The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Investigating the Effect of Paralogs on Microarray Gene-Set Analysis

André Faure

· October 2008 ·

Department of Molecular and Cell Biology
University of Cape Town

Submitted in fulfilment of the requirements for the degree of Master of Science

Supervised by
Nicola Mulder and Cathal Seoighe

University of Cape Town

This work was supported by the National Bioinformatics Network (NBN) and the University of Cape Town (UCT). The views and conclusions in this document are those of the author and should not be interpreted as representing the views of the sponsoring institutions.

Abstract

In order to interpret the results obtained from a microarray experiment, researchers often shift focus from analysis of individual differentially expressed genes to analyses of sets of genes. These gene-set analysis (GSA) methods use previously accumulated biological knowledge from databases such as the Gene Ontology (GO) or KEGG to group genes into sets based on their annotations. They aim to rank these gene sets in a way that reflects their relative importance in the experimental situation in question. The objective is that this approach reveals sets of genes with subtle but coordinated behaviour implicating specific biological processes or pathways in the response under study.

Several GSA methods have been proposed and debates have ensued on the statistical foundations of the different approaches and the various hypothesis tests used. In particular, criticism has been directed at methods that rely on a strict cut-off to determine significant genes and those that assume genes are expressed independently.

We show that paralogs, which typically have high sequence identity and similar molecular functions also exhibit high correlation in their expression patterns. This, together with the fact that the calculation of gene-set significance by all GSA methods is influenced by the number of genes in the gene set, means that sets with high numbers of paralogs are ranked in a biased manner that reflects more the redundant and dependent nature of paralogs than any biological phenomenon.

We investigate the extent of this confounding factor common to all GSA methods and propose a solution in the form of Indygene, a web tool that reduces a supplied list of genes to one that is independent with respect to pairwise paralogous relationships. We use the tool to reanalyse previously published microarray datasets to determine the utility of this pre-processing step.

University of Cape Town

Contents

University of Cape Town

University of Cape Town

University of Cape Town

List of Figures

University of Cape Town

List of Tables

University of Cape Town

List of Abbreviations and Acronyms

%ID global percentage sequence identity

ALL acute lymphoid leukaemia

AML acute myeloid leukaemia

BLAST Basic Local Alignment Search Tool

BLOSUM Blocks Substitution Matrices

cDNA complementary DNA

CDS coding sequence

CVID Common Variable Immune Deficiency

DAG directed acyclic graph

DDC duplication-degeneration-complementation

DNA deoxyribonucleic acid

ES Enrichment Score

FDR false discovery rate

FWER family-wise error rate

GCRMA GC robust multi-array average

GO Gene Ontology

GSA gene-set analysis

GSEA Gene Set Enrichment Analysis

HSP high-scoring sequence pairs

University of Cape Town

KEGG Kyoto Encyclopaedia of Gene and Genomes

LeFE Learner of Functional Enrichment

MGD Mouse Genome Database

MISP maximum independent set problem

ML maximum likelihood

MM mismatched probe

mRNA messenger RNA

MSigDB:C1 MSigDB cytogenetic gene sets

MSigDB:C2 MSigDB functional gene sets

MSigDB Molecular Signatures Database

NASC Nottingham Arabidopsis Stock Centre

PAM Accepted Point Mutation

PAML Phylogenetic Analysis by ML

PBMC peripheral blood mononuclear cells

PCR polymerase chain reaction

PGE2 Prostaglandin E2

PM perfect matched probe

probeID Affymetrix probe set identifier

RNA ribonucleic acid

rRNA ribosomal RNA

SAFE significance analysis of function and expression

SAGE Serial Analysis of Gene Expression

SAM Significance Analysis of Microarrays

SGD *Saccharomyces* Genome Database

UniProtKB UniProt Knowledgebase

University of Cape Town

Acknowledgements

I thank my supervisors Dr. Nicky Mulder and Prof. Cathal Seoighe for giving me the opportunity to be involved in the dynamic research fields of computational biology and bioinformatics. Their insight and helpful guidance gave me motivation and support throughout the entire research experience. A special thank you to Natasha Wood and all other members of the UCT Computational Biology research group for providing an enjoyable atmosphere that enabled stimulating discussions. Thanks also to Rodger Duffett whose technical skills are responsible for the lab's excellent computing infrastructure. Lastly I would like to thank my family and close friends for their unwavering encouragement.

Trieste, October 2008

University of Cape Town

Chapter 1

Background

University of Cape Town

1.1 DNA Microarrays

Gene expression is the process whereby genetic information stored in a stable form such as DNA is used as a template to produce functional gene products from the genes that encode them. The actions and properties of each cell type at a particular instant are largely determined by the diversity and concentration of these expressed gene products. By regulating gene expression the cell can control the concentrations of these gene products and therefore their level of activity in the cell. This mechanism of gene regulation helps to characterise distinct cell types and allows them to respond to changes in their environment (Lodish et al. 2001).

The process of gene expression consists of multiple steps, each of which provides the opportunity for regulation and therefore has the ability to affect the quantity of the resulting gene product. Molecular mechanisms affecting the amount of a particular messenger RNA (mRNA) start with transcription initiation control e.g. modulation of the levels and/or activities of activators and repressors and changes in chromatin structure. Other downstream mechanisms include regulation of RNA processing and nuclear transport (Lodish et al. 2001). Although the rate of synthesis of specific protein gene products is subject to mRNA degradation and translational and post-translational control mechanisms, measuring mRNA transcript abundance in a collection of cells provides a convenient estimate of their levels. By quantifying mRNA transcript levels in cells from different tissues and

under different conditions, one can gain insight into the biological mechanisms underlying those differences (Simon et al. 2003).

DNA microarrys are tools for quantifying the types and amounts of mRNA transcripts present in a sample of cells at a particular time point. Different types of microarrays exist, but they generally consist of a solid support surface on which strands of polynucleotides, or probes, have been attached at pre-determined positions. mRNA isolated from a specimen is then converted to form labelled polynucleotides, or targets, which are then washed over the microarray. The labelled targets then hybridise to probes possessing sufficient Watson-Crick complementarity to them, forming heteroduplexes. After washing the excess sample off the solid surface, only labelled target that is bound to its complementary probe should remain. Measuring the intensity of the target label at each probe on the microarray provides an estimate of the relative quantity of mRNA in the specimen and thus the level of expression of each corresponding gene (Simon et al. 2003). Advances in fabrication technology and techniques have made current microarrays' microminiaturized and highly parallel format possible. This together with the rapidly growing number of fully sequenced genomes has resulted in devices with the ability to measure the expression levels of all the genes in an entire genome (Heller 2002).

The two DNA microarray systems most commonly used today are printed, or spotted, microarrays and the Affymetrix GeneChip system (Affymetrix Inc. 2008). These technologies differ in many respects including cost, target preparation and results analysis. However, each technology has individual benefits depending on the specific application for which it is intended (Knudsen 2004).

University of Cape Town

1.1.1 Printed Microarrays

These microarrays consist of probes of complementary DNA (cDNA), PCR product or oligonucleotides that have been printed on a microscope slide by a robotic spotter. Each probe is complementary to a unique gene and is normally bound to the surface of the slide by a poly-lysine or poly-amine coating (Knudsen 2004). Because cDNA probes are generally hundreds of bases long, hybridization conditions are relatively specific and cross-reactivity is limited. However, the robotic printing process introduces substantial variability in the size and shape of the spots, and the distribution of labelled sample across the face of the

array is often neither uniform nor consistent across different arrays. This makes comparisons of gene expression levels across arrays difficult. This problem is characteristic of all printed microarrays, independent of the spotted probe type. To overcome some of this interarray variability, two samples labelled with different fluorescent dyes e.g. Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green), are usually co-hybridized on the same array. In this case, the second sample may either be a reference for use on all arrays to control for experimental variability, or represent a specimen of biological interest. The intensities of the two different fluorescence frequencies corresponding to the two samples can then be measured using separate laser sources (Simon et al. 2003). For the aforementioned reasons, absolute levels of gene expression are not normally determined with this method, however a benefit of this technology is that it provides a great deal of flexibility in the choice of arrayed elements. Smaller, customized microarrays can be designed for specific investigations (Knudsen 2004).

1.1.2 Affymetrix GeneChip Microarrays

In microarrays produced by Affymetrix, short 25mer oligonucleotides are synthesised directly onto silicon chips using a photolithographic process (Fodor et al. 1991). In this iterative process, each step involves using a mask to control the light actuated synthesis and attachment of a single nucleotide to anchors or growing oligonucleotide chains at specific positions on the chip. This results in probes on GeneChip arrays being more homogenous compared to those on printed microarrays, reducing interarray variability and enabling estimates of the absolute value of gene expression (Simon et al. 2003). Therefore a single sample is usually hybridized to GeneChips – comparisons of two samples require the use of two separate microarrays.

However, the relatively short 25mer oligonucleotide probes result in substantial cross-hybridization. Affymetrix has attempted to deal with this problem by providing multiple probe pairs for each target transcript. For each probe that is a perfect match (PM) to its target sequence, Affymetrix also includes a mismatched probe (MM) that is identical to the PM except for a single nucleotide mismatch located directly in the middle of the 25-base probe sequence. The manufacturers argue that a better estimate of the intensity due to hybridisation to the true target transcript is obtained by subtracting the signal intensity at the mismatched probe from that at the perfectly matched probe (Affymetrix Inc. 2008).

University of Cape Town

1.1.3 Microarray Data Pre-Processing

After the hybridisation reaction and stimulation of the array with a laser, an image file is created that stores the fluorescence intensities at different pixel locations on the array. Image analysis is the process whereby an intensity value for each spot or feature on the array is extracted from this pixel-level information using computer software. It consists of a number of steps including gridding to locate the spot positions, segmentation or separation of each spot from the background, foreground intensity extraction and background correction. Typical foreground intensity extraction measures involve taking the mean or median of the pixel intensities (Simon et al. 2003). Background correction is the process whereby the foreground intensity at each spot is corrected for non-specific binding and auto-fluorescence. The intensity attributable to these non-biological effects is usually estimated by taking the background signal between spots. However the most common approach is to subtract a globally, or regionally estimated background as opposed to a local value that has the potential to introduce more noise into data than is eliminated (Knudsen 2004).

Importantly, an array normalisation procedure needs to be carried out before gene expression values can be compared between arrays. Intensity imbalances between different RNA samples occur because of non-biological reasons and different normalisation methods may need to be applied depending on the specific effects present on the arrays under study (Simon et al. 2003). The aim is to assign similar expression values to genes that are truly non-differentially expressed across arrays. This can be accomplished by adjusting the array expression values according to intensities observed in biologically stable housekeeping genes or spiked control genes that are artificially introduced into the sample (Knudsen 2004). The normalisation algorithms then use these genes as the basis for performing the adjustment and range from simple linear or global normalisations to combination location and intensity normalisations.

1.2 Identifying Differentially Expressed Genes

After the requisite steps, including microarray image analysis, quality control and data normalisation have been completed, interarray gene expression levels can be compared.

University of Cape Town

An important application of DNA microarray technology deals with identifying genes that are differentially expressed between prespecified classes of arrays. The classification of expression measurements from each microarray experiment could be according to differences in experimental circumstances, tissue source or some other biological condition of interest. The goal is to obtain a list of genes that are responsible for the biological differences between the classes. When the genes in the list have known molecular functions, they can help to explain the molecular underpinnings of the class representing the samples under study. On the other hand, a gene with unknown function's occurrence in the list of differentially expressed genes can help to characterise its function (Simon et al. 2003). Here we discuss the situation of two classes, where the aim is to identify genes that have higher expression in one class compared to another.

Suppose there are J_1 and J_2 microarrays representing experimental replicates in class 1 and 2 respectively. The gene expression measurements for a particular gene in class 1 and 2 can be summarised by the means of their class values, $\bar{x}_1$ and $\bar{x}_2$. An early approach used the fold change of these mean expression values to quantify a gene's expression change between two classes (Lee et al. 1999). Choosing a twofold difference to be significant would correspond to identifying genes satisfying $|\bar{x}_1 - \bar{x}_2| \geq 1$ as differentially expressed, where the typically used base 2 logarithmic transformation of expression data has been used. However, this approach has a high probability of falsely declaring genes to be differentially expressed (Miller et al. 2001). This is because using fold change alone does not incorporate information about the variability of expression values in each class and is therefore not valid for making statistical inferences concerning differential expression (Allison et al. 2006).

University of Cape Town

1.2.1 Statistical Hypothesis Tests

The general starting point for these statistical hypothesis tests is to assume that gene expression measurements in both classes are sampled from the same underlying population. This is referred to as the null hypothesis. The procedure is then to determine the probability of obtaining the observed differences in class measurements, or those more extreme, under this null hypothesis. This probability is quantified with a so-called P-value that can be used to control the proportion of the time the null hypothesis is rejected when it is true. The lower this P-value, the more statistically significant the data is said to be and the more confident one can be in rejecting the null hypothesis and declaring the gene as differentially

expressed.

There have been various test-statistics proposed for such a situation and their differences originate from their assumptions about the underlying population distributions. The most common of which is the two-sample t-statistic (Simon et al. 2003),

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{J_1} + \frac{1}{J_2})}} \tag{1.1}$$

where s_p is an estimate of the pooled within-class variability,

$$s_p^2 = \frac{(J_1 - 1)s_1^2 + (J_2 - 1)s_2^2}{J_1 + J_2 - 2} \tag{1.2}$$

which can be used when equal variance between classes is assumed. s_1^2 and s_1^2 are the sample variances for class 1 and 2 respectively. When equal variance between classes cannot be assumed, Welch's t-statistic can be used, where the denominator in Equation 1.1 is replaced with $\sqrt{s_1^2/J_1 + s_2^2/J_2}$. The above t-statistic and its variants can be viewed as the ratio of between-class to within-class variability of gene expression values. Because of the cost and difficulty in performing large numbers of microarray experiments, the number of replicates in each class is often low. This leads to poor estimates of within-class variability and can make small fold changes statistically significant when it is underestimated. A simple solution is to only consider genes with fold change at least 2, thereby preventing low P-values largely occurring as a result of inaccurate variance estimates (Knudsen 2004). Significance Analysis of Microarrays (SAM) uses a similar approach, where a small constant is added to the denominator of the test statistic to circumvent this issue (Tusher et al. 2001).

Other methods base the within-class variance estimation not only on a single gene's measurements, but include variance estimates from all genes present on the array. This approach, called variance shrinkage, uses all the data simultaneously thereby capitalising on the highly parallel nature of microarrays. Several methods have been developed (Baldi and Long 2001, Cui et al. 2005), which all seem to work reasonably well and lead to more powerful testing when class replicates are low (Allison et al. 2006).

The above tests all implicitly assume that the gene expression data used follows the nor-

University of Cape Town

mal distribution. Although deviations from the normal distribution have been shown to be small for microarray data (Giles and Kipling 2003), an alternative nonparametric approach to determining differential expression is a permutation test. The permutation t-test compares the t-statistic, as in Equation 1.1, to the distribution of the t-statistic, t^* obtained after randomly permuting the class labels of the $J_1 + J_2$ microarrays (Simon et al. 2003). The two-sided P-value from the permutation t-test is estimated by

$$P\text{-value} = \frac{1 + \#\text{ of random permutations where } |t^*| \geq |t|}{1 + \#\text{ of random permutations}} \tag{1.3}$$

or, when all permutations can be enumerated,

$$P\text{-value} = \frac{\#\text{ of permutations where } |t^*| \geq |t|}{\frac{(J_1+J_2)!}{(J_1!J_2!)}} \tag{1.4}$$

Another nonparametric permutation test that can be used is the Wilcoxon rank-sum test, in which the gene expression values are replaced with their ranks (Hollander and Wolfe 1999). The P-value obtained from this test is the same as that obtained when the permutation t-test is performed on the ranked data. Although an advantage of this test is that it is relatively insensitive to extreme values, it can also be insensitive to real differences between the two classes. This, together with the availability of modern computers able to handle the permutations present in the permutation t-test, usually make this the preferred alternative (Simon et al. 2003).

University of Cape Town

1.2.2 Multiple Hypothesis-Testing

The effect of multiple testing needs to be considered when hypothesis tests are applied across all genes in a microarray experiment to determine which, if any are differentially expressed. Under the null hypothesis of no genes being differentially expressed, an average of 5% of genes will still be identified as significant using a P-value cut-off of $\alpha = 0.05$. With the number of genes on a microarray often on the order of thousands, this can translate into an unacceptable number of false positives (Knudsen 2004). Initially developed methods, such as the Bonferroni correction, were designed to control the family-wise error

rate (FWER). These methods limit the probability of falsely rejecting one or more null hypotheses to below the α-value across the entire experiment (Simon et al. 2003).

However, controlling the FWER is a very conservative approach and most biologists are willing to accept some erroneous conclusions if this allows for discoveries to be made (Allison et al. 2006). The false discovery rate (FDR) is the average proportion of false positives amongst the genes identified as being differentially expressed. Benjamini and Hochberg (1995) first coined the term FDR and developed a method for controlling it at a specified level. After ranking the genes by their P-value from lowest to highest and starting at the top of the list, all genes are accepted as differentially expressed where

$$P\text{-value} \leq \frac{i}{m}q \tag{1.5}$$

where i is the number of genes accepted so far, m is the total number of genes tested and q is the desired FDR. The FDR can also be estimated by permutation as is done in the SAM method (Tusher et al. 2001). Here the class labels are permuted and each hypothesis test, for example a two-sample t-test, is repeated for all genes. The number of individual null hypotheses rejected using the permuted data, divided by the number of individual null hypotheses rejected for the unpermuted data gives an estimate of the FDR at the chosen α-value.

