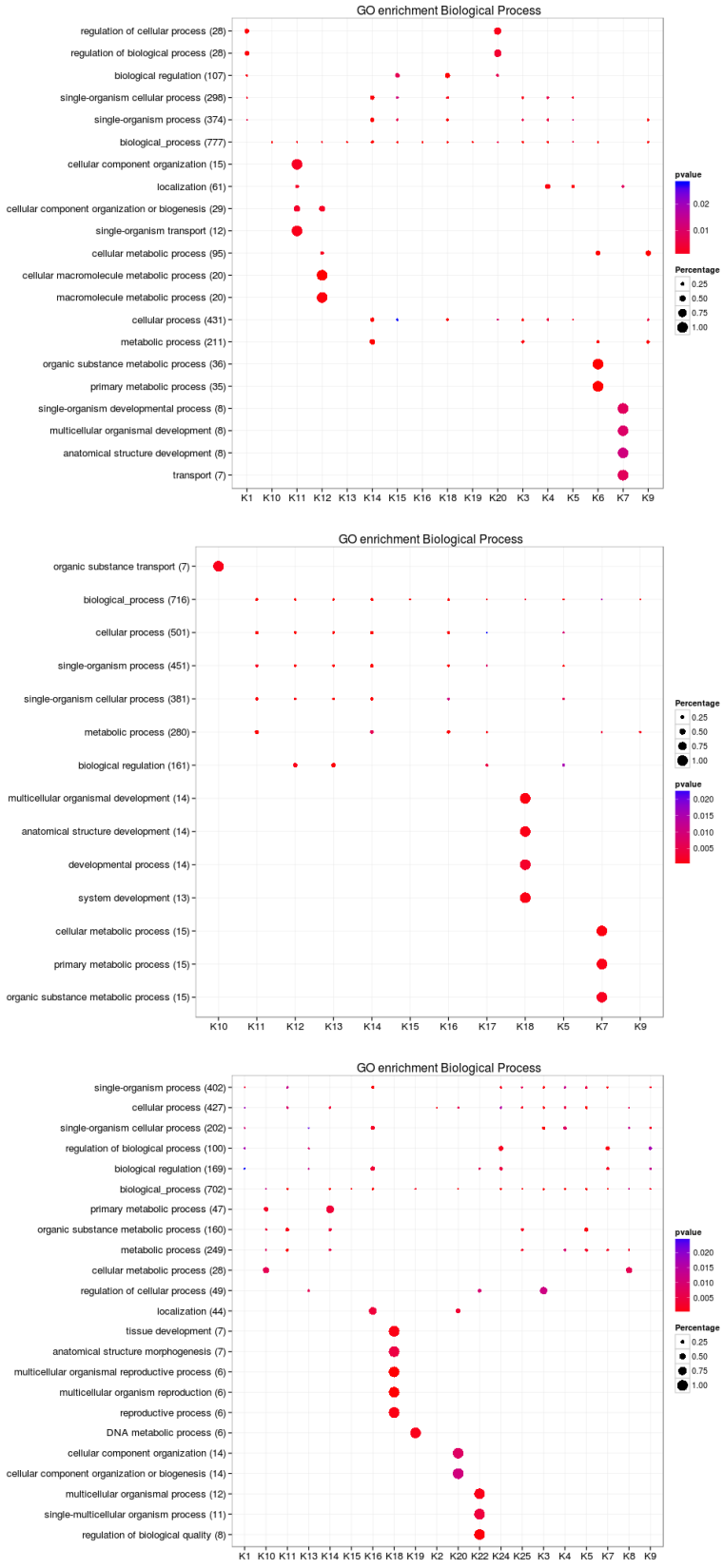


Figure S6.14. Biological process mapping of Rat species for clustered genes of model2 identified by BCC(top), HIM(middle) and MDI(bottom).

