RESAMPLING-BASED TESTS OF FUNCTIONAL CATEGORIES IN GENE EXPRESSION STUDIES

William T. Barry

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2006

Approved by:

Advisor: Fred A. Wright
Co-advisor: Andrew B. Nobel
Reader: Lawrence L. Kupper
Reader: Mayetri Gupta
Reader: Charles M. Perou

ABSTRACT

William T. Barry: Resampling-based tests of functional categories in gene expression studies
(Under the direction of Dr. Fred A. Wright and Dr. Andrew B. Nobel)

DNA microarrays allow researchers to measure the coexpression of thousands of genes, and are commonly used to identify changes in expression either across experimental conditions or in association with some clinical outcome. With increasing availability of gene annotation, researchers have begun to ask global questions of functional genomics that explore the interactions of genes in cellular processes and signaling pathways. A common hypothesis test for gene categories is constructed as a *post hoc* analysis performed once a list of significant genes is identified, using classically derived tests for 2x2 contingency tables. We note several drawbacks to this approach including the violation of an independence assumption by the correlation in expression that exists among genes. To test gene categories in a more appropriate manner, we propose a flexible, permutation-based framework, termed SAFE (for Significance Analysis of Function and Expression).

SAFE is a two-stage approach, whereby gene-specific statistics are calculated for the association between expression and the response of interest and then a global statistic is used to detect a shift within a gene category to more extreme associations. Significance is assessed by repeatedly permuting whole arrays whereby the correlation between all genes is held constant and accounted for. This permutation scheme also preserves the relatedness of categories containing overlapping genes, such that error rate estimates can

be readily obtained for multiple dependent tests. Through a detailed survey of gene category tests and simulations based on real microarray, we demonstrate how SAFE generates appropriate Type I error rates as compared to other methods. Under a more rigorously defined null hypothesis, permutation-based tests of gene categories are shown to be conservative by inducing a special case with a maximum variance for the test statistic. A bootstrap-based approach to hypothesis testing is incorporated into the SAFE framework providing better coverage and improved power under a defined class of alternatives. Lastly, we extend the SAFE framework to consider gene categories in a probabilistic manner. This allows for a hypothesis test of co-regulation, using models of transcription factor binding sites to score for the presence of motifs in the upstream regions of genes.

ACKNOWLEDGMENTS

First, I would like to thank my advisors, Dr. Fred Wright and Dr. Andrew Nobel, for their constant support and guidance in writing this dissertation, and for also being excellent mentors in helping me grow as a researcher and biostatistician. I appreciated the insightful comments and suggestions of my committee members, Dr. Mayetri Gupta, Dr Larry Kupper, and Dr. Charles Perou, and also their willingness to offer both instruction and guidance throughout my education.

Finally, I would like to thank my family, friends, and fellow colleagues who have stood beside me as I have pursued this endeavor. Their love and kindness has made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area under the curve
CDF	Cumulative density function
DNA	Deoxyribonucleic acid
FDR	False discovery rate
FWER	Familywise error rate
GO	Gene Ontology
LR	Likelihood ratio
Pfam	Protein family database
PSWM	Position-specific weight matrix
ROC	Receiver operating characteristic
SAFE	Significance analysis of function and expression
TF	Transcription factor

1 Introduction and Literature Review

1.1 Introduction

Recent advances in high-throughput biotechnologies have led to the development of experimental methods for simultaneously measuring the expression of multiple genes, at either the transcriptional (Schena et al. 1995) or translational level (Honore et al. 2004). In particular, DNA microarray technology has found the widest application, extending across many areas of biology and medicine. With nucleotide sequences representing thousands of genes affixed onto a single slide, microarrays are able to obtain a snap-shot of transcription across much of the genome for one or more biological samples, and have been constructed for many diverse organisms. These technologies, along with other large-scale efforts, have allowed researchers to ask more global questions of functional genomics (Kohane et al. 2003) that extend the biological knowledge obtained for single genes to that of groups of genes and their interactions in cellular processes and signaling pathways.

In applying microarrays to the study of functional genomics, most experimental designs can be broadly characterized as one of two types. The first are discriminant analyses of either biological samples or gene expression profiles (Eisen et al. 2001), where many traditional methods of supervised and unsupervised learning have been implemented. These include hierarchical clustering (Eisen et al. 2001), self organized maps (Golub et al. 1999), and support vector machines (Brown et al. 2001) along with novel methods, such as biclustering across both genes and samples (Kluger et al. 2003). The second

popular use of microarrays is the identification of differential expression among the set of genes represented on the array (Schena et al. 1995). Although these studies often employ classical methods for testing the associations of gene expression, statistical considerations are needed for the high dimensionality of the data where thousands of genes are being measured over a much smaller number of samples, typically numbering in the tens, or at most hundreds, of arrays.

While it is important to address the differential expression of genes individually, most biological phenomena and human diseases are thought to occur through the interactions of multiple genes, via signaling pathways or other functional relationships. As the understanding of cellular processes has grown, descriptions of gene function have accumulated in databases of annotation that extend across the known genome for one or multiple species. For example, one of the first databases of known genes, SWISS-PROT, provides a set of keywords for each gene based on a taxonomy that includes pathways, diseases and general biological processes (Boeckmann et al. 2003). Gene annotation has also been presented in more complicated structures, such as the hierarchical vocabularies generated by the Gene Ontology Consortium (Ashburner et al. 2000). With the biological information assembled into curated vocabularies, one can group genes together based on a shared keyword or function. Thus, research questions are beginning to shift from the activity of genes individually to that of broader functional groups of genes, and the coexpression measured by microarray technologies provides a unique opportunity to design hypothesis tests to answer these questions.

Herein, we will examine some of the standard statistical methodologies utilized in differential expression experiments, and develop a series of methodologies for address-

ing research questions involving functional categories of genes. The remainder of this chapter provides a detailed description of the common microarray technologies and some techniques for processing gene expression data. The statistical methods that have been used to conduct hypothesis tests of differential expression are reviewed, along with issues regarding multiple comparison. Modern databases for the annotation and functional characterization of known genes are summarized, and a recent class of "gene-list enrichment" tests is briefly described.

In Chapter 2, a general framework for conducting hypothesis tests of gene categories is presented with a distinct nomenclature for describing the multivariable expression data and also for sets of functional categories. Within this framework a permutation-based approach to hypothesis testing is proposed and implemented in an example dataset involving several different types of comparisons. Chapter 3 more closely surveys the different methods of testing gene categories that have been proposed in the literature. For each distinct method, the underlying null hypotheses are explicitly derived since little consideration has been given in the literature. We then use these null hypotheses and simulations based on real microarray data to illustrate shortcomings in these methods and to suggest a broader null hypothesis for functional categories. A bootstrap-based method is suggested as being able to test this broader null without parametric assumptions and is shown to be less conservative than permutation in this setting. Improved coverage and power are presented via simulation and a real microarray setting. In Chapter 4 the concept of a functional category is extended to a more probabilistic definition to incorporate uncertainty in gene annotation. This extension provides a novel method for studying transcriptional regulation of DNA sequences. Models are defined based

on methodologies for transcription factor motif discovery, and used to score the non-coding sequences around genes for the presence of known motifs. From this we calculate the posterior probability of a gene's membership in a function category of transcription factor targets, and test for concerted differential expression in microarray data. Lastly, these methods are extended to consider the interactions of transcription factors, and to update estimates of binding sites based on new co-expression data.

1.2 Microarray technology

Over the past decade, a number of different DNA microarray technologies have been developed that allow researchers to assay gene expression across either the human genome or the genome of several model organisms (Brown and Botstein 1999). Broadly speaking, microarrays measure gene expression in mRNA samples by reverse-transcribing a labeled target sample and hybridizing it to a series of probes that have been affixed to chips in a specified grid. Protocols vary in the manner in which target samples are labeled and in how probes are designed to correspond to known transcripts. The preprocessing of microarray data into estimates of expression are highly platform specific, but will typically involve the following steps: 1) quantifying hybridization from the intensity of scanned images, 2) spatial and/or global normalization of arrays, 3) model-based estimation of expression from either sets of probes for a single transcript, or ratios of probes from different samples, 4) the potential filtering of lowly expressed genes or outlying samples. The usual output of such preprocessing steps is a rectangular matrix of expression estimates for a given set of genes and samples. Details about the chip design and data

preprocessing steps are given below for two of the most common array types: spotted cDNA microarrays and high-density oligonucleotide arrays.

In spotted cDNA arrays, first introduced by Schena et al. (1995), robotics is used to adhere specified probes onto a glass slide. Probes are usually nucleotide sequences that are a few hundred base pairs in length and which have been individually amplified by PCR from bacterial clones. This allows researchers to design customized arrays to include the parts of a species' genome that are of interest. Commercially prepared arrays are also available from companies such as Agilent Technologies which provide a standard platform that cover a large proportion of the genome of interest. Because of the unknown efficiency in immobilizing a probe to a particular spot, arrays have been designed to measure expression in two mRNA samples labeled separately with the red Cy5 and green Cy3 dyes. It is common for a reference sample to be used as one of the samples for all arrays in a given experiment, although other designs have been proposed that use chips in a more efficient manner by balancing samples across arrays and using dye-swaps (Kerr and Churchill 2001). Appropriate methods for the normalization of cDNA have been suggested in literature. Dudoit et al. (2002) suggested using LOESS normalization within the print-tips for robotically spotting arrays, while Wolfinger et al. (2001) proposed a linear mixed model with random effects for array and dyes with interactions. With these and other preprocessing steps, cDNA microarray data is presented as either individual expression estimates, or ratios between the two channel intensities. The following section will describe testing procedures that have been proposed for both data structures.

High-density oligonucleotide arrays are another popular form of gene expression technology, and arrays for many different species have been made commercially available by

Affymetrix. Probes consist of short oligonucleotide sequences (usually 25 base pairs in length) that are synthesized directly to glass slides using a photolithographic process (Kohane et al. 2003). This technique can produce chips with hundreds of thousands of different probes affixed which allows multiple probes to be designed for a single transcript (and are collectively termed a "probeset" by Affymetrix). A probeset typically consists of anywhere from five to twenty probe-pairs that correspond to distinct sequences within the transcript. Each probe-pair consists of a "perfect match" (PM) probe and a "mismatch" (MM) probe where a single base change switch is made in the 13th position of the probe. Different models have been proposed for estimating expression from a probeset, with considerable debate as to whether MM probes appropriately represent the non-specific hybridization to the short oligomers. Li and Wong (2001) proposed several models that contain multiplicative parameters for every probe, termed "probe sensitivity indexes", that represent the rate at which hybridization occurs, and use either the PM information only, the difference in PM and MM measurements, or both. Chu et al. (2004) proposed a similar set of linear mixed models for log-transformed intensities, and Irizarry et al. (2003) proposed using quantile normalization and robust fitting of an additive model on the log scale to obtain expression estimates from the PM data in oligonucleotide arrays.

1.3 Multiple testing of differential expression

In many applications of microarray data, the experimenter seeks to identify statistically significant associations between the expression profiles of genes and another variable related with each array, such as a sample group assignment, an experimental factor, or

survival time. We will refer to this additional variable as the "response" regardless of whether it is an observation of a random variable, or a fixed constant determined by the experimental design. The most common methods for analyzing expression data proceed in a gene-specific manner, using a statistical model to relate the response to the expression of each gene. In the earliest publications of cDNA spotted arrays, a hard threshold for fold change was suggested as the criterion for considering significant differential expression (Chen et al. 1997; Schena et al. 1995). However, such tests are non-statistical in that they ignore the amount of variability that exists in the expression data. Subsequently, more appropriate tests have been employed in two-sample comparisons, including the parametric Student's t-test (Galitski et al. 1999) and the non-parametric Wilcoxon rank sum test (Troyanskaya et al. 2002). More complex models have been suggested for particular microarray types, including mixed models that combine normalization and testing into a single step (Wolfinger et al. 2001) and a Bayesian model for the ratios of expression particular to cDNA arrays (Newton et al. 2001). In each of these methods, the association of each gene's expression to the response is considered separately; however, "shrinkage"-based methods are becoming popular in which improved estimates are obtained from considering the entire dataset (Cui et al. 2005; Hu and Wright 2005). A permutation-based method has been proposed by Tusher et al. (2001) that employs a modified t-statistic in two-sample comparisons. By adding an estimated variance inflation factor to the denominator of all statistics, this approach effectively down-weights genes that are lowly expressed, and thereby shows an improvement in the expected number of false discoveries among the genes significantly associated with the response.

Once a test statistic has been chosen, the primary statistical obstacle is accounting

Table 1: Possible outcomes from m hypothesis tests when the true states of being either null or alternative are fixed and known.

	Accept	**Reject**	**Total**
Truly Null	U	V	m_0
Truly Alternative	T	S	m_1
	W	R	m

for the number of comparisons needed to test all genes. In the multiple testing literature, the outcomes of the m tests are usually delineated as falling into one of four types, as shown in Table 1 (Benjamini and Hochberg 1995).

The random variables U and S represent the two kinds of correct conclusions that are made, while V is the number of false positives (Type I errors) and T is the number of false negatives (Type II errors) that occur. Two parameters that are often used to describe error when conducting multiple tests are the family-wise error rate (FWER) and the false discovery rate (FDR). Different methods have been proposed for either controlling or estimating one of these error rates in analyses containing multiple tests.

The FWER is defined as the probability of having at least one Type I error among the rejected hypotheses, $Pr(V \geq 1)$. Classically, a Bonferroni correction is employed as a single-step p-value adjustment, where for the ith test $\tilde{p}_i = \min(m \cdot p_i, 1)$. This provides conservative control of the FWER regardless of the correlation structure among the tests

Ge et al. (2003). Holm (1979) suggested a similar step-down procedure that applies successively less stringent adjustments to the ordered values, $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m}$

$$\tilde{p}_{r_i} = \max_{l:1,\ldots,i} \left[\min((m-l+1) \cdot p_{r_l}, 1) \right] \tag{1.1}$$

Westfall and Young (1989) proposed a resampling-based procedures for controlling the FWER when correlation exists among the hypothesis tests, that defines the adjusted p-values as

$$\tilde{p}_{r_i} = \max_{l:1,\ldots,i} \left[Pr(\min_{h:l,\ldots,m} P_{r_h} \leq p_{r_l} | m_1 = 0) \right]. \tag{1.2}$$

Even though this definition conditions on the fact that all genes are truly null ($m_1 = 0$), strong control of the FWER was proved for any realization of null and alternative hypotheses (Westfall and Young 1993). For each of these controlling procedures, a corresponding estimate of the FWER exists for every p-value cut-off to a rejection region.

The FWER error rate is often criticized as being too stringent a criterion when rejecting more than a few hypotheses. For this reason, methods that focus on the FDR have received much attention in the microarray literature where thousands of genes are tested simultaneously. The FDR was originally defined by Benjamini and Hochberg (1995) to be the expected rate of false positives among the rejected hypotheses $E[\frac{V}{R}]$ where in order to be finite, the ratio $\frac{V}{R}$ is defined to be zero when $R = V = 0$.

$$FDR = E\left[\frac{V}{R} | R > 0 \right] Pr(R > 0) \tag{1.3}$$

A second definition termed the "positive" false discovery rate (pFDR) considers the expectation alone, and has a direct Bayesian interpretation when the hypotheses are treated as random (Storey 2003) . In many applications the probability of no rejections

is small so the difference between these alternative definitions is negligible. Linear step-up procedures to control the FDR were proposed by Benjamini and Hochberg (1995) for independent tests and then by Benjamini and Yekutieli (2001) for correlated tests. To estimate the positive FDR of a given rejection region Storey and Tibshirani (2003) proposed several methods based on the following formulation, and applied the term "q-value" as the following error rate of a p-value and its corresponding estimate

$$\begin{aligned} q(p) &= \inf_{\{\Gamma: p \in \Gamma\}} pFDR(\Gamma) \\ \hat{q}(p) &= \min_{\{p_\Gamma \geq p\}} \left(\frac{\hat{m}_0 \cdot p_\Gamma}{\#\{p_i \leq p_\Gamma\}} \right) \end{aligned}$$

where Γ is the rejection region applied marginally to all hypothesis tests. It should be noted that the FDR controlling and estimating procedures can be applied to either parametrically derived p-values or empirical p-values obtained from resampling. Because of correlation in gene expression, the resampling-based procedures for estimating error rates have been shown to be more powerful in example microarray datasets (Ge et al. 2003; Reiner et al. 2003).

1.4 Gene categories

Over the past few decades the biological knowledge obtained from conventional biochemical and genetic studies have been accumulated in different public databases. As an example of one of the earliest endeavors, the SWISS-PROT database was established in 1986 to provide detailed description of protein sequences in a standard nomenclature (Boeckmann et al. 2003). Now containing over 230,000 entries, SWISS-PROT provides

sequence information along with names, species of origin, and references for every entry. In addition SWISS-PROT provides a set of keywords, based on a taxonomy that includes pathways, diseases and general biological processes. SWISS-PROT also provides cross references to other gene classifications, like that of InterPro and the Protein Families (Pfam) databases. Pfam has used multiple sequence alignment and hidden Markov models to identify 8296 "protein families" that share homology-based domains in their protein amino acid sequence (Sonnhammer et al. 1997). From these sources of information, a functional category can be formed by the set of genes which share a annotation feature, such as a SWISS-PROT keyword or a Pfam domain.

More recently, the Gene Ontology Consortium (GO) has developed a comprehensive vocabulary of gene annotation that is separated into three domains of classification: Biological Process, Cellular Component, and Molecular Function (Ashburner et al. 2000). In each domain, the ontology is structured as a directed acyclic graph (DAG), with a hierarchy of terms that vary from broad levels of classification (*e.g.* 'DNA Metabolism') down to more narrow levels (*e.g.* 'leading strand elongation'), as represented in Figure 1. For each GO term, a functional category is generally defined as containing the set of genes annotated directly to the node or to any terms that occupy descendant nodes in the ontology (Ashburner et al. 2000; Zhou et al. 2002). For example, from the subset of the Biological Process ontology shown in Figure 1, the mouse gene Lig1 would be in categories for 'DNA ligation', 'DNA recombination', 'DNA repair', 'DNA-dependent DNA replication', 'DNA replication' and the parent node 'DNA metabolism'.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database that further details the interaction of genes by the signaling pathways the gene products are involved

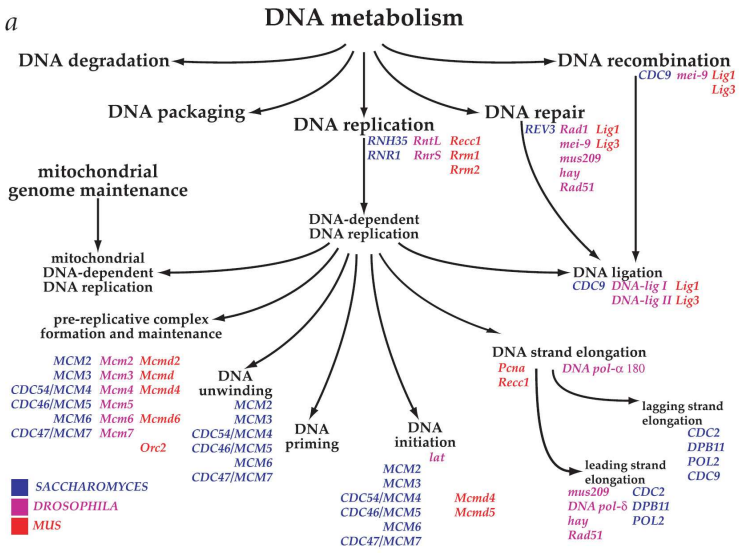

Figure 1: Example of the structure of Gene Ontology from Ashburner et al. (2000). A subset of the Biological Process DAG is shown with gene members from 3 different species

in (Kanehisa 1997). While a KEGG pathway contains considerably more information than mere membership of the genes, testing procedures are not available for the complex interactions of proteins networks, and as such whole or partial pathways may be reduced to functional categories as a way of examining their general associations to a response of interest in a microarray dataset.