University of Cape Town

1.3 Gene-Set Analysis

Obtaining a list of genes that can be confidently declared differentially expressed is usually the starting point of a difficult and complicated process of interpretation. The aim is to translate the results from a differential expression analysis into useful knowledge about the relevant biological differences in a microarray experiment. In an attempt to find patterns in such a resulting list, researchers initially sifted through relevant publications and gene annotations manually, to find biological themes associated with the genes therein (Kayo et al. 2001). This can be arduous and time-consuming, with researchers reportedly spending approximately 200 hours on the task in one case (Hosack et al. 2003).

The creation of the Gene Ontology (GO) (Ashburner et al. 2000), with its consistent and

machine-readable gene and gene product annotation system, enabled the development of computer software able to perform this type of analysis automatically. A flood of tools for this purpose have been developed in recent years, all of which are based on the principle of grouping genes into sets, based on their annotations (Khatri and Draghici 2005). These gene-set analysis (GSA) methods then aim to rank these sets in a way that reflects their relative importance regarding the observed gene expression changes. The genes annotated to the biological themes identified as significant can then be duly investigated and assayed in a wet-lab environment. Apart from the obvious time- and labour-saving benefits of this automatic approach, the incorporation of an independent representation of previously accumulated biological knowledge into the analysis has proven to be powerful (Allison et al. 2006). Shifting the focus from individual genes to sets of genes has also been shown to identify biological themes more consistent across independent studies than results from single-gene analyses (Subramanian et al. 2005).

Although biologists are usually interested in identifying differential expression (Allison et al. 2006), and GSA is often used to interpret results from such an analysis, GSA has also been used to interpret results from expression cluster analysis. Clustering methods are used to form subgroups of objects where each subgroup, or cluster, contains objects that are more similar to each other than objects in different clusters. Unsupervised clustering methods are used to discover patterns in a dataset where no additional information about the existing structure is known (Simon et al. 2003). Because of the high dimensionality of microarray data these methods have been useful and extensively employed in the visualisation and analysis of gene expression data (Alizadeh et al. 2000, Tamayo et al. 1999). When genes are clustered according to their expression profiles, the resulting clusters contain genes whose expression patterns are most similar to each other with respect to a given similarity metric. Although the resulting subgroups are largely dependent on the specific clustering algorithm and parameters used, the hope is that the clusters obtained reflect a common role or related function of the member genes. GSA has be used to highlight biological themes shared by genes that have been clustered together, based on their expression profiles (Hosack et al. 2003). Others have also used GSA to analyse data from the Serial Analysis of Gene Expression (SAGE) (Velculescu et al. 1995) and data generated from proteomics research (Hosack et al. 2003).

Although the use of annotations from the GO for GSA has been popular since its inception, other databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG)

University of Cape Town

(Kanehisa and Goto 2000) and the Molecular Signatures Database (MSigDB) (Subramanian et al. 2005) have also been used. In the following subsection, we give a brief overview of the different biological databases and annotation sources that have been used to form gene sets in GSA.

1.3.1 Biological Databases for GSA

The Gene Ontology project was started to address the need for consistent descriptions of gene products in different organisms. The researchers from three different model organism databases: Flybase (*Drosophila*), *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD), upon discovering high levels of orthology and functional conservation between genes in different species, recognised the need for a common language of annotation. It has subsequently become the 'gold standard' for annotating genes and gene products, expanding to include information from many other genome repositories and enabling dynamic maintenance and interoperability between them (Harris et al. 2004). The ability to group genes into gene sets based on their GO annotations has also made it a routine choice for use in GSA (Yang et al. 2008).

An ontology consists of a formal set of well-defined terms with well-defined relationships that describe a domain of knowledge. The GO consists of three independent ontologies that serve as a representation of biological knowledge about the roles of genes and gene products in the cell. Terms in the molecular function ontology describe the biochemical activity of a gene product, for example 'catalytic activity, 'transporter activity' or 'binding'. The biological process ontology contains terms referring to the multi-step biological objective to which a gene or gene product contributes, by the involvement of its molecular function. Examples include 'growth' and 'response to stimulus'. Lastly, terms in the cellular component ontology refer to the place in the cell where the gene product is localised or active, for example 'ribosome' or 'nuclear membrane' (Ashburner et al. 2000).

The terms are structured in a directed acyclic graph (DAG) with more specific terms represented as children of more general terms. The DAG structure allows each term to have more than one child or parent, where the parent-child relationships can be of type 'is_a' or 'part_of' (Harris et al. 2004). Terms in the biological process ontology can also be connected to other biological processes, molecular functions or biological qualities by the

University of Cape Town

'regulates', 'positively_regulates' or 'negatively_regulates' relationships (Gene Ontology Consortium 2008). These relationships and their transitive properties allow complex vocabulary systems to be developed using reason-based structures (Ashburner et al. 2000). They also have important implications for GSA, as annotation of a gene to a particular GO term implies that it is also annotated to all parents of that term. This is referred to as the 'true path rule' (Gene Ontology Consortium 2008). A useful consequence of this rule is that the level of detail for a GSA can be adjusted by restricting the focus to terms at a particular level in the GO DAG. GO SLIMs, available for numerous organisms, have been created for this purpose. They consist of high-level terms hand-picked from each ontology by experienced biologists and can be used in GSA to obtain a broad overview of the important biological themes in a microarray experiment.

Although the GO 'biological process' ontology contains pathway-related terms, it is not intended to represent details of these reaction pathways (Ashburner et al. 2000). This type of information is available in pathway databases such as KEGG (Kanehisa and Goto 2000), GenMAPP (Gladstone Institutes 2008) and Biocarta (BioCarta Inc. 2008), which contain manually drawn maps and descriptions of molecular interactions and reaction networks, showing the roles of the annotated gene products. Gene sets for GSA have been formed by common annotations to terms in these databases.

Other examples include genes sets based on cytogenetic bands, or chromosomal proximity, shared cis-regulatory motifs, transcription factor binding sites or protein domains from InterPro (Nam and Kim 2008). Moreover, the Molecular Signatures Database (MSigDB) (Broad Institute 2008) contains five major collections of human gene sets. They were developed for use with the Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) software, but are freely available for download and use with other GSA tools. The collections include chromosomal position, regulatory motif and selected GO term gene sets. The remaining two consist of sets of genes co-regulated in microarray studies, and a collection of curated gene sets from online pathway databases, publications in PubMed and knowledge of domain experts. There are a wide variety of options for gene-set definitions for GSA and the choice depends on the specific application and biological question being asked.

University of Cape Town

1.3.2 GSA Methods

Having discussed the sources of information used to construct gene sets, we now turn our attention to the specifics of GSA methods themselves. This section details the various statistical approaches proposed for exploiting gene-set information in the interpretation of microarray data. Using the classification system proposed by Goeman and Buhlmann (2007), GSA methods can be separated into three broad categories. The initially developed and most popular methods start from a list of differentially expressed genes and use a statistical hypothesis test to determine which gene sets are over- or underrepresented in this list (Khatri and Draghici 2005). Methods in the second category attempt to find patterns in the entire vector of P-values generated from a differential expression analysis. The third category of methods start from the raw expression data and base their hypothesis tests on the genes in each gene set in isolation, usually without regard for other genes on the microarray (Goeman and Buhlmann 2007). We provide details and discuss shortcomings of selected methods from each category in the following subsections.

University of Cape Town

1.3.2.1 Strict Cut-Off Methods

These GSA methods aim to determine whether a gene set is over- or underrepresented among the genes satisfying a certain criterion in a microarray experiment consisting of m genes. Typically, the criterion is based on a strict P-value or FDR cut-off for significant differential expression. Given this list of differentially expressed genes, with length m_D, and the list of genes in a gene set, with length m_G, it is possible to constuct a 2×2 contingency table as indicated in Table 1.1.

	In gene set ($\in$GO category)	Not in gene set ($\notin$GO category)	Total
Differentially expressed gene	m_{GD}	m_{G^cD}	m_D
Non-differentially expressed gene	m_{GD^c}	$m_{G^cD^c}$	m_{D^c}
Total	m_G	m_{G^c}	m

Table 1.1: A 2×2 contingency table for assessing overrepresentation of a gene set among the differentially expressed genes in a microarray analysis.

The P-value for the overrepresentation of a gene set among the differentially expressed genes can be calculated using a statistical test for independence. The null hypothesis,

H_0, states that the m_D differentially expressed genes are selected at random from the m genes on the microarray. An equivalent formulation of the null hypothesis is that there is no association between gene-set membership and propensity to be differentially expressed. Fisher's exact test gives the exact probability of obtaining a value at least more extreme than the observed m_{GD}, for the random variable M_{GD}, under this null hypothesis. The test is based on the hypergeometric distribution where the probability of observing a particular value for the random variable M_{GD} is given by,

$$\mathrm{P}(M_{GD} = x) = \frac{\binom{m_G}{x}\binom{m_{G^c}}{m_D - x}}{\binom{m}{m_D}} = \frac{m_G! m_{G^c}! m_D! m_{D^c}!}{m! x! (m_G - x)! (m_G - x)! (m_{D^c} - m_G + x)!} \quad (1.6)$$

The P-value for overrepresentation is then calculated by summing the probabilities of each possible realisation of M_{GD} at least as great as the value observed, given the marginal totals in Table 1.1 (Ewans and Grant 2005). Alternatively, the less biologically interesting situation of underrepresentation of a gene set can be assessed by considering values at least as small as the observed value for M_{GD}. Popular tools such as GoMiner (Zeeberg et al. 2003), EASEonline (Hosack et al. 2003) and FatiGO (Al-Shahrour et al. 2004) perform this type of analysis using Fisher's exact test. The hypergeometric distribution can be approximated by the binomial distribution and the normal distribution under certain conditions, although this only offers advantages in terms of computational tractability. Tools such as CLENCH (Shah and Fedoroff 2004)and GO::TermFinder (Boyle et al. 2004) offer alternative statistical tests based on these approximate null distributions. Another alternative to the Fisher's exact test is the χ^2 test which can be used when the expected values of the cells in Table 1.1 are not too low or unequally distributed, given the marginal totals. GOstat (Beissbarth and Speed 2004), GoSurfer (Zhong et al. 2004) and the NetAffx GO Mining Tool (Cheng et al. 2004) offer this test. Whatever the sample sizes and observed 2×2 contingency table cell frequencies, Fisher's exact test, based on the exact null distribution of M_{GD} given by the hypergeometric distribution, is the recommended option for assessing overrepresentation (Rivals et al. 2007).

Allison et al. (2006) offer criticism for the methods in this category which test for overrepresentation of a gene set in the list of differentially expressed genes. They argue that by

University of Cape Town

focussing only on the significant genes satisfying an arbitrary threshold, information about the continuous evidence supporting differential expression is lost. Pan et al. (2005) show that the choice of threshold can severely influence the biological conclusions drawn from analyses using these methods. More fundamentally, they criticize the statistical models themselves, which take the gene rather than the case as the sampling unit, relying on gene randomisation to assess significance. Goeman and Buhlmann (2007) point out that this is inconsistent from a microarray experimental design viewpoint, where a replication of the experiment involves a new sample of subjects, which are subjected to the same measurements, not a new sample of genes from the same subjects. The models are also inappropriate as they are based on the highly unrealistic assumption that gene transcripts are expressed independently. Goeman and Buhlmann (2007) illustrate the consequences of the violation of this assumption by way of a simulation experiment. The simulation shows that gene-set P-values are often falsely significant when they contain genes with highly correlated expression patterns. For this reason, tests that use sample randomisation to assess statistical significance are widely acknowledged as more appropriate (Nam and Kim 2008) as this results in more realistic P-values and less false positives.

University of Cape Town

1.3.2.2 Methods That Use the Entire Vector of P-Values

The methods in this category were developed in response to limitations in the strict cut-off methods described in the previous section. Issues such as their reliance on and sensitivity to a pre-defined threshold for differential expression are problematic, especially when no differentially expressed genes are determined. For this reason, these methods instead aim to find distribution patterns of gene-set members in the sorted gene list resulting from a differential expression analysis.

The method proposed by Al-Shahrour et al. (2005), divides the list into K partitions and then uses the Fisher's exact test to determine which of the M gene sets are overrepresented in each partition. This approach is similar to that discussed in the previous section, except that significant gene sets are determined for each partition and not just for the partition of significantly differentially expressed genes. Therefore $K \times M$ tests are carried out as opposed to the M tests in the traditional approach. Using a recommended K between 20 and 50, this results in a considerable multiple testing problem, which is addressed by calculating FDR-adjusted P-values (Benjamini and Yekutieli 2001). The adjusted values,

representing overrepresentation in the K-th partition, are then plotted on a XY-graph against the value of the statistic, or P-value for differential expression, corresponding to the partition. The authors mention that gene sets found to be significant across a range of partitions will appear clustered together in the graph and can be more trusted as representing the underlying biology in the microarray experiment (Al-Shahrour et al. 2005). Another related approach proposed by Breitling et al. (2004), sorts the list of genes from a microarray experiment according to their fold-change. It then iterates over the list, from highest to lowest fold-change, calculating a score based on the probability of encountering the number of gene-set members observed in the list so far. The minimum score for a particular gene-set is called the PC-value and its statistical significance is determined by comparing this value to those obtained after randomly permuting the gene list a large number of times.

These two methods only address the issues related to using a strict cut-off and still base the calculation of significance on gene randomisation, similarly to the methods in the previous section. Gene set enrichment analysis (GSEA) (Subramanian et al. 2005) ranks the genes in a microarray experiment according to their expression correlation with a two-class phenotype. The ranked gene list of the form $L = \{g_1, \ldots, g_m\}$ is then used to calculate an enrichment score (ES), which reflects the enrichment of genes in a gene set G towards the extremes of the list. Gene sets whose genes are non-randomly distributed in L, and therefore have a higher ES, are expected to be more related to the gene expression differences between the two phenotype classes. The ES is calculated by stepping down the list L and increasing a weighted Kolmogorov-Smirnov statistic when a gene in G is found and decreasing it otherwise. The ES is defined as the maximum value that the running sum, $|P_{hit} - P_{miss}|$ attains, where

$$\begin{aligned} P_{hit}(G,i) &= \sum_{\substack{g_j \in G \\ j \leq i}} \frac{|r_j^p|}{m_R}, \qquad \text{where} \qquad m_R = \sum_{g_j \in G} |r_j^p| \\ P_{miss}(G,i) &= \sum_{\substack{g_j \notin G \\ j < i}} \frac{1}{(m - m_G)} \end{aligned} \tag{1.7}$$

and where $r(g_j) = r_j$ is the expression correlation of gene j with the two-class phenotype,

University of Cape Town

p is an exponent to control the weight of the step and m_G is the total number of genes in the gene set. With $p = 1$, the ES is incremented by an amount dependent on each gene's normalised correlation with the phenotype or, in other words, its position in the list. This ensures that a relatively uninteresting gene set with genes that are mostly unchanged, and therefore concentrated towards the middle of the list L, do not receive a large ES. With $p = 0$, the ES reduces to the traditional Kolmogorov-Smirnov statistic, where ES is incremented by an amount only dependent on the gene set size i.e. $1/m_G$ for a gene in G or $1/(m - m_G)$ for a gene not in G.

To assess the statistical significance of the ES, GSEA uses sample randomisation, rather than gene randomisation as in the previously mentioned methods. Each P-value is calculated by comparing the obtained ES to its estimated null distribution, which is obtained by permuting the class labels many times and recalculating the ES for each permutation. The authors state that this method of determining significance preserves the complex gene-gene correlation structure in the data and produces more biologically reasonable P-values (Subramanian et al. 2005).

The significance analysis of function and expression (SAFE) (Barry et al. 2005) is a similar, but more general approach to GSEA, providing a framework for gene set hypothesis testing. It starts by generating an ordered list of genes based on so-called 'local' statistics, such as ordinary t-statistics, and then uses 'global' statistics to detect a shift of the genes in a gene set towards the extremes of the list. Examples of possible global statistics include the Wilcoxon rank sum or the Kolmogorov-Smirnov statistic. As in GSEA, SAFE uses sample randomisation to assess the significance of the global statistics.

GSEA has been shown to produce interesting and biologically relevant results (Mootha et al. 2003), even in cases where no genes were found to be differentially expressed ($\alpha = 0.05$) after a multiple testing correction was applied. The authors of this method also show the power of incorporating independently obtained biological information into the analysis of micorarray data by using the tool to highlight biological themes consistent across multiple studies that share little similarity at the individual gene level (Subramanian et al. 2005). However, a number of authors have criticised the approach. Damian and Gorfine (2004) first recognised that the scores assigned to gene sets by GSEA are affected by the presence or absence of other gene sets. For instance, some gene sets are 'penalised' as a result of the presence of other gene sets containing highly differentially expressed genes. In

University of Cape Town

a review, Allison et al. (2006) subsequently referred to this situation as a 'zero-sum-game', where the weight of evidence supporting differential expression of one gene set is judged relative to the other background gene sets.