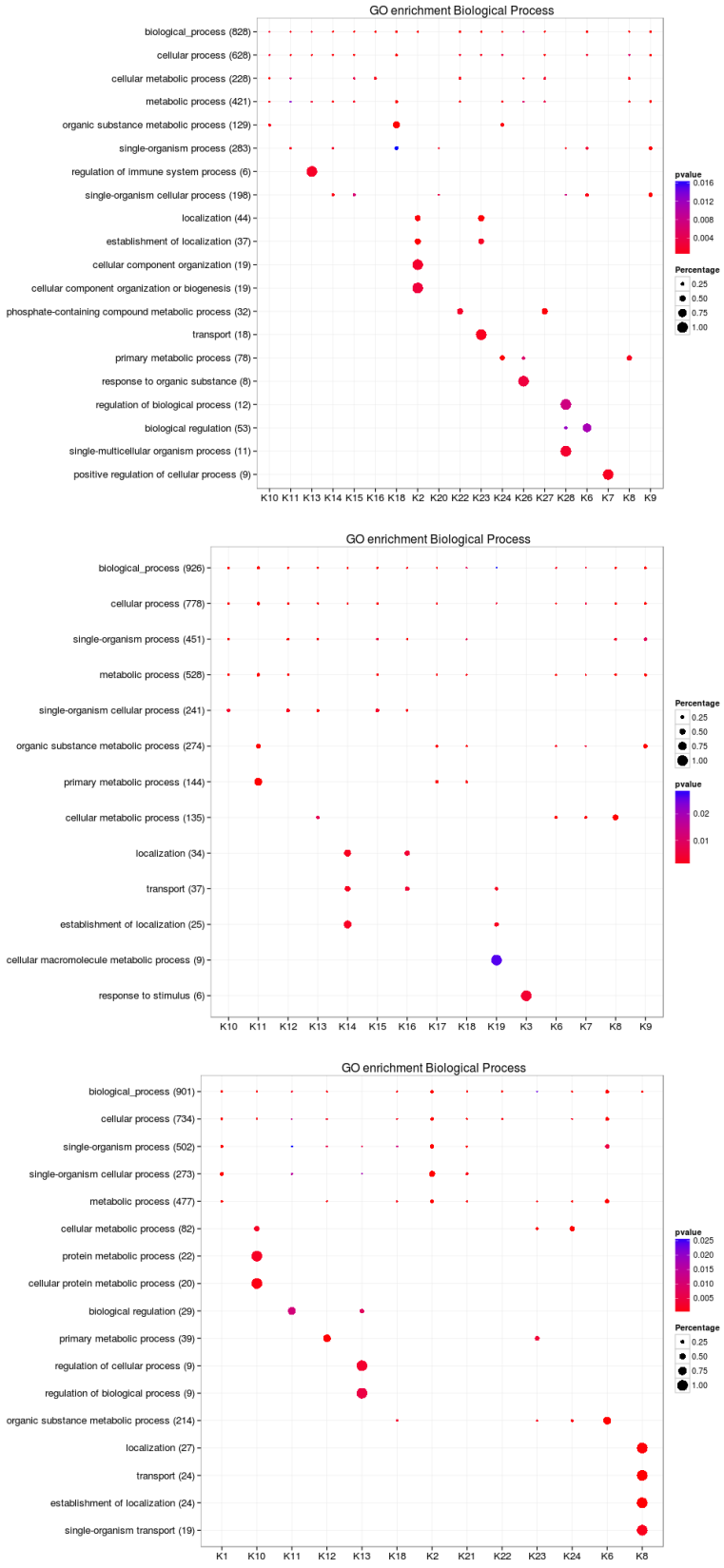


Figure S6.15. Biological process mapping of Rat species for clustered genes of model3 identified by BCC(top), HIM(middle) and MDI(bottom).

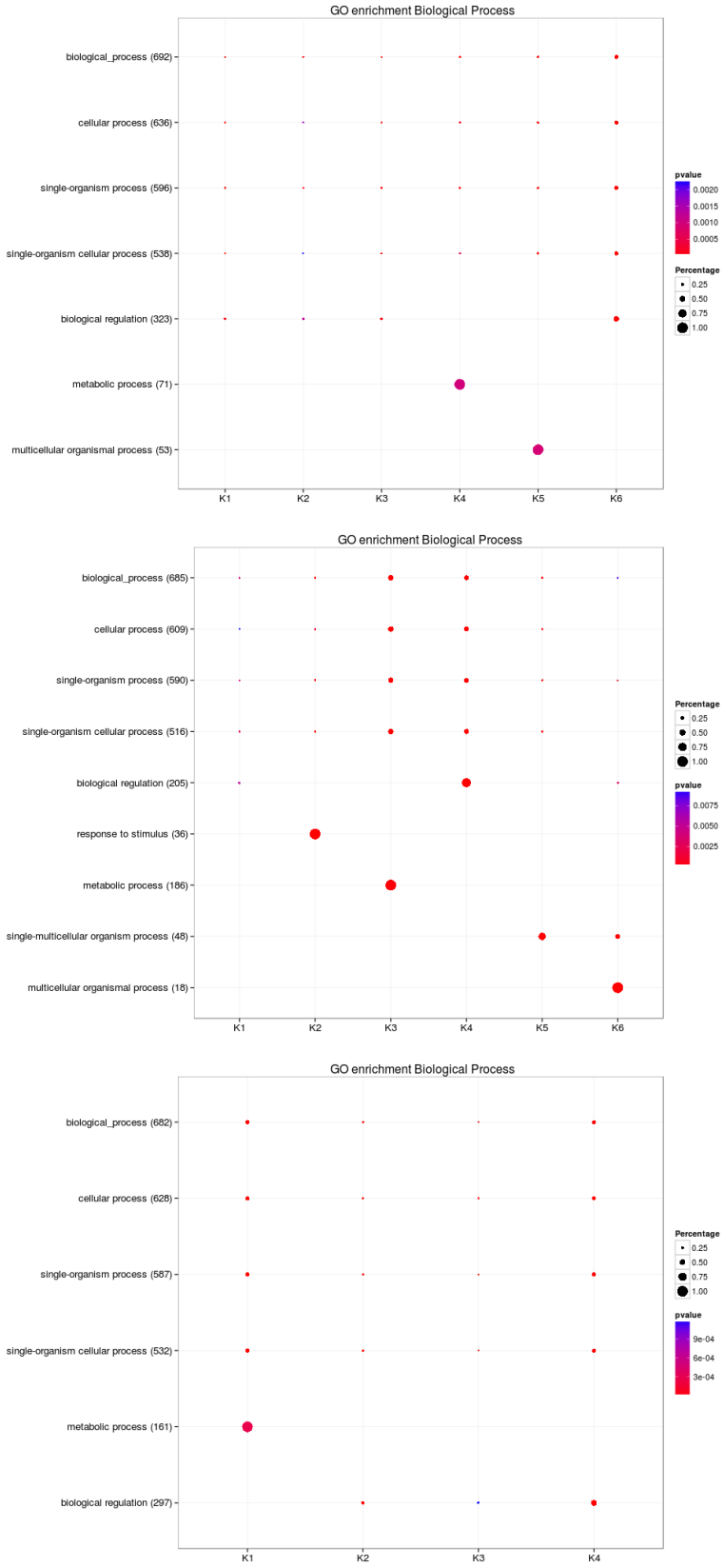


Figure S6.16. Biological process mapping of Data2(Prostate 2) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