Traditional gene-specific approaches for testing differential expression do not readily provide a way of making inferences about functional categories. Originally, researchers have informally perused lists of significant genes for functional annotations in order to make judgments about the underlying biology (Tusher et al. 2001). This is a subjective

process at best, and frequently the list of significant genes is too long to develop a parsimonious understanding of the role of biological function.

A number of publications and software packages in the last three years have proposed simple hypothesis tests for the differential expression of gene categories, in which a secondary analysis is performed once the list of significant genes has been determined. The most common method looks for over-representation, or "enrichment", of the category within the gene-list using techniques traditionally employed in the analysis of contingency tables (*e.g.* Fisher's Exact Test). Draghici et al. (2003) and Kim and Falkow (2003) were two of the first publications to describe the tests for over and provide tools for conducting tests on lists of genes: Onto-Express and LARK respectively. Subsequently, a series of online tools have also been developed including GOStat from Beißbarth and Speed (2004), FatiGO (Al-Shahrour et al. 2004), EASE (Hosack et al. 2003), and FuncAssociate (Berriz et al. 2003). Several other softwares have been developed that can also display the tests of over-representation across the DAG structure of a GO ontology: MAPPfinder (Doniger et al. 2003), GoMiner (Zeeberg et al. 2003), GoSurfer (Zhong et al. 2004), and GO Tree Machine (Zhang et al. 2004). In all of these software packages, testing for over of a keyword is done by appealing to standard sampling theory. Assume a total of m genes are on the array, and g of them are annotated to the term of interest. The p-value for having x genes make a gene-list of length k is derived from the hypergeometric distribution as

$$P(X \geq x|m, g, k) = \sum_{i=x}^{min(g,k)} \frac{\binom{g}{i}\binom{m-g}{k-i}}{\binom{m}{k}} \tag{1.4}$$

Many of the softwares also use Binomial, χ^2, or Normal approximations in conducting

the traditional tests of the difference in proportions, and in some, tests are also conducted by permuting the gene assignments of categories (Berriz et al. 2003; Zhong et al. 2004). In this way, the random sampling of genes assumed in the parametric tests is induced, but the relatedness of overlapping categories is accounted for in the estimated error rates for multiple testing. Other parametric tests have been proposed that use a more continuous measure of gene-specific significance (*e.g.* Boorsma et al. (2005); Goeman et al. (2004); Kim and Volsky (2005)), and permutation-based tests have been proposed using similar statistics (Mootha et al. 2003; Virtaneva et al. 2001). The gene-list enrichment tests have been criticized for having ill-defined null hypotheses (Allison et al. 2006) and for making assumptions inappropriate for microarray data (Barry et al. 2005), but are increasingly becoming a default tool for testing functional categories in differential expression studies.

A full discussion of the various hypothesis testing methodologies and their associated assumptions will be given in the following chapters.

1.5 Resampling-based tests

In many statistical applications, it is necessary to develop procedures which do not depend on any parametric assumptions about the observed data. The field of non-parametric statistics has sought to identify quantities whose distributions under a null hypothesis are not restricted by as many assumptions of how the data are derived; examples include rank-based statistics and other values that compare empirical distribution functions of the data in various ways. For many complex problem, no such distribution-free quantities may exist for the association of interest. If instead a statistic is chosen that will depend

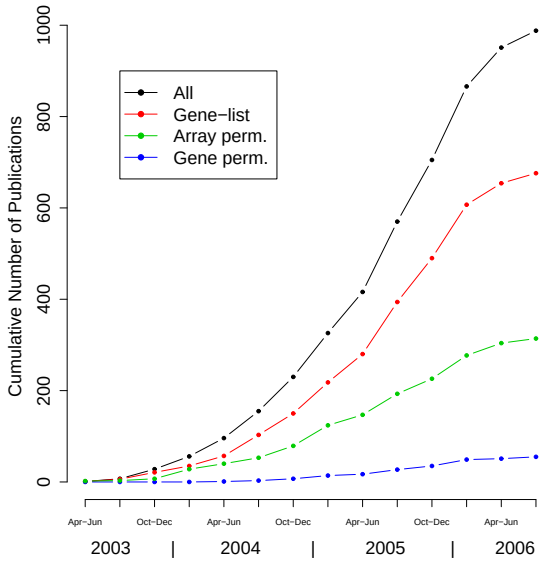

Figure 2: The cumulative number of citations of gene category tests plotted quarterly since 2003. Results are shown for 'gene-list' methods, resampling based methods that permute either gene- or array-assignments, and the union of the three sets (see Chapter 3 for a full discussion of these methodologies). Citations were obtained from ISI Web of Knowledge on 8/15/06.

on some parametric assumptions of the data, one may be able to use resampling in order to understand its underlying distribution. By recalculating the quantity from replicate datasets inferences can be made about certain properties of the underlying distribution. The two most common resampling-based tests can be broadly categorized as *permutation* where resampling observations without replacement allows a null hypothesis to be induced, and *bootstrap* methods where resampling the data with replacement can produce interval estimates around the observed statistic. Over the past few decades, advancements in technology have allowed these computationally-intensive methods to be

widely implemented in statistical applications.

1.5.1 Permutation testing

Permutation of observed data was originally proposed by R.A Fisher in the 1930s as a theoretical argument for justifying the t-distribution in a two-sample location problem, and has been utilized in deriving the null distribution for many non-parametric statistics (Hollander and Wolfe 1999). Specifically, if a statistic is written as some function of independent units of the observed data, $t_{obs} = T(x_1, \ldots, x_n)$, an empirical p-value can be simply obtained from the $n!$ reorderings of the data, $x_1^*, \ldots, x_n^*$

$$p = \frac{\#\text{ of permutations where } T(x_1^*, \ldots, x_n^*) \geq t_{obs}}{n!} \tag{1.5}$$

For many experimental designs, like the two-sample comparison, there will be fewer than $n!$ unique values T^* can take, which leads to a more discrete distribution of empirical p-values. Also, for large n it is often times sufficient to approximate p with a smaller number of randomly selected permutations

$$p \doteq \frac{1 + \sum_{k=1}^{K} I(t_k^* \geq t_{obs})}{K + 1} \tag{1.6}$$

Under this definition, p follows the discrete uniform distribution for the null hypothesis induced via permutation. For many uses of permutation tests, the induced null may not be expressly stated nor confirmed as pertaining to the research question of interest.

Examples of the use of permutation in the microarray literature extend from differential expression (Tusher et al. 2001) and corrections for multiple testing (Dudoit et al. 2003; Tusher et al. 2001), to validating unsupervised classification methods like principal

component analysis (Landgrebe et al. 2002), and similarity scores for gene categories (Rahnenführer et al. 2004). When applying permutation-based methods to microarray analysis, it is important to recognize what null hypothesis is induced by the randomization scheme, and whether it is appropriate for the given task (Allison et al. 2006).

1.5.2 Bootstrap testing

The general bootstrap method was proposed by Efron (1979) and is based on the presumption that the observed data is generated from an unknown probability model, F, as depicted in Figure 3 as adapted from Efron and Tibshirani (1998). If one defines $\theta = T(F)$ as a parameter of interest that is some function of the underlying distribution of the data, the plug-in principle suggests that a simple estimate of θ can be obtained from the empirical distribution function, $\hat{F}$, that is a corresponding estimate of F. In order to make inference on θ from $\hat{\theta} = T(\hat{F})$, resamples of the data are drawn from $\hat{F}$ yielding replicates of the statistic $\{\hat{\theta}^*\}$.

Many different methods have been proposed for using the bootstrap resamples to build confidence intervals for θ. If a normal approximation is assumed for the statistic, the replicate values can be used to generate bias and variance estimates for a confidence interval (Efron 1979, 1981). When a reasonable estimate for the variance of the statistic is available, confidence intervals can be generated from studentized versions of the statistic (Efron 1981). Percentile intervals use quantiles of the resampled statistics to estimate the limits such that results are completely insensitive to monotonic transformations of the statistic. Adjusted quantiles have been prosed in the "BCa" method to account for any

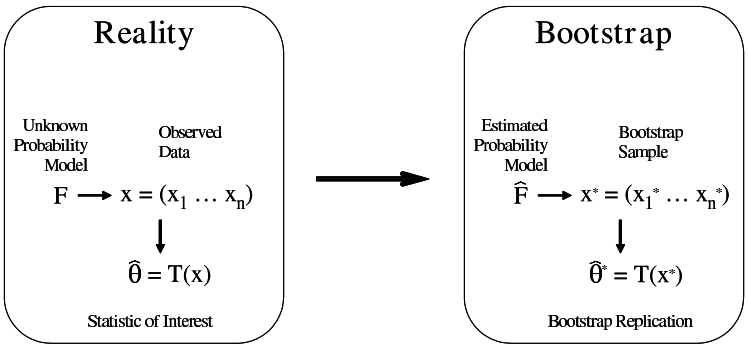

Figure 3: Schematic of the bootstrap philosophy recreated from page 87 of Efron and Tibshirani (1998). In order to know the properties of a test statistic when there is an unknown probability model, F, that generates the observed data, $\mathbf{x}$, resamples taken from the empirical distribution of the data gives replicates of the statistic that allow one to approximate its distribution.

biases in the statistic (Efron 1981). Improvements to these basic bootstrap intervals can be made by "double bootstrap methods" where the bias of using resamples of the observed data is measured by resampling a second time from the bootstrap replicates (Beran 1987). For all interval estimates that can be obtained from bootstrap methodologies, there exists a corresponding hypothesis test that looks for the inclusion of null value of the statistic, $\theta_0 = E_{H_0}[\hat{\theta}]$ in the interval. The proper coverage of any of these intervals may not be precise for small n because the discreteness of $\hat{F}$ might prevent it from being a good estimate of F, and smoothing methods may be employed to improve performance (Polansky and Schucany 1997).

Bootstrap algorithms have been proposed in the microarray literature as a means for the cross-validation of classification studies (Braga-neto and Dougherty 2004), and for

multiple testing issues with differential expression studies (Tsai et al. 2003); however, the effects of resampling in the high-dimensional space of microarrays must be carefully considered in any new application (Troendle et al. 2004).

2 Testing categories by structured permutation

2.1 Introduction

With the understanding of gene function now extending across much of the genome, investigators are increasingly turning from questions about individual genes to those about cellular processes involving groups of genes. In microarray experiments that measure the association between gene expression and some response of interest, this is translated into constructing hypothesis tests for the differential expression observed across any number of functional categories. As detailed in the previous chapter, several publications have proposed examining functional categories after a gene-by-gene analysis has been performed (Al-Shahrour et al. 2004; Beißbarth and Speed 2004; Berriz et al. 2003; Doniger et al. 2003; Draghici et al. 2003; Hosack et al. 2003; Kim and Falkow 2003; Zeeberg et al. 2003; Zhang et al. 2004; Zhong et al. 2004). Each of these methods tests for the over-representation of a functional category within a list of significant genes through use of the hypergeometric distribution (see equation 1.4) or an approximation thereof. However, there are several disadvantages in applying these methods to microarray data. First, they rely in an inherent way on the gene-specific analysis that generated the significant list, and are sensitive to the criteria used to determine the cutoff for inclusion in the list. Moreover, by merely testing over-representation these methods fail to consider a gene's relative position in (or out) of the ranked list. If genes belonging to a functional

category show a consistent but modest association to the response of interest, they may fail to reach the criteria for inclusion in the gene list when issues like multiple testing are accounted for. In this case, the accumulation of effects across a category would go unnoticed when examining only membership in the list. A much bigger concern with gene-list enrichment tests is that do not take into account the possible correlation among genes within and outside a category. For categories with highly correlated genes, the true Type I error will be substantially higher than the reported p-value, resulting in anti-conservative tests. These drawbacks suggest the importance in finding an improved method of testing gene categories. In the following chapter a framework is presented for testing the associations of a functional category of genes in a more valid manner.

2.2 The SAFE framework

In order to assess the differential expression of gene categories, we propose a flexible, permutation-based framework, termed SAFE (for Significance Analysis of Function and Expression). SAFE extends and builds on an approach first employed in Virtaneva et al. (2001) for a two-sample microarray comparison of cancer subtypes. More recently, a similar method was proposed for a comparison of diabetes subtypes (Mootha et al. 2003). A two-stage approach is employed to assess the significance of a gene category. First, gene-specific statistics are calculated that measure the association between expression and the response of interest. Hereafter, we will refer to these as *local* statistics. Then a larger-scale *global* statistic is constructed as a function of the local statistics, with the goal of detecting a shift within a gene category to more extreme values, as compared to all

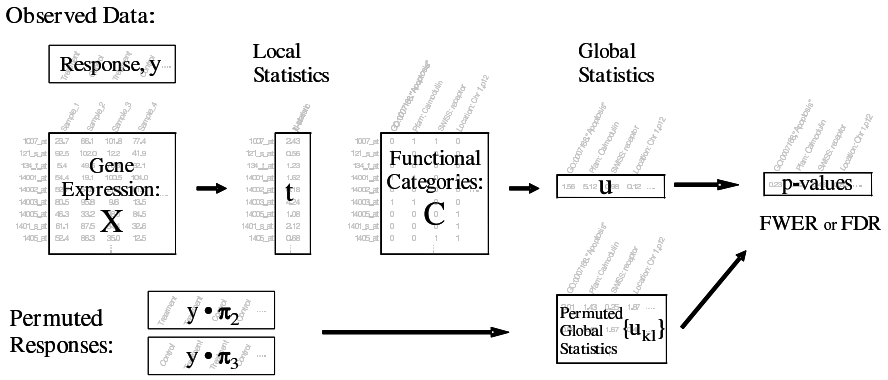

Figure 4: Schematic for the significance analysis of function and expression (SAFE). The observed data consist of a matrix of normalized expression estimates, $\mathbf{X}$, a response vector, $\mathbf{y}$, and gene category assignments defined *a priori* in a matrix $\mathbf{C}$. For the observed and permuted data, gene-specific local statistics and category-specific global statistics are computed such that p-values are obtained for each category along with estimated error rates.

other genes. The significance of the global statistics is assessed by repeatedly permuting the array assignments and recomputing local and global statistics. In this manner, the correlation between all genes is maintained by holding the gene expression data constant. Furthermore, the relationships among categories which contain overlapping genes will be preserved, which is important for multiple testing considerations.

The SAFE procedure is described in detail in the following sections and a schematic is provided in Figure 4. It generalizes and extends the method of Virtaneva et al. (2001) in two critical respects: 1) SAFE naturally encompasses a wide variety of experimental designs and response vectors, and 2) appropriate methods of error rate estimation can be applied directly in the permutation scheme. A series of informative plots are proposed for visualizing the differential expression seen within significant category.

2.2.1 The observed data

The following notation is introduced for describing DNA microarray data and gene category tests. In the following chapter we will demonstrate how this general form allows the variety gene category tests proposed in the literature to be presented in a unified way. Let the observed expression data for m genes and n samples be given by the matrix $\mathbf{x}$, where the expression of the i-th gene in the j-th sample is x_{ij}. For the expression values of the i-th gene, the row vector that corresponds is given as $\mathbf{x}_{i*}$, and for the j-th sample, the column vector is written as $\mathbf{x}_{*j}$. The term "gene" is used to generically identify a row of $\mathbf{x}$ but can also correspond to a probe or probeset for a transcript, depending on the array platform and pre-processing steps. Therefore, a single gene might be represented by different transcripts and appear as multiple rows of $\mathbf{x}$. Extensions of SAFE are proposed in Chapter 4 that would give an appropriate way to account for the multiple representation of a gene on an array. We will generally assume that suitable normalization and other data pre-processing steps as described in Chapter 1 (*cf.* Dudoit et al. (2002); Li and Wong (2001)) have been performed. The relevant sample information is represented by the response vector $\mathbf{y}$, where each element, y_j, can be a group assignment based on the experimental design or a continuous measure. For some experimental designs y_j may be more than a scalar value, as seen in the survival analysis performed in section 2.3.3.

Prior to SAFE analysis, a collection of functional categories of interest must be specified. When a total of L categories are under examination, the gene membership can be stored in a $m \times L$ matrix of indicators, where $c_{ih} = 1$ if gene i belongs to category h and $c_{ih} = 0$ otherwise. Thus, the data for a SAFE analysis is contained in the three objects,

$\mathbf{X}$, $\mathbf{y}$, and $\mathbf{C}$.

2.2.2 Statistics and permutation

Two statistics must be specified in SAFE: The first is based on the experimental design, and termed a "local" statistic $t = T(\mathbf{x}_{i*}, \mathbf{y})$, measuring the association between the expression profile of gene i and the response vector. In a study where $y_j \in \{0, 1\}$ denotes one of two experimental conditions, one might use either a t-statistic, a non-parametric statistic, or some other measure for comparing $\{x_{ij} : y_j = 0\}$ and $\{x_{ij} : y_j = 1\}$ (*e.g.* fold-change.) As genes in the same category might exhibit changes in either direction, a two-sided local statistic such as the absolute value of a t-statistic would be the natural choice in a exploratory analysis unless the underlying biological suggests a concerted direction of differential expression in a category of interest.

The global statistic assesses how the distribution of local statistics within a category differs from local statistics outside the category. For a given category, h, the statistic $u = U(t_1, \ldots, t_m;\ \mathbf{c}_l)$ measures some difference between the local statistics of genes within category, namely $\{t_i : c_{ih} = 1\}$, and the local statistics of genes in the complement of the category, namely $\{t_i : c_{ih} = 0\}$. Typically little is known about the joint density of the local statistics. For this reason we favor rank-invariant choices for U, such as the Wilcoxon rank sum (Virtaneva et al. 2001) as likely to retain reasonable power under a variety of experimental designs.