Goeman and Buhlmann (2007) explain these observations by clarifying the differences between various GSA methods in a formal manner. Firstly, a distinction is made between GSA methods based on their null hypotheses. They are either competitive, H_0^{comp}, or self-contained, H_0^{self}, where their general formulations are given by,

$$H_0^{comp} \quad : \quad \text{The genes in } G \text{ are as often differentially expressed as the genes in } G^c.$$

$$H_0^{self} \quad : \quad \text{No genes in } G \text{ are differentially expressed.}$$

where G represents the gene set of interest and G^c its complement. Methods of the strict cut-off variety are competitive in that they compare the relative enrichment of gene-set members in the list of differentially expressed genes, to the complementary or background gene list. This is an equivalent formulation to that given for H_0^{comp} above. The authors show that it is a natural choice for methods testing a competitive null hypothesis to determine statistical significance by gene randomisation. On the other hand, assessing significance using sample randomisation is the intuitive alternative for methods testing a self-contained null hypothesis. GSA methods testing a self-contained null hypothesis will be discussed in the following subsection. GSEA is a hybrid GSA method in that its choice of a Kolmogorov-Smirnov-like statistic is motivated by a competitive null hypothesis, whereas it determines significance of each ES using sample randomisation. Goeman and Buhlmann (2007) offer this as a reason for its low power in some instances. They also offer strategies for transforming existing GSA methods, which test competitive null hypotheses, to ones that test self-contained null hypotheses. The latter GSA methods avoid issues relating to the relative scoring of gene sets and the problematic assumptions made when performing gene randomisation to assess statistical significance.

University of Cape Town

1.3.2.3 Methods That Model the Raw Expression Data Directly

The main distinguishing factor between methods in this section and the preceding one is the lack of two separate steps in the methodologies of the former. These methods start from the raw expression data, as opposed to first obtaining a ranked list of genes on which to perform *post hoc* hypothesis testing.

The SAM procedure uses a t-like statistic to test whether an individual gene in a microarray experiment is differentially expressed. SAM-GS, developed by Dinu et al. (2007), is an extension of the SAM procedure to identify gene sets showing significant differential expression. The null hypothesis is self-contained and states that a gene set is not differentially expressed across a two-class phenotype. Given a gene set G, where $\bar{x}_1(j)$ and $\bar{x}_2(j)$ are defined as the average levels of expression for gene j in classes 1 and 2 respectively, the $SAMGS$ test statistic is,

$$SAMGS = \sum_{i=1}^{|G|} d_i^2 \tag{1.8}$$

where,

$$d(j) = \frac{\bar{x}_1(j) - \bar{x}_2(j)}{s(j) + s_0} \tag{1.9}$$

The 'gene-specific scatter' $s(j)$ is a pooled standard deviation over the two classes and s_0 is a small positive constant to circumvent issues related to underestimations of variability (see 1.2.1). The $SAMGS$ test statistic gives a summary of the standardised differences of all the genes in a gene set. In accordance with its self-contained null hypothesis, SAM-GS evaluates significance by way of sample randomisation, where a P-value is calculated by comparing the test statistic to its null distribution obtained by permuting the microarray class labels many times (Dinu et al. 2007). Importantly, SAM-GS was developed to detect bidirectional gene expression changes. Therefore, a significant P-value merely indicates that the genes in the gene set exhibit substantial expression change between the two phenotype classes without distinguishing between differentially up- or down-regulated genes. Although many of the methods already discussed are intended to detect gene sets that

University of Cape Town

are regulated in only one direction, some methods including strict cut-off methods and GSEA can easily be altered to test for gene sets with bidirectional gene expression changes (Saxena et al. 2006).

Other methods in this category which test a self-contained null hypothesis include Global Test (Goeman et al. 2004) and a related approach called ANCOVA Global Test (Mansmann and Meister 2005). Global Test tests whether subjects, or microarray experiments, with similar gene expression profiles have similar class labels, based on a logistic regression model. Applied to a gene set, it tests how well the expression profiles of the member genes are able to predict the class labels. The Global Test is versatile in that it can be applied in diverse microarray experimental design situations including two-class, multi-class, continuous and survival outcome types (Goeman et al. 2005). Another useful feature of this method is that it can be applied to the gene set comprising all the genes on the microarray as an initial quality check. If the result of the test is not significant, it is unlikely that there are many differentially expressed genes present (Goeman et al. 2004). Compared to Global Test, ANCOVA Global Test has the roles of class labels and gene expression profiles exchanged in regression models, and its authors point out that it performs better than Global Test in certain situations (Mansmann and Meister 2005).

Liu et al. (2007) performed a comparative evaluation of the three aforementioned GSA methods using a simulation experiment and three real-world microarray datasets. All three methods display similar performance, except SAM-GS exhibits slightly higher power with regard to highly significant, and therefore highly interesting, gene sets. SAM-GS however has the comparative disadvantage of only being applicable in the situation of a two-class phenotype (Liu et al. 2007). A drawback common to all GSA methods using sample randomisation is that a large number of permutations is needed to obtain low P-values. It may also be impossible to obtain P-values below the $\alpha = 0.05$ significance level in microarray experiments with a low number of replicates in each class (Goeman et al. 2004).

Two other GSA methods which fall into this category are PathwayRF (Pang et al. 2006) and the Learner of Functional Enrichment (LeFE) (Eichler et al. 2007), which both use machine learning approaches to analyse gene expression data in terms of gene sets. The field of machine learning is involved with the development of computer software that automatically improves with experience. Both PathwayRF and LeFE use random forest, which is a type

University of Cape Town

of classifier whose purpose is to place items (in this case microarray experiments) into groups based on their attributes (in this case quantitative gene expression information). The random forest algorithm builds an ensemble of decision trees based on training gene expression data and a measure of performance related to its ability to correctly predict the class labels of unseen microarrays. Each tree consists of nodes representing a test of an attribute of an item, and branches corresponding to a possible value of the attribute. The tree can then be used to classify a microarray experiment not in the training data by sorting it down the decision tree from its root to a leaf that provides the resulting classification (Mitchell 1997). The random forest's resulting classification and error rate is obtained by aggregating information from all the decision trees. According to the authors of LeFE, this ensemble approach has favourable characteristics such as low bias and low variability despite the inherently noisy nature of gene expression data (Eichler et al. 2007).

LeFE uses a permutation t-test to compare the predictive ability of the genes in a gene set to that of other randomly selected genes on the array. In this sense it is competitive in nature and establishes statistical significance by gene randomisation. However, useful features of the LeFE algorithm include the fact that it assigns importance scores to each gene in a gene set. This information can be used to guide subsequent gene-level research once an interesting gene set has been found. The LeFE authors also show that their approach is not susceptible to gene set size bias as in PathwayRF where larger gene sets tend to be ranked higher than smaller gene sets (Eichler et al. 2007).

However arguably the most significant distinguishing factor in these machine learning approaches is that they can capture complex nonlinear relationships that may exist between genes in a gene set. Gene products often interact in complicated ways that cannot be elucidated by simply considering the coordinated up- or down-regulation of a group of genes. The decision trees at the heart of the random forest algorithm can capture relationships where, for instance, gene product A and B need to be down-regulated before gene product C's up-regulation has a significant effect on the biological phenotype or class outcome (Eichler et al. 2007). Such relationships are known to occur in molecular biological pathways, but it remains to be determined if the relationships hypothesised by tools such as LeFE actually exist in reality.

University of Cape Town

1.4 Paralogs

The concepts of 'analogy' and 'homology' both refer to the presence of similarities between compared items. In biology, the word 'homolog' was first introduced to refer to the idea of an archetypal body plan in vertebrates, as opposed to analogous body parts that simply possess the same function (Owen 1848). After the publication of Darwin's seminal *Origin of Species* (Darwin 1859), the observation of these homologies was used as evidence in support of evolution (Huxley 1860). Subsequently, the word 'homolog' has been used to denote genes sharing a common evolutionary origin and Fitch (1970) made the distinction between two specific types of homologs, terming them 'orthologs' and 'paralogs'. Orthologs are genes in different genomes that originate from a single ancestral gene in their last common ancestor, and are therefore separated by a speciation event. On the other hand, paralogs are genes related by a duplication event (Koonin 2005) and although they need not necessarily be present in the same genome, we restrict our interpretation to this scenario.

Both paralogy and orthology statements are based on computationally determined sequence similarity between genes, or sequence or structural similarity between the proteins they encode. It is important to note that these statements are in fact inferences, because they imply the occurrence of evolutionary events that are unobservable. However, such inferences from phylogenetic analyses have proven useful in the interpretation of the vast amount of data generated in the post-genomic era and are fundamental to the field of evolutionary genomics (Koonin 2005). Orthologs typically have equivalent biological functions in different organisms and this property has been used extensively to make predictions about gene function. The likely function of an uncharacterised gene of interest can be inferred based on sequence similarity to a gene for which functional information is known (Eisen 1998).

University of Cape Town

1.4.1 Evolution by Gene Duplication

Paralogs, or gene duplicates, also often have related biological functions. The formation of paralogous genes can occur by the duplication of single genes, chromosomal regions, or whole genomes leading to polyploidisation. It is recognised that the most common fate of duplicated genes is deletion or degradation of one paralog (Wolfe and Shields 1997) as the redundant copy may result in undesirable increased dosage effects and be an unnecessary

burden on the cellular machinery of the nucleosome. However, benefits of retaining a gene in duplicate are clear in situations where amplification of the gene product is desirable, such as ribosomal RNA (rRNA) genes. Apart from cases of immediate benefit, Fisher (1928) recognised the significance of these duplications in terms of their contribution to evolutionary innovation. Thereafter, Ohno (1970) provided a coherent explanation of how gene duplication could lead to the formation of novel functions and argued that this process is the major driving force behind the evolution of genomes. Although the principle in its purest form has been challenged, it states that after gene duplication, one of the paralogs performs the ancestral function while the other is free to undergo otherwise detrimental mutations, eventually leading to functional novelty (neofunctionalisation).

Force et al. (1999) put forward an alternative mode of evolution by gene duplication called subfunctionalisation. This is also referred to as the duplication-degeneration-complementation (DDC) model, where both paralogs undergo complementary loss of gene subfunctions in such a way that each gene retains an aspect of the original ancestral gene function. The result is that either the two genes are able to complement or substitute for each other, or that they diverge completely to perform unrelated functions. In addition, there are numerous examples where the regulatory sequences of paralogs have diverged, allowing the specialisation of each gene's spatial and/or temporal expression programme. This process has been demonstrated by Hittinger and Carroll (2007) who compared the activity of the paralogs *GAL1* and *GAL3* in *Saccharomyces cerevisiae* to that of the unduplicated bi-functional gene in *Kluyveromyces lactis*. Apart from performing different functional aspects of the inferred ancestral gene, the duplication has also allowed the independent regulatory optimisation of each paralog, in a process the authors term 'adaptive conflict resolution'.

He and Zhang (2005), in their analysis of yeast protein interaction and human gene expression data, realised that neither the neofunctionalisation nor the subfunctionalisation models alone are sufficient in explaining the functional divergence of duplicated genes. They put forward the notion of subneofuntionalisation, which consists of an initial stage of rapid subfunctionalisation that is often followed by a prolonged period of substantial neofunctionalisation. This more complex hybrid model explains the initial retention of paralogs despite their redundancy and early selective constraints (Kondrashov et al. 2002), while at the same time accounting for the high numbers of new functions observed.

University of Cape Town

1.4.2 Sequence Analysis Methods for Identifying Paralogs

The sheer volume of data generated by recent genome sequencing efforts has led to a demand for sophisticated analyses of biological sequences. Many methods for sequence analysis such as phylogenetic tree reconstruction, sequence alignment and RNA secondary structure analysis use probabilistic modelling approaches (Durbin et al. 1998). In this subsection we very briefly discuss computational methods based on pairwise alignment that can be used to find homologous sequences.

The first step in determining whether two DNA or protein sequences have evolved from a common progenitor involves their alignment. Global alignment algorithms, such as the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), align sequences over their entire length, whereas local alignment algorithms, such as the Smith-Waterman algorithm (Smith and Waterman 1981), align only sub-sequences of each sequence. Because genetic material is changed over time and generations through mutations, the alignments usually need to take into account non-identical matches such as substitutions, insertions and deletions (indels) in either sequence. The optimal alignment sought is the one that reflects the evolutionary relationships between the sequences most accurately. This is done using a scoring system that takes into account the probability of each residue change resulting in a total score that reflects the likelihood that the alignment was produced as a consequence of divergence from a common ancestor (Ewans and Grant 2005). An example of a simple scoring system for DNA sequences is:

$$SCORE = (\text{the number of matches}) - (\text{the number of mismatches and indels}) \quad (1.10)$$

However, for protein sequences it is more complicated with scoring schemes using substitution matrices such as PAM (Accepted Point Mutation) and BLOSUM (Blocks Substitution Matrices), which reflect the likelihood of different amino acid substitutions having occurred based on a trusted dataset. The statistical significance of the best alignment obtained can then be determined by calculating the probability that the alignment arose by chance. When the sequences are long, the number of possible alignments can be prohibitively high to list exhaustively and dynamic programming approaches, such as those used in the Smith-Waterman or Needleman-Wunsch algorithms, are more feasible. Even though

University of Cape Town

these offer advantages in terms of improved time complexity, when a sequence is queried against a large database of sequences to find homologs, the search may still not produce results within an acceptable amount of time. BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) uses a heuristic technique to limit the search to sequences that appear to be the most promising, as well as efficient estimations of statistical significance. This has resulted in BLAST's extremely good performance, its replacement of the forerunner tool FASTA (Lipman and Pearson 1985) and made it a powerful and widely used tool in the medical and biological sciences.

BLAST is commonly used to find paralogs or orthologs of a given sequence and E-values obtained from a query, which indicate the probability of such alignments occurring by chance, are often quoted in an analysis. However, Koski and Golding (2001) point out that the hit sequence with the lowest E-value does not necessarily indicate the closest evolutionary neighbour to the query sequence in the database. Therefore a low score in a BLAST run is not sufficient to imply evolutionary proximity and further phylogenetic analysis is normally needed to make such inferences.

University of Cape Town

1.4.3 Coexpression of Paralogs

It is well known that paralogs show a high degree of functional similarity, as mentioned previously. Large-scale automatic annotations of gene products to functional terms in databases such as the GO have long exploited this fact. Intuitively, one would also expect that at least part of the up-stream regulatory sequence of a duplicated gene be copied along with the rest of the coding part. Lee et al. (2002) found that the number of regulatory elements shared between paralogs increases with protein identity. This suggests that gene regulatory sequences are often co-duplicated with their coding regions. In terms of one form of the subfunctionalisation mode of evolution, after gene duplication the regulatory roles of each paralog differentiate and are refined according to their respective subfunctions. Although initially, and especially in cases where increased dosage requirements or the redundant backup of important genes is necessary, they may have correlated expression patterns. van Noort et al. (2004) studied the gene expression networks of *Saccharomyces cerevisiae* and their analysis revealed a correlation between the fraction of coexpressed paralogs and their sequence similarity.

We investigate gene expression correlation of paralogous genes in *Arabidopsis thaliana* in the following chapter with the view to investigating their effect on the results obtained from GSA. Evidence suggesting that paralogs exhibit three-fold redundancy i.e. in sequence, expression and function, has led us to suspect that paralogs may influence these results as GSA often involves comparisons between two of these factors.

University of Cape Town

Chapter 2

Coexpression of Paralogs

University of Cape Town

Although *Arabidopsis thaliana* is a small uninteresting-looking plant that is commonly found growing as a weed, it possesses characteristics such as a small genome and rapid life-cycle that have made it a widely-used model organism for researchers in the plant sciences. It was the first plant genome to be sequenced (Kaul et al. 2000) and much research has been done to attribute functional information to its genes and proteins. Here we use gene and protein sequence data together with a large collection of gene expression experiments to determine the extent to which paralogs in *Arabidopsis* have correlated expression patterns. We then investigate the relationship between this correlation and paralog sequence similarity and evolutionary time.

Although our analysis is restricted to this one organism as a case study, it is not unreasonable to expect comparable results in other eukaryotes. The results here also represent the starting point of our subsequent analyses where we investigate the effects of paralogs on the results from GSA. All custom scripts, unless otherwise specified, were written in the Python programming language (van Rossum and Drake 2001).

2.1 Materials and Methods

2.1.1 Paralog Prediction

We obtained a FASTA format file containing all amino acid sequences in the *Arabidopsis thaliana* proteome from UniProt Knowlegebase (UniProtKB) (UniProt 2008). UniProtKB is an expertly curated and largely non-redundant database for protein information. To determine candidate paralogs in *Arabidopsis*, we first formatted the FASTA protein sequence database (`formatdb` program) and then ran an all-against-all BLAST (`blastall` program, `blastp` option) locally using the complete proteome as query file. We chose a BLAST expectation value (*E*-value) cut-off of 10^{-5} (see below for justification) and used the XML output format. Other default parameters used include the BLOSUM62 amino acid substitution matrix for alignment scoring. BLAST uses a local sequence alignment algorithm and this often results in multiple high-scoring sequence pairs (HSPs) for two compared protein sequences. Reciprocal hits are another source of duplicate information. We took this into account when parsing the BLAST output, producing a maximum of one record per protein pair.

BLAST *E*-values are not a measure of the overall similarity between sequences, because they are probabilistic values based on the likelihood of observing local matches between sub-sequences and are also affected by factors such as database size and sequence length. We therefore obtained a global percentage sequence identity measure (%ID) for candidate paralogs by performing a global alignment using an implementation of the Needleman-Wunsch algorithm (`needle` program) from the EMBOSS (Rice et al. 2000) software suite. A wrapper script was used to run the `needle` program with default parameters.

As a simple test to determine whether the chosen BLAST *E*-value cut-off 10^{-5} was appropriate for elucidating the majority of paralogs over a suitably large protein sequence %ID range, we repeated the above analysis using 1000 randomly selected amino acid sequences from the *Arabidopsis* proteome. The %ID frequency distribution for candidate paralogs was then compared for two different *E*-value cut-offs of 10^{-5} and 10 (default value).