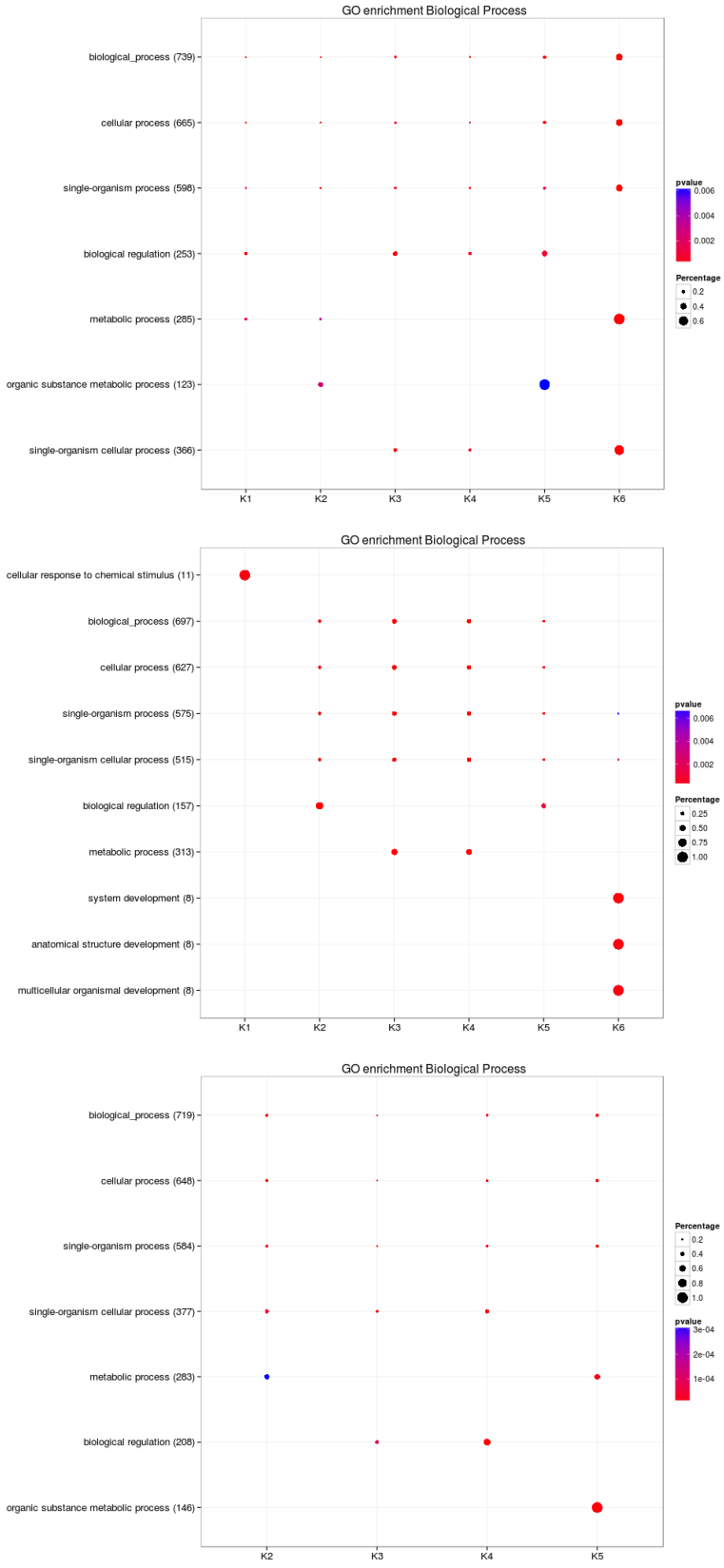


Figure S6.17. Biological process mapping of Data3(Breast 1) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

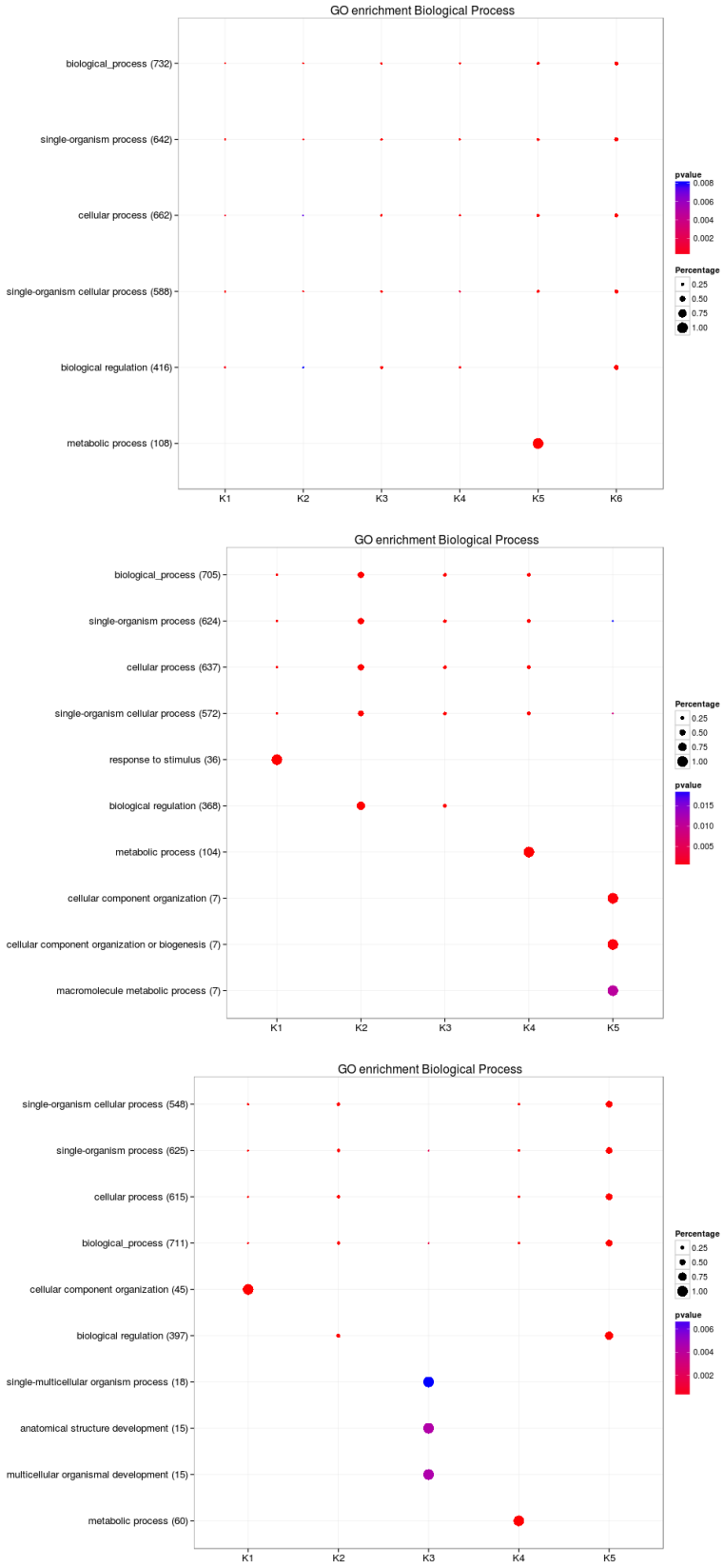


Figure S6.18. Biological process mapping of Data4(Breast 2) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