The significance of the global statistic for each functional category is assessed through a group $\Pi = \{\pi_1, \ldots, \pi_K\}$ of permissible permutations of the response vector. The per-

mutations in Π reflect the underlying experimental design, including pairing of samples, blocking, or other sampling-based constraints. For many experimental designs, all $n!$ permutations are permissible, although fewer equivalent permutations of the response vector may exist (as in the two-sample problem). For datasets of even modest size, it may not be computationally feasible to use all permutations, and the elements of Π are chosen as a random sample from all permissible permutations. The elements of Π can be represented as permutations of the integers $\{1, \ldots, n\}$, so that Π is stored an $n \times K$ matrix. We will restrict π_1 to be the identity permutation, corresponding to the observed order of the response vector.

For each gene and each permutation $\pi_k \in \Pi$, let $t_{ik} = T(\mathbf{x}_{i*}, \mathbf{y} \cdot \pi_k)$ be the value of T when the response is permuted according to π_k. Here $\mathbf{y} \cdot \pi = (y_{\pi(1)}, \ldots, y_{\pi(n)})$ is a re-ordering of the components of $\mathbf{y}$ according to π. Let $\mathbf{u}$ be the $K \times L$ matrix with entries u_{kh} for the h-th functional category under permutation π_k. Permutation-based p-values are computed for each category as $p_h = K^{-1} \sum_{k=1}^{K} I\{u_{kh} \geq u_{1h}\}$, with $I\{\cdot\}$ denoting the indicator function. By restricting π_1 in this manner, the empirical p-value will appropriately follow a discrete uniform distribution under permutation.

2.2.3 Error rate estimation and plots

As in gene-specific analyses of microarray data, it is important to correct for multiple testing when a set of gene categories are considered. In addition to computing empirical p-values as described above, the permutation scheme can also be used to compute resampling-based estimates of the FWER (Westfall and Young 1989) or the FDR (Storey

and Tibshirani 2003; Yekutieli and Benjamini 1999) for the set of categories that fall within a given rejection region. First, the matrix of global statistics is converted into a $K \times L$ matrix of empirical p-values with elements

$$p_{kl} = \frac{1}{K} \sum_{h=1}^{K} I\{u_{hl} \geq u_{kl}\} \tag{2.1}$$

In this way, every column, and thus every category, has empirical p-values that range from $\frac{1}{K}$ to 1. If we define a rejection region by the interval, $[0, p]$, the Westfall-Young estimate of the FWER can be written as

$$\widehat{FWER}_{WY}(p) = \max_{l:p_l \leq p} \left[\frac{1}{K} \sum_{k=1}^{K} I\left(\min_{h:p_h \geq p_l} p_{kh} \leq p_l \right) \right] \tag{2.2}$$

Thus each p-value that occurs in the rejection region (indexed by l in equation 2.2) is compared to the minimum permuted p-value of all categories less significant. Then, the maximum of these comparisons is taken as the FWER estimate as part of the step-down procedure.

To estimate the FDR through resampling, Yekutieli and Benjamini (1999) proposed the following statistic for a similarly defined rejection region.

$$\widehat{FDR}_{YB}(p) = \min_{l:p_l \geq p} \left[\frac{1}{K-1} \sum_{k=2}^{K} \left(\frac{\hat{V}_k(p_l)}{\hat{V}_k(p_l) + \hat{S}(p_l)} \right) \right] \tag{2.3}$$

The functions $\hat{V}_k(\cdot)$ and $\hat{S}_k(\cdot)$ correspond to estimates of the number of true and false positives as presented in Table 1, and are defined as $\hat{V}_k(p) = \sum_{l=1}^{L} I(p_{kl} \leq p)$ and $\hat{S}(p) = \hat{V}_1(p) - \frac{1}{K-1} \sum_{k=2}^{K} \sum_{l=1}^{L} I(p_{kl} \leq p)$. The minimum is taken among the categories less than or equal to rejection region as part of the step-up procedure common to FDR estimation and control.

Storey and Tibshirani (2003) has also proposed a resampling-based method for estimating the FDR. In addition to defining a rejection region, another region is required that is thought to contain almost entirely true null hypotheses, $[p_0, 1]$.

$$\widehat{pFDR}_{ST}(p) = \min_{l:p_l \geq p} \left[\frac{W_1(p_0) \cdot \frac{1}{K-1} \sum_{k=2}^{K} R_k(p_l)}{\frac{1}{K-1} \sum_{k=2}^{K} W_k(p_0) \cdot R_1(p_l)} \right] \tag{2.4}$$

where $R_k(p) = \sum_{l=1}^{L} I(p_{kl} \leq p)$ and $W_k(p) = \sum_{l=1}^{L} I(p_{kl} \geq p)$ also represent estimates of the corresponding unknown outcomes given in Table 1.

Non-resampling based error estimates, such as the traditional FDR step-up procedure by Benjamini and Hochberg (1995) and the basic q-value estimate (Storey and Tibshirani 2003) can be readily applied to $\{p_h\}$. However, these methods may be less appropriate for the unknown dependence among categories. Permutation enables control of multiple-testing error rates among correlated tests without the need to adopt overly conservative procedures (*e.g.* Benjamini and Yekutieli (2001)). Permutation-based control of the FWER exploits positive correlation among the global statistics for categories with overlapping genes, while a Bonferroni threshold in this case will be highly conservative. In our examples using the GO ontologies, the dependence between some categories (nodes) is very strong, as many related categories contain identical or nearly identical sets of genes.

In addition to a p-values and error rate estimates, the significance of each category can be presented in the form of a SAFE-plot. For category h, the SAFE-plot displays the empirical cumulative distribution function (eCDF) of the ranked local statistics $\{t_i : c_{ih} = 1\}$. A category that contains many genes that are more differentially expressed on average will have higher ranked local statistics, and therefore show a right-ward shift

in the eCDF from the diagonal. In cases where an absolute value is taken to create a two-sided local statistic, such as $|t|$ in the two-sample comparison, ranking genes by the untransformed statistic will reveal the directions of differential expression for individual genes in the category. Labeled tick marks along the top of the graph allow the investigator to observe the genes most responsible for a categories significance. When gene categories have additional structural relationship such as the hierarchy of GO ontologies, we find it is also useful to display the SAFE significance results within a graphical representation of the structure. For GO, SAFE results can be plotted across the directed acyclic graph to identify the relationships among significant categories.

2.3 Examples from a microarray dataset

To demonstrate the applicability and flexibility of SAFE, gene category analyses were conducted for several responses in a study of human lung carcinomas by Bhattacharjee et al. (2001). A total of 202 lung specimens were assayed with hgu95Av2 oligonucleotide arrays (Affymetrix, Santa Clara, CA). The data consisted of 16 normal tissues and 186 tumors, sub-classified as adenocarcinomas ($n = 139$), pulmonary carcinoids ($n = 20$), small-cell lung carcinomas ($n = 6$), and squamous cell lung carcinoma ($n = 21$). Additional clinical information, including survival times, were available for 125 of the adenocarcinomas. Our significance analyses focused on three comparisons: (1) a two-sample comparison of normal versus cancerous samples; (2) an ANOVA model comparing cancer subtypes; and (3) a survival analysis within the adenocarcinoma subgroup.

CEL files for the 202 hgu95Av2 arrays were obtained from http://www.pnas.org and

expression estimates were obtained from the dChip v1.3 software from Li and Wong (2001). In keeping with the terminology above, each hgu95Av2 probeset is referred to as a "gene" even though in many cases multiple probesets are known to correspond to the same gene. Arrays were normalized by quadratic scaling to an artificial array of median expressions for each gene (Yoon et al. 2002). Genes were filtered out when called absent by the Affymetrix MAS5.0 algorithm in more than half the samples of every tissue type. These preprocessing steps resulting in expression estimates for 202 microarrays and 7299 genes.

Each SAFE analysis involved a common set of functional categories derived from GO and Pfam. Annotations for the hgu95Av2 array are available in the NetAffx (Liu et al. 2003) format from http://www.affymetrix.com. GO gene categories sets were generated from the hierarchical structure of an ontology in the standard manner (Ashburner et al. 2000; Beißbarth and Speed 2004; Zeeberg et al. 2003), using simple algorithms to create the $\mathbf{C}$ matrix of indicators required for the SAFE analysis. The 7299 expressed genes had a total of 3860 GO nodes and 1811 Pfam domains linked to them. In order to retain power in this example, only categories of a sufficient size are considered: including 120 cellular component nodes having at least 10 expressed genes, and 207 biological process nodes and 132 molecular function nodes having at least 40 expressed genes. Pfam gene categories were limited to the 176 domains annotated to at least 10 expressed genes.

For each response vector, an appropriate local statistic was chosen, the Wilcoxon rank sum was used as the global statistic,

$$u = \sum_{i=1}^{m} c_i \cdot \mathrm{Rank}(t_i) \tag{2.5}$$

and $K = 10,000$ permutations was randomly generated for the set of arrays corresponding to the response of interest. Permutation p-values were calculated for each category, along with the Westfall-Young FWER estimate and the Benjamini-Yekutieli FDR estimate (Westfall and Young 1989; Yekutieli and Benjamini 1999). All significant categories in the rejection region with an estimated FDR ≤ 0.10 are reported in Table 2. This demonstrates that significant results are achievable in SAFE, even when explicitly accounting for multiplicity of tests far greater in number than previous reports have considered (Berriz et al. 2003; Mootha et al. 2003; Zeeberg et al. 2003; Zhong et al. 2004).

Table 2: Significant GO and Pfam gene categories for the three comparisons in the Bhattacharjee et al. (2001) lung carcinoma study. For each response, the largest subset of all categories with a FDR ≤ 0.1 is reported along with the corresponding FWER estimates.

Category ID and Name	**Size**	p**-value**	$\widehat{FDR}$	$\widehat{FWER}$
Normal versus Cancer				
GO:0016460, 'Myosin II'	10	0.0004	0.066	0.157
GO:0000786, 'Nucleosome'	19	0.0004	0.066	0.157
Pfam:PMP22_Claudin	11	0.0005	0.066	0.188
ANOVA among subtypes				
GO:0007010, 'Cytoskeleton org. and biogen.'	128	0.0003	0.064	0.125
GO:0007017, 'Microtubule-based process'	67	0.0005	0.064	0.194
GO:0006996, 'Organelle org. and biogen.'	153	0.0005	0.064	0.194
GO:0016043, 'Cell org. and biogenesis'	283	0.0007	0.064	0.253
GO:0009117, 'Nucleotide metabolism'	82	0.0007	0.064	0.253
GO:0007028, 'Cytoplasm org. and biogen.'	175	0.0011	0.087	0.358
GO:0006164, 'Purine nucleotide biosynth.'	45	0.0016	0.099	0.459
Survival of adenocarcinomas				
GO:0005643, 'Nuclear pore'	30	0.0002	0.034	0.084
GO:0046930, 'Pore complex'	30	0.0002	0.034	0.084

2.3.1 Two-sample comparison

As a first examination of differential expression, a two-sample comparison was made between normal and tumor samples using the absolute value of the Welch t-statistic as the local statistic. Using the SAFE nomenclature, where $y_j = 1$ if the array corresponded to a tumor sample, and $y_j = 0$ if it corresponded to a normal sample, the local statistic for the i-th gene is written as

$$t_i = \frac{|\bar{x}_{i,1} - \bar{x}_{i,0}|}{\sqrt{\frac{s^2_{i,1}}{n_1} + \frac{s^2_{i,0}}{n_0}}} \tag{2.6}$$

where $n_c = \sum_{j=1}^{n} I(y_j = c)$, $\bar{x}_{i,c} = \frac{1}{n_c} \sum_{j=1}^{n} x_{ij} \cdot I(y_j = c)$, and $s^2_{i,c} = \frac{1}{n_c - 1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i,c})^2 \cdot I(y_j = c)$ $c = 0, 1$. Observed values ranged from very close to 0 to 18.4. Under 10,000 permutations of the array assignments, 1235 genes (17% of all tests) achieved a minimum empirical p-value 0.0001. With such dramatic differences between normal and tumor tissue producing a long list of differentially expressed genes, obtaining useful biological conclusions requires a broader perspective.

Among the 635 functional categories we considered in SAFE, three categories had $p \leq 0.0005$ and met the criteria for inclusion in Table 2: the cellular component nodes, GO:001640 'Myosin II' and GO:0000786 'Nucleosome', and also the Pfam domain 'PMP22 Claudin.' SAFE-plots display the relative extent and direction of differential expression observed for the sets of genes in these categories (Figure 5). Of the 10 expressed genes annotated to 'Myosin II,' 9 were substantially under-expressed in the tumor samples compared to normal ($p = 0.0004$). In contrast, the GO term 'Nucleosome' had 16 of 19 genes over-expressed in the tumor samples ($p = 0.0004$). Of the 11 genes annotated to 'PMP22 Claudin,' 4 were substantially over-expressed in cancer and 6 were substantially

under-expressed, ($p = 0.0005$).

These results demonstrate the various directions of differential expression that can be detected in a two-sample SAFE analysis. Since no overlap in gene membership occurs among the three categories, they can be separate findings. Several of the genes that are present in these categories have been associated with other forms of cancer, however the SAFE results suggest that families related genes may be dis-regulated in cancer.

The roles of myosin-related and cell-motility genes have long been studied in cancer and metastasis. A novel myosin family gene, *MYO18B*, was recently shown to be inactivated in approximately 50% of lung cancers (Nishioka et al. 2002). The nucleosome genes we observed to be overexpressed in cancer were primarily histone family genes; acetylation of histones has been linked to *MYO18B* inactivation and lung cancer (Tani et al. 2004). Over of Claudin-4, as observed here, has been linked to metastatic breast and pancreatic cancers (Michl et al. 2003; Nichols et al. 2004). By examining entire gene categories instead of individual genes, we are able to identify a manageable number of gene categories warranting further hypothesis and study.

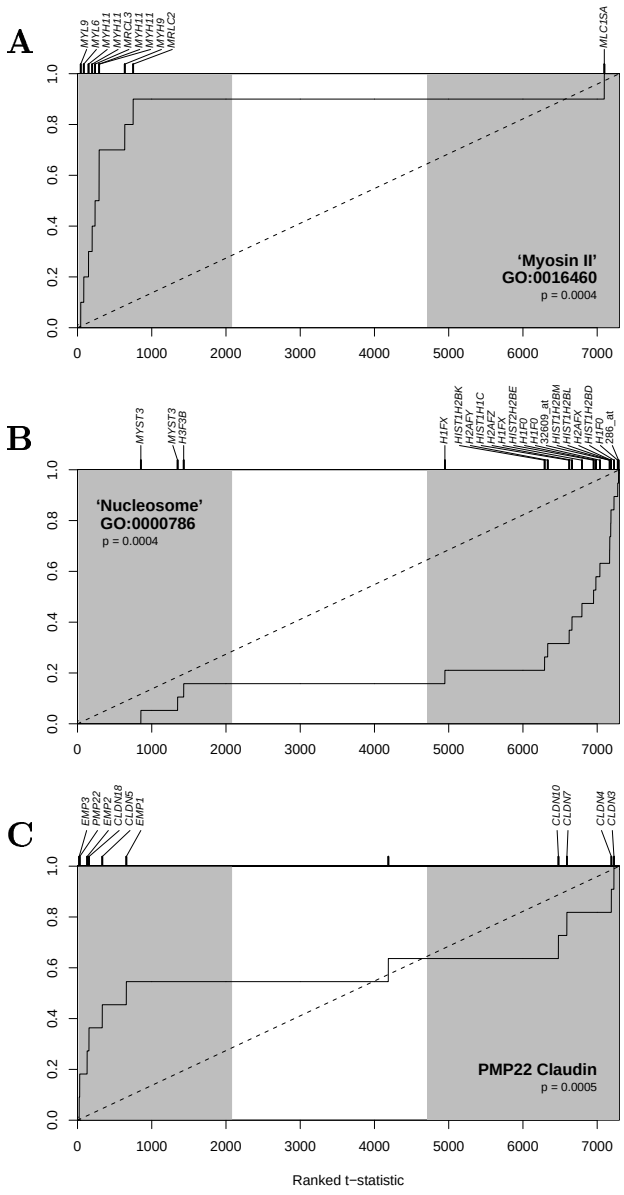

Figure 5: SAFE-plots for significant categories in normal versus tumor. Welch t-statistics were computed for all expressed genes. The shaded region represents the 5% tail area of the empirically derived null ($|t| > 2.26$). The empirical CDF for a gene category is plotted (solid line) against the ranks of all genes (dashed line). Tick marks above each plot display the location of genes within a category. Several genes are represented by more than one hgu95Av2 probeset. Significant gene categories can show consistent (A) under, or (B) over in tumor versus normal, or (C) bidirectional differential expression.

2.3.2 ANOVA

To look for differences in gene expression among the four cancer subtypes, the standard ANOVA F-statistic was used. A scaled F-statistic can be defined in the SAFE nomenclature using $y_j \in \{1, 2, 3, 4\}$ for each of the four tumor classifications

$$t_i = \frac{\sum_{c=1}^{4} n_c(\bar{x}_{i,c} - \bar{\bar{x}}_i)^2}{(\sum_{j=1}^{n}(x_{ij} - \bar{\bar{x}}_i)^2 - \sum_{c=1}^{4} n_c(\bar{x}_{i,c} - \bar{\bar{x}}_i)^2)} \tag{2.7}$$

where $n_c = \sum_{j=1}^{n} I(y_j = c)$, $\quad \bar{x}_{i,c} = \frac{1}{n_c}\sum_{j=1}^{n} x_{ij} \cdot I(y_j = c) \quad c = 1, 2, 3, \text{and } 4$, and $\bar{\bar{x}}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$. For a total of 2689 genes (37% of all tests) the observed local statistic achieved the minimum empirical p-value ($p = 0.0001$). The substantial differences in expression profiles between cancer subtypes provided the basis for successful discrimination in the original report (Bhattacharjee et al. 2001). Here we employ SAFE to establish which functional categories consistently differ in expression across cancer subtypes.

Eight biological process nodes (having p-values ≤ 0.0019) met the criterion of $\widehat{FDR} \leq 0.1$ for inclusion in Table 2. By viewing the location of the significant categories in the hierarchical structure of the ontology (Figure 6) it is apparent that they fall into two distinct families: 'Cell organization and biogenesis' (GO:0009117), and 'Nucleotide metabolism' (GO:0016043). The plot also illustrates that a broader category can be more significant than any of the nodes beneath it, due to the aggregation of gene effects across different descendants. These results add biological interpretability to the cluster analyses and gene-specific analyses from the original report.