University of Cape Town

2.1.2 Calculation of Expression Correlation

The aim here is to compare expression information of paralogs at the gene level and therefore it was first necessary to convert the protein-pair data to that specified in terms of gene-pairs. Although infrequent, it sometimes occurs that the same protein sequence is encoded by multiple genes within the same genome and, because of alternative splicing in eukaryotes, single genes often give rise to multiple protein isoforms (Lodish et al. 2001). To account for this we used information from UniProt entries to assign gene names to each protein pair and remove duplicate and self-matching gene entries from the list of candidate paralogs.

We then used Affymetrix GeneChip (microarray) data from the Nottingham Arabidopsis Stock Centre's (NASC) AffyWatch service (Craigon et al. 2004) to determine whether gene paralogs exhibit similarity in their expression patterns. The data consist of gene expression measurements from over 1500 ATH1 GeneChips used in diverse experiments and made publicly available by NASC between 2002 and 2005. After removal of outlier arrays, multiple array normalisation was carried out using the GCRMA (GC robust multi-array average) method (Wu et al. 2004), which takes into account the stronger hydrogen bonding of Guanine/Cytosine nucleotide pairs compared to Adenine/Thymine pairs. We calculated expression correlation values for all pairs of genes in the list using this normalised meta-dataset. When more than one Affymetrix probe set identifier (probeID) was available for a particular gene, we attempted to select the most reliable one based on probeID suffix descriptions. As not all known genes are represented and probed on the ATH1 GeneChips, it was not possible to calculate expression correlation values for all candidate paralogs.

To quantify gene expression correlation, we used Spearman's rank correlation coefficient (Spearman's ρ), which is a non-parametric measure of correlation that is robust to outlying observations (Simon et al. 2003). Unlike Pearson correlation, Spearman's ρ does not make assumptions about the normality of the compared gene expression variables, but on the other hand is not as conservative as Kendall's τ correlation. For the calculations we used a custom script and the RPy package (Moreira et al. 2008) to enable use of the necessary statistical functions in the R Programming Language (R Foundation 2008).

University of Cape Town

2.1.3 Evolutionary Distance Between Paralogs

We used the method of (Goldman and Yang 1994) implemented in the `codeml` program from the PAML (Phylogenetic Analysis by Maximum Likelihood) program package (Yang 1997) to estimate the number of synonymous substitutions per synonymous site (d_S) for paralogs. This maximum-likelihood (ML) method requires an explicit model of codon substitution and chooses parameters for this model that maximise the likelihood of observing the two paralog's sequences. The model and parameters obtained are then used to estimate d_S. Because `codeml` takes as input the compared paralogs' coding sequences (CDSs) aligned with respect to their aligned protein sequences, we first obtained a FASTA format file containing all *Arabidopsis* CDSs from the EMBLCDS database (Kulikova et al. 2007). A wrapper script was then used to run the necessary programs and append d_S values to the list of paralog pairs. However it was not possible to calculate d_S values for all paralog pairs as CDSs were not available for all protein sequences initially obtained from UniProtKB.

2.2 Results and Discussion

2.2.1 Paralog Prediction

The structure and function of proteins is directly determined by their amino acid sequence. This together with the fact that the genetic code is degenerate makes amino acid sequence similarity a more sensitive indicator of homology than nucleotide sequence similarity. We determined candidate paralogs in the model organism *Arabidopsis* using its entire proteome and the two-step procedure described in Section 2.1.1. The procedure involved the all-against-all comparison of 35007 *Arabidopsis* protein sequences from UniProtKB and the global alignment and scoring of 982254 pairs with BLAST E-values below a cut-off threshold of 10^{-5}. Although this is a commonly used BLAST threshold, we conducted a simple test to confirm that it reliably elucidates protein pairs covering a substantial global sequence identity range. After repeating the two-step procedure with E-value cut-offs of 10 and 10^{-5} for 1000 randomly selected proteins, we found that their %ID frequency distributions were approximately equal for %ID > 20 with 304 and 293 paralogs found in each case respectively.

University of Cape Town

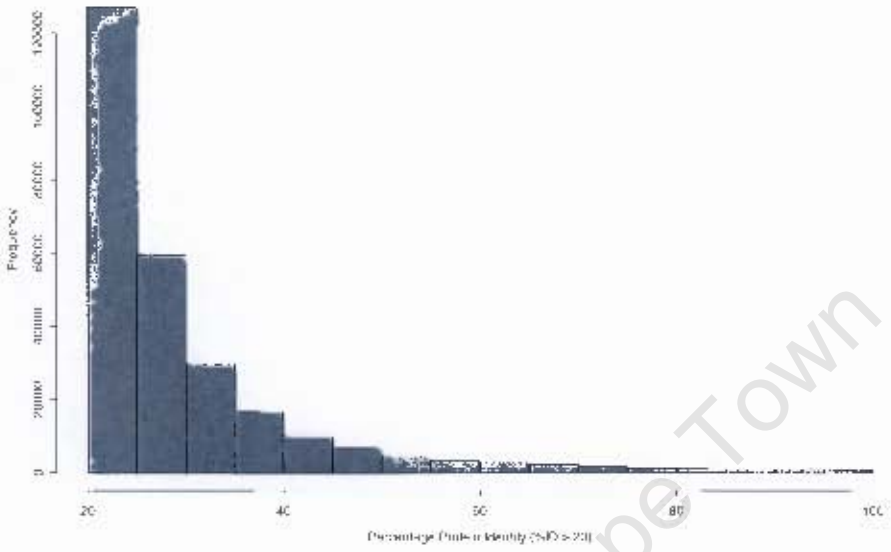

Figure 2.1: Percentage protein identity distribution of candidate gene paralogs in *Arabidopsis*. Only paralog pairs with %ID > 20 could be reliably predicted.

The distributions were very different below this threshold, with 6255 and 703 paralogs found in each case respectively. The relatively relaxed E-value cut-off of 10 results in many protein pairs with low %ID not found with the cut-off of 10^{-5}. However we are interested in candidate paralogs with relatively high sequence identity, as these pairs tend to be truly homologous more often than their lower scoring counterparts. Rost (1999) found that 90% of protein pairs with greater than 30%ID were homologous, whereas below 25%ID less than 10% were. They used the term 'twilight zone' to describe the 20-35%ID region, in which paralogy cannot be definitively declared and below which there is an explosion in the number of false positives. We therefore proceed with the analysis by considering only those protein pairs with %ID > 20, despite the apparent arbitrary dichotomisation between paralogs above and below this threshold.

After attributing gene names to all protein pairs and removing redundant gene matches and those without associated gene name information, 677473 gene pairs remained. Figure 2.1 shows the percentage sequence identity distribution for those candidate paralogs with %ID > 20. As expected, the majority of paralogs have low %ID, but surprisingly 236 gene pairs with 100% protein identity were found.

University of Cape Town

2.2.2 Calculation of Expression Correlation

We investigated the coexpression of paralogs using a large *Arabidopsis* microarray gene expression dataset from NASC. As described in Section 2.1.2, gene expression values across multiple experiments were used to calculate correlation values for each pair of paralogs. We were able to calculate correlation values for 409944 gene pairs (not all paralogs found were represented on the microarray platform used) and Figure 2.2 shows the mean expression correlation for paralogs at different levels of protein identity. A clear trend can be seen where gene expression correlation of paralogs tends to increase with increasing protein sequence similarity. On average, paralogs with 90-100% protein sequence identity have a strong correlation ($\bar{\rho} > 0.5$) in their gene expression patterns. Figure 2.3 displays a similar increasing trend in terms of the fraction of highly coexpressed ($\rho > 0.5$) paralogs at increasing protein sequence identity intervals.

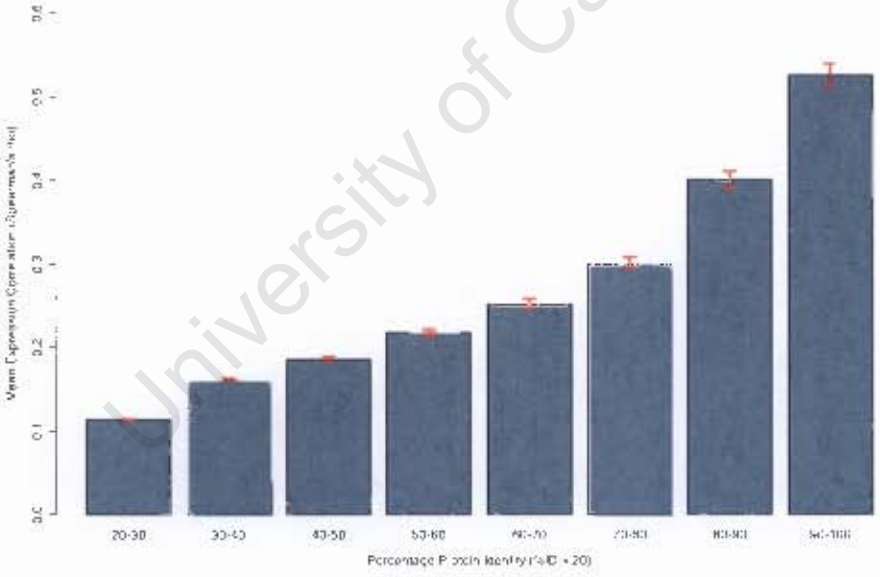

Figure 2.2: Mean expression correlation (Spearman's ρ) of gene paralogs in *Arabidopsis* at various protein sequence identity levels where %ID > 20. Error bars indicate the standard error of the estimated mean values.

As these expression correlation values are based on over 1500 individual microarray experiments, the results presented here are highly statistically significant and provide evidence in support of the notion that the regulatory and coding sequences of paralogs tend to

University of Cape Town

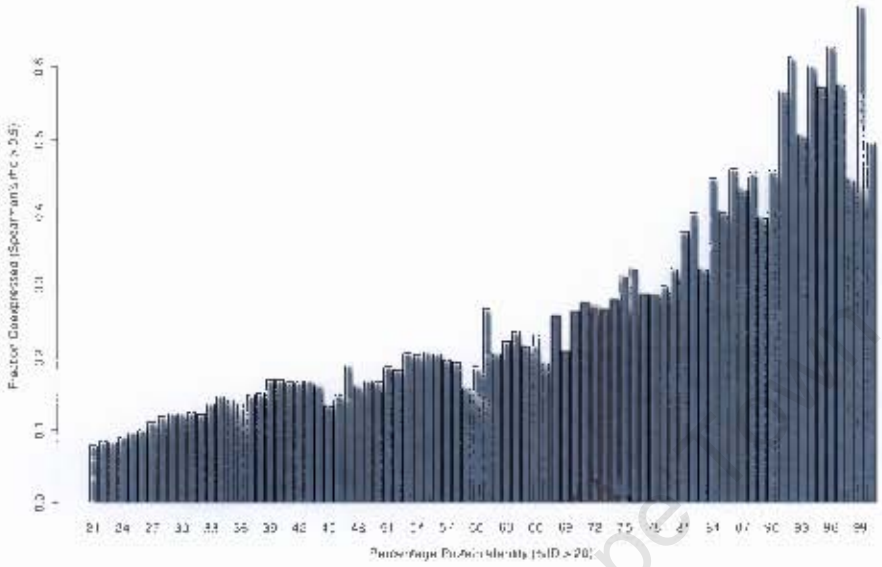

Figure 2.3: Fraction of coexpressed (Spearman's $\rho > 0.5$) gene paralogs in *Arabidopsis* at various protein sequence identity levels.

co-diverge (see Section 1.4.3). The extent of the observed correlation in the expression patterns of paralogs also warrants further investigation in terms of their effect on results from microarray GSA (see Chapter 4).

2.2.3 Evolutionary Distance Between Paralogs

We estimated the evolutionary time since the formation of paralogs by inferring the number of synonymous substitutions per synonymous site that have occurred to bring about the observed differences in their nucleotide sequences. Synonymous, or silent, nucleotide substitutions are substitutions that do not alter the encoded functional protein sequence. Because (for the most part) they are not associated with a fitness cost to the organisms carrying these mutations, they are more readily retained than nonsynonymous substitutions and the rate at which they occur is normally similar across different genes in a genome. They can therefore be used as a molecular clock to estimate the evolutionary time that has elapsed since the duplication and divergence of two paralogous sequences (Kafatos et al. 1977).

University of Cape

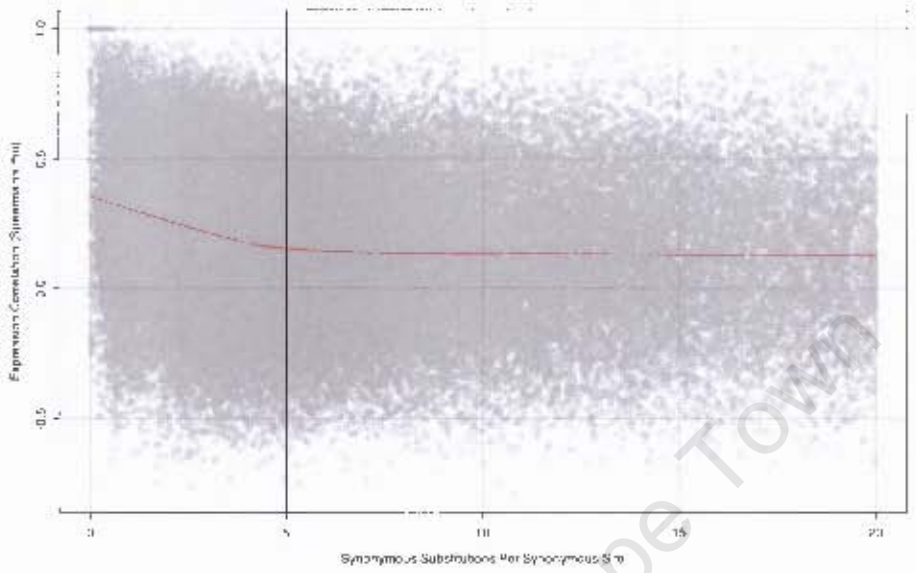

Figure 2.4: Expression correlation (Spearman's ρ) of gene paralogs in *Arabidopsis* as a function of the evolutionary distance (d_S) between them.

We calculated the number of synonymous substitutions per synonymous site (d_S) for paralogs in *Arabidopsis* using their corresponding coding sequences (CDSs) and the procedure described in Section 2.1.3. Figure 2.4 shows expression correlation values of paralogs as a function of d_S. The smoothed red line shows a locally-weighted polynomial regression technique (`lowess` function in R) applied to the data. We focus on the region where $d_S < 5$ as large values of d_S are not reliably estimated. From this region it can be seen that the expression patterns of paralogs tend to diverge over time. We make the comparison between 'recent' and 'older' paralogs explicit in Figure 2.5, which shows two notched box-and-whisker plots for paralogs where $d_S < 1$ and $d_S \geq 1$ respectively. The notches of the two plots do not overlap and this indicates that the two medians differ significantly and there is strong evidence to suggest that the expression patterns of 'recent' paralogs tend to be more similar than those of 'older' paralogs (Wilcoxon rank sum test P-value $< 2.2 \times 10^{-16}$).

In Figure 2.4 it is also interesting to note that a few 'recent' paralogs have very dissimilar expression patterns with $\rho < -0.5$. The rapid divergence in their expression patterns shortly after duplication is consistent with the subnecofunctionalisation mode of evolution (see Section 1.4.1).

University of Cape Town

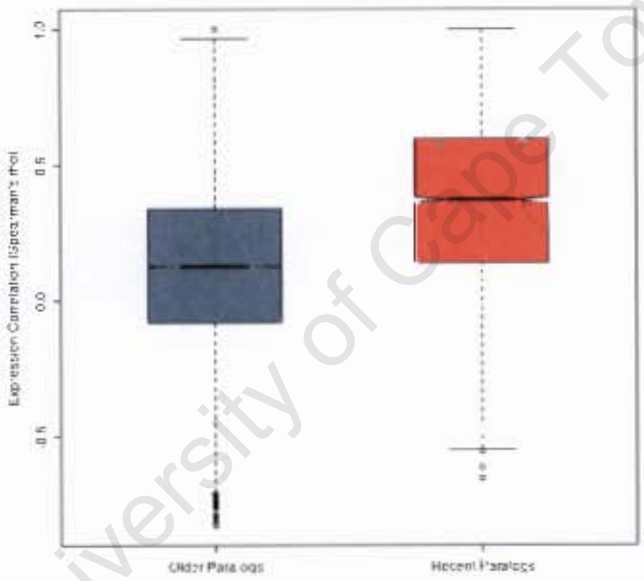

Figure 2.5: Box-and-whisker plots comparing the expression correlation (Spearman's ρ) of 'recent' and 'older' gene paralogs in *Arabidopsis*. 'Recent' and 'older' gene paralogs are defined by their estimated synonymous substitution distance where $d_S < 1$ and $d_S \geq 1$ respectively.

University of Cape Town

Chapter 3

Indygene Tool

University of Cape Town

Despite the fact that paralogs tend to have correlated expression patterns (see Chapter 2), many GSA methods either explicitly or implicitly assume that all genes in the microarray dataset under study are expressed independently of one another. In this chapter we discuss the Indygene tool, which reduces a supplied list of genes to one without paralogous relationships, where the goal is to proceed with GSA thereafter. By using a paralog-reduced gene list that more realistically satisfies the above-mentioned independence assumption, the hope is that undesirable effects are diminished and more biologically relevant GSA results are obtained. In addition, GSA results obtained using the reduced gene list could be used to verify and assess the plausibility of results obtained using the original gene list.

We evaluate three different graph theoretic algorithms able to perform the reduction at the heart of the Indygene tool and compare their performances in order to determine the most suitable candidate. We then discuss the tool's back-end processing system and web interface with respect to a typical run using an example dataset. In Chapter 4 we apply the tool to previously published gene expression datasets to determine its utility.