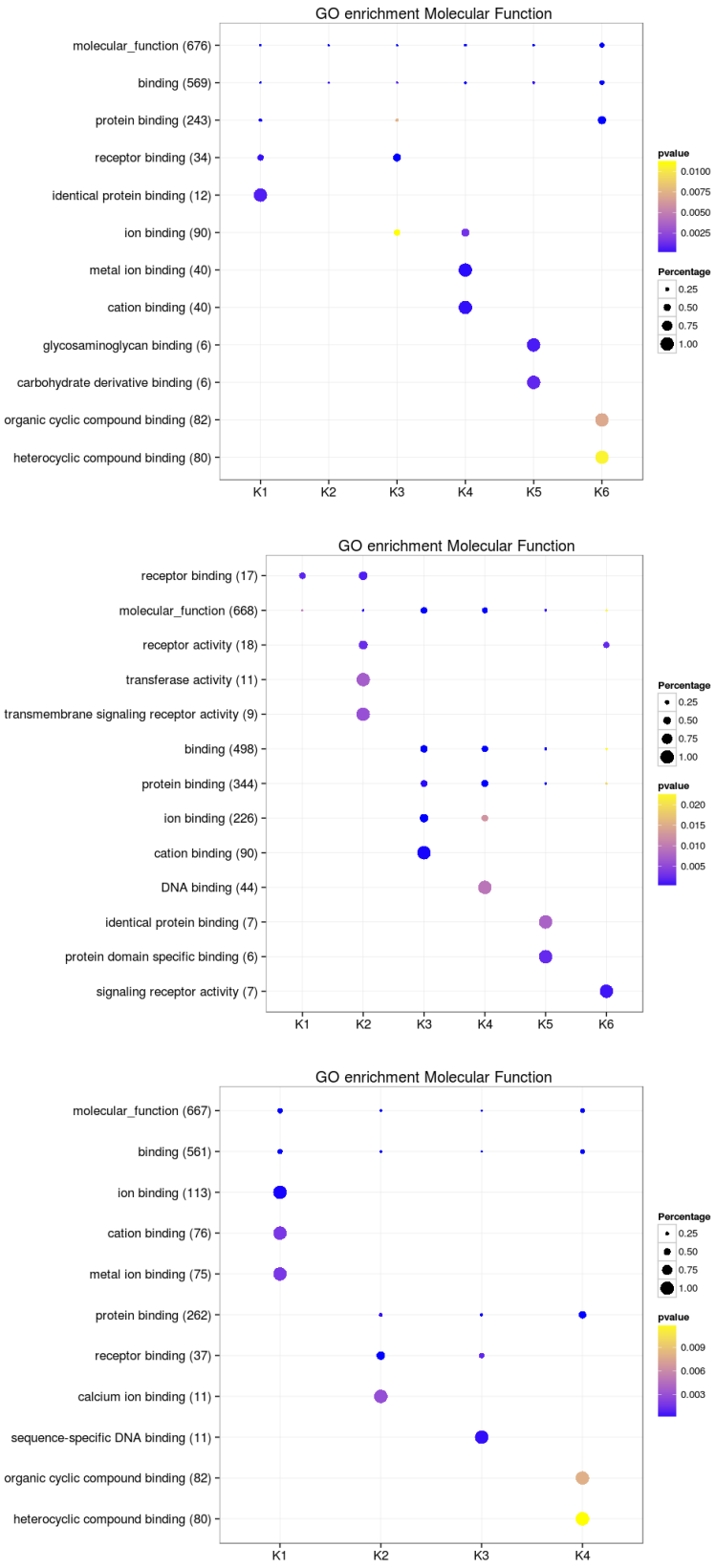


Figure S6.19. Molecular function mapping of Data2(Prostate2) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

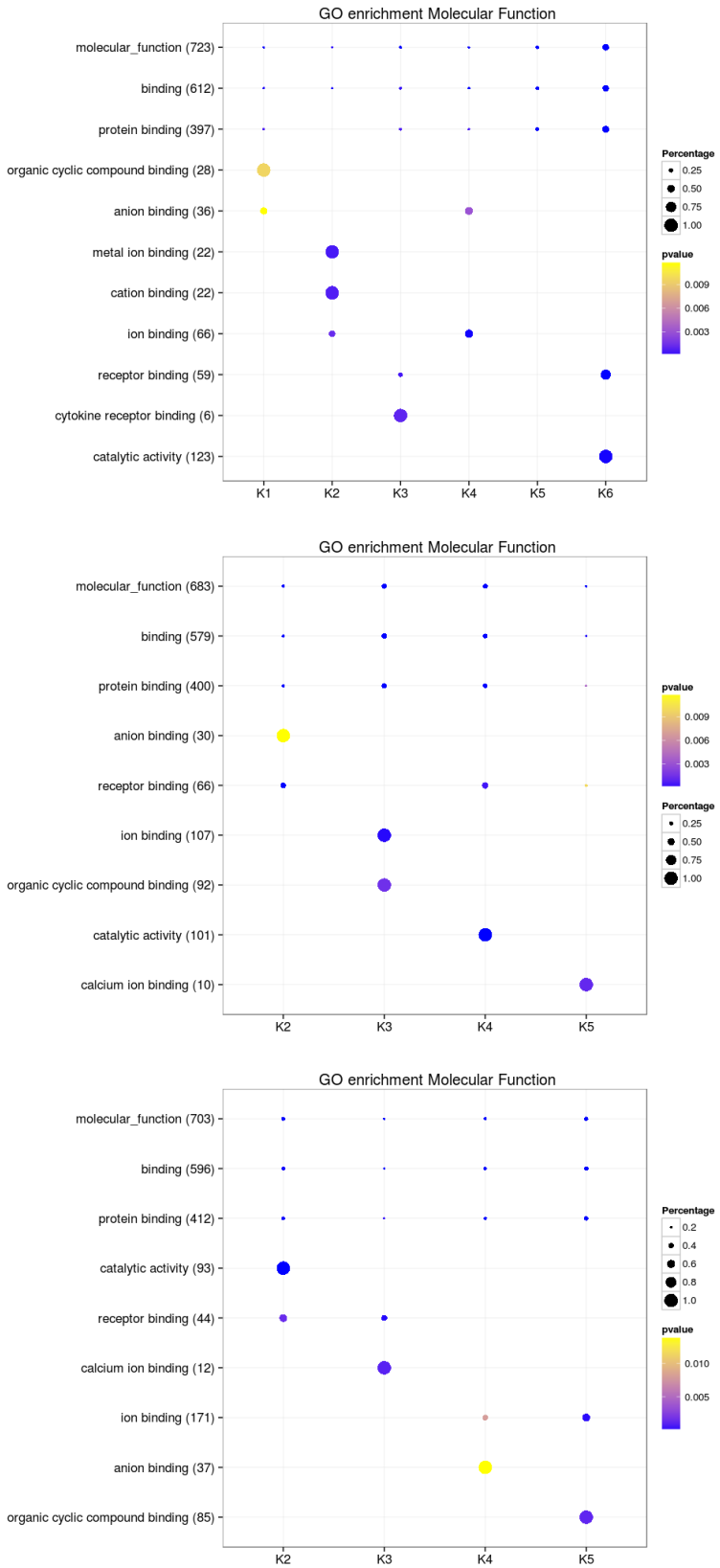


Figure S6.20. Molecular function mapping of Data3(Breast1) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