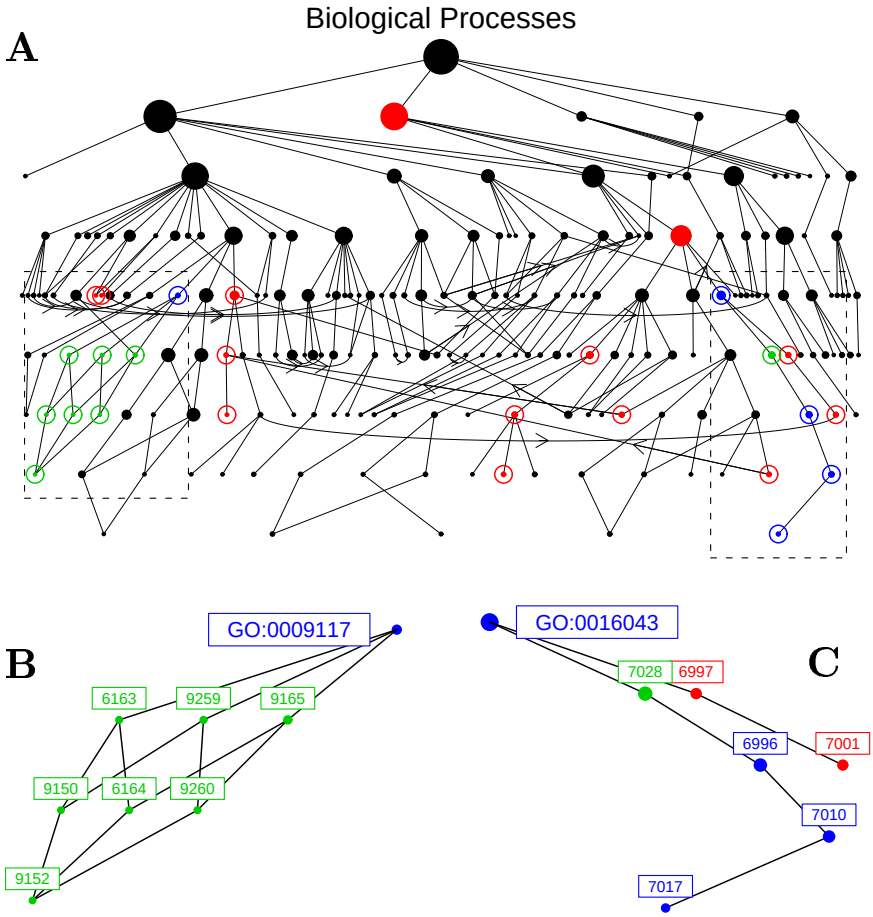

Figure 6: SAFE results displayed across the DAG structure for Biological Process nodes. In Gene Ontology, nodes can have multiple parents, and for lateral or upward edges, arrows are drawn from parent to child to indicate the lineage. The area of each node is proportional to the number of genes in the corresponding category. Nodes are colored by statistical significance: blue ($p < 0.001$) green ($0.001 \leq p < 0.01$), or red ($0.01 \leq p < 0.1$). Two distinct sub-graphs containing all significant nodes (blue or green) are expanded in the figure: (B) nodes under 'Nucleotide metabolism' GO:0009117 and (C) nodes under 'Cell organization and biogenesis' GO:0016043.

2.3.3 Survival analysis

Censored survival data were available for 125 subjects with adenocarcinomas, with 71 observed deaths and 54 censored observations. The association between a gene's expression and survival was assessed with a univariate Cox proportional hazard model (Cox 1972). Let y_j represent the censored failure time for the j-th array, using the pair of values $y_j = \{t_j, d_j\}$ with t_j measuring the time to event and letting $d_j = 1$ if a death occurred, and $d_j = 0$ if the corresponding subject was censored. In the Cox model, regression coefficients are estimated by the maximum of the partial likelihood

$$\hat{\beta}_i = \sup_{\beta_i} L(\beta_i) = \sup_{\beta_i} \prod_{j=1}^{n} \frac{d_j \cdot \exp(x_{ij} \cdot \beta_i)}{\sum_{r \in Risk(t_j)} \exp(x_{ir} \cdot \beta_i)} \tag{2.8}$$

where $Risk(t_j)$ refers to the riskset for that time consisting of all subjects for whom a death or censored outcome had not yet been observed. Although $\hat{\beta}_i$ does not have a closed form, the log likelihood is strictly concave, and can thus be solved quickly for all genes using Newton-Raphson iteration or a bisection algorithms.

The local statistic is the Wald-type statistic

$$t_i = \frac{|\hat{\beta}_i|}{\widehat{se}(\hat{\beta}_i)} \tag{2.9}$$

where the standard error of the regression estimate is approximated by the observed information of the partial likelihood

$$\widehat{se}(\hat{\beta}_i) = \left(\frac{-\partial^2}{\partial \beta_i^2} \log L(\beta_i)_{|\beta_i = \hat{\beta}_i} \right)^{-\frac{1}{2}} \tag{2.10}$$

The resulting Z-like statistics ranged from 0 to 3.98. While 496 expressed genes had a gene-specific p-value less than 0.05 ($|z| \geq 1.96$), none was significant after multiple-testing

correction (all the common FDR and FWER estimates presented in this proposal were greater than 0.2). The data provide an example where standard gene-specific approaches fail to provide useful conclusions because no effects are strong enough to pass the multiple-testing criterion. We then applied the SAFE approach, which is sensitive to the aggregate effect of genes with related biological functions.

After accounting for multiple testing, two related GO cellular component nodes were significant (Table 2): GO:0005643 'Nuclear pore,' and GO:0046930 'Pore complex.' However, the nodes for 'Nuclear pore' and 'Pore complex' contain an identical set of 30 genes and should be considered a single finding ($p = 0.0002$). Likewise, the parental node, 'Nuclear membrane,' was marginally significant ($p = 0.0012, \widehat{FDR} = 0.106$) but shared 30 of 51 genes with the other nodes. An additional SAFE-plot for the genes unique to 'Nuclear membrane' (not shown) indicates that only the nuclear pore genes are associated with survival.

Although the original report (Bhattacharjee et al. 2001) found a relationship between survival and a cluster-defined adenocarcinoma subclass ($p = 0.005$), this result is stronger, remarkably specific in its biological implications, and offers new directions for exploration in the biology of cancer progression and survival. We note that the role of nuclear transport in cancer (Kau et al. 2004) and cancer aggressiveness (Agudo et al. 2004) has been the subject of recent attention in the literature.

2.4 Discussion

These examples demonstrate the applicability of SAFE to a variety of experimental designs and measures of gene-specific differential expression. It is further observed that significant categories can be found both when many gene-specific associations are observed across the array or with few significant genes.

Although both SAFE and the *gene-set enrichment analysis* (GSEA) proposed by Mootha et al. (2003) are two-stage procedures that employ array permutation, there are two distinctions to be made. First, GSEA uses a Kolmogorov-Smirnov type global statistic that looks for any general difference between the empirical CDFs of category and complement local statistics. In doing so, this method has been criticized for being sensitive to departures from the null that do not necessarily reflect increased association of expression and response values in the category (Damian and Gorfine 2004). For instance, a category containing local statistics that are very non-significant but similar in magnitude (*e.g.,* t-statistics all close to 0 in a two-sample experiment) will also be rejected by GSEA.

Secondly, we note that SAFE calculates permutation-based p-values using a separate null permutation distribution for each category (*i.e.*, column of $\mathbf{u}$; equation 2.1), rather than pooling all the values in $\mathbf{u}$ into a single null distribution. In contrast, GSEA uses pooling to compute a FWER-adjusted p-value for the largest Kolmogorov-Smirnov statistic, after scaling the statistics based upon differing category sizes. However, such standardization methods ignore the unknown correlation among local statistics and can therefore produce unequal null distributions among the categories. The inadequate stan-

dardization of global statistics provides a strong rationale against pooling in SAFE. Examining the permutation distributions of several exemplary Wilcoxon statistics that have been standardized (Figure 7) demonstrates the instances in the example data where the global statistics remain improperly scaled. In this circumstance, a p-value generated from the pooled null distribution will not control the Type I error of a given category properly, and can differ from the nominal p-value by a factor of 10 or more. Although pooling within SAFE meets the technical requirements for weak control of the FWER (Westfall and Young 1989), inadequate standardization will reduce power for most categories.

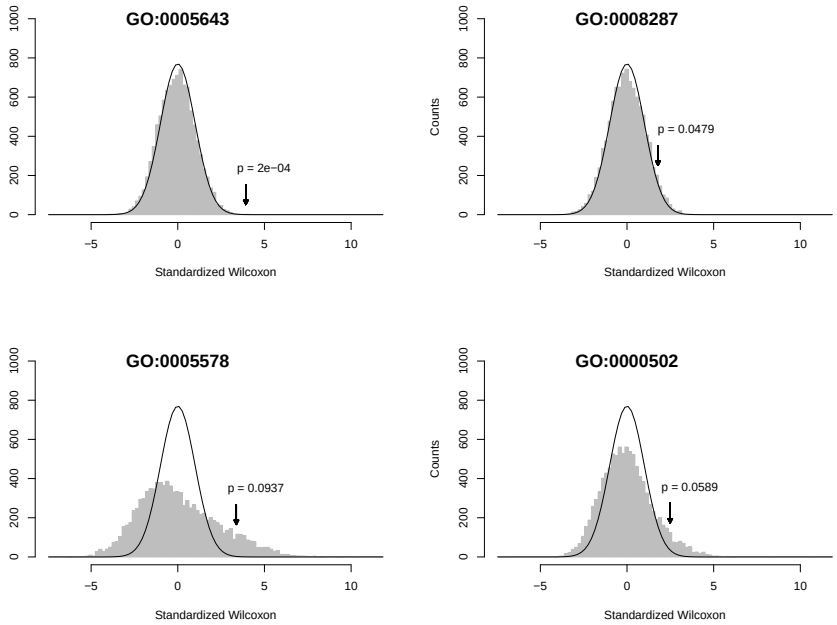

Figure 7: Null distributions for standardized global statistics for 4 GO cellular component categories in the analysis of survival among adenocarcinomas. A scaled normal density is overlaid on each histogram as the asymptotic distribution of a standardized Wilcoxon. Empirical p-values were calculated for the observed statistic relative to the permuted nulls for each category, and are invariant to standardization of the Wilcoxon statistics. The upper panels show good agreement between the theoretical and empirical null distributions for (A) the most significant category and (B) a marginally significant category. The lower panels (C) and (D) display poorly standardized statistics that have greater variance than the theoretical distribution, and would thus have inflated Type I errors for pooling-based p-values ($p = 0.0133$ and $p = 0.0352$, respectively).

3 A comparison of gene category tests

3.1 Introduction

Beginning with Virtaneva et al. (2001), a number of publications have proposed tests for assessing the association between response and gene categories. The most commonly employed tests are designed to begin with a list of significant genes. A secondary analysis then looks for over-representation, or *enrichment*, of genes within the category on the gene-list, using Fisher's Exact Test or other tests of 2 x 2 contingency tables (see Section 1.4 for a list of methods and softwares). Other approaches examine the significance of genes using more direct comparisons of the gene-specific measures of DE, thereby avoiding any need for intervening gene lists. In these methods, tests are constructed either for an average difference of gene-specific statistics (Boorsma et al. 2005; Kim and Volsky 2005), or using classical rank-based procedures for two-sample comparisons (Barry et al. 2005; Ben-shaul et al. 2005; Mootha et al. 2003). Herein we describe how the existing gene category methods can also be broadly sorted according to the null hypotheses they test against.

Gene category testing is now widely performed in a range of fields and the results of such analyses are frequently reported without independent verification. However, we argue that category testing has not yet been placed on firm statistical foundation. As pointed out in a recent review by Allison et al. (2006), even fundamental issues such as a

formal definition of the underlying null hypothesis and a proper demonstration of Type I error have not been provided for many of the various methods in the literature.

3.1.1 Contributions

In this chapter we provide a careful and rigorous analysis of gene category testing by first defining a general framework. Presenting and contrasting existing methods in this manner allows us to identify two distinct classes that are defined by the null hypotheses they assume or induce. Several shortcomings of these methods are revealed through derivation and through simulations from an example dataset. We then propose an alternative approach to hypothesis testing that can overcome these shortcomings so that it provides more power while demonstrating proper coverage under the null hypothesis. This approach can also be applied to a wider set of experimental designs.

As a first class of gene category test, many methods implicitly assume the gene-specific test statistics are independent and identically distributed (*i.i.d.*). However, a casual inspection of the microarray literature indicates that the assumption of independence is violated in the vast majority of cases. In simulations generated from real microarray data, we illustrate how correlation in expression causes these methods to be extremely anti-conservative, leading to a large number of false discoveries. As another approach to category testing, permutation of the expression data has been proposed as a means of inducing suitable null hypotheses. In these methods, the choice of sampling unit greatly influences the outcome (Breslin et al. 2004). We describe how permutation of gene assignments merely induces the same null in class 1 tests of there being *i.i.d.* gene-

specific statistics. Conversely, array permutation methods constitute a second class of gene category tests, having been proposed with the stated intention of preserving the correlation in expression observed among genes (Barry et al. 2005; Mootha et al. 2003). We next define an important property of gene-specific statistics that is necessary for proper coverage under array permutation. When this property is met, the induced null hypothesis is that gene-specific test statistics are dependent, yet approximately identically distributed according to no association with the response. Gene category methods that rely on this null are shown to provide better coverage in simulated data.

In defining these two classes of tests, we propose that a broader null hypothesis is warranted for gene categories tests, allowing for both dependent coexpression of genes and also varied degrees of association between gene expression and response. Interestingly, array permutation approaches can be quite conservative under certain forms of this null. The conservativeness can be explained in part through an analytical argument which shows that the maximum variance of the category-wide test statistic occurs under the special case induced by array permutation. We present a simple and powerful bootstrap-based approach that allows for the more general null hypothesis to be tested. Finally, we demonstrate the utility of this new method in a breast cancer dataset, and discuss several other advantages that the bootstrap-based tests have over array permutation procedures.

3.2 A general framework for gene category tests

3.2.1 Notation and framework

To describe the variety of gene category tests in a unified way, we will continue to refer to the observed expression data as $\mathbf{x}$, and the response as $\mathbf{y}$ as in Section 2.2.1. When we regard an unrealized expression matrix as a collection of random variables, we will use uppercase versions of the standard notation, *i.e.*, $\mathbf{X}$, X_{ij}, $\mathbf{X}_{i*}$ and $\mathbf{X}_{*j}$.

To more easily derive the properties of tests, a single gene category will be represented by a subset $C \subseteq \{1, \ldots, m\}$ such that $i \in C$ if and only if gene i is a member of the category. The size of a category C will be denoted by $m_C = \sum_{i=1}^{m} I(i \in C)$. For any category C, the complementary set of genes will be denoted by $\bar{C}$ and be of size $m_{\bar{C}} = m - m_C$.

We also adopt the terminology in the previous chapter, where hypothesis tests of gene categories can be viewed as a two-stage procedure (see Box 1). In the first stage, a *local statistic* measures the association between the expression profile of each gene and the response. We denote the local statistic of gene i by $T_i = T(\mathbf{X}_{i*}, \mathbf{y})$ and let t_i be the corresponding value from observed data. The function $T(\cdot)$ is typically chosen in accordance with the experimental design and scientific goal of the study. In a two-condition experiment, one could use a t-statistic or average fold change, while in more complex experimental designs, for example censored time-to-event data, a local statistic derived from the Cox proportional hazard model may be used to test for an association between gene expression and patient outcome. For many common experiments, T will estimate or be related to an gene-specific parameter that captures the association between

response and expression. In the example local statistics for a two-condition experiment that are given above, the related parameters would be a scaled difference and a ratio of population means, respectively. Properties of local statistics are examined more fully in Section 3.5.3.

In the second stage of a gene category test, a *global statistic* is used to compare the local statistics of genes within a category C to those in the complement. We denote the global statistic for category C by $U = U(T_1, \ldots, T_m : C)$, and in the following sections describe the functional forms of $U(\cdot)$ that relate to methods of testing gene categories that have already been proposed. Existing methods focus on either detecting a difference in the proportion of genes called significant, or detecting a shift in the average local statistic within the category versus its complement. Through describing the global statistics these methods employ, and the way in which p-values are obtained, it can be seen that two distinct classes of gene category tests exist. These classes are defined by the underlying null hypotheses that are either assumed or induced by resampling-based procedures.

Box 1: Common elements of gene category tests

Gene category tests are typically two-stage procedures requiring the following statistics:

- A *local statistic* that measures the association between response (*e.g.* experimental condition) and expression of each gene.
- A *global statistic* that compares the local statistics within a category to those of its complement.

Two classes of hypothesis tests are typically designed for each global statistic:

1. Parametric or rank-based procedures that assume independent and identically distributed local statistics, or gene permutation methods that induce the same null.
2. Array permutation methods which induce a null that maintains the correlation structure among genes while removing all associations to the response.

Error rate controlling or estimating procedures address the multiple comparisons involved in simultaneously testing a number of different gene categories.

3.3 A Survey of gene category test statistics

Gene category test statistics can be generally be represented as looking for a change in the DE of genes within a category relative to the genes in its complement. In a number of the gene category publications, hypothesis tests are designed from traditional methods for comparing two random samples of data. In these proposals, though, we note that the null hypothesis has not be explicitly defined, and it is rarely discussed whether the necessary assumptions are met in gene expression data. In the following section, we will demonstrate that a particular null hypothesis is assumed by a variety of gene category tests. The tests that fall into this class vary in terms of the global statistics that are chosen, and whether exact or approximate distributions are used to determine p-values, but can be collectively stated as follows.

Definition 1. *Class 1 gene category tests are defined by the assumed or induced null hypothesis. For local statistics $T_1, \ldots, T_m$, the null can be stated as*

$$H_0: \ T_1, \ T_2, \ \ldots \ , \ T_m \ \text{ are } i.i.d \text{ with } \ T_i \sim F \tag{3.1}$$

where F can take any general form, but is typically thought to correspond to there being no association between expression and the response of interest.

The global statistics that have been proposed can be classified as "categorical" when a list of significant genes has been previously identified by a gene-specific analysis, and "continuous" when a more direct measure of DE is available for each gene. Two global

statistics are presented below for each case, and a brief description is given of the corresponding non-resampling based Class 1 tests. We will focus on one-sided forms of the tests because in most applications one is only interested in categories showing increased association with the response relative to what is seen across the array.

3.3.1 A survey of the global test statistics

Categorical Test Statistics

Gene-list enrichment methods have developed as a *post hoc* means of testing a category once genes with significant amounts of DE have been identified. Let R denote the rejection region for the local statistics that produces the significant gene list. R might be determined independently from the data, (*e.g.,* from quantiles of a central t-distribution), or in a data-dependent manner (*e.g.,* from methods to control the FWER or FDR for the multiple comparison of m genes).

Gene-list enrichment tests only consider the dichotomous outcomes of the m gene-specific hypothesis tests, $I\{T_i \in R\}$. The differential expression within C and $\bar{C}$ is therefore summarized by a 2×2 contingency table (Figure 8).