3.1 Materials and Methods

3.1.1 Comparison of Greedy Algorithms for the MISP

Consider a graph G representing a list of m genes and the paralogous relationships between them as vertices and edges respectively. A number of graph theoretic algorithms can be used to find approximate solutions to the maximum independent set problem (MISP) applied to G. We evaluated three such algorithms: GRAND, GMAX and GMIN, all of which use a greedy strategy (see Section 3.2.1). The simplest algorithm, GRAND, randomly removes vertices with non-zero degree until the resulting sub-graph is independent. GMAX is similar to GRAND, however instead of randomly removing vertices, a vertex of maximum degree is removed at each step. GMIN differs from the preceding two algorithms in that it selects a vertex of minimum degree to retain at each step. The selected vertex and all of its adjacent vertices are then removed from the remaining graph. The process is repeated until G becomes empty and the retained vertices form an independent set.

To evaluate the performance of these algorithms we implemented them in custom Python scripts and applied them to real-world data relevant to the intended application. The data comprised lists ranging in length from 500 to 10000 randomly selected *Arabidopsis* genes and we indicated paralogous relationships between gene pairs if their pre-calculated global protein sequence identity was $> 20\%$ (see Section 2.1.1). This data was then represented as a graph using the adjacency list data structure. In an adjacency list, each of the m vertices (in this case genes) in a graph is allocated a unique index in the list. The indices of all vertices adjacent to a particular vertex (in this case paralogous genes) are then listed at the corresponding index in the list. An alternative data structure is the $m \times m$ symmetric adjacency matrix which uses a '1' to indicate an edge between two vertices, represented by their row/column numbers, and '0' otherwise. The adjacency list was chosen for this application, because it is more efficient when dealing with relatively sparse graphs, as are generated with this type of data. Because the algorithms are heuristic in nature and therefore often yield different solutions when applied to the same dataset more than once, we ran each algorithm 10 times on each dataset and recorded the resulting independent set sizes and computation times in each case.

University of Cape Town

3.1.2 Indygene Back-End Processing

The input to the Indygene back-end processing system is a job file, which contains the user-inputted gene list and other meta-information including the organism name, gene identifier type, timestamp and user details. The job files are created by the web interface software (see Section 3.1.3) upon submission by a user. The processing system is shown in Figure 3.1 and begins with the parsing of the oldest job file in the job queue. If necessary, Affymetrix probe set data is used to convert microarray probeIDs to their corresponding gene names. The gene names are then compared with each other and with information from UniProtKB to ensure that each gene name/symbol is valid and unique.

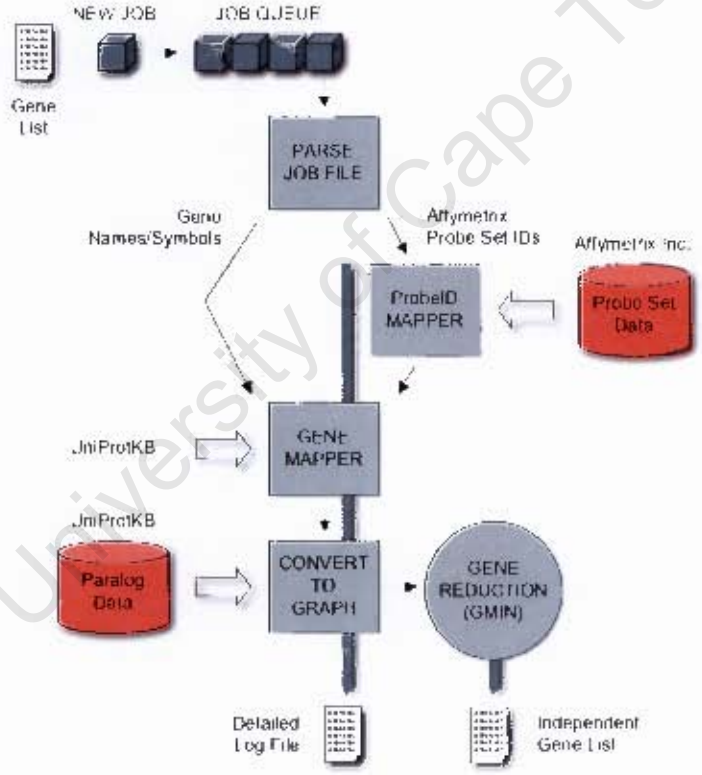

Figure 3.1: Flow diagram of the Indygene back-end processing system.

Pre-computed protein similarity information for a selected set of organisms is used to construct an adjacency list representation of the gene graph. The method of Section 2.1.1 was used to compute global protein sequence identity values (%ID) for gene pairs in different genomes in advance. This data together with a user-selected %ID paralog cut-off

University of Cape Town

is used to determine whether an edge exists between two genes in the graph. The greedy GMIN algorithm is then used to reduce the gene graph to a subset of vertices such that no edges remain. Finally, this solution is used to construct an independent gene list file that can be downloaded by the user via the user interface.

Simultaneously to the above process, a log file is created that details the status of each identifier in the original gene list and any modifications that have been made to it. This includes information about unrecognised gene/probeID identifiers or identifiers with no associated protein in UniProtKB, redundant or synonymous identifiers and Affymetrix probeIDs excluded from the analysis based on the fact that they target more than one gene. Paralogs associated with each gene are also included in the log file as this information could possibly be used in other analyses. A log summary is created and included together with the full log in a log file that can be downloaded by the user via the web interface.

3.1.3 Indygene Web Interface

The Indygene front-end consists of a web application that was developed using the Web.Py framework (WebPy 2008). The controller handles requests for three dynamic pages, namely 'Tool', 'Output' and 'About', each of which is generated using the Web.Py templating system. The default 'Tool' page provides a form for submitting a new job to Indygene and requires the user to specify the following:

- Job title typically consisting of the user's name or a relevant description
- Organism under study selected from a list of available taxa
- Criterion for paralogy based on minimum protein sequence identity
- Gene identifier type (gene names/symbols or Affymetrix probe set identifiers)
- If relevant, Affymetrix microarray platform selected from a list of those available
- File containing a list of gene identifiers to be reduced

The controller checks that all required fields have been supplied and if so, constructs a job file from this information and saves it in a directory of queued jobs, where it is eventually

University of Cape Town

processed as discussed in Section 3.1.2. The user is then redirected to the 'Output' page that displays a unique job ID number for future reference if the job is not yet complete. This page automatically refreshes periodically and when the job is complete it provides links for the user to download the reduced gene file and log file. The default version of the 'Output' page provides a search facility enabling users to recall the results from a previously-submitted job using its corresponding job title or job ID. The 'About' page provides information about the Indygene tool and its purpose and gives contact details for the authors.

University of Cape Town

3.2 Results and Discussion

3.2.1 Comparison of Greedy Algorithms for the MISP

The genes in a gene list and the paralogous relationships between them can be represented in a graph G as vertices and edges respectively. Reducing a list of genes to one without paralogous relationships is equivalent to finding an independent set in G, which is a subset of its vertices with no edges. Typically such graphs contain many independent sets of different sizes, but in view of the cost and time involved in generating gene expression data, we are interested in obtaining the largest independent set possible so as to retain the maximum amount of information for further analysis. This is the optimisation version of the independent set problem, called the maximum independent set problem (MISP), which attempts to find the largest independent set in G. MISP is known to be an NP-complete problem (Karp 1972) and therefore there are no efficient algorithms to calculate its exact solution in a reasonable amount of time.

Heuristic algorithms for optimisation problems usually involve a sequence of steps, where each step involves a set of local choices, each leading to a global solution. Greedy algorithms take the locally optimal solution at each step and while they do not always yield globally optimal solutions, they can provide useful approximations (Cormen et al. 2001). We consider three greedy algorithms that provide approximate solutions to the MISP, namely GRAND, GMAX and GMIN (see Section 3.2.1). If $\alpha(G)$ is the size of the maximum independent set in G and $d(v)$ is the degree of vertex v, Caro (1979) and Wei (1981) both independently showed that

$$\alpha(G) \geq \sum_{v \in V} 1/[d(v) + 1] \tag{3.1}$$

which has subsequently been referred to as the Caro-Wei theorem (Sakai et al. 2003). Caro (1979) found that GMIN outputs an independent set of size at least the above bound and later Griggs (1983) proved the same for GMAX. For a graph G with degree bounded by Δ, Halldorsson and Radhakrishnan (1997) proved that GMIN outputs an independent set of size at least $3\alpha(G)/(\Delta + 2)$. Also, by showing that GMAX can output two-vertex solutions when applied to graphs that are complete bipartite less a single perfect matching, they proved that the upper limit for its guaranteed lower bound is $2\alpha(G)/(\Delta + 1)$. Therefore GMIN's guaranteed lower bound on independent set size, specified in terms of Δ, is greater than that of GMAX.

University of Cape Town

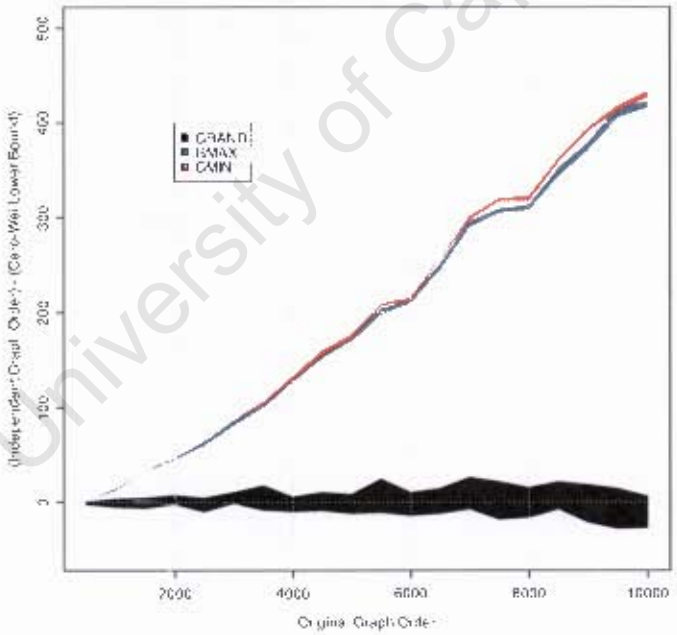

Figure 3.2: Graph order before and after the application of three greedy algorithms for the MISP to random *Arabidopsis* gene graphs of differing sizes. We indicate the solution order range over 10 replications in each case. The ordinate shows the number of genes by which the independent graph order exceeds the lower bound given by the Caro-Wei theorem.

We compared the practical performance of the three algorithms by applying them to gene

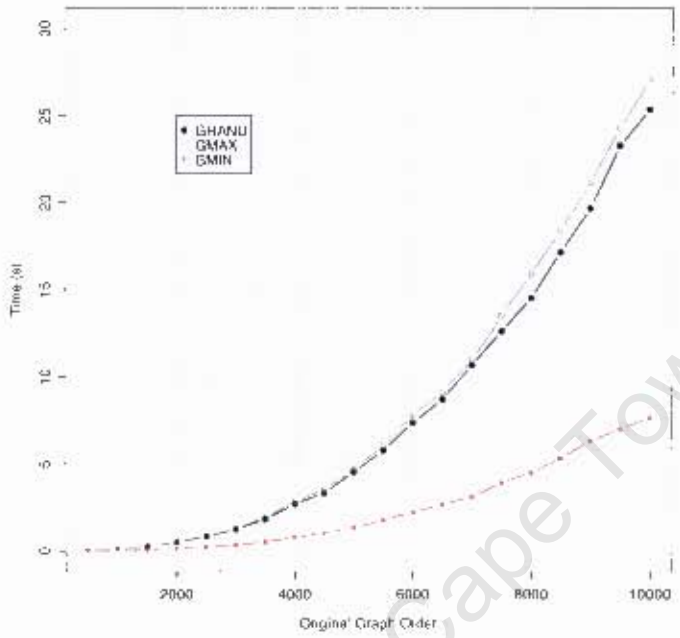

Figure 3.3: Mean computation times for three greedy algorithms for the MISP applied to random *Arabidopsis* gene graphs of differing sizes. We indicate the mean calculation time over 10 replications in each case.

graphs created using randomly generated lists ranging in length from 500 to 10000 randomly selected *Arabidopsis* genes (see Section 3.2.1). Figure 3.2 and Figure 3.3 show the resulting independent set sizes and computation times respectively. In Figure 3.2 we plot the number of genes by which each resulting independent set exceeds the lower bound given by the Caro-Wei theorem. GRAND tends to produce independent graphs with order near this lower bound and performs the worst of the three algorithms in this respect. Both GMIN and GMAX improve on solutions from GRAND by hundreds of genes when the graph order is high, with GMIN finding solutions at least as large as those found by GMAX. In terms of computational time, Figure 3.3 shows that GMIN is the most time-efficient algorithm. We therefore adopted an optimised version of this algorithm in the Indygene tool.

University of Cape Town

3.2.2 An Example Indygene Run

Here we describe a typical session with the Indygene web application by submitting the complete list of 22812 probeIDs from the Affymetrix *Arabidopsis* (ATH1) GeneChip microarray platform for reduction. We used the default criterion for paralogy of 20% protein sequence identity and submitted the list of probeIDs to Indygene via the form on the 'Tool' page shown in Figure 3.4. At the time of writing, the following taxa were available for selection: *Arabidopsis*, human, mouse and rat. In addition, the following Affymetrix microarray platforms were available: Arabidopsis ATH1 Genome Array, Arabidopsis Genome Array, Human Genome Focus Array, Human Genome U133 Array Plate Set, Human Genome U133 Plus 2.0 Array, Human Genome U133 Set, Human Genome U133A 2.0 Array, Human Genome U95 Set.

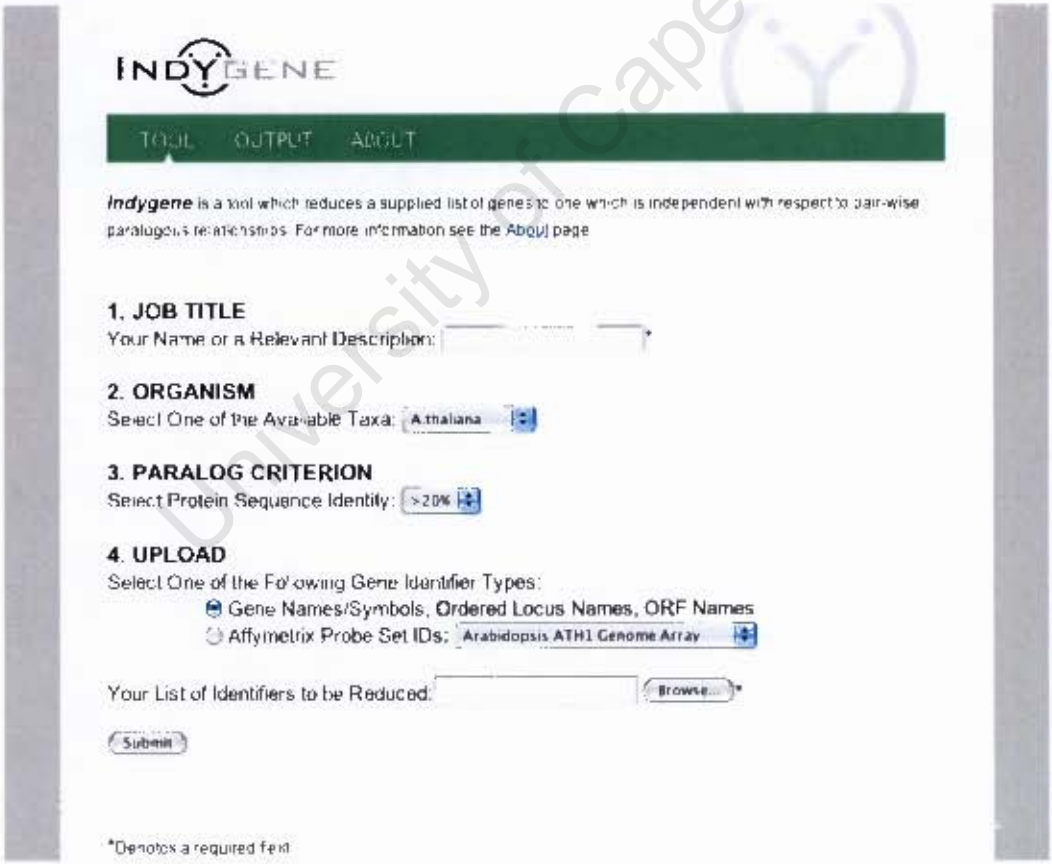

Figure 3.4: Indygene 'Tool' page showing the form used to submit a gene list for processing.