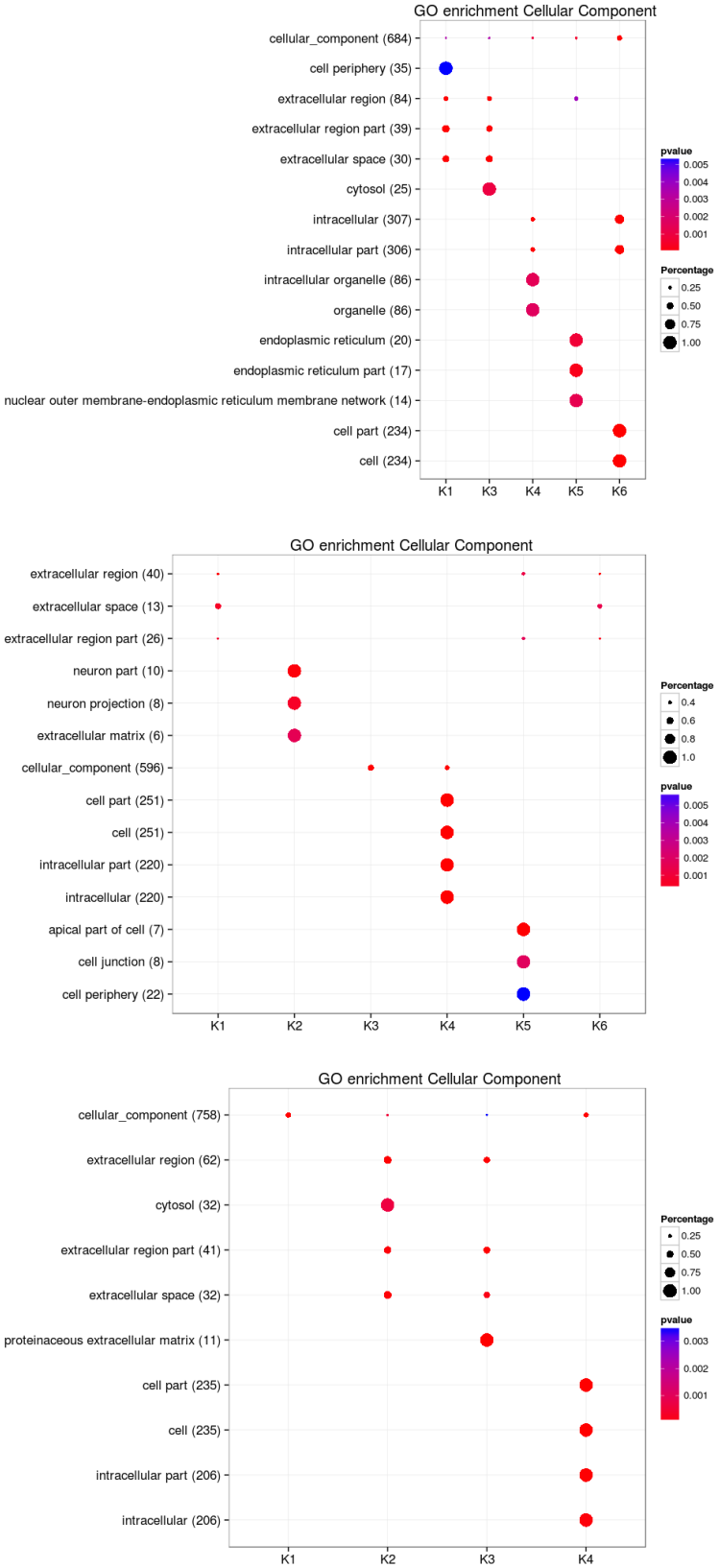


Figure S6.21. Cellular component mapping of Data2(Prostatet2) for clustered genes that were identified by BCC(top), HIM(middle) and MDI(bottom).

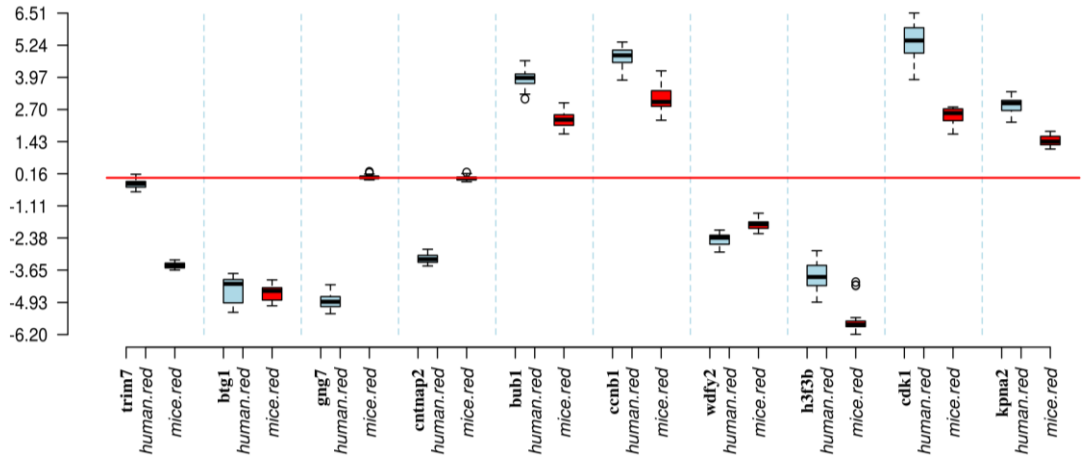


Figure S7.1. Expressions of the top ten homogeneous genes identified by eBayes across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DEs. The horizontal line in red represents zero DE.