The traditional contingency table tests that have been proposed for gene category analysis include the χ^2 test of homogeneity, Fisher's Exact test, and minor variants of these. In the classical derivation of these tests, the Bernoulli variables $I\{T_1 \in R\}, \ldots, I\{T_m \in R\}$ are assumed to be independent with the probabilities of rejection $P(T_i \in R) = \pi_C$ for $i \in C$ and $P(T_i \in R) = \pi_{\bar{C}}$ for $i \in \bar{C}$, respectively. The tests then

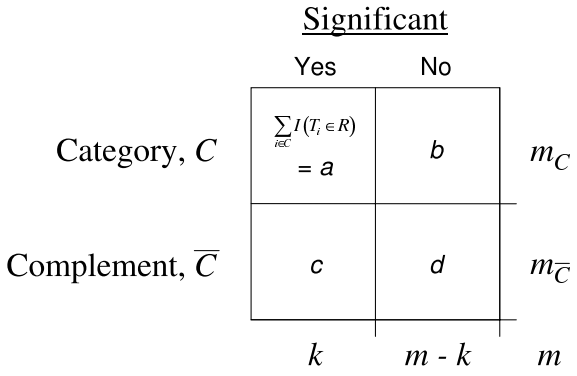

Figure 8: The results from a gene-specific analysis as given in a 2×2 table for a category versus its complement. The size of the two gene sets, given by m_C and $m_{\bar{C}}$ respectively, are assumed to be fixed quantities. The complete table can then be determined by knowing the number of rejections in the category and either the total number of rejections, k, or the number of rejections in the complement.

look for departures from $\pi_{\bar{C}} = \pi_C$, where all indicators would be *i.i.d.* It is worthwhile to note that the Class 1 null (3.1) is sufficient but not necessary for the dichotomous outcomes to be *i.i.d.* under a given R. However, (3.1) guarantees the categorical null holds for any possible choice of rejection region.

In several of the gene-list enrichment software packages the χ^2 test of homogeneity is proposed as an approximate test for large categories (Beißbarth and Speed 2004; Draghici et al. 2003). The one-sided version of this test is equivalent to the difference in proportions test proposed originally by Pearson (1911), where the global statistic can be written as

$$U_P = \hat{\pi}_C - \hat{\pi}_{\bar{C}} = \frac{1}{m_C} \sum_{i \in C} I\{T_i \in R\} - \frac{1}{m_{\bar{C}}} \sum_{i' \in \bar{C}} I\{T_{i'} \in R\}. \tag{3.2}$$

By the central limit theorem, the two proportions are asymptotically Gaussian for large

m_C and $m_{\bar{C}}$, and a Z-test is performed on a standardized form of U_P.

Fisher's Exact Test is more commonly applied in gene-list methods, and is noted to be a conditional test based on the total number of rejected hypotheses, $K = \sum_{i=1}^{m} I\{T_i \in R\}$. Once K is established, the global statistic can be represented as the number of genes in the category that are rejected,

$$U_F = \sum_{i \in C} I\{T_i \in R\} \tag{3.3}$$

and an exact one-sided p-value is obtained from the hypergeometric distribution. This p-value is conditional on K, and using it in an unconditional hypothesis test will lead to slightly conservative results, particularly when the category is small (Yates 1984). Depending on how the gene-list is determined, it is not always clear whether it is appropriate to condition on K, but exact tests are often favored in order to handle small categories. For moderately sized categories, we note there will be little difference between the exact conditional and approximate unconditional tests.

Continuous Test Statistics

It is also possible to directly compare the associations of expression to response without first using a gene-specific test to dichotomize the local statistics. Several of the more recently proposed gene category tests are designed in this manner. In particular, if one is interested in the average amount of DE seen in C relative to that of $\bar{C}$ than a straight forward global statistic for this comparison is the average difference in local statistics

between the two sets.

$$U_D = \frac{1}{m_C} \sum_{i \in C} T_i - \frac{1}{m_{\bar{C}}} \sum_{i' \in \bar{C}} T_{i'}. \tag{3.4}$$

Hypothesis tests of U_D have been proposed by two similar methods. In one, a t-test is performed after standardizing by the pooled sample variance of local statistics (Boorsma et al. 2005), while in the second method a Z-test is done after U_D is scaled by the overall standard deviation in local statistics (Kim and Volsky 2005). For a typical category where $m_C \ll m$, the variance estimates in both methods will be reasonable close, and similar results will be obtained because of the large number of degrees of freedom of the t-distribution ($df = m - 2$).

When using the global statistic in (3.4), the results will be sensitive to the chosen form of the local statistics (*e.g.,* deciding between a t-statistic or its corresponding p-value), and may not be robust to skew or outlying observations. Rank-based global statistics avoid both of these shortcomings, as they are invariant to monotone transformations of the local statistics. The Wilcoxon rank sum test has been implemented in its classical form in the software GOStat (Beißbarth and Speed 2004). In the absence of ties the global statistic is written as

$$U_W \;=\; \sum_{i \in C} \mathrm{Rank}(T_i) \tag{3.5}$$

Under the Class 1 null hypothesis, the discrete CDF of U_W is known once m_C and $m_{\bar{C}}$ are specified. In this case a hypothesis test can be implemented using tables of exact p-values, or through a Z-test based on a standardized form U_W that under independence will be asymptotically correct for large categories.

The rank-based Kolmogorov-Smirnov test has also been implemented in a gene cat-

egory test which can also be characterized as testing against the class 1 null (Ben-shaul et al. 2005). However, the Kolmogorov-Smirnov type statistic has been criticized in gene category testing for being sensitive to departures that do not necessarily reflect increased amount of DE in the category (Damian and Gorfine 2004); for example, a category with no DE but with local statistics that all happen to be very close to one another would be identified as significant by these tests. For this reason, we will restrict our focus to the tests of average differences when considering continuous global statistics.

3.4 The effect of correlation on Class 1 tests

In this section we more closely examine the assumption of independent local statistics, and its failure to hold in gene expression data. First, correlation in expression is defined and related to correlation in local statistics. Decompositions of the variances of global statistics demonstrate the effect this dependency has on Class 1 hypothesis tests. A simulation study based on real microarray data exhibits the extreme anti-conservative behavior of these tests in the presence of realistic levels of correlation in expression.

3.4.1 Correlations in expression and local statistics

Let $\rho^X_{i,i'} = \text{Corr}(X_{ij}, X_{i'j})$ be the population correlation between genes i and i'. For experimental designs with independent arrays, a natural estimate of $\rho^X_{i,i'}$ is the observed Pearson sample correlation coefficient

$$r_{i,i'} = \frac{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 \cdot \sum_{j=1}^{n}(x_{i'j} - \bar{x}_{i'})^2}} \tag{3.6}$$

where $\bar{x}_i = n^{-1} \sum_{j=1}^{n} x_{ij}$.

The distributions of global statistics under the Class 1 null hypothesis are noted to be more directly effected by the correlation between local statistics, namely $\rho^T_{i,i'} = \text{Corr}(T_i, T_{i'})$. In the special case that T takes a linear form $T(\mathbf{X}_{i*}, \mathbf{y}) = \sum_{j=1}^{n} a(y_j) \cdot X_{ij}$ for some function $a(\cdot)$, a simple calculation shows that $\rho^T_{i,i'} = \rho^X_{i,i'}$. An example of a linear local statistic would be an unscaled difference in sample means, *e.g.*, fold change on the log-scale; this choice of local statistic is appropriate if the logarithm is a variance-stabilizing transformation of expression data.

In general, the relationship between the correlations $\rho^X_{i,i'}$ and $\rho^T_{i,i'}$ does not have a closed analytic form, although it can often be shown numerically to be monotone and quite linear. Monte Carlo simulations of gene expression data demonstrate this linear relationship holds in several standard experimental designs and corresponding measures of DE including t-statistics for two-condition studies and for simple linear regressions (Figure 9). When linearity holds, (3.6) is also a good estimate of $\rho^T_{i,i'}$.

3.4.2 Correlation and Variance Inflation

The effect that the $\frac{m \cdot (m-1)}{2}$ pairwise correlations will have on some of the Class 1 gene category tests can be seen by expanding the variances of particular global statistics. Here, we derive the true variances of the continuous global statistics, U_D and U_W, and show how they are greater than what occurs under the *i.i.d.* assumption when categories have positively correlated gene members. For the categorical global statistics, U_F and U_P, the variance is also inflated in the presence of positively correlated categories, but is not as

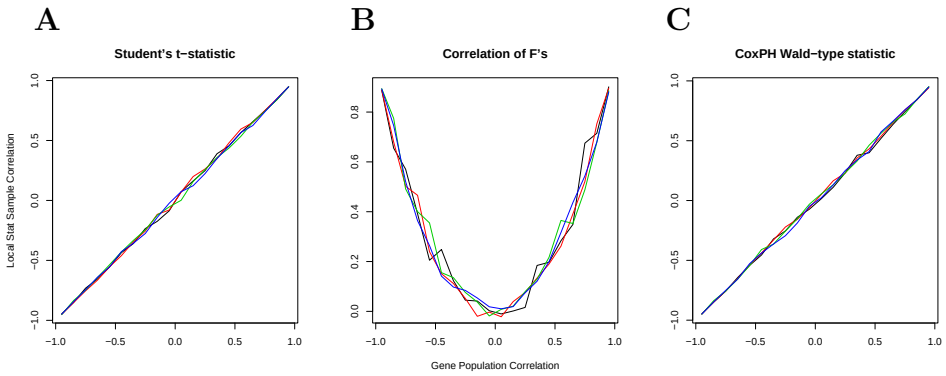

Figure 9: Correlations in expression and local statistic were generated by Monte Carlo simulation of Gaussian expression for two genes in several experimental designs: (**A**) Student's t for a two-sample comparison; (**B**) F statistic for an ANOVA with 4 groups; (**C**) Cox-proportional hazard model for relating expression to exponentially distributed survival and censoring times. In each design, the variance of expression in the second gene ranged from 1 to 10 times greater, and data was simulated for $n = 40$ arrays.

easily presented because of its dependency on both the underlying distribution of local statistics T and also the rejection region R.

For the average difference global statistic, U_D, the true variance will differ from that under the *i.i.d.* null in class 1 tests by three additive terms

$$\mathrm{Var}[U_D] = \mathrm{Var}_{i.i.d.}[U_D] + \frac{m_C - 1}{m_C}\rho_C + \frac{m_{\bar{C}} - 1}{m_{\bar{C}}}\rho_{\bar{C}} - \rho_{C,\bar{C}} \tag{3.7}$$

$$\text{where} \qquad \rho_C = \frac{1}{m_C \cdot (m_C - 1)} \sum_{i \in C} \sum_{\substack{i' \in C \\ i' \neq i}} \rho^T_{i,i'} \tag{3.8}$$

$$\rho_{\bar{C}} = \frac{1}{m_{\bar{C}} \cdot (m_{\bar{C}} - 1)} \sum_{i \in \bar{C}} \sum_{\substack{i' \in \bar{C} \\ i' \neq i}} \rho^T_{i,i'} \tag{3.9}$$

$$\rho_{C,\bar{C}} = \frac{1}{m_C \cdot m_{\bar{C}}} \sum_{i \in C} \sum_{i' \in \bar{C}} \rho^T_{i,i'}. \tag{3.10}$$

The additional terms in the variance are the average pairwise correlation within the category (3.8), within the complement (3.9), and across the two gene sets (3.10). Moreover, the variance implied by (3.1) (given by $\text{Var}_{i.i.d.}[U_D]$ in the above equation) is inversely proportional to m_C. Thus, for fixed values of the average correlations, the proportional variance inflation $\frac{\text{Var}[U_D]}{\text{Var}_{i.i.d.}[U_D]}$ will become more pronounced in larger categories. We note that ρ_C can vary greatly across categories while $\rho_{\bar{C}}$ and $\rho_{C,\bar{C}}$ will close to the average correlation across the array, which is nearer to zero in most datasets. Because of this, categories exhibiting positive correlation will have a U_D global statistic with greater variance than what is assumed under (3.1) leading to an anti-conservative Class 1 test.

For the Wilcoxon rank sum global statistic, it is difficult to relate the effect correlation in expression will have on the exact test of U_W which is based on the discrete distribution of the ranked sum. Nonetheless, solving for the variance of the statistic provides indirect evidence that the distribution will be misspecified when independence is violated, and also relates to the improper standardization of U_W that occurs in the class 1 approximate Z-test. In the following theorem, $\text{Var}[U_W]$ is derived in the special case of jointly Gaussian local statistics with any correlations $\{\rho^T\}$. The pdf and cdf of a univariate and bivariate Gaussian distribution are denoted by ϕ, Φ and ϕ_2, Φ_2 respectively.

Theorem 1. *Let $T_1, \ldots, T_m$ be identically distributed random variables that follow a multivariate Gaussian distribution with unit variances and pairwise correlations $\{\rho^T_{ij}\}$. Then for some category, $C \subset \{1, \ldots, m\}$, the variance of $U_W = \sum_{i \in C} \text{Rank}(T_i)$ is given by*

$$Var[U_W] = \frac{1}{2\pi} \sum_{i \in C} \sum_{j \in C} \sum_{k \notin C} \sum_{l \notin C} \sin^{-1} \left(\frac{\rho_{ij}^T + \rho_{kl}^T - \rho_{jk}^T - \rho_{il}^T}{\sqrt{(2 - 2\rho_{ik}^T) \cdot (2 - 2\rho_{jl}^T)}} \right) \tag{3.11}$$

Proof: The variance of U_W can be decomposed into covariances between pair-wise comparisons of local statistics through use of the Mann-Whitney form of the statistic as follows

$$\begin{aligned} \text{Var}[U_W] &= \text{Var}\Big[\sum_{i \in C} \text{Rank}(T_i) \Big] \\ &= \text{Var}\Big[\frac{m_C \cdot (m_C + 1)}{2} + \sum_{i \in C} \sum_{i' \in \bar{C}} I\{T_i > T_{i'}\} \Big] \\ &= \sum_{i \in C} \sum_{j \in C} \sum_{k \notin C} \sum_{l \notin C} \text{Cov}[I\{T_i > T_k\}, I\{T_j > T_l\}] \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} &\text{Cov}[I\{T_i > T_k\}, I\{T_j > T_l\}] \\ &= E[I\{T_i > T_k\} \cdot I\{T_j > T_l\}] - E[I\{T_i > T_k\}] \cdot E[I\{T_j > T_l\}] \\ &= \Pr(\{T_k - T_i < 0\} \cap \{T_l - T_j < 0\}) - \Pr(T_k - T_i < 0) \cdot \Pr(T_l - T_j < 0) \end{aligned} \tag{3.13}$$

Note that Each pair of differences in local statistics follows a centered bivariate Normal distribution

$$\begin{bmatrix} T_k - T_i \\ T_l - T_j \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 - 2 \cdot \rho_{ik}^T & \rho_{ij}^T + \rho_{kl}^T - \rho_{il}^T - \rho_{jk}^T \\ \rho_{ij}^T + \rho_{kl}^T - \rho_{il}^T - \rho_{jk}^T & 2 - 2 \cdot \rho_{jl}^T \end{bmatrix} \right). \tag{3.14}$$

Therefore each term in (3.12) may be evaluated as follows

$$
\begin{aligned}
&\mathrm{Cov}[I\{T_i > T_k\}, I\{T_j > T_l\}] \\
&= \Phi_2\Big(0, 0; \rho = \frac{\rho^T_{ij} + \rho^T_{kl} - \rho^T_{jk} - \rho^T_{il}}{\sqrt{(2 - 2\rho^T_{ik}) \cdot (2 - 2\rho^T_{jl})}}\Big) - \Phi(0) \cdot \Phi(0) \qquad (3.15) \\
&= \int_{-\infty}^{0} \int_{-\infty}^{0} \phi_2(x, y; \rho)\, dx\, dy - \frac{1}{4} \\
&= \int_{-\infty}^{0} \int_{-\infty}^{\frac{-\rho^2 z_2}{\sqrt{1-\rho^2}}} \phi_2(z_1, z_2; \rho = 0)\, dz_1\, dz_2 - \frac{1}{4} \qquad (3.16) \\
&= \int_{0}^{\infty} r \cdot \exp\left(-\frac{r^2}{2}\right) dr \cdot \int_{\pi}^{\frac{3\pi}{2} + \sin^{-1}(\rho)} \frac{1}{2\pi}\, d\theta - \frac{1}{4} \qquad (3.17) \\
&= \frac{1}{4} + \frac{\sin^{-1}(\rho)}{2\pi} - \frac{1}{4} = \frac{\sin^{-1}(\rho)}{2\pi} \qquad (3.18)
\end{aligned}
$$

where in (3.16) we have used the transformation $z_1 = \frac{x - \rho y}{\sqrt{1-\rho^2}}$, $z_2 = y$ and then in (3.17) the transformation $z_1 = r\cos\theta$, $z_2 = r\sin\theta$ □

Despite these analytical solutions for U_D and U_W under a special case, in general the relationship between $\{\rho^X_{i,i'}\}$ and $\{\rho^T_{i,i'}\}$, and also the relationship between $\{\rho^T_{i,i'}\}$ and the variance of the global statistic is unknown. Thus, we have conducted a simulation using real microarray data to quantify the improper Type I error rates when applying Class 1 tests to gene expression data that is correlated but has no association to the response of interest.

3.4.3 A Simulation Study

A simulation of a two-condition experiment was constructed from a subset of the lung carcinoma microarray data from Bhattacharjee et al. (2001). 100 adenocarcinoma samples

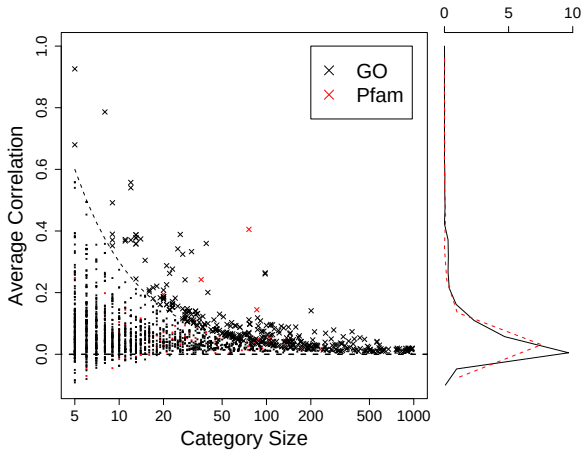

Figure 10: Scatterplot and histograms of the 1823 within-category correlations for the adenocarcinomas samples in the simulation study. 95% of GO and 88% of Pfam categories showed positive correlation on average. The dashed line separates categories with $m_C \cdot \rho_C > 2.5$.

were arbitrarily selected and for which expression estimates were available for 7299 genes (see Chapter 2 for the microarray pre-processing steps). 1823 GO and Pfam categories were identified with at least 5 members among the expressed genes. The within-category average pairwise sample correlations ranged from -0.09 to 0.93, with more than 86% of the categories exhibiting more correlation than what is seen on average across the entire array ($\bar{r} = 0.012$). A scatter plot of average correlation versus size is given in Figure 10, and is representative of what is seen in most datasets. This general increase in correlation within categories reflects the findings that coexpression among genes is highly linked to function (Lee et al. 2004).