The processing and reduction of the gene list to 11918 independent probeIDs took approx-

University of Cape Town

LOG SUMMARY (TOTALS):

17932 Identifiers OK (STATUS=0)
3577 WARNING:No gene product for identifiers (STATUS=1); Included in analysis
341 WARNING:Identifiers redundant (STATUS=2); Excluded from analysis
962 WARNING:Affymetrix ProbeID targets more than one gene (STATUS=3); Excluded from analysis
0 WARNING:Affymetrix ProbeID not recognized (STATUS=4); Excluded from analysis

FULL LOG:

Input Identifier	Gene	Status	Comments	Paralogs
244901_at		1		
244902_at		1		
244903_at		1		
244903_at		2	244903_at has been duplicated	
244903_at		2	244903_at has been duplicated	
244904_at		1		
244905_at		1		
244906_at		1		
244907_at		1		
244908_at		1		
244909_at		1		
244910_s_at	AT2G07686	1		
244911_at	AT2G07683	1		
244912_at	AT2G07783	1		
244913_at		1		
244914_at	AT2G07682	1		
244915_s_at	AT2G07682	2	244915_S_AT and 244914_AT refer to the same gene	
244916_at	AT2G07682	2	244916_AT and 244914_AT refer to the same gene	
244917_at		1		
244918_at		1		
244919_at	AT2G07768	1		
244920_s_at	AT2G07751	0		
244921_s_at		3		
244922_s_at	AT2G07674	2	244922_S_AT and 244923_S_AT refer to the same gene	
244923_s_at	AT2G07674	0		
245027_at	AT2G26550	0		AT1G69720
245028_at	AT2G26570	0		AT5G42880,AT1G45545,AT5G55860,AT5G66030,AT4G33390
245029_at	AT2G26580	0		
245030_at	AT2G26620	0		AT3G62110,AT1G02460,AT5G48140,AT2G43860,AT1G60590,AT3G16850,AT1G19170,AT4G18180,A
245031_at	AT2G26360	0		AT1G14140,AT5G09470,AT3G20240,AT5G15640,AT5G01340,AT2G33820
245032_at	AT2G26630	0		
245033_at	AT2G26380	0		AT3G20820,AT4G28560,AT1G80080,AT1G66830,AT3G05360,AT3G25020,PGIP2,AT3G12610,AT3G4
245034_at	AT2G26390	0		AT2G25240,AT1G64010,AT2G14540,AT1G47710,AT2G35590,AT2G35580,AT2G35570,AT3G45220
245035_at	AT2G26400	0		AT5G43850
245036_at	AT2G26410	0		AT4G23060,AT5G03960,AT3G09710,AT3G52290,AT5G35670,AT4G00820,AT3G59690,AT5G07240,A
245037_at	AT2G26420	0		AT3G56960,AT1G10900,AT3G07960,AT1G77740,AT1G60890,AT3G09920,AT2G41210,AT1G21980
245038_at	AT2G26560	0		AT3G63200,AT2G39220,AT3G54950,AT5G43590
245039_at	AT2G26600	0		AT5G20390,AT2G01630,AT3G55780,AT4G14080,AT5G18220,AT4G26830,AT2G16230,AT4G18340,A
245040_at	AT2G26520	0		
245041_at	AT2G26530	0		AT3G62630,AT2G15760
245042_at	AT2G26540	0		
245043_at	RCY1	0		
245044_at	PETM	1		
245045_at	AT2G26590	0		

Figure 3.5: Excerpts from an Indygene log file resulting from the submission of the list of Affymetrix probeIDs from the ATH1 GeneChip microarray. Grey bars indicate sections of the log file that have been omitted.

University of Cape Town

imately 30 seconds. Excerpts from the resulting log file are shown in Figure 3.5. The first column of the tab-separated full log, titled 'Input Identifier', gives the gene identifiers (in this case Affymetrix probeIDs) as supplied by the user. The second column, titled 'Gene' gives the corresponding gene symbols when available. Each identifier in the input file is given a status code, shown in the third column, where a value of '0' indicates a recognised, valid and unique identifier for which there is a known corresponding protein product. An explanation and course of action for each of the other status codes (or warnings) is given in the log summary. The duplicate occurrences of identifier '244903_at' were given status '2' and removed from the analysis. Also, because the probeIDs '244914_at', '244915_s_at' and '244916_at', all target the At2g07682 gene, only one of these ('244914_at') was retained. An explanation of this is given in the fourth column titled 'Comments'. The fifth and final column gives a list of the paralogs identified based on the user-selected %ID

cut-off to determine paralogy.

The 'Output' page provides links enabling the download of the above log file and the reduced gene list file. The latter contains gene identifiers in their originally submitted format ready for use in conjunction with any preferred GSA tool.

University of Cape Town

Chapter 4

Reanalysis of Previously Published Datasets

University of Cape Town

In Chapter 2 we showed that paralogs tend to have correlated expression patterns and we argued that their presence in microarray gene expression data is therefore likely to affect results from GSA. In Chapter 3 we developed Indygene, which reduces the number of paralogous relationships in a microarray dataset. Here we use this tool to answer two important questions. Firstly, do paralogs in reality significantly affect results from GSA? If so, this also means that removing paralogs before performing GSA should significantly alter the relative ranking of gene set results obtained. Secondly, do these novel results represent plausible hypotheses regarding the biological processes underlying the response under study, in a way that is particular to the paralog-reduced dataset?

To answer the first question we investigated whether performing a gene reduction on a real-world microarray dataset using Indygene alters the results from subsequent GSA significantly more so than similar sized random gene reductions. We used a permutation testing procedure, which involves repeating the GSA many times in order to obtain a level of confidence associated with our result. Ideally this procedure should be carried out for various GSA approaches, however it becomes computationally expensive for sophisticated methods. We therefore restricted our analysis to a relatively simple strict cut-off method from the category of GSA approaches discussed in Section 1.3.2.1.

To answer the second question we reanalysed previously published gene expression datasets

using three different procedures: GoMiner, GSEA and SAM-GS. The rationale was to establish whether performing GSA on paralog-reduced datasets could reveal novel and biologically relevant themes not otherwise found to be significant. Our choice of tools represents a cross-section of currently available GSA methods and the results of our analyses are presented in order of their increasing statistical conservativeness (see Section 1.3.2.1 and the corresponding subsections below for a review of each method). When performing GSA there are a number of parameters that are under the control of the investigator. To ensure impartiality in our choice of dataset, gene set definitions and significance threshold we adopted the choices of the investigators who performed the initial analyses on the originally published datasets.

With respect to these comparisons, it should also be noted that any discrepancies highlighted between different sets of results are anecdotal and not intended to show definitive benefits or drawbacks of either approach.

University of Cape Town

4.1 Materials and Methods

4.1.1 Statistical Significance of GSA Results Using Indygene

Alonso et al. (2003) used the Affymetrix GeneChip platform to examine the gene expression patterns in *Arabidopsis* seedlings and apices, and determined 628 genes whose expression levels were significantly altered after treatment with exogenous ethylene. Similarly to these authors, we performed GSA on this dataset using Fisher's exact test to rank terms in the GO Biological Process ontology by their overrepresentation in these genes compared to the rest of the genes on the microarray. This GSA approach belongs to the category of strict cut-off methods discussed in Section 1.3.2.1. Apart from issues related to this method's assumption that genes are expressed independently, Alexa et al. (2006) noted that the complex structure of the GO also introduces dependencies among GO terms in the DAG. At present there is no consensus on the most appropriate way to deal with this issue when performing GSA, so to circumvent it we restricted our analysis to Plant GO SLIM terms, which represent an orthogonal collection of high-level biological themes particularly relevant in the context of the plant cell. Although many existing software tools perform this type of GSA, we ran the analysis locally to enable the required number of randomisations

described below.

We then used Indygene to remove pair-wise paralogous relationships with protein sequence identity $> 30\%$ and repeated the above GSA. To quantify the differences between the two resulting ordered lists of GO terms i.e. before and after the reduction, we used a ranked correlation measure (Kendall's τ). Although researchers normally focus on the few statistically significant or highly ranked GO terms towards the top of the list, we considered the entire list so as to incorporate information about the change in relative ranking of all GO terms.

We determined the statistical significance of this difference by comparing the above correlation test statistic to the null distribution of correlation values resulting from all possible similar-sized gene reductions. This is a nonparametric significance testing procedure known as a randomisation test. Because the number of distinct gene reductions was prohibitively large we used Monte Carlo sampling, which considers a fixed number of randomly generated reductions instead of enumerating all possibilities. One thousand random 'samples' were used to generate an estimate of the correlation null distribution.

University of Cape Town

4.1.2 Reanalysis of GSA Datasets Using Indygene

We reanalysed previously published gene expression datasets used by the authors of GoMiner, GSEA and SAM-GS to demonstrate the utility of their proposed GSA methods. We selected these three tools, each representing one of the three major categories of GSA methods discussed in Section 1.3.2, to determine Indygene's usefulness across a broad range of methods. Using the original datasets we compared the GSA results obtained before and after removing paralogous relationships with protein sequence identity $> 30\%$ using Indygene.

Using GoMiner, we reanalysed the Common Variable Immune Deficiency (CVID) gene expression dataset published by Zeeberg et al. (2005). The authors used custom microarrays to measure the gene expression response to CD3 and CD28 antigens/antibodies in peripheral blood mononuclear cells (PBMC) from one CVID patient and six healthy donors. By comparisons to the healthy donor controls, they identified 57 genes that were significantly differentially expressed in the cells from the CVID patient. Using this information, we submitted the original and paralog-reduced gene lists for analysis using the High-Throughput

GoMiner web interface. We also used GoMiner to reanalyse the human airway epithelial cell transcriptome dataset of Spira et al. (2004). The study involved the gene expression profiling of epithelial cell samples obtained at bronchoscopy from 85 subjects, 23 of which were healthy and had never smoked. The authors identified 2382 genes that were expressed in all of these healthy never-smokers. To find GO Biological Process terms enriched in these genes, we once again used the High-Throughput GoMiner web interface to submit the original and paralog-reduced gene lists for GSA.

Using the Java GSEA Desktop Application, we reanalysed the five different gene expression datasets covered in the article by Subramanian et al. (2005). The first dataset comprised mRNA expression profiles of lymphoblastoid cells from 15 males and 17 females, in which the authors aimed to identify cytogenetic gene sets (MSigDB:C1) and functional gene sets (MSigDB:C2) enriched in either gender. The second study involved the identification of targets of the transcription factor p53, which regulates the cell cycle in response to various cellular stress signals including DNA damage, thereby suppressing tumorogenesis. They used NCI-60 cancer cell lines to find functional gene sets (MSigDB:C2) enriched in the expression patterns of 17 classified as possessing the wild-type *p53* gene when compared to that of 33 classified as carrying mutations in the gene (Olivier et al. 2002), and vice-versa. Thirdly, they used GSEA and cytogenetic gene sets (MSigDB:C1) to find positions of frequent chromosomal alteration in acute lymphoid leukaemia (ALL) or acute myeloid leukaemia (AML). The dataset consisted of expression patterns obtained from 24 ALL patients and 24 AML patients (Armstrong et al. 2002). Lastly, datasets from two independent studies were used to determine whether GSEA could identify functional gene sets (MSigDB:C2) correlated with clinical outcome in lung cancer. The Boston (Bhattacharjee et al. 2001) and Michigan (Beer et al. 2002) studies measured gene expression levels in tumour samples from 62 and 86 patients with lung adenocarcinomas respectively, indicating patient survival outcome as either 'good' or 'poor'.

Using R code of the SAM-GS procedure made available by Dinu et al. (2007), we reanalysed the 'p53 status' dataset of Subramanian et al. (2005) using the original and paralog-reduced gene lists. Futher details of this gene expression dataset are as indicated above.

University of Cape Town

4.2 Results and Discussion

4.2.1 Statistical Significance of GSA Results Using Indygene

The Indygene tool, discussed in Chapter 3, aims to remove the minimum number of genes from a gene list that make it devoid of all paralogous relationships, which tend to be associated with correlated expression patterns (see Chapter 2). If the presence of paralogs does indeed affect results from GSA in a biased manner, the elimination of paralagous relationships in a gene expression dataset should lead to significantly different GSA results. Here we determine whether performing a paralog-reduction prior to GSA tends to generate results that are significantly different from those obtained after a random reduction of the same number of genes.

Apart from its industrial importance, made apparent by the fact that it is the most produced organic compound in the world (McCoy et al. 2006), ethylene is also an important compound in biology. In humans, exposure to low concentrations results in mild effects limited to a pleasant odour and a state of euphoria, but in plants it acts as a potent hormone with wide-ranging physiological effects. It is involved in important developmental processes such as disease/wounding resistance and leaf/flower senescence (Johnson and Ecker 1998). Alonso et al. (2003) measured the genome-wide expression changes in plants in response to ethylene. We used Indygene to investigate whether the removal of paralogous relationships significantly influences the results from GSA applied to this dataset.

We performed GSA on the original microarray dataset as explained in Section 4.1.1 and compared the resulting list of GO SLIM terms from the Biological Process ontology to that obtained after reducing the dataset by 6126 genes using Indygene. The obtained correlation value of $\tau = 0.65$ quantifies the difference between the ranking of terms in the two lists. To determine whether this difference was statistically significant and not merely related to the removal of a large amount of genes, we estimated the null distribution for τ using the Monte Carlo sampling procedure described in Section 4.1.1 (see Figure 4.1).

When compared to this null distribution, a nonparametric P-value ≈ 0.007 was obtained, indicating that the presence of paralogs can significantly affect results from GSA. In other words, a paralog reduction as performed by Indygene can result in a significantly different GSA term ranking, not simply attributable to gene removal alone. Although this strength-

University of Cape Town

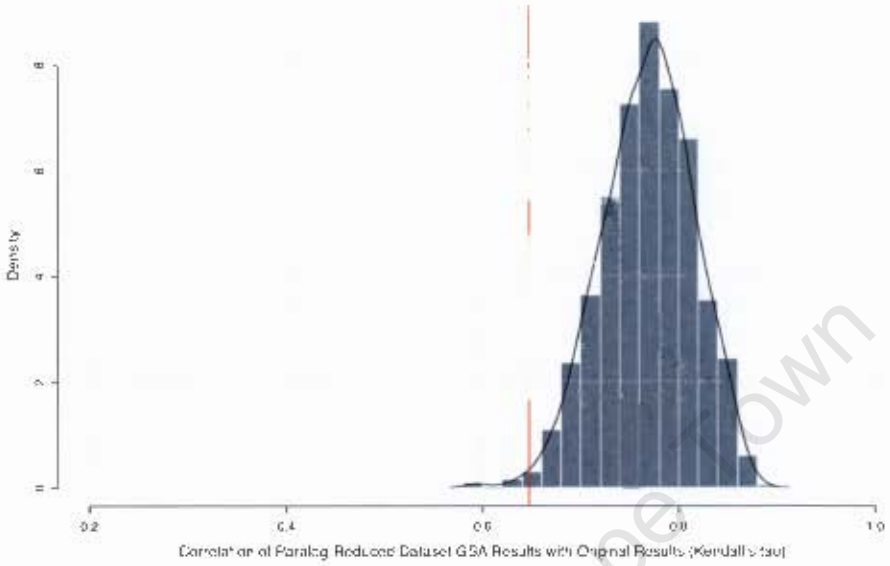

Figure 4.1: Estimated null distribution for τ used to determine whether the paralog-reduced dataset (red vertical line at $\tau = 0.65$) produces significantly different GSA results. The abscissa gives Kendall's correlation (τ) between the ranked GO term lists before and after randomly reducing the dataset by 6126 genes. The black line indicates the approximate probability density function of the null distribution, estimated using a Gaussian smoothing kernel.

ens the argument for using Indygene before performing GSA, the biological validity of such a novel ranking and the novel hypotheses it provides is uncertain. This is the subject of the following subsections.

University of Cape Town

4.2.2 Reanalysis of GSA Datasets Using Indygene

4.2.2.1 Reanalysis of GoMiner Dataset Using Indygene

The GoMiner tool is a member of the category of GSA methods that test for overrepresentation of functional terms in a set of differentially expressed (or otherwise interesting) genes when compared to its corresponding background set. The methods in this category use Fisher's Exact Test, or a variant of this test based on an approximation to the hypergeometric distribution, and are the least statistically conservative, explicitly assuming that genes are expressed independently. There are numerous tools that are based on similar

Original dataset: unique GSA results	Reduced dataset: unique GSA results
GO:0008219 - cell death	GO:0009225 - nucleotide-sugar metabolic process
GO:0016265 - death	GO:0042632 - cholesterol homeostasis
GO:0031529 - ruffle organization and biogenesis	GO:0006890 - retrograde vesicle-mediated transport Golgi to ER
GO:0048259 - regulation of receptor mediated endocytosis	GO:0000059 - protein import into nucleus docking
GO:0045045 - secretory pathway	GO:0006692 - prostanoid metabolic process
GO:0044275 - cellular carbohydrate catabolic process	GO:0006693 - prostaglandin metabolic process
GO:0006006 - glucose metabolic process	GO:0006183 - GTP biosynthetic process
GO:0048193 - Golgi vesicle transport	GO:0007368 - determination of left right symmetry
GO:0030832 - regulation of actin filament length	GO:0008543 - fibroblast growth factor receptor signaling pathway
GO:0007018 - microtubule-based movement	GO:0009799 - determination of symmetry
GO:0016052 - carbohydrate catabolic process	GO:0009855 - determination of bilateral symmetry
GO:0001508 - regulation of action potential	GO:0030520 - estrogen receptor signaling pathway
GO:0043067 - regulation of programmed cell death	GO:0046039 - GTP metabolic process
GO:0006879 - iron ion homeostasis	
GO:0042981 - regulation of apoptosis	
GO:0007265 - Ras protein signal transduction	
GO:0030032 - lamellipodium biogenesis	
GO:0009894 - regulation of catabolic process	
GO:0032940 - secretion by cell	
GO:0007010 - cytoskeleton organization and biogenesis	
GO:0040017 - positive regulation of locomotion	
GO:0051272 - positive regulation of cell motility	
GO:0006402 - mRNA catabolic process	
GO:0006471 - protein amino acid ADP-ribosylation	
GO:0008064 - regulation of actin polymerization ...	
GO:0030036 - actin cytoskeleton organization ...	
GO:0048468 - cell development	
GO:0005996 - monosaccharide metabolic process	
GO:0006996 - organelle organization and biogenesis	

Table 4.1: GoMiner GSA results indicating GO Biological Process terms significantly overrepresented amongst the genes expressed in airway epithelial cells from never-smokers. Only those terms exclusive to the results obtained from either the original or paralog-reduced list are shown. A P-value cut-off of $\alpha = 0.05$ was used to determine significance.

represented amongst the expressed genes. Also, Mollerup et al. (2002) found that estrogen receptors are expressed in normal lung tissue in both sexes, enabling the 'estrogen receptor signalling pathway' (GO:0030520) to respond to the hormone, which is required for the promotion of lung function by the maintenance of alveoli (Massaro and Massaro 2004).