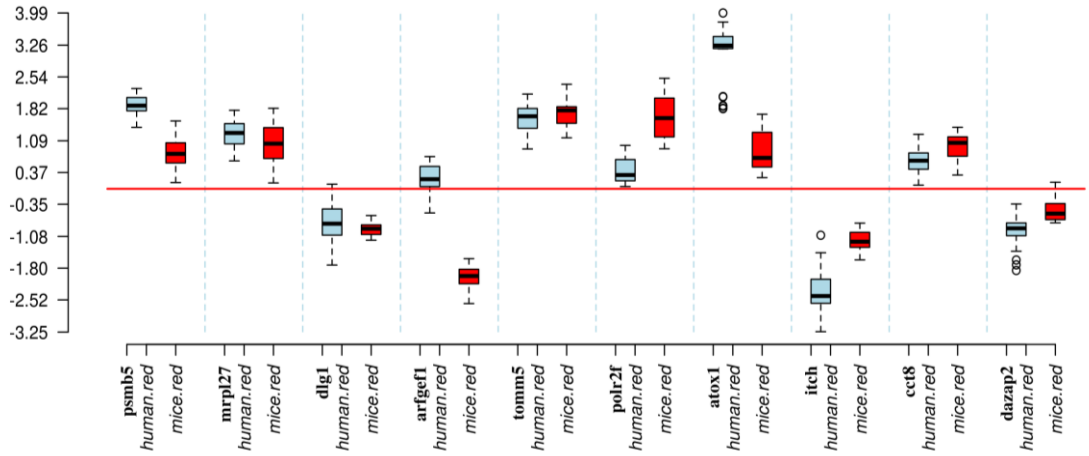


Figure S7.2. Expressions of the top ten homogeneous genes identified by SAMr across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DEs. The horizontal line in red represents zero DE.

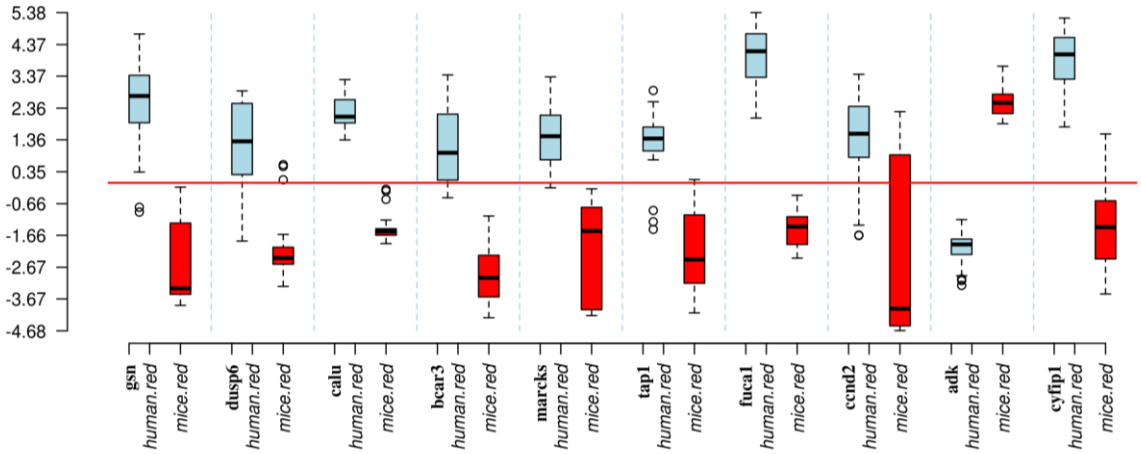


Figure S7.3. Expressions of the top ten homogeneous genes identified by Cyber-T across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DEs. The horizontal line in red represents zero DE.

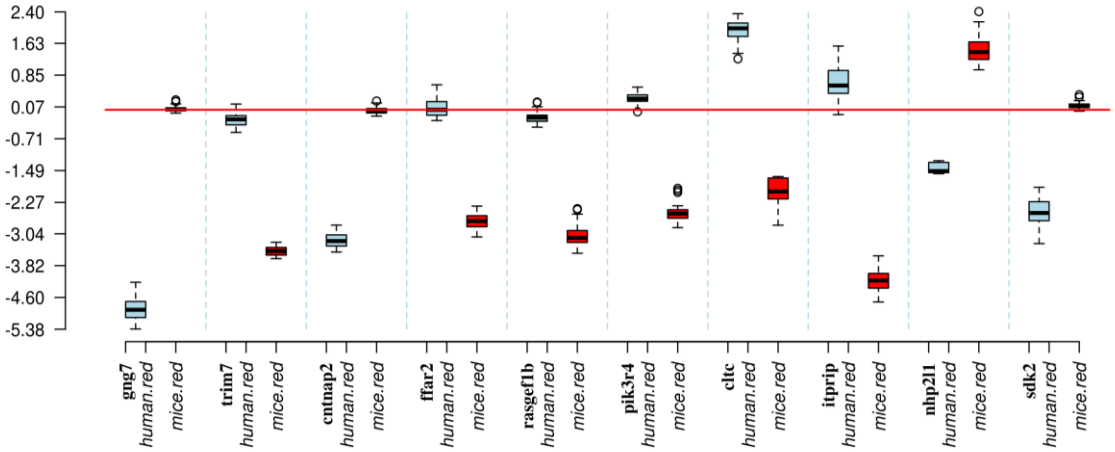


Figure S7.4. Expressions of the top ten heterogeneous genes identified by eBayes across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DE. The horizontal line in red represents zero DE.