500 randomly selected response vectors were generated for a two-condition experiment with equal sample size $n_1 = n_2 = 50$. In this way, there is no association between expression and experimental condition, and thus no category is deemed to have greater DE. In this scheme, we note that the expression matrix is held constant across simulations and the sample gene-gene correlations $\{r_{i,i'}\}$ remain fixed.

For each realization of the response vector, the absolute value of a pooled-variance t-statistic is used as the local statistic, and U_F, U_P, U_D, and U_W were calculated. For the Fisher's Exact Test statistic U_F and the difference in proportions U_P, the rejection region is set at the 0.95 quantile of the t_{98}-distribution. For each global statistic and each category, parametric tests yield a nominal p-value for every realized response vector. Histograms of the nominal p-values pooled across all categories and all realizations demonstrate the extreme non-uniformity of p-values under the induced null hypothesis, indicating the poor performance of Class 1 tests (Figure 11).

The average coverage of these tests is estimated by the proportion of simulated p-values that fall below a given α level. For each global statistic, the corresponding Class 1 test becomes more anti-conservative as one considers smaller p-values cut-offs for significance (Table 3). To illustrate how this behavior also affects the family-wise error rate among the $L = 1823$ categories, we applied a Bonferroni correction to a nominal α level for the different p-values. In this case, the true FWER seen across 500 simulations is defined as

$$FWER = \frac{1}{500} \sum_{i=1}^{500} I \left\{ \sum_{j=1}^{L} I \left\{ p_{i,j} < \frac{\alpha}{L} \right\} > 0 \right\} \tag{3.19}$$

where $p_{i,j}$ is the p-value for category j under realization i. Since there is substantial

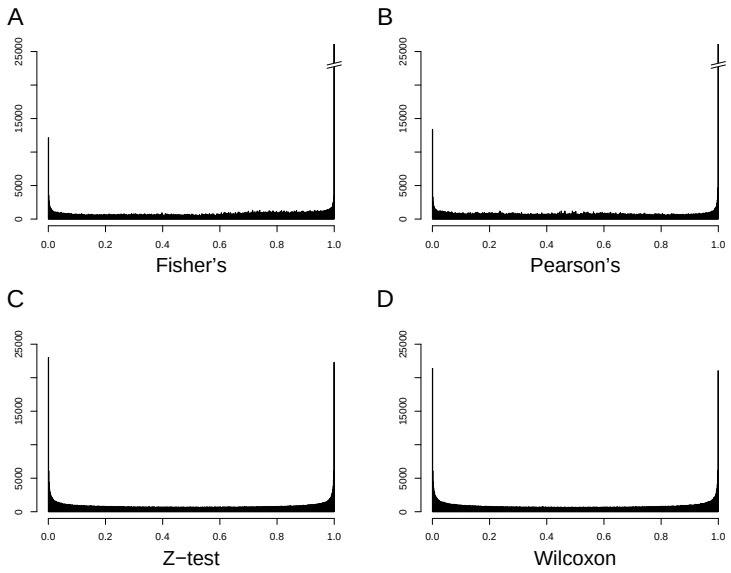

Figure 11: Histograms of p-values (1823 categories and 500 simulations) for the gene-list enrichment tests ((**A**) Fisher's Exact (**B**) and Pearson's difference in proportions), and for the average difference tests ((**C**) Z-test and (**D**) Wilcoxon rank sum). The large number of small and large p-values demonstrate the over dispersion that occurs in positively correlated gene categories from incorrect estimates of the variance.

overlap in the membership of gene categories with annotations like Gene Ontology, the use of Bonferroni threshold should be conservative in controlling the FWER. Therefore it might be thought to provide some protection against the nominally anti-conservativeness of Class 1 tests. However, for each global statistic, the minimum p-value passed the Bonferroni threshold in the majority of simulations (U_F : 0.772, U_P : 0.906, U_D : 0.926, and U_W : 0.916), illustrating the error rate is far greater than the target level. The extreme anti-conservativeness of the class 1 tests of all four global statistics suggests a different approach is needed to conduct valid gene category tests.

3.5 Class 2 tests and permutation

3.5.1 Defining the null hypothesis in class 2 tests

In the above sections, we demonstrate how the null hypothesis of class 1 tests (3.1) is violated by the correlation in gene expression. For this reason, a second class of gene category tests is warranted that can identify increases in differential expression within a category while accounting for correlation.

Definition 2. *Class 2 gene category tests are defined by the assumed or induced null hypothesis. For local statistics $T_1, \ldots, T_m$, the desired null is stated as*

$$H_0: \ T_1, \ T_2, \ \ldots \ , \ T_m \ \text{ are identically distributed with } \ T_i \ \dot{\sim} \ F \tag{3.20}$$

where F can take any general form, but is typically thought to correspond to there being no association between expression and the response of interest.

If all pairwise correlations in local statistics were known, $\{\rho^T_{i,i'}\}$, the true variances of the average difference statistic, U_D, and Wilcoxon rank sum statistic, U_W, will be as given in (3.7) and (3.11) respectively. From this, approximate Z-tests could be performed on standardized global statistics. In absence of knowing true correlations, a certain type of resampling-based tests can be used to induce the class 2 null.

3.5.2 Permutation-based gene category tests

Several gene category tests have proposed using permutation as a means of obtaining empirical p-values. In applying permutation to the data matrix of gene expression, methods have chosen the independent sampling unit to either be the expression profile of genes (*i.e.,* row permutation) or of arrays (*i.e.,* column permutation). It is important to note that the choice of sampling unit will directly effect the induced null hypothesis, and has been shown to dramatically influence the outcome of gene category tests (Breslin et al. 2004). Here, we present the induced null hypothesis of each in more detail and illustrate how array permutation methods are uniquely able to induce the class 2 null when the form of local statistics are chosen appropriately.

Gene permutation

Several permutation-based methods have proposed randomly reordering the rows of the data matrix (Ashburner et al. 2000; Pavlidis et al. 2004; Zhong et al. 2004). In this setup, the collection of local statistics remains unchanged while the category assignments are randomized. This resampling scheme is noted to induce the null hypothesis in (3.1) with reassigned local statistic following the empirical distribution of the observed values $\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m} I\{t_i \leq t\}$. The randomization also removes all correlation among the reassigned local statistics, and therefore will give results that are approximately equal to the *i.i.d.* tests. This was confirmed in the simulations presented above, with test results of the four global statistics under row permutation being equally anti-conservative when

a sufficient number of resamples is taken for the desired α-level of the test. Therefore, these methods do not offer any improvements in coverage over the non-resampling based class 1 tests.

Array permutation

The second manner in which permutation has been implemented is through reordering the column vectors of expression, reflecting that an array constitutes the independently sampled unit. This design has been implemented in GSEA for a Kolmogorov-Smirnov type global statistic (Mootha et al. 2003), and in SAFE for a Wilcoxon rank sum type global statistic (Barry et al. 2005). Array permutation procedures are applicable to experimental designs where reassigning samples or response information effectively removes the association of interest. As noted previously (Barry et al. 2005), this form of permutation does not change the observed correlation in expression among genes, such that the Class 2 null hypothesis is induced if the local statistics are identically distributed. In order to more fully describe a necessary property of local statistics required to induce this null (3.20), we revisit the process of selecting an appropriate form of $T(\cdot)$ for a certain experimental design.

3.5.3 δ-dependent local statistics

In most settings where gene category testing is performed, investigators are also interested in examining some gene-specific association to the response of interest. For many common

differential expression experiments, an unknown gene-specific parameter δ_i can be defined that meaningfully captures this association. In order to conduct a gene-specific analysis of differential expression, $T(\cdot)$ is chosen as a measure that can be used in a hypothesis test against a null value for the $\{\delta_i\}$. As illustration, consider a two-condition experiment where the response vector y_j takes values in the set $\{1, 2\}$, indicating the sample condition of the array. If the expression of gene i has expectation μ_{1i} and μ_{2i} under the two conditions and common variance σ_i^2, then the underlying association of interest in these experiments could be presented as the scaled difference in means

$$\delta_i = \frac{\mu_{1i} - \mu_{2i}}{\sigma_i \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.21}$$

where $n_k = \sum_{j=1}^{n} I\{y_j = k\}$ (Hu and Wright 2005). In this case, the gene-specific test of interest is $H_{0,i} : \delta_i = 0$ and the pooled-variance t-statistic is a natural choice of local statistic (Galitski et al. 1999). When gene expression is Gaussian the local statistic follows a central t-distribution under the null hypothesis. Other choices of $T(\cdot)$ that may also be appropriate for gene-specific testing of $\delta_i = 0$ as defined in (3.21), *e.g.,* a Wilcoxon rank sum statistic.

In general, a function $T(\cdot)$ is a proper choice of test statistic for a null of the form, $H_{0,i} : \delta_i = d$, when the distribution $F(T_i \mid \delta_i = d)$ is known and does not depend on any nuisance parameters. We note that some distributional properties may still require specification. For instance, with the Student's t-statistic proposed above, the distribution is known for gene expression data that is Gaussian once the degrees of freedom $n_1 + n_2 - 2$ is specified. When the distribution of $T(\cdot)$ can be specified in this manner for any choice of d, we refer to it as being *δ-determined*. This property is noted to also be important in the

theory of interval estimation and pivotal quantities. If the CDF $F(T_i \mid \delta_i = d)$ is known and does not depend on nuisance parameters, it can always be used as a pivotal quantity to construct a confidence set for δ_i by inverting the rejection region of the corresponding hypothesis test (Casella and Berger 2002).

Being $\delta-$determined is also important when conducting gene category tests, so that differences in nuisance parameters do not influence the comparison of a category against its complement. We illustrate the ramifications of this by returning the two-condition experiment and the gene-specific parameter from (3.21). Under this definition of δ, the individual means and variances of expression are considered nuisance parameters. Suppose that for each gene one directly uses the modified t-statistic from the SAM software (Tusher et al. 2001) as the local statistic. This statistic contains a constant in the denominator that effectively penalizes lowly-expressed genes in order to improve the FDR for lists of rejected genes. The SAM t-statistic is not $\delta-$determined because its distribution will depend on the means and variances of expression. Consider a category consisting of mainly highly-expressed genes (*e.g.,* "housekeeping" genes). Even if no genes were differentially expressed across conditions, and thus no category should be considered special in this regard, genes in the category would often appear amongst the most-significant genes in a ranked list. The category would thus appear to be significant. Categories with lowly-expressed genes would experience the opposite effect, and would be unlikely to be considered significant even under the alternative hypothesis.

When $\delta-$determined statistics are chosen, the null induced by array permutation, where every gene-specific association takes a null value, $H_0 : \delta_1 = \ldots = \delta_m = d$ can be stated in terms of the Class 2 null (3.20) where $F = F(T \mid d)$. For the remainder of the

paper, we will only consider local statistics that are $\delta-$determined, or approximately so when n is large.

3.5.4 Simulated coverage of class 2 tests

Array permutation can be employed to construct a Class 2 test of each of the global statistic presented above, and are evaluated through the simulation study. In this case, the tests are ensured to be of proper size, since both the randomization procedure in the simulation and array permutation employ the same sampling schemes. We confirmed this by obtaining empirical p-values for each category and each realization of the response vector, but due to computational restrictions the minimum possible empirical p-value was 0.001. The slight error in coverage that are noted for U_P, U_Z and U_W reflect sampling variability. The Class 2 Fisher's Exact Test results are somewhat conservative, and can be attributed to the numerous tied global statistics that occur in small categories, and which produced highly discretized p-values that will be conservative at a given α-level.

3.6 A more general null for gene category tests

In writing the Class 2 null hypothesis (3.20) induced by array permutation, we note a second potential shortcoming of the existing gene category methods. Both classes of procedures assume a null hypothesis under which the marginal distribution of every local statistic is identically distributed. However, the overall goal of gene category testing is to establish whether or not an relative increase in the amount of differential expres-

Table 3: The ratio of realized Type I error rates over different α levels for the Class 1 (parametric) and Class 2 (array permutation) tests of each global statistics. Results are from 500 randomizations of a subset of the adenocarcinoma sample from Bhattacharjee et al. (2001) into a two-condition experiment $n_1 = n_2 = 50$.

	Fisher's		**Pearson's**		**Z-test**		**Wilcoxon**	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
$\alpha = 0.1$	1.19	0.40	1.32	1.01	1.82	1.02	1.86	1.02
$\alpha = 0.01$	3.40	0.21	3.49	1.02	5.92	1.02	5.83	1.03
$\alpha = 0.001$	13.3	0.14	14.7	1.06	25.2	1.06	23.5	1.02
$\alpha = 1e-4$	65.3	-NA-	72.6	-NA-	130	-NA-	116	-NA-
$\alpha = 1e-5$	366	-NA-	432	-NA-	759	-NA-	669	-NA-
$\alpha = 1e-6$	2219	-NA-	2918	-NA-	4880	-NA-	4183	-NA-

sion is observed. For example, if a fraction of the genes on the array are differentially expressed to an identical degree while the remaining genes non-differentially expressed, than any category with the same proportions should be considered no different than the complementary set of genes. However, the array permutation null is violated in this case.

Based on this example, we propose the following less restrictive null hypothesis. Instead of requiring all to have a common level of differential expression, we allow each gene to fall into one of $K \leq m_C$ strata, where each has a different degrees of association with the response of interest. Formally, we restrict the gene-specific parameters δ_i,

$i = 1, \ldots, m$, to belong to a finite set $\{d_1, d_2, \ldots, d_K\}$. Let $\beta_{C,k} = m_C^{-1} \sum_{i \in C} I\{\delta_i = d_k\}$ denote the proportion of genes in C that belong to the k-th stratum, and let $\beta_{\bar{C},k}$ denote the corresponding proportion of genes in the complement with the same amount of DE. The null hypothesis is that the proportions of genes in each of the K strata is the same for both gene sets:

$$H_0 : \quad \beta_{C,k} = \beta_{\bar{C},k} = \beta_k \quad k = 1, \ldots, K \tag{3.22}$$

The null in (3.22) allows a broad variety of associations with the response of interest, while maintaining the overall goal of gene category tests. This formulation can also be thought of as ensuring the empirical distribution of gene-specific parameters will be identical between the two sets. Last, we note that the Class 1 and Class 2 nulls become special cases with $K = 1$ stratum.

To define a set of alternative hypotheses of interest, we restate that a functional category of interest is one with more overall differential expression among its constituent genes than what is seen across the array. Thus, a natural alternative to (3.22) is when the average DE of genes in the category is greater than in the complementary set. In our notation, this can be written as

$$H_A : \quad \sum_{i=1}^{K} \beta_{C,k} \cdot d_k > \sum_{i=1}^{K} \beta_{\bar{C},k} \cdot d_k \tag{3.23}$$

where the Wilcoxon rank sum, U_W, will continue to be a well suited global statistic for identifying increased amounts of differential expression in a robust manner. Other alternatives to (3.22) exist, where the category's eCDF of gene-specific parameters is different than that of the complement, but without DE being greater on average. But these alternatives relate to the earlier criticism of Kolmogorov-Smirnov type tests as

being of less biological interest.

In the following subsections we will describe simple bootstrap-based tests that are compatible with the stratified null (3.22) and which approximately maintain the correlation structure of the expression data. Distributional properties of U_W are derived under (3.22) that demonstrate both its utility in the bootstrap and also a reason they have improved coverage over Class 2 tests when under the stratified null. The simulation study is adapted to create the more general null in order to quantify the improvements with real gene categories, and to also demonstrate increases in power under defined alternatives (3.23).

3.6.1 Defining the bootstrap-based tests

Standard bootstrap methodology is based on the assumption that the observed data can be divided into independent units that are derived from an unknown probability model. When the statistic of interest is sufficiently regular, resampling from the empirical distribution of the observed data enables one to form confidence intervals without parametric assumptions (Efron and Tibshirani 1998).

For most experimental designs in the analysis of microarray data, the independent sampling unit is the joint vector $\{\mathbf{x}_{*j}, y_j\}$ containing both the m gene expression measurements and response information for a sample. In order to approximate the unknown probability model of the data, we resample the joint vectors with replacement. Let $\mathbf{b} = (b_1, \ldots, b_n)$ be a resampling vector whose elements are independent and uniformly distributed over the integers $\{1, \ldots, n\}$. Associated with $\mathbf{b}$ is a resampled response

$\mathbf{y}^{*b} = (y_{b_1}, \ldots, y_{b_n})$, and a resampled expression matrix in which the measurements of gene i are given by $\mathbf{x}_{i\cdot}^{*b} = (x_{ib_1}, \ldots, x_{ib_n})$. From the resampled data, local statistics $t_i^{*b} = T(\mathbf{x}_{i\cdot}^{*b}, \mathbf{y}^{*b})$, and a global statistic $u^{*b} = U(t_1^{*b}, \ldots, t_m^{*b} : C)$ may be calculated in the usual way. Let B denote the total number of bootstrap samples.

We use standard methods to generate bootstrap confidence intervals for the parameter $\theta = E[U]$, where U is suitably chosen so that the expectation is known under the stratified null H_0 in (3.22). The corresponding hypothesis test determines if $\theta_0 = E_{H_0}[U]$ falls in the constructed interval.

In the following theorem, we show that the expectation of the Wilcoxon global statistic U_W under the stratified null hypothesis is the same as in the classical *i.i.d.* setting, $\theta_0 = \frac{m_C \cdot (m+1)}{2}$, regardless of K and the constants $d_1, \ldots, d_K$.

Let $T_1, \ldots, T_m$ be absolutely continuous random variables having densities in an indexed family $\{f(t;\delta) : \delta \in D\}$. In particular, we assume that the distribution of T_i has density $f_{T_i}(t) = f(t;\delta_i)$ for some sequence of indices $\delta_1, \ldots, \delta_m$. For any T_i and T_j with $\delta_i = \delta_j$, we assume that the joint distribution is symmetric, *i.e.,* $f_{T_i,T_j}(t_1, t_2) = f_{T_i,T_j}(t_2, t_1)$ for all t_1, t_2

The following elementary lemma will be useful in evaluating the moments of U_W.

Lemma 1. *Let T_1 and T_2 be distributed as $f(t;\delta_1)$ and $f(t;\delta_2)$ and assume that* $\Pr(T_1 = T_2) = 0$. *Define $\mu(\delta_1, \delta_2) \equiv E[I\{T_1 > T_2\}]$, then $\mu(\delta_1, \delta_2) = 1 - \mu(\delta_2, \delta_1)$ and $\mu(\delta_1, \delta_2) = \frac{1}{2}$ when $\delta_1 = \delta_2$.*

Let K be the number of strata of differentially expressed genes present on the array,

so that for each $i = 1, \ldots, m$ the index δ_i is contained in the fixed set $D = \{d_1, \ldots, d_K\}$. Let $\beta_k = \frac{1}{m}\sum_{i=1}^{m} I\{\delta_i = d_k\}$ be the proportion of genes on the array that are in the k-th strata (see Section 3.6 for more detail).