Two of the other remaining terms (GO:0009799 'determination of symmetry', GO:0009855 'determination of bilateral symmetry') also seem to represent plausible biological processes given the importance of airway symmetry. These results show that performing GSA on a paralog-reduced expression dataset using the GoMiner (strict cut-off) approach can yield novel and biologically relevant terms not otherwise elucidated. As GoMiner and other equivalent GSA methods make the same explicit assumption regarding the independent expression of gene transcripts, it is reasonable to expect similar findings with these other tools.

University of Cape Town

4.2.2.2 Reanalysis of GSEA Datasets Using Indygene

GSEA is a popular tool amongst biologists and its use in the analysis of human gene expression data has been recommended in recent reviews of GSA methods (Allison et al. 2006, Nam and Kim 2008). This tool starts from a list of genes ranked according to their association with the microarray class labels and then attempts to find gene sets distributed in a non-random fashion throughout this list, particularly concentrated towards its extremes. GSEA falls into the category of GSA methods that use the entire vector of P-values from a differential expression analysis and although it does not explicitly assume gene expression independence, it does make use of a competitive null hypothesis (see Section 1.3.2.2). This may cause GSEA to rank gene sets containing paralogous relationships in a biased manner. We compared the differences between the results from the reanalysis of five different GSEA gene expression datasets before and after eliminating paralogous relationships using Indygene. The five datasets cover a diverse range of topics, namely gender-specific expression differences in lymphoblastoid cells, p53 status in cancer cell lines, classification of acute leukaemias and two lung cancer outcome studies (see Section 4.1.2). Subramanian et al. (2005) used these datasets to show GSEA's ability to detect subtle but coordinated expression changes in sets of related genes defined by MSigDB (see Section 1.3.1). MSigDB consists of five major collections of human gene sets, but we only make use of two of these: MSigDB:C1, which contains gene sets based on chromosomal location and MSigDB:C2, which contains gene sets based on common roles in metabolic/signalling pathways or coregulation in response to chemical/genetic perturbations. The results of our analyses, which were obtained using the same significance threshold as these authors, are shown in Table 4.2.

On average, the numbers of significant gene sets obtained with GSEA using the original and Indygene-reduced datasets were similar (44 and 38 respectively across the five datasets). This makes intuitive sense as a gene set could potentially become less significant when paralogous genes that are interesting are removed, but on the other hand, the removal of paralogous genes that are uninteresting could result in greater significance. This is worth noting, because one might expect that the removal of a large number of genes would result in a significant reduction in statistical power and therefore also the number of significant gene sets.

The original and paralog-reduced results for the lymphoblast cell lines dataset were similar,

University of Cape Town

Original dataset: unique GSA results	Reduced dataset: unique GSA results
[1] Lymphoblast cell lines: **-Enriched in males:** TGFBETA_C2_UP **-Enriched in females:**	**[1] Lymphoblast cell lines:** **-Enriched in males:** CROONQUIST_IL6_STROMA_UP **-Enriched in females:** CHESLER_HIGHEST_FOLD_RANGE_GENES BHATTACHARYA_ESC_UP
[2] p53 status in NCI-60 cell lines: **-Enriched in p53 wild type:** P53HYPOXIAPATHWAY HSP27PATHWAY MMS_HUMAN_LYMPH_HIGH_24HRS_UP P53PATHWAY KANNAN_P53_UP P53_BRCA1_UP RADIATION_SENSITIVITY	**[2] p53 status in NCI-60 cell lines:** **-Enriched in p53 wild type:**
[3] Acute leukaemias: **-Enriched in ALL:**	**[3] Acute leukaemias:** **-Enriched in ALL:** chr13q14
[4] Lung cancer outcome (Boston study): **-Enriched in poor outcome:** TGFBETA_C1_UP HDACI_COLON_TSA_DN	**[4] Lung cancer outcome (Boston study):** **-Enriched in poor outcome:** CANCER_UNDIFFERENTIATED_META_UP MARSHALL_SPLEEN_BAL TRNA_SYNTHETASES EGF_HDMEC_UP AMINOACYL_TRNA_BIOSYNTHESIS ZELLER_MYC_UP HDACI_COLON_BUT16HRS_DN ZHAN_MULTIPLE_MYELOMA_SUBCLASSES_DIFF MYC_TARGETS MENSE_HYPOXIA_UP SMITH_HTERT_UP DOX_RESIST_GASTRIC_UP BASSO_REGULATORY_HUBS
[5] Lung cancer outcome (Michigan study): **-Enriched in poor outcome:** TGFBETA_C1_UP HSA00010_GLYCOLYSIS_AND_GLUCONEOGENESIS GLYCOLYSIS GLUCONEOGENESIS MENSE_HYPOXIA_UP VEGFPATHWAY ROME_INSULIN_2F_UP INSULIN_SIGNALING BHATTACHARYA_ESC_UP VANTVEER_BREAST_OUTCOME_GOOD_VS_POOR_DN GLYCOLYSIS_AND_GLUCONEOGENESIS ZUCCHI_EPITHELIAL_DN HYPOXIA_REVIEW	**[5] Lung cancer outcome (Michigan study):** **-Enriched in poor outcome:**

Table 4.2: GSEA results of five diverse gene expression datasets showing gene sets significantly enriched in the phenotype indicated. Functional gene sets (MSigDB:C2) were used in all cases, except for the leukaemia dataset where cytogenetic gene sets (MSigDB:C1) were used. Only those sets exclusive to the results obtained from either the original or paralog-reduced list are shown. A threshold of FDR ≤ 0.25 was used to determine significance.

University of Cape Town

with the latter revealing one gene set enriched in males and two in females, which did not occur with the former. According to MSigDB:C2, the gene set attributed to work by Crooncquist et al. (2003) indicates "genes upregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those co-cultured with bone marrow stromal cells". The relevance of this gene set and the sets based on studies by Chesler et al. (2005) and Bhattacharya et al. (2004) to gender-specific expression differences is not clear. The significance of these gene sets may be artifactual and due to confounding factors such as a gender-biased sampling programme in these studies.

No significantly enriched gene sets unique to the paralog-reduced 'p53 status' dataset were found. However, the one significantly enriched gene set unique to the paralog-reduced 'acute leukaemias' dataset corresponds to the 13q14 cytogenetic location ('MSigDB:C1:-chr13q14') containing the *RB* gene, which is often deleted or translocated in patients with AML, but rarely in ALL (Tanaka et al. 1999). This evidence confirms the importance of this gene set regarding expression differences between acute leukaemia subclasses.

Unlike the 'Michigan lung cancer outcome' study, analysis of its Boston counterpart yielded many significantly enriched gene sets unique to the paralog-reduced dataset and a number of these are plausible contributers to the poor outcome observed. 'MSigDB:C2:CANCER_UNDIFFERENTIATED_META_UP' is a gene set comprised of 69 genes commonly upregulated in undifferentiated cancer. Undifferentiated cancers tend to be more malignant than well-differentiated cancers, possibly explaining the association between this gene set and a poor survival outcome. Also, the 'MSigDB:C2:ZELLER_MYC_UP' and 'MSigDB:C2:MYC_TARGETS' gene sets contain genes that are up-regulated, or otherwise responsive, to *Myc*. The *Myc* protein is a transcription factor that stimulates the expression of many genes involved in cell-cycle progression. Its overexpression has also been associated with many types of cancer (Lodish et al. 2001). Furthermore, Berns et al. (1996) and Grotzer et al. (2001) found that high *Myc* expression is correlated with a poor outcome in patients with breast and brain cancers respectively. It is conceivable that a similar relationship exists in the case of lung cancer. Explanations for the other significantly enriched gene sets unique to the paralog-reduced dataset are not apparent, but could provide novel hypotheses for future investigation.

University of Cape Town

4.2.2.3 Reanalysis of SAM-GS Dataset Using Indygene

SAM-GS is a member of the category of GSA methods that model the raw expression data directly (see Section 1.3.2.3) and was recently developed by Dinu et al. (2007) amid concerns about the statistical foundations of other popular GSA methods, such as GoMiner and GSEA. It has not drawn criticism aimed at its statistical approach to the same extent that these other methods have, but it has not yet garnered much of a following amongst biologists either. This may be due to its testing procedure that uses a self-contained null hypothesis, which is invariably more powerful than those based on competitive null hypotheses (Goeman and Buhlmann 2007) (see also Section 1.3.2.2). Essentially SAM-GS tests whether a gene set is significantly associated with the phenotype of interest using gene expression information from the genes in that gene set alone. This can result in a long list of significant gene sets containing many, or *all* gene sets (Dinu et al. 2007), which might hinder biological interpretation in the same way that a lengthy list of genes from a differential expression analysis can. Therefore despite their rigorous statistical formulations, methods such as SAM-GS may be off the mark in terms of biologists' requirements.

Original dataset: unique GSA results	Reduced dataset: unique GSA results
APOPTOSIS	ADIP_VS_FIBRO_DN
APOPTOSIS_GENMAPP	BCNU_GLIOMA_MGMT_48HRS_UP
APOPTOSIS_KEGG	BRCA1_SW480_DN
CELLCYCLEPATHWAY	BREAST_CANCER_ESTROGEN_SIGNALING
CHEMICALPATHWAY	DAC_PANC50_UP
FSH_HUMAN_GRANULOSA_UP	DNA_DAMAGE_SIGNALING
G1PATHWAY	DRUG_RESISTANCE_AND_METABOLISM
HSA05219_BLADDER_CANCER	G2PATHWAY
HSP27PATHWAY	HSA05040_HUNTINGTONS_DISEASE
IL4PATHWAY	OXSTRESS_BREASTCA_UP
P53_BRCA1_UP	PARP_KO_UP
RACCYCDPATHWAY	PASSERINI_APOPTOSIS
SA_FAS_SIGNALING	SA_DIACYLGLYCEROL_SIGNALING
	SHEPARD_NEG_REG_OF_CELL_PROLIFERATION

Table 4.3: SAM-GS results indicating functional gene sets (MSigDB:C2) significantly enriched in the expression patterns of NCI-60 cancer cell lines with wild-type *p53*, compared to those of *p53* mutants. Only those terms exclusive to the results obtained from either the original or paralog-reduced list are shown. A threshold of FDR ≤ 0.001 was used to determine significance.

To benchmark the performance of SAM-GS against GSEA, Dinu et al. (2007) reanalysed the 'p53 status' dataset of Subramanian et al. (2005) discussed in Section 4.2.2.2. They used SAM-GS to analyse this dataset and identify MSigDB (see Section 1.3.1) gene sets exhibiting bi-directional expression change across a two-class phenotype, defined by the presence or absence of the wild-type *p53* gene. The results of our SAM-GS analysis using

University of Cape Town

the original and paralog-reduced datasets are shown in Table 4.3. Here we focus on the most important gene sets satisfying the stricter significance threshold of FDR ≤ 0.001, as the original authors' threshold of FDR ≤ 0.01 resulted in about one hundred significant terms in each case.

As discussed in Section 4.1.2, the transcription factor p53 plays an important role in the cellular response to DNA damage. Because cells in cancerous tissue possess mutations in their DNA, up-regulation of *p53* and other downstream 'MSigDB:C2:-DNA_DAMAGE_SIGNALING' genes is to be expected for such cells with the wild-type. The wild-type *p53* protein also has the ability to arrest cells with damaged DNA at particular points in the cell-cycle to avoid copying of these errors and provide time for their repair (Lodish et al. 2001). This anti-proliferative effect of wild-type p53 is evident in the significance of the 'MSigDB:C2:SHEPARD_NEG_REG_OF_CELL_PROLIFERATION' gene set, described as containing human genes whose orthologs in zebra fish negatively regulate cell proliferation. Another interesting significantly enriched gene set unique to the paralog-reduced dataset is the 'MSigDB:C2:DRUG_RESISTANCE_AND_METABOLISM' gene set. The findings of Bunz et al. (1999), which seem to corroborate this result, show that the presence of mutations affecting *p53* in human cancer cells renders them resistant to certain drugs used in cancer therapy.

These results show that despite the statistical validity of the procedure used in SAM-GS, removing paralogous relationships in the data can elucidate novel and biologically plausible hypotheses in the form of significantly enriched gene sets not found in the original dataset. Although we have only performed this analysis for SAM-GS, these results indicate that Indygene is likely to have utility when used in conjunction with other GSA methods in the same category.

University of Cape Town

Chapter 5

Concluding Remarks

University of Cape Town

GSA often represents the first attempt to make biological sense of the data obtained from a microarray experiment. These methods offer a powerful approach as they leverage large amounts of previously obtained scientific knowledge (gene sets) with the goal of generating hypotheses regarding new data. This can proceed in a self-reinforcing way, as results from one microarray GSA can become part of the background knowledge used to perform another. Promising hypotheses from such analyses are usually investigated by means of further experimentation and therefore the accuracy of results from GSA has a direct impact on the amount of time, effort, money and success associated with the overall study. We investigated the effect of paralogs on the results from GSA as we suspected that their presence might affect the ability to accurately identify the most important sets of genes for subsequent research.

As expected, we found that paralogs tend to have correlated expression patterns. This is at odds with the explicit assumption of gene expression independence made by many GSA methods and we found that this contradiction significantly affects results from these analyses. To study this issue we developed the Indygene tool, which efficiently removes paralogous relationships from a given dataset and we found that such a reduction, performed prior to GSA, has the ability to generate novel and biologically plausible hypotheses not otherwise obtained. This was demonstrated for three different GSA approaches (GoMiner, GSEA, SAM-GS) when applied to the reanalysis of previously published microarray datasets and suggests that the Indygene tool has utility when used in conjunction

with a broad range of methods.

However Indygene should be regarded as a temporary alternative in the absence of more sophisticated approaches for dealing with the dependencies between genes in such analyses. Instead of avoiding invalid assumptions by performing *ad hoc* gene removal, new GSA methods could attempt to model gene-gene dependencies directly. We argue that future GSA methods should behave in such a way that the weight of evidence implicating a particular biological process based on the coordinated expression change of the participating members is greater when they are evolutionarily distinct as opposed to when they are related. This may warrant a Bayesian statistical approach to GSA where the prior probability that paralogs exhibit coordinated expression change is greater than that of unrelated genes.

Also, it may be useful to adopt a systems biological approach to GSA where the role and relative contribution of each gene in a biological pathway or process is taken into account. Traditional approaches treat all genes in a gene set as equally important indicators of the (in)activation of a particular biological process. However from studies of biological pathway dynamics and control we know that this is not the case and it may be beneficial to somehow incorporate this information into the analysis of expression data.

Finally, the authors of GSA tools often fail to explicitly state the assumptions of the models that they use and this can result in researchers missing important caveats of the tools. To prevent confusion and misinterpretation of GSA results, model assumptions should be explicitly stated and more effort should be made to determine their real-world biological validity.