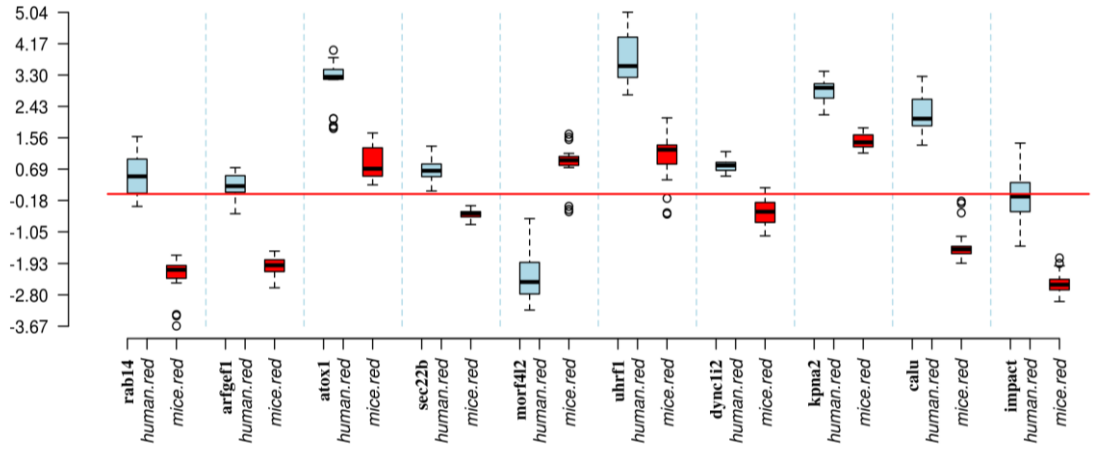


Figure S7.5. Expressions of the top ten heterogeneous genes identified by SAMr across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DE. The horizontal line in red represents zero DE.

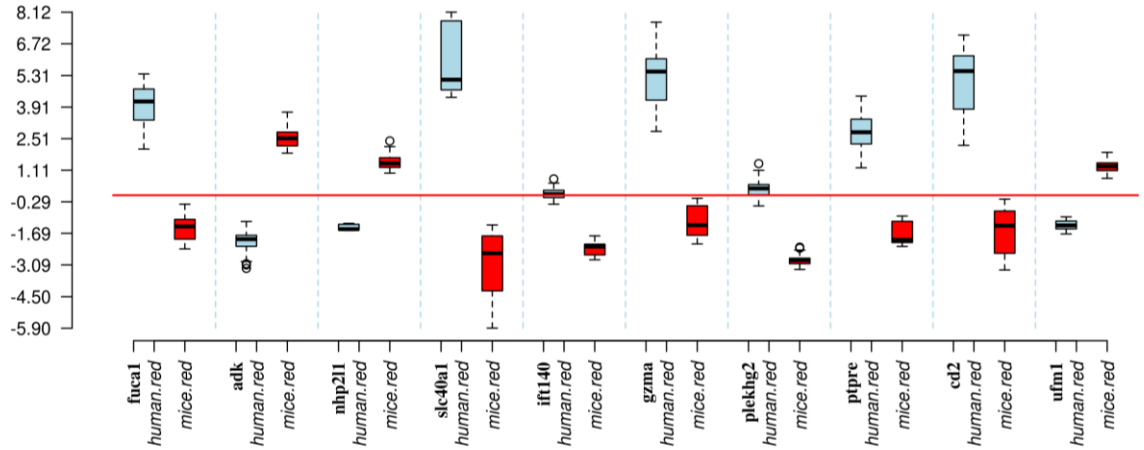


Figure S7.6. Expressions of the top ten heterogeneous genes identified by Cyber-T across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The vertical axis represents DE. The horizontal line in red represents zero DE.

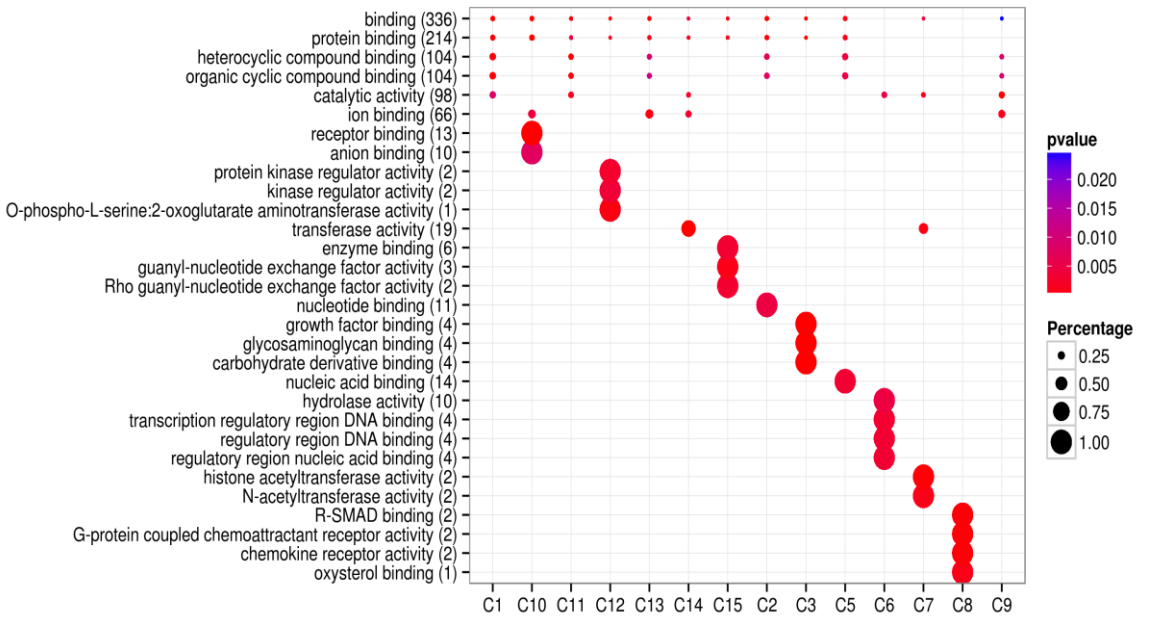


Figure S7.7. Molecular function mapping for clustered homogeneous DEGs identified by MSG for the mice species. C1 – C15 represent clusters identified by Mclust.