Theorem 2. *For a category* $C \subseteq \{1, \ldots, m\}$ *where* $\frac{1}{m_C}\sum_{i\in C} I\{\delta_i = d_k\} = \frac{1}{m}\sum_{i=1}^{m} I\{\delta_i = d_k\} = \beta_k$ *for every stratum, then the expectation of* U_W *is*

$$E[U_W] = \frac{m_C \cdot (m+1)}{2}. \tag{3.24}$$

Proof: The expectation of U_W may be calculated as follows by decomposing the $m_C \cdot m_{\bar{C}}$ pairwise comparison of T's into K^2 different terms involving $\mu(d_k, d_{k'})$.

$$\begin{aligned}
E[U_W] &= E\Big[\sum_{i\in C}\mathrm{Rank}(T_i)\Big] = E\Big[\frac{m_C\cdot(m_C+1)}{2} + \sum_{i\in C}\sum_{j\notin C} I\{T_i > T_j\}\Big] \\
&= \frac{m_C\cdot(m_C+1)}{2} + \sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{\substack{i\in C\\ \delta_i = d_k}}\sum_{\substack{j\notin C\\ \delta_j = d_{k'}}} \mu(d_k, d_{k'}) \\
&= \frac{m_C\cdot(m_C+1)}{2} + \sum_{k=1}^{K}\sum_{k'=1}^{K} m_C\cdot\beta_k\cdot m_{\bar{C}}\cdot\beta_{k'}\cdot\mu(d_k, d_{k'}) \\
&= \frac{m_C\cdot(m_C+1)}{2} + m_C\cdot m_{\bar{C}}\Big[\sum_{k=1}^{K}\frac{\beta_k^2}{2} + \sum_{k'<k}\beta_k\cdot\beta_{k'}\cdot\big[\mu(d_k, d_{k'}) + \mu(d_{k'}, d_k)\big]\Big] \\
&= \frac{m_C\cdot(m_C+1)}{2} + m_C\cdot m_{\bar{C}}\Big[\sum_{k=1}^{K}\frac{\beta_k^2}{2} + \sum_{k'<k}\beta_k\cdot\beta_{k'}\Big] \\
&= \frac{m_C\cdot(m_C+1)}{2} + \frac{m_C\cdot m_{\bar{C}}}{2}\Big[\sum_{k=1}^{K}\beta_k\Big]^2 \\
&= \frac{m_C\cdot(m_C+1)}{2} + \frac{m_C\cdot m_{\bar{C}}}{2} = \frac{m_C\cdot(m+1)}{2}
\end{aligned}$$

We note that the last expression does not depend on the number of strata K, the proportion of genes in each $\{\beta_1, \ldots, \beta_K\}$, nor the degrees of association $\{d_1, \ldots d_K\}$. Further-

more, it equals the expectation of the Wilcoxon rank sum under the traditional Class 1 null (3.1) □

Theorem 2 holds regardless of the dependence structure among local statistics. Since the expectation of U_W is fixed under any form of the stratified null, a hypothesis test can be conducted by determining whether the null value is contained in an appropriately defined confidence interval. Similar derivations for the global statistics U_Z and U_P can demonstrate they also have a fixed expectation of 0 under (3.22). By contrast, the expectation of the global statistic employed in Fisher's Exact test depends on the K gene-specific parameters, and the expectation of the Kolmogorov-Smirnov type global statistic used in Mootha et al. (2003) depends on both the gene-specific parameters and the correlation structure among local statistics. Thus, standard bootstrapped confidence intervals can not be used to conduct hypothesis tests for these global statistics. The Wilcoxon global statistic U_W is still favored as a robust statistic that avoids the arbitrariness of choosing a rejection region for the gene-list methods. Thus, in the remaining sections it will be the only global statistic considered.

In order to test the null in (3.22) against the one-sided alternatives described in (3.23), we produce a confidence interval for U_W whose lower bound L_α is an estimate of the α quantile of the unknown distribution of U_W. The associated test rejects H_0 when $\theta_0 < L_\alpha$. A basic procedure for producing a confidence interval via bootstrap resampling is the percentile method (Efron 1979). In this case the lower bound is simply the sample α-percentile of the resampled values $\{u^{*b}\} : L_\alpha = u^*_{(B\cdot\alpha)}$. The percentile method is

straightforward to compute and invariant under monotone transformations of the global statistics. However, its coverage is often poor, especially when the sample size is small (Efron 1987) due to the difficulty of estimating the tail distribution of the global statistic. The slight anti-conservativeness of the resulting test is reflected in simulations below.

Alternatively, if one assumes that the distribution of the global statistic is approximately Gaussian, a confidence interval can be generated using common bootstrap-based estimates of the moments of U_W. The resulting one-sided confidence interval has a lower bound given by

$$L_\alpha = \bar{u}^* - \widehat{se}^*(U) \cdot t_{n-1,1-\alpha}. \tag{3.25}$$

where

$$\bar{u}^* = \frac{1}{B}\sum_{b=1}^{B} u_b^* \quad \text{and} \quad \widehat{se}^*(U) = \left[\frac{\sum_{b=1}^{B}(u_b^* - \bar{u}^*)^2}{B-1}\right]^{\frac{1}{2}} \tag{3.26}$$

In (3.5) we note that the Wilcoxon global statistic U_W is the sum of $m_C \cdot (m - m_C)$ pairwise comparisons of local statistics. When the average correlation between terms is not extreme and m_C is large, approximate normality of U_W follows from the Central Limit Theorem. Histograms of resampled global statistics confirm that the approximation to the Gaussian distribution is appropriate for the large number of genes considered in most microarray experiments. One advantage the t-interval has over the percentile interval is that the maximum attainable significance is not bounded by the number of resamples taken. Our simulations suggest that $B = 200$ arrays are typically sufficient for estimating the first two moments.

3.6.2 Coverage Under a Simulated Null

The coverage of permutation- and bootstrap-based tests of U_W under the stratified null hypothesis was examined using randomization of the lung cancer dataset from Section 3.4.3. Several null hypotheses were investigated with $K = 2$ classes of genes. In each the gene-specific parameters in (3.21) took one of two values, 0 or $d > 0$. To artificially generate differentially expression in the i-th gene, the expression values were first standardized to have unit variance; then $d \cdot \sqrt{1/n_1 + 1/n_2}$ was added to the measurements x_{ij} with $y_j = 1$. Simulations were run with three levels of DE, $d = 1,\ 3, \text{and}\ \ 5$, and also for three proportions of DE, $\beta = \frac{1}{5},\ \frac{1}{3},$ and $\frac{1}{2}$. For each proportion, β, a subset of non-overlapping categories were selected such that $\beta \cdot m_C$ and $\beta \cdot m_{\bar{C}}$ are integers. This resulted in 41 categories being considered for $\beta = \frac{1}{5}$, 40 categories for $\beta = \frac{1}{3}$, and 34 categories for $\beta = \frac{1}{2}$. The categories exhibited a wide range of correlation, reflective of that seen across all categories.

For each of 1000 randomizations of tumor status, array permutation and bootstrap-based hypothesis tests were conducted using 2500 permutations and resamples, respectively. Coverage was determined by comparing the empirically derived p-values to various α levels (Figure 12). For $\alpha = 0.05$, the bootstrap coverage was only slightly greater and remained relatively unchanged regardless of β and d, whereas the coverage of permutation testing dropped dramatically as β and d diverged from 0. For $d = 3$ and $\beta = \frac{1}{3}$, the minimum empirical p-value under permutation was 0.012, so the estimated coverage for any $\alpha < 0.012$ would be zero (Figure 12C).

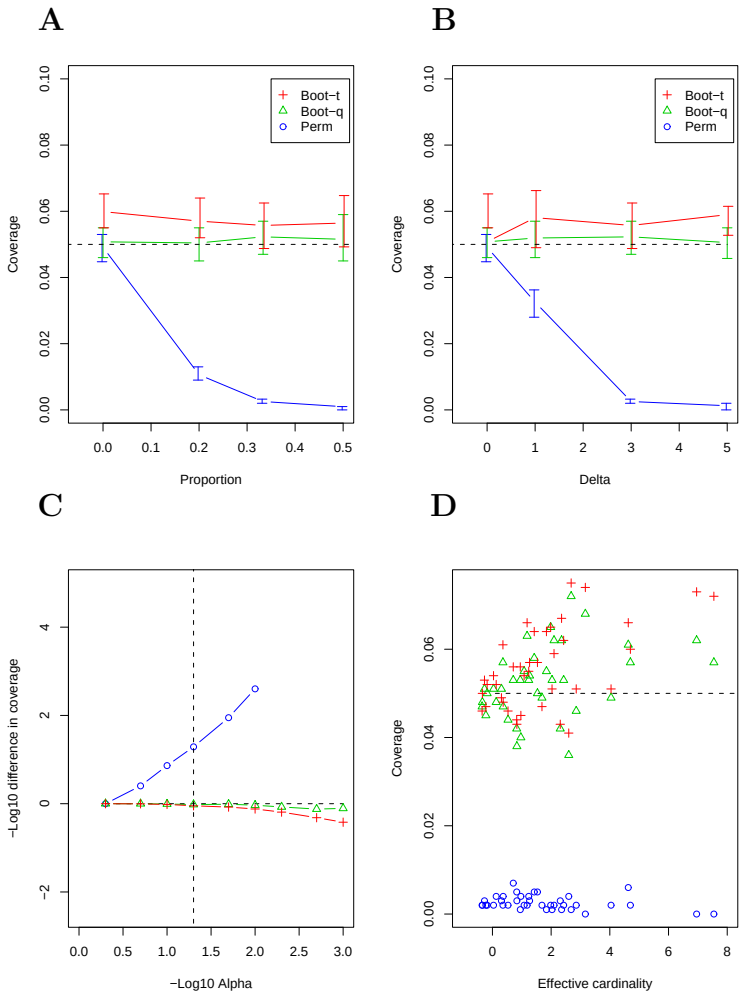

Figure 12: Performance of bootstrap- and permutation-based SAFE tests under different null hypotheses. The average coverage of a category is shown for (**A**) four different proportions of DE and (**B**) for four different levels of DE. (**C**) The coverage at different α levels is shown for $d = 3$ and $\beta = \frac{1}{3}$, and (**D**) the coverage for each category is plotted against the effective cardinality.

These findings were also confirmed using simulated data from an independent Gaussian model for a two-condition experiment, although not demonstrated here. In the simulation above, the bootstrap methods, while slightly anti-conservative, maintained their approximately correct coverage regardless of the null hypothesis being induced. However, in simulated expression data with a smaller sample size of $n = 20$, the anti-conservativeness of the percentile-based bootstrap method becomes more pronounced at smaller α. Since many microarray datasets can be of this size, the bootstrap Student's t-interval is suggested as the preferred approach.

3.6.3 Proof of improper coverage under permutation

The poor performance of permutation-based testing can be attributed to the fact, noted above, that a null is induced under which the local statistics are approximately identically distributed (3.20). We show in the following theorem that, for suitably correlated Gaussian local statistics, the variance of the Wilcoxon global statistic U_W is maximized under the $K = 1$ null in (3.20). Since the variance of U_W under the stratified null (3.22) will be smaller, and will in fact decrease as genes become more differentially expressed, the array permutation-based tests will tend to be conservative, as is seen in Figure 12.

The following lemma regarding the bivariate Gaussian distribution will be useful for the theorem.

Lemma 2. *For the bivariate normal distribution, the following is true for the function* $f(x, y) = \Phi_2(x, y; \rho) - \Phi(x) \cdot \Phi(y)$*:*

1. $f(0, 0)$ *is a global maximum when* $\rho > 0$

2. *$f(0,0)$ is a global minimum when $\rho < 0$*

3. *$f(x,y) = 0$ when $\rho = 0$*

Proof: The first derivatives of $f(x,y)$ are

$$\begin{aligned}
\frac{\partial f}{\partial x}(x,y) &= \frac{\partial}{\partial x}\left(\Phi_2(x,y;\rho) - \Phi(x)\cdot\Phi(y)\right) \\
&= \phi(x)\cdot\Phi\Big(\frac{y-\rho x}{\sqrt{1-\rho^2}}\Big) - \phi(x)\cdot\Phi(y) \qquad (3.27) \\
&\propto \Phi\Big(\frac{y-\rho x}{\sqrt{1-\rho^2}}\Big) - \Phi(y) \qquad (3.28)
\end{aligned}$$

and $\frac{\partial f}{\partial y}$ has an analogous form due to symmetry. Since Φ is a strictly increasing function, setting the derivatives equal to zero leads to the following equations

$$\begin{aligned}
y - \rho x &= \sqrt{1-\rho^2}\cdot y \\
x - \rho y &= \sqrt{1-\rho^2}\cdot x \qquad (3.29)
\end{aligned}$$

for which $\{x = 0, y = 0\}$ is the only solution when $\rho \neq 0$. Since $(0,0)$ is the only stationary point, a second derivative test can be used to determine whether it is a global minimum or maximum (Thomas and Finney 1992). The second derivatives can be solved as follows

$$\begin{aligned}
\frac{\partial f}{\partial x^2}(x,y) &= \phi'(x)\left[\Phi\Big(\frac{y-\rho x}{\sqrt{1-\rho^2}}\Big) - \Phi(y)\right] + \phi(x)\cdot\phi\Big(\frac{y-\rho x}{\sqrt{1-\rho^2}}\Big)\cdot\frac{-\rho}{\sqrt{1-\rho^2}} \\
\frac{\partial f}{\partial y^2}(x,y) &= \phi'(y)\left[\Phi\Big(\frac{x-\rho y}{\sqrt{1-\rho^2}}\Big) - \Phi(x)\right] + \phi(y)\cdot\phi\Big(\frac{x-\rho y}{\sqrt{1-\rho^2}}\Big)\cdot\frac{-\rho}{\sqrt{1-\rho^2}} \\
\frac{\partial f}{\partial x\partial y}(x,y) &= \phi(x)\cdot\left[\phi\Big(\frac{y-\rho x}{\sqrt{1-\rho^2}}\Big)\cdot\frac{1}{\sqrt{1-\rho^2}} - \phi(y)\right] = \frac{\partial f}{\partial y\partial x}(x,y) \quad \text{by symmetry}
\end{aligned}$$

At the point $\{x = 0, y = 0\}$ the derivatives are equal to

$$\begin{aligned}
\frac{\partial f}{\partial y^2}(0,0) = \frac{\partial f}{\partial x^2}(0,0) &= \left[0 + \phi(0)^2 \cdot \frac{-\rho}{\sqrt{1-\rho^2}}\right] \\
&= \phi(0)^2 \cdot \frac{-\rho}{\sqrt{1-\rho^2}} \qquad (3.30) \\
\frac{\partial f}{\partial x \partial y}(0,0) = \frac{\partial f}{\partial y \partial x}(0,0) &= \phi(0) \cdot \left[\phi(0) \cdot \frac{1}{\sqrt{1-\rho^2}} - \phi(0)\right] \qquad (3.31)
\end{aligned}$$

and the discriminant takes the form

$$\begin{aligned}
D(0,0) &= \frac{\partial f}{\partial x^2}(0,0) \cdot \frac{\partial f}{\partial y^2}(0,0) - \frac{\partial f}{\partial x \partial y}(0,0)^2 \\
&= \left(\phi(0)^2 \cdot \frac{-\rho}{\sqrt{1-\rho^2}}\right)^2 - \left(\phi(0) \cdot \left[\phi(0) \cdot \frac{1}{\sqrt{1-\rho^2}} - \phi(0)\right]\right)^2 \\
&= \phi(0)^4 \left(\frac{\rho^2}{1-\rho^2} - \frac{(1-\sqrt{1-\rho^2})^2}{1-\rho^2}\right) \\
&= \phi(0)^4 \cdot 2 \cdot \frac{\sqrt{1-\rho^2} - (1-\rho^2)}{1-\rho^2} \qquad (3.32)
\end{aligned}$$

Since $\sqrt{1-\rho^2} > (1-\rho^2)$ for all non-zero $\rho \in (-1, 1)$, the discriminant is strictly positive, proving that either a minimum or a maximum must exist. From the second derivatives in (3.30), one can show that $f(0,0)$ is a minimum when $\rho < 0$ and a maximum when $\rho > 0$. Lastly $f(x,y)$ is exactly 0 when $\rho = 0$ by independence □

In order to prove $\mathrm{Var}[U_W]$ is a maximized when $K = 1$, we must place the following restriction on the correlations among local statistics.

Definition 3. *For local statistics $T_1, \ldots, T_m$ with correlations $\{\rho_{ij}^T\}$, a category $C \subseteq \{1, \ldots, m\}$ will be called* correlation dominant *if for every $\{i,j\} \in C$ and $\{k,l\} \notin C$ it is true that $\rho_{ij}^T \geq \rho_{ik}^T$ and $\rho_{kl}^T \geq \rho_{kj}^T$, so that all correlations within the category and within the complement are greater than those across the two gene sets.*

Theorem 3. *Let $T_1, \ldots, T_m$ be random variables that follow a multivariate Gaussian distribution with means $\delta_1, \ldots, \delta_m$, unit variances and correlations $\{\rho^T_{ik}\}$. For a correlation dominant gene category C, the variance of U_W has a global maximum at $\delta_1 = \delta_2 = \ldots = \delta_m = d$.*

Proof: The variance of U_W can be decomposed into the covariances given in (3.12) as described in Theorem 1, but unlike (3.14), the paired differences in local statistics now follow a non-central bivariate normal distribution with marginal means $\delta_k - \delta_i$ and $\delta_l - \delta_j$. From (3.13) each covariance term can be written as

$$\mathrm{Cov}[I\{T_i > T_k\}, I\{T_j > T_l\}] = \Phi_2\Big(\delta_k - \delta_i, \delta_l - \delta_j; \rho\Big) - \Phi(\delta_k - \delta_i) \cdot \Phi(\delta_l - \delta_j) \quad (3.33)$$

where ρ is defined as in (3.15). We consider in turn several cases.

When $i = j$ and $k = l$, ρ is proportional to $2 - 2 \cdot \rho^T_{ik}$, which is positive quantity except when the genes are perfectly correlated which is ruled out by the definition of a correlation dominant category. From Lemma 2, (3.33) is maximized when $\delta_i = \delta_k$. Since this is true for all $\{i, k\}$ pairs of category and complement genes, a global maximum of the summed covariances will occur when all local statistics have the same mean.

When $i = j$ and $k \neq l$, ρ is proportional to $1 + \rho^T_{ij} - \rho^T_{jk} - \rho^T_{il}$ and will be greater than 0 for a correlation dominant category such that a maximum occurs when $\delta_i = \delta_k = \delta_l$. An analogous argument holds for when $i \neq j$ and $k = l$.