University of Cape Town

Bibliography

Affymetrix Inc. Affymetrix. Online, 2008. URL `http://www.affymetrix.com`. 2, 3

F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/btg455. 13

F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, 21(13):2988–2993, 2005. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bti457. 14, 15

A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/btl140. 46

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000. ISSN 0028-0836 (Print). doi: 10.1038/35000501. 9

D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, 2006. ISSN 1471-0056 (Print). doi: 10.1038/nrg1749. 5, 6, 8, 9, 13, 17, 53

J. M. Alonso, A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen, P. Shinn, D. K. Stevenson, J. Zimmerman, P. Barajas, R. Cheuk, C. Gadrinab, C. Heller, A. Jeske, E. Koesema, C. C. Meyers, H. Parker, L. Prednis, Y. Ansari, N. Choy, H. Deen, M. Geralt, N. Hazari, E. Hom, M. Karnes, C. Mulholland, R. Ndubaku, I. Schmidt, P. Guzman, L. Aguilar-Henonin, M. Schmid, D. Weigel, D. E. Carter, T. Marchand, E. Risseeuw, D. Brogden,

University of Cape Town

A. Zeko, W. L. Crosby, C. C. Berry, and J. R. Ecker. Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301(5633):653–657, 2003. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1086391. 46, 49

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990. ISSN 0022-2836 (Print). doi: 10.1006/jmbi.1990.9999. 24

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, 2002. ISSN 1061-4036 (Print). doi: 10.1038/ng765. 48

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000. ISSN 1061-4036 (Print). doi: 10.1038/75556. 8, 10, 11

A. Babiker, O. Andersson, D. Lindblom, J. van der Linden, B. Wiklund, D. Lutjohann, U. Diczfalusy, and I. Bjorkhem. Elimination of cholesterol as cholestenoic acid in human lung by sterol 27-hydroxylase: evidence that most of this steroid in the circulation is of pulmonary origin. *J Lipid Res*, 40(8):1417–1425, 1999. ISSN 0022-2275 (Print). 51

P. Baldi and A. D. Long. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6): 509–519, 2001. ISSN 1367-4803 (Print). 6

W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9): 1943–1949, 2005. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bti260. 16

D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8(8):816–824, 2002. ISSN 1078-8956 (Print). doi: 10.1038/nm733. 48

T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bth088. 13

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995. 8

University of Cape Town

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist*, 29(4):1165–1188, 2001. 14

E. M. Berns, J. G. Klijn, M. Smid, I. L. van Staveren, M. P. Look, W. L. van Putten, and J. A. Foekens. TP53 and MYC gene alterations independently predict poor prognosis in breast cancer patients. *Genes Chromosomes Cancer*, 16(3):170–179, 1996. ISSN 1045-2257 (Print). 55

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, 98(24):13790–13795, 2001. ISSN 0027-8424 (Print). doi: 10.1073/pnas.191502998. 48

B. Bhattacharya, T. Miura, R. Brandenberger, J. Mejido, Y. Luo, A. X. Yang, B. H. Joshi, I. Ginis, R. S. Thies, M. Amit, I. Lyons, B. G. Condie, J. Itskovitz-Eldor, M. S. Rao, and R. K. Puri. Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood*, 103(8):2956–2964, 2004. ISSN 0006-4971 (Print). doi: 10.1182/blood-2003-09-3314. 55

BioCarta Inc. BioCarta - Charting Pathways of Life. Online, 2008. URL `http://www.biocarta.com`. 11

E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bth456. 13

R. Breitling, A. Amtmann, and P. Herzyk. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5:34, 2004. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-5-34. 15

Broad Institute. GSEA | MSigDB. Online, 2008. URL `http://www.broad.mit.edu/gsea/msigdb/index.jsp`. 11

F. Bunz, P. M. Hwang, C. Torrance, T. Waldman, Y. Zhang, L. Dillehay, J. Williams, C. Lengauer, K. W. Kinzler, and B. Vogelstein. Disruption of p53 in human cancer cells alters the responses to therapeutic agents. *J Clin Invest*, 104(3):263–269, 1999. ISSN 0021-9738 (Print). doi: 10.1172/JCI6863. 57

Y. Caro. New results on the independence number. *Tel-Aviv University*, pages 75–79, 1979. 39, 40

University of Cape Town

J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. Kimbrough, M. S. Cline, G. Liu, R. Shigeta, D. Kulp, and M. A. Siani-Rose. NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, 20(9): 1462–1463, 2004. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bth087. 13

E. J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, D. W. Threadgill, K. F. Manly, and R. W. Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*, 37(3):233–242, 2005. ISSN 1061-4036 (Print). doi: 10.1038/ng1518. 55

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001. 39

D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*, 32(Database issue):D575–7, 2004. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkh133. 28

P. A. Croonquist, M. A. Linden, F. Zhao, and B. G. Van Ness. Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, N-ras-activating mutations, and coculture with bone marrow stromal cells. *Blood*, 102(7):2581–2592, 2003. ISSN 0006-4971 (Print). doi: 10.1182/blood-2003-04-1227. 55

X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 2005. ISSN 1465-4644 (Print). doi: 10.1093/biostatistics/kxh018. 6

D. Damian and M. Gorfine. Statistical concerns about the GSEA procedure. *Nat Genet*, 36(7):663; author reply 663, 2004. ISSN 1061-4036 (Print). doi: 10.1038/ng0704-663a. 16

C. Darwin. On the origin of species. *London*, 1859. 21

I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:242, 2007. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-242. 18, 48, 56

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998. 23

G. S. Eichler, M. Reimers, D. Kane, and J. N. Weinstein. The LeFE algorithm: embracing the complexity of gene expression in the interpretation of microarray data. *Genome Biol*, 8(9):R187, 2007. ISSN 1465-6914 (Electronic). doi: 10.1186/gb-2007-8-9-r187. 19, 20

University of Cape Town

J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, 1998. ISSN 1088-9051 (Print). 21

W. J. Ewans and G. R. Grant. *Statistical Methods in Bioinformatics*. Statistics for Biology and Health. Springer, 2005. 13, 23

R. A. Fisher. The possible modification of the response of the wild type to recurrent mutations. *Am. Nat.*, 62:115–26, 1928. 22

W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–106, 1970. 21

S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773, 1991. ISSN 0036-8075 (Print). 3

A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999. ISSN 0016-6731 (Print). 22

R. J. Garrison, W. B. Kannel, M. Feinleib, W. P. Castelli, P. M. McNamara, and S. J. Padgett. Cigarette smoking and HDL cholesterol: the Framingham offspring study. *Atherosclerosis*, 30(1):17–25, 1978. ISSN 0021-9150 (Print). 51

Gene Ontology Consortium. the Gene Ontology. Online, 2008. URL `http://geneontology.org`. 11

P. J. Giles and D. Kipling. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19(17):2254–2262, 2003. ISSN 1367-4803 (Print). 7

Gladstone Institutes. GenMAPP - Home Page. Online, 2008. URL `http://www.genmapp.org`. 11

J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. ISSN 1460-2059 (Electronic). doi: 10.1093/bioinformatics/btm051. 12, 14, 17, 56

J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004. ISSN 1367-4803 (Print). 19

J. J. Goeman, J. Oosting, A.-M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957, 2005. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bti267. 19

University of Cape Town

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5):725–736, 1994. ISSN 0737-4038 (Print). 29

Google Inc. Google Scholar. Online, 2008. URL `http://scholar.google.com`. 51

J. Griggs. Lower bounds on the independence number in terms of the degrees. *J. Combin. Theory*, Ser. B(34):22–39, 1983. 40

M. A. Grotzer, M. D. Hogarty, A. J. Janss, X. Liu, H. Zhao, A. Eggert, L. N. Sutton, L. B. Rorke, G. M. Brodeur, and P. C. Phillips. MYC messenger RNA expression predicts survival outcome in childhood primitive neuroectodermal tumor/medulloblastoma. *Clin Cancer Res*, 7(8):2425–2433, 2001. ISSN 1078-0432 (Print). 55

M. Halldorsson and J. Radhakrishnan. Greed is good: approximating independent sets in sparse and bounded-degree graphs. *Algorithmica*, 18(1):145–163, 1997. 40

M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–61, 2004. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkh036. 10

X. He and J. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164, 2005. ISSN 0016-6731 (Print). doi: 10.1534/genetics.104.037051. 22

M. J. Heller. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4:129–153, 2002. ISSN 1523-9829 (Print). doi: 10.1146/annurev.bioeng.4.020702.153438. 2

C. T. Hittinger and S. B. Carroll. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163):677–681, 2007. ISSN 1476-4687 (Electronic). doi: 10.1038/nature06151. 22

M. Hollander and D. Wolfe. *Nonparametric Statistical Methods*. Wiley, 2nd edition edition, 1999. 7

D. A. Hosack, G. J. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biol*, 4(10):R70, 2003. ISSN 1465-6914 (Electronic). doi: 10.1186/gb-2003-4-10-r70. 8, 9, 13

University of Cape Town

T. H. H. Huxley. The origin of species. *Westminst. Rev.*, 17:541–70, 1860. 21

P. Johnson and J. Ecker. THE ETHYLENE GAS SIGNAL TRANSDUCTION PATHWAY: A Molecular Perspective. *Annual Reviews in Genetics*, 32(1):227–254, 1998. 49

F. C. Kafatos, A. Efstratiadis, B. G. Forget, and S. M. Weissman. Molecular evolution of human and rabbit beta-globin mRNAs. *Proc Natl Acad Sci USA*, 74(12):5618–5622, 1977. ISSN 0027-8424 (Print). 32

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000. ISSN 0305-1048 (Print). 10, 11

R. Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, 43:85–103, 1972. 39

S. Kaul, H. Koo, J. Jenkins, M. Rizzo, T. Rooney, L. Tallon, T. Feldblyum, W. Nierman, M. Benito, X. Lin, et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. 26

T. Kayo, D. B. Allison, R. Weindruch, and T. A. Prolla. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc Natl Acad Sci U S A*, 98(9):5093–5098, 2001. ISSN 0027-8424 (Print). doi: 10.1073/pnas.081061898. 8

P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bti565. 9, 12

S. Knudsen. *Guide to Analysis of DNA Microarray Data*. John Wiley and Sons, Inc., second edition edition, 2004. 2, 3, 4, 6, 7

F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Selection in the evolution of gene duplications. *Genome Biol*, 3(2):RESEARCH0008, 2002. ISSN 1465-6914 (Electronic). 22

E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39: 309–338, 2005. ISSN 0066-4197 (Print). doi: 10.1146/annurev.genet.39.073003.114725. 21

L. B. Koski and G. B. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, 2001. ISSN 0022-2844 (Print). doi: 10.1007/s002390010184. 24

T. Kulikova, R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. P. G. Pastor,

University of Cape Town

S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, 35(Database issue): D16–20, 2007. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkl913. 29

C. K. Lee, R. G. Klopp, R. Weindruch, and T. A. Prolla. Gene expression profile of aging and its retardation by caloric restriction. *Science*, 285(5432):1390–1393, 1999. ISSN 0036-8075 (Print). 5

T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1075090. 24

D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985. ISSN 0036-8075 (Print). 24

Q. Liu, I. Dinu, A. J. Adewale, J. D. Potter, and Y. Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8:431, 2007. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-431. 19

H. Lodish, A. Berk, and S. L. Zipursky. *Molecular Cell Biology*. W.H. Freeman, 2001. 1, 28, 55, 57

U. Mansmann and R. Meister. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med*, 44(3):449–453, 2005. ISSN 0026-1270 (Print). doi: 10.1267/METH05030449. 19

D. Massaro and G. D. Massaro. Estrogen regulates pulmonary alveolar formation, loss, and regeneration in mice. *Am J Physiol Lung Cell Mol Physiol*, 287(6):L1154–9, 2004. ISSN 1040-0605 (Print). doi: 10.1152/ajplung.00228.2004. 52

M. McCoy, M. Reisch, A. H. Tullo, P. L. Short, J.-F. Tremblay, and W. J. Storck. Production: growth is the norm. *Chem. Eng. News*, 84(28):59–68, 2006. 49

R. A. Miller, A. Galecki, and R. J. Shmookler-Reis. Interpretation, design, and analysis of gene array expression experiments. *J Gerontol A Biol Sci Med Sci*, 56(2):B52–7, 2001. ISSN 1079-5006 (Print). 5

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. 20

S. Mollerup, K. Jorgensen, G. Berge, and A. Haugen. Expression of estrogen receptors alpha and beta in human lung tissue and cell lines. *Lung Cancer*, 37(2):153–159, 2002. ISSN 0169-5002 (Print). 52

University of Cape Town

V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273, 2003. ISSN 1061-4036 (Print). doi: 10.1038/ng1180. 16

W. Moreira, G. R. Warnes, and L. Gautier. RPy home page. Online, 2008. URL `http://rpy.sourceforge.net/`. 28

D. Nam and S.-Y. Kim. Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–197, 2008. ISSN 1477-4054 (Electronic). doi: 10.1093/bib/bbn001. 11, 14, 53

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970. ISSN 0022-2836 (Print). 23

S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970. 22

M. Olivier, R. Eeles, M. Hollstein, M. A. Khan, C. C. Harris, and P. Hainaut. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat*, 19(6):607–614, 2002. ISSN 1098-1004 (Electronic). doi: 10.1002/humu.10081. 48

R. Owen. On the archetype and homologies of the vertebrate skeleton. *London: Murray*, 1848. 21

K.-H. Pan, C.-J. Lih, and S. N. Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc Natl Acad Sci U S A*, 102(25):8961–8965, 2005. ISSN 0027-8424 (Print). doi: 10.1073/pnas.0502674102. 14

H. Pang, A. Lin, M. Holford, B. E. Enerson, B. Lu, M. P. Lawton, E. Floyd, and H. Zhao. Pathway analysis using random forests classification and regression. *Bioinformatics*, 22 (16):2028–2036, 2006. ISSN 1460-2059 (Electronic). doi: 10.1093/bioinformatics/btl344. 19

I. D. Pavord and A. E. Tattersfield. Bronchoprotective role for endogenous prostaglandin E2. *Lancet*, 345(8947):436–438, 1995. ISSN 0140-6736 (Print). 51

R Foundation. The R Project for Statistical Computing. Online, 2008. URL `http://www.r-project.org/`. 28

P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277, 2000. ISSN 0168-9525 (Print). 27

University of Cape Town

I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007. ISSN 1460-2059 (Electronic). doi: 10.1093/bioinformatics/btl633. 13

B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering Design and Selection*, 12(2):85–94, 1999. 30

S. Sakai, M. Togasaki, and K. Yamazaki. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126(2-3):313–322, 2003. 40

V. Saxena, D. Orgill, and I. Kohane. Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res*, 34(22):e151, 2006. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkl766. 19

N. H. Shah and N. V. Fedoroff. CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, 20(7):1196–1197, 2004. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bth056. 13

R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Statistics for Biology and Health. Springer, 2003. 2, 3, 4, 5, 6, 7, 8, 9, 28

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981. ISSN 0022-2836 (Print). 23

A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, and J. S. Brody. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci U S A*, 101(27):10143–10148, 2004. ISSN 0027-8424 (Print). doi: 10.1073/pnas.0401422101. 48, 51

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005. ISSN 0027-8424 (Print). doi: 10.1073/pnas.0506580102. 9, 10, 11, 15, 16, 48, 53, 56

P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96 (6):2907–2912, 1999. ISSN 0027-8424 (Print). 9

K. Tanaka, M. Arif, M. Eguchi, S. X. Guo, Y. Hayashi, H. Asaoku, T. Kyo, H. Dohy, and N. Kamada. Frequent allelic loss of the RB, D13S319 and D13S25 locus in myeloid malignancies with deletion/translocation at 13q14 of chromosome 13, but not in lymphoid malignancies. *Leukemia*, 13(9):1367–1373, 1999. ISSN 0887-6924 (Print). 55

University of Cape Town

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, 2001. ISSN 0027-8424 (Print). doi: 10.1073/pnas.091062498. 6, 8

UniProt. The universal protein resource (UniProt). *Nucleic Acids Res*, 36(Database issue): D190–5, 2008. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkm895. 27

V. van Noort, B. Snel, and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, 5 (3):280–284, 2004. ISSN 1469-221X (Print). doi: 10.1038/sj.embor.7400090. 24

G. van Rossum and F. L. Drake. *Python Reference Manual*. Virginia, USA, 2001. URL `http://www.python.org`. 26

V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995. ISSN 0036-8075 (Print). 9

WebPy. Web.py: Think about the ideal way to write a web app. write the code to make it happen. Online, 2008. URL `http://webpy.org`. 38

V. Wei. A lower bound on the stability number of a simple graph. Technical report, Bell Laboratories Technical Memorandum 81-11217-9, Murray Hill, NJ, 1981, 1981. 39

K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997. ISSN 0028-0836 (Print). doi: 10.1038/42711. 21

Z. Wu, R. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–918, 2004. 28

D. Yang, Y. Li, H. Xiao, Q. Liu, M. Zhang, J. Zhu, W. Ma, C. Yao, J. Wang, D. Wang, Z. Guo, and B. Yang. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, 24(2): 265–271, 2008. ISSN 1460-2059 (Electronic). doi: 10.1093/bioinformatics/btm558. 10

Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556, 1997. ISSN 0266-7061 (Print). 29

B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003. ISSN 1465-6914 (Electronic). 13

B. R. Zeeberg, H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D. W. Kane, M. Reimers, R. M. Stephens, D. Bryant, S. K. Burt, E. Elnekave, D. M. Hari, T. A. Wynn,

University of Cape Town

C. Cunningham-Rundles, D. M. Stewart, D. Nelson, and J. N. Weinstein. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, 6:168, 2005. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-6-168. 47, 51

S. Zhong, K.-F. Storch, O. Lipan, M.-C. J. Kao, C. J. Weitz, and W. H. Wong. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl Bioinformatics*, 3(4):261–264, 2004. ISSN 1175-5636 (Print). 13

University of Cape Town

strict cut-off approaches (see Section 1.3.2.1), however GoMiner has proven to be one of the most popular tools with over 400 academic citations to date (Google Inc. 2008). In an updated version of the tool (High-Throughput GoMiner), Zeeberg et al. (2005) demonstrated the tool's utility by applying it in the interpretation of a novel dataset including the gene expression response of a patient with CVID (see Section 4.1.2). Once again employing GoMiner, we reanalysed the CVID dataset using the original and paralog-reduced gene lists. However these gene lists were nearly identical, because the relatively low number of 3968 unique genes detected on the printed microarrays possessed few paralogous relationships. Therefore the two sets of results obtained using GoMiner were essentially equivalent with no significant discrepancies.

As a more substantial comparison, we used the human airway epithelial cell transcriptome dataset of Spira et al. (2004). The authors used GoMiner to find terms in the GO Molecular Function ontology that were associated with genes expressed in the airway epithelial cells of healthy never-smokers, and therefore also associated with these cells' normal functioning (see Section 4.1.2). We repeated the analysis using the original and paralog-reduced gene lists and terms from the GO Biological Process ontology. Using the same P-value cut-off of $\alpha = 0.05$ used by Spira et al. (2004), over 130 GO terms were found to be significant in both cases. Table 4.1 shows only those terms exclusive to the results obtained from either the original or paralog-reduced list.

A number of interesting GO biological process terms were only found to be significant when GoMiner was applied to the paralog-reduced dataset. Their role in the normal functioning of airway epithelial cells seems plausible given the supporting evidence in the literature. For instance Babiker et al. (1999) found that the human lung plays an important role in maintaining 'cholesterol homeostasis' (GO:0042632) by the elimination of cholesterol as cholestenoic acid. Also, it is well known that smoking is associated with increased HDL cholesterol levels (Garrison et al. 1978), possibly explained by the malfunctioning of this homeostatic process caused by exposure to tobacco smoke.

Prostaglandins are a major product of airway epithelium and different types are involved in functions such as bronchodilation and bronchoconstriction. Namely, endogenous Prostaglandin E2 (PGE2) has been found to control the latter (Pavord and Tattersfield 1995). It is therefore not surprising that 'prostaglandin metabolic process' (GO:0006693) and its parent term 'prostanoid metabolic process' (GO:0006692) were found to be over-

University of Cape Town