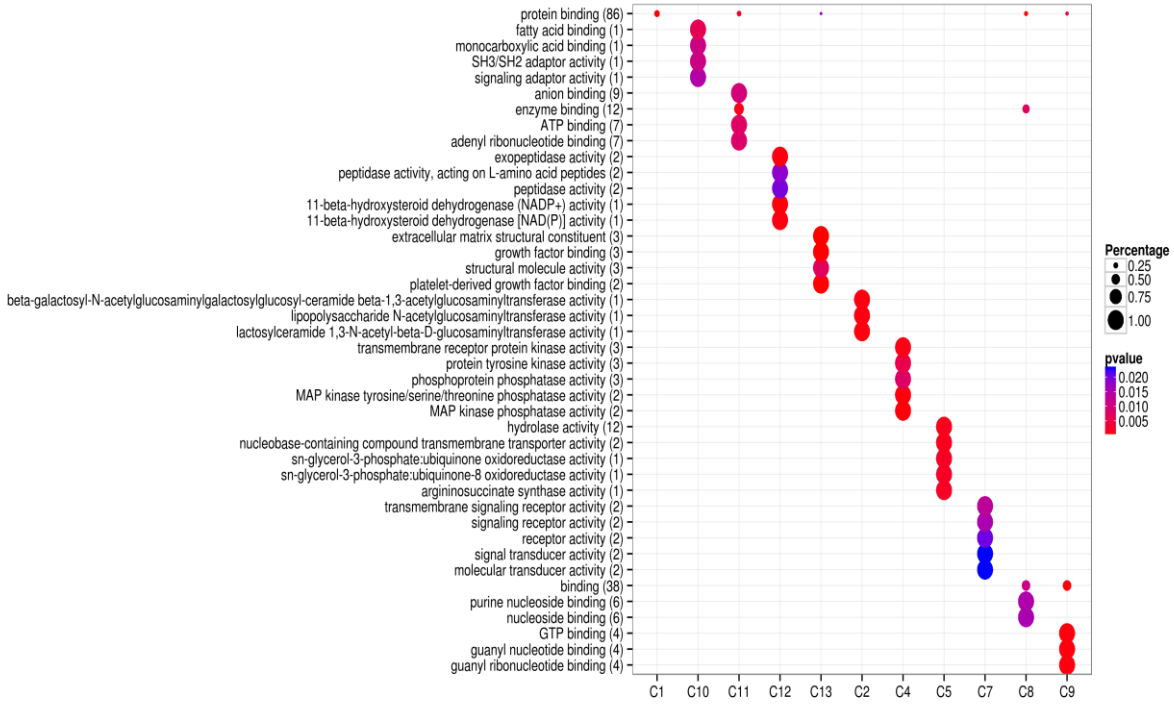


Figure S7.8. Molecular function mapping for clustered heterogeneous DEGs identified by MSG for the human species. C1 – C15 represent clusters identified by Mclust.

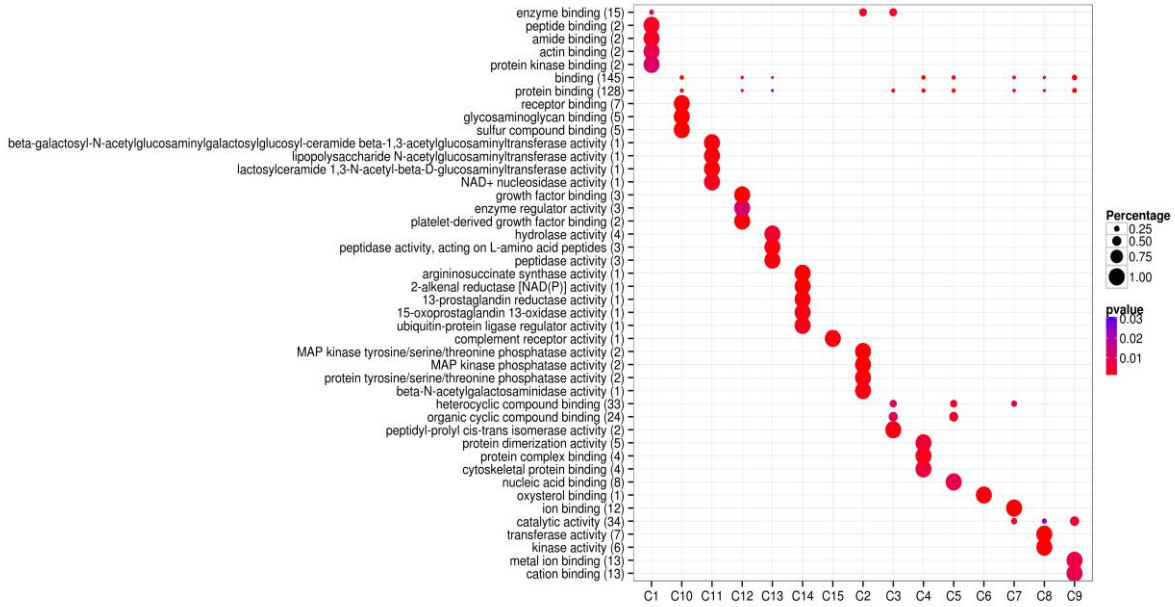


Figure S7.9. Molecular function mapping for clustered heterogeneous DEGs identified by MSG for the mice species. C1 – C15 represent clusters identified by Mclust.

List of submitted paper:

Conference papers:

Yang, Z., **Alwatban, A.,** Everson, R. and Yang, Z.R., 2014. Multi-Scale Gaussian Mixtures for Cross-species Study. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1).

Yang, Z., **Alwatban, A.** and Yang, Z.R., 2014, October. A mean pattern model for integrative study—Integrative self-organizing map. In *Biomedical Engineering and Informatics (BMEI), 2014 7th International Conference on* (pp. 643-648). IEEE.

Al-Watban, A., Yang, Z.H., Everson, R. and Yang, Z.R., 2012, April. A novel data mining approach for differential genes identification in small cancer expression data. In *Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on* (pp. 1-6). IEEE.

Al-Watban, A.S. and Yang, Z.R., 2012, January. Bimodal gene prediction via gap maximisation. In *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)* (p. 163).

Journal Papers:

Wappett, M., Dulak, A., Yang, Z.R., **Al-Watban, A**., Bradford, J.R. and Dry, J.R., 2016. Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs. *BMC genomics*, *17*(1), pp.1-15.

Papers to be submitted soon

Multi-scale Gaussian Mixtures for Differentially Expressed Genes in Cross-species Studies
Abdullatif Alwatban, ZiHua Yang , Suhaib Mohammed, Richard Everson, Zheng Rong Yang.

Temporal species diversity discovery of pre-implantation embryonic development using dual-scale Gaussian mixture
Abdullatif Alwatban, Louisa Knocker, Zheng Rong Yang