For $i \neq j$ and $k \neq l$, either ρ will be positive if $(\rho^T_{ij} + \rho^T_{kl}) > (\rho^T_{jk} + \rho^T_{il})$ so that (3.33) is maximized when $\delta_k = \delta_i$ and $\delta_l = \delta_j$, or ρ will be exactly 0 if $(\rho^T_{ij} + \rho^T_{kl}) = (\rho^T_{jk} + \rho^T_{il})$ and (3.33) will be constant. This inequality of summed correlations is again guaranteed for correlation dominant categories.

This proves a global maximum for $\text{Var}[U_W]$ is achieved at $\delta_1 = \delta_2 = \ldots = \delta_m = d$ since only in this case will every covariance term in (3.12) be either maximized, or a constant. This situation corresponds to the assumption of identically distributed local statistics in the Class 2 null (3.20) that is the special case in the stratified null (3.22) when $K = 1$ $\square$.

Although the above theorem required that a dominance restriction be placed on the correlation structure of local statistics. However, many for non-correlation dominant categories with a positive average within category correlation, $\text{Var}[U_W]$ can also be shown numerically to be greater when $K = 1$ than in any $K = 2$ stratified null. This was illustrated in different categories of genes and in expression data from real microarray datasets.

3.6.4 Power under simulated alternatives

In order to assess the relative power of the bootstrap tests over array permutation, alternative hypotheses were specified that corresponded to the criterion in (3.23), and were induced in the randomized adenocarcinoma data. To achieve increases in DE in some or all of the strata, a constant was added or multiplied to the gene-specific parameters described in section 3.6.2. More precisely, if $\{\delta_i^0 : i \in C\}$ are the gene-specific parameters under H_0, we consider H_1 to be of the form of either $\{\delta_i = c + \delta_i^0 : i \in C\}$ or $\{\delta_i = c \cdot \delta_i^0 : i \in C\}$. In this way, power curves of each of the resampling-based tests can be displayed by varying c. Figure 13 illustrates the effects when c is applied in an additive

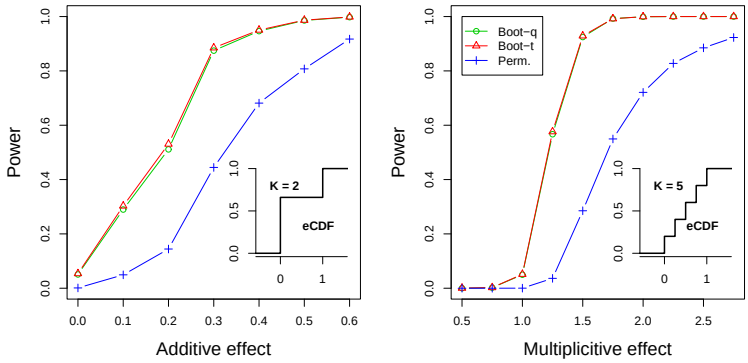

Figure 13: Average power of permutation and bootstrap based gene category tests as a function of the scaling constant c. Results based on randomized microarray data and real GO categories. (**A**) $K = 5$ classes, with $\{d_k\}$ equally spaced between 0 and 1, and (**B**) $K = 2$ classes of genes with 1/3 differentially expressed at $d = 1$ (as given by the CDF in the inset graphs). Both scenarios exhibit more power against the alternative with the bootstrap tests.

manner for $K = 2$ stratum with DE and non-DE genes, and in a multiplicative manner for an example with $K = 5$ stratum. The results demonstrate the improved power of the bootstrap methods over array permutation.

3.7 Analysis of a survival microarray dataset

The breast cancer survival datasets from Chang et al. (2005) is used to illustrate the power and utility of bootstrap-resampling as compared to array permutation. A total of $n = 295$ breast cancer samples were analyzed on Agilent microarrays, and normalized

gene expression estimates were obtained for a subset of $m = 11176$ genes that were annotated to at least one of 1348 GO terms (details on normalization, filtering, and formation of gene categories are omitted, but available from the principle investigators). Survival times and clinical covariates were available for each array. A Wald-type statistics from the univariate Cox proportional hazard model was used to test the association between expression and patient outcome.

For the permutation- and bootstrap-based tests, the Wilcoxon rank-sum was the global statistic with results obtained from 1000 permutation/bootstrap resamples of the data. The p-values produced by the bootstrap percentile- and t-intervals were in good agreement across the set of categories (Spearman rank correlation > 0.999), suggesting that the distributions of resampled global statistics had roughly Gaussian tails. The permutation test also showed good agreement with the bootstrap (rank correlation of 0.977 with bootstrap results), but a distinct difference in the number of categories passing certain levels of significance was observed (Table 4). The improved power of the bootstrap methods is apparent from the increased number of significant categories. Moreover, we have established that the increase in significant categories is far greater than could be induced by the slight anti-conservativeness of the bootstrap approach expected for the large sample size. The minimal possible p-value of the permutation and bootstrap-quantile tests are limited by the 1000 resamples that were taken of the data. The bootstrap t-interval does not have this restriction, and 28 categories were observed to pass the conservative Bonferroni threshold for $\alpha = 0.05$. Because of the iterative nature of the solution to the Cox-proportional hazard model, taking additional resamples of the dataset quickly becomes computationally taxing, and would be prohibitive when trying

Table 4: The number of significant GO categories is given for various α levels (uncorrected for multiple testing) when tests are conducted via the permutation, bootstrap-quantile, or bootstrap-t method.

	Perm	**Boot-Quant**	**Boot-t**
$\alpha = 0.1$	195	222	220
$\alpha = 0.05$	129	157	160
$\alpha = 0.01$	56	72	85
$\alpha = 0.005$	36	63	73
$\alpha = 0.001$	12	40	48
$\alpha = 3.7e-5^*$	-NA-	-NA-	28

* Bonferroni cutoff

to control the FWER across such a large number of categories.

3.8 Discussion

We have used the SAFE framework as presented in Chapter 2 to describe the different methods proposed for testing differential expression within a gene category. By stating the Class 1 and Class 2 null hypotheses behind these tests we illustrate their shortcomings and propose a novel bootstrap-based approach that uniquely allows for genes within the category and complement to be correlated and have different levels of differential expression.

Because of the extreme anti-conservativeness we have demonstrated for the popular gene-list methods, we feel it is important to survey their use in the literature and explore for possible errors that have been reported from positively correlated categories. The simulation method in Section 3.4.3 can be extended to other datasets to estimate the FDR or FWER for a given set of rejected hypotheses.

For the newly proposed bootstrap-based tests, future work is warranted for identifying a correction that can remove the slight anti-conservativeness seen in datasets with few arrays. The asymptotic behavior of global test statistics should be explored in more detail for the limiting cases of large m (which is common) and large n (which is almost never as large as m). Furthermore, resampling-based estimates of the FWER and FDR could be developed from the bootstrapped global statistics that could account for positively correlated categories in a manner like the Westfall and Young and the Benjamini-Yekutieli procedures (Benjamini and Yekutieli 2001; Westfall and Young 1993).

As a last but very important advantage to the bootstrap-based procedure, we note that by resampling with replacement it is uniquely capable of incorporating covariate information in a sensible manner. In permutation testing, by inducing a null that breaks the association between the response and expression, the covariate information can no longer relate to both variables. Yet, there may be no obvious reason to couple the covariate to either the response or expression, and the corresponding null should be fully recognized for either choice. Conversely, by resampling the sample information jointly, the bootstrap allows the relationship between all three to be maintained. To illustrate this point, we have implemented a multivariate Cox model to the breast cancer survival dataset that includes clinical covariates previously identified to be associated

with patient survival time. Estrogen Receptor status and tumor grade were reported by the original authors as being significantly associated with patient survival (Chang et al. 2005). By including the significant clinical covariates in the model, we can test for changes in expression that are significantly associated with survival over and above the clinical effects. For this particular example no categories were found to be more significant than their univariate test results, but this point may be potentially important in other experiments with complex designs and multiple covariates.

4 SAFE and transcription factor binding sites

4.1 Introduction

Over the past decade, genome sequencing projects and high throughput biotechnologies have led to an overabundance of publicly available information about gene composition, regulation, and function. These data have provided vast insight into the fundamental processes of transcription and translation from DNA to proteins and other functional components of cells. As the information is coalesced into structured databases that extend across genomes and species, the task of extracting biological insight from various sources has become a "informatics" challenge requiring both proper models for the complex underlying biology, valid estimation and hypothesis testing mechanisms, and the appropriate means of computing and interpreting of the volume of results obtained from data mining and exploratory analyses.

With the near completion of the human genome project, along with other efforts to identify entire genomes of model organisms, the opportunity exists to search for patterns in the nucleotide sequence that may relate to biological function. In recent years, sequence information has been made available through databases like the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu/). The identification of gene-coding sequences is complicated in humans and eukaryotic organisms by the presence of introns and alternative splice variants, but developed algorithms have found over 30,000 open reading

frames in the genome. While the coding regions of genes make up only a small fraction of the entire genome (1.5%), it is known that the surrounding regions of genes are involved in cellular processes that control the activation of expression through the binding of protein complexes termed transcription factors (TFs). Experimental procedures first allowed investigators to directly assay the binding sites of a single TF *in vitro*, and identify motifs through alignment procedures (Funk et al. 1992). These techniques become both costly and inefficient when considering multiple motifs or binding information from high throughput technologies like microarrays and the yeast-two hybrid system. Consequently, computational techniques have been developed for the discovery of multiple motifs in longer upstream regions of implicated genes (Liu et al. 2002). Further evidence suggests that gene regulation occurs through the complex interaction of multiple TF protein complexes that have been jointly referred as "cis-regulated modules", and several methods have been proposed for their discovery (Gupta and Liu 2005; Thompson et al. 2004; Wasserman et al. 2000). Once identified, the TF binding sites may be deposited into public repositories including the JASPAR (Sandelin et al. 2004) and TRANSFAC databases (Matys et al. 2003).

4.1.1 Motif discovery literature

Over the past two decades several approaches have been proposed for finding an unknown motif among a set of sequences that have been implicated as being co-regulated by a transcription factor. A first method for finding motifs of a fixed length, w, utilized the expectation maximum (EM) algorithm in finding maximum likelihood estimates from

a mixture model (Bailey and Elkan 1994). Briefly, the sequences data is reduced to all w-length oligomers, $X_1 \ldots X_n$, which are assumed to be distributed as one of two product-multinomials. True motifs are parameterized in a position-specific manner: $\Theta = (\theta_1, \ldots, \theta_w)$, with θ_i representing the frequencies of the four bases at the i-th position; background sequences have a single vector of base frequencies for all positions, θ_0. A mixing parameter for the two states, λ, is also in the model. The complete likelihood has the n sequences and unobserved indicators, $Z_1 \ldots Z_n$, for whether the oligomer is a motif or background.

$$\log L(\Theta, \theta_0, \lambda \mid X, Z) = \sum_{i=1}^{n} Z_i \log(\lambda \cdot p(X_i \mid \Theta) + (1 - Z_i) \log((1 - \lambda) \cdot p(X_i \mid \theta_0)) \quad (4.1)$$

In successive implementations of the EM algorithm in the software MEME, several improvements have been made to the model, including restricting the $\{Z_i\}$ such that overlapping motifs are excluded, and adding an additional constant to the frequency estimates that is equivalent to adding Dirichlet prior to the multinomial parameters (Bailey and Elkan 1994).

Concurrent to the development of MEME, Lawrence et al. (1993) proposed treating the same situation as a missing data problem that can be solved in a Bayesian manner through Gibbs sampling. It was also found that the model could be improved by adding a 3rd-order Markov chain to the background parameters, θ_0. Liu et al. (1995) proposed a more general prior for the missing indicators of motif-background status with a *motif abundance ratio*, β, that is similar in interpretation to the mixing parameter in MEME. Recently, such methods have been extended to looking for multiple motifs simultaneously, and for allowing variable widths in motifs (Jenson et al. 2004).

While the above models search for motifs that are common patterns in a set of implicated sequences, it is also important to be able to score the presence of a motif, once identified, in a DNA sequence. The authors of the MEME algorithm also designed a software, MAST (Bailey and Gribskov 1998), for computing p-values for both multiple motifs and sequences that are based on the match score proposed by Staden (1990). In the Bayesian framework, several scores can be derived from the posterior of a given sequence (Liu et al. 2002) that take the form of an entropy (or Kullback-Leibler) distance.

4.1.2 Contributions

Here we address in detail how several models in the discovery of binding sites are adapted to scoring upstream sequences for the presence of known motifs. By expressing this information as a posterior probability, sequence scores taken across the human genome are used to generate functional categories of genes potentially regulated by a known TF. We further note that the models of motif occurrences can be expanded to look for the joint presence of TFs that may occur in cis-regulating modules. Next, a hypothesis testing mechanism is detailed that looks for increased differential expression in microarray experiments as evidence of transcription regulation. Lastly, we propose using the combined information of upstream sequences and mRNA expression as a means of improving estimates of the TF binding site. Analysis are performed on both simulated and real datasets in order to demonstrate the validity and capability of this model-based framework.

4.2 Models for TF binding motifs

DNA sequences are made up by four nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). It is observed that in the human genome, DNA sequences are highly conserved among individuals, particularly in gene-coding regions. Because of this uniformity, the consensus sequence of genes can be represented by strings of the 4 letters listed above with only relatively rare exceptions for polymorphisms and *de novo* mutations. However, in some areas of DNA and protein sequence analysis (*e.g.* TF binding sites and homologous protein domains) inherent variation is observed among a set of implicated sequences.

When representing a transcription factor binding site, a consensus sequence is less adequate since many similar base-pair combinations are capable of binding to a single protein complex. As a partial solution, a more complete dictionary of 15 letters has been defined by IUPAC for every observable subset of the 4 bases (Lathe 1986). While this nomenclature can somewhat describe the site-specific variability of a binding motif, it does not provide a means for representing unequal probabilities of the occurrence of bases. To describe a binding site in a more quantitative manner, a *position specific weight matrix* (PSWM) is constructed once a set of implicated oligomers have been identified experimentally as TF binding sites and properly aligned. The matrix consists of the frequency counts of each base at each position, and is thus $4 \times w$ in size for a w-length motif (Figure 14). A second popular representation of a set of aligned binding sites has been termed a sequence logo (Schneider and Stephens 1990), which scales the frequencies in each position by their information content and displays the motif in a bar graph using

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 5 & 3 & 4 & 5 & 13 & 0 & 17 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 4 & 1 & 15 & 2 & 1 & 1 \\ 8 & 7 & 0 & 0 & 0 & 17 & 0 & 0 & 0 & 11 & 16 & 16 & 0 & 0 & 0 & 14 & 0 & 0 & 1 & 2 \\ 2 & 6 & 13 & 12 & 4 & 0 & 0 & 0 & 17 & 0 & 0 & 0 & 15 & 14 & 13 & 2 & 0 & 0 & 14 & 1 \\ 2 & 1 & 0 & 0 & 0 & 0 & 0 & 17 & 0 & 6 & 1 & 1 & 1 & 2 & 0 & 0 & 2 & 15 & 1 & 13 \end{bmatrix}$$

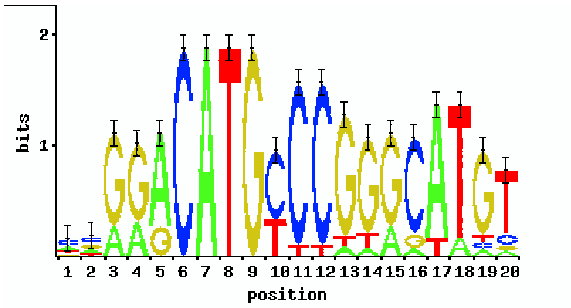

Figure 14: Position specific weight matrix of the motif for the p53 protein complex based on 17 identified binding sites (Funk et al. 1992). The corresponding sequence logo is shown as provided from the JASPAR database

the common base letters ordered by their respective prevalence.

4.2.1 Notation

To denote the sequence and motif information, we will consider a set of m sequences presented as $\mathcal{S} = \{\mathbf{s}_1 \dots \mathbf{s}_m\}$. For our purposes, the sequence data can be considered to be of equal length L . For a given sequence, $\mathbf{s}$, let the notation s_i be the i-th nucleotide in the sequence and $\mathbf{s}_{[a:b]}$ represent the fragment from position a to b. As a basic statistical model for the position specific weight matrix, Lawrence and Reilly (1990) originally suggested a multinomial distribution for the observed counts at each site. If one assumes

independence among the sites, s_1 to s_w

$$(s_1, \ldots, s_w) \sim \text{ProductMultinomial}(\Theta) \tag{4.2}$$

with $\Theta = (\theta_1, \ldots, \theta_w)$ such that $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,4})$ represents the probabilities of observing each of the four bases at the i-th position. When a set of aligned sequences provide a PSWM, the parameters of a known motif are given as, $\theta_{i,j} = \frac{n_{i,j}}{N}$ where $N = \sum_j n_{ij}$.

In order to model the background sequence that does not contain motifs, an independent multinomial model (IND) can be applied to each position with $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,4})$. Under the assumption that true motifs will be rare when m and L are large, θ_0 can be computed from the overall proportion of nucleotides in $\mathcal{S}$. In order to capture more of the basic structure in the non-coding regions of the genome, a higher-order Markov chain model (MCN) can also be utilized with transition probabilities θ_0, generated from all observed oligomers of the appropriate length (Jenson et al. 2004).

In the models that have been derived in the motif discovery literature, a parameter is also given for the *motif abundance ratio* describing the frequency of occurrence in the implicated sequences. For our purposes we will define as a similar parameter β to be the probability of a motif beginning at any randomly selected site. In the following models, β will either be considered fixed and known, or treated in a Bayesian manner using a prior distribution.

In scoring sequences for the presence of a known motifs, we will define several potential alternative hypotheses to contrast against the null that no motifs are present ($H_0 : \mathbf{s}$ is generated entirely from θ_0). In generating scores for upstream sequences from a given

model, we will focus on either a likelihood ratio statistic,

$$LR(\mathbf{s}) = \frac{\Pr(\mathbf{s} \mid H_A)}{\Pr(\mathbf{s} \mid H_0)} \tag{4.3}$$

or a related Bayes Factor when priors are given for certain parameters. To create gene categories from these scores, we will consider the posterior probability of the alternative hypothesis

$$\begin{aligned} \Pr(H_A \mid \mathbf{s}) &= \frac{\Pr(\mathbf{s} \mid H_A) \cdot \Pr(H_A)}{\Pr(\mathbf{s} \mid H_A) \cdot \Pr(H_A) + \Pr(\mathbf{s} \mid H_0) \cdot \Pr(H_0)} \\ &= \left[1 + \frac{1 - P_A}{P_A} \cdot LR(\mathbf{s})^{-1}\right]^{-1} \end{aligned} \tag{4.4}$$

which is a monotonic increasing function of $LR(\mathbf{s})$ that will also depend on a model-selection prior $\Pr(H_A) = 1 - \Pr(H_0) = P_A$. In the following subsections, scores are derived for three different types of models.

4.2.2 Single-site models

A basic model for the presence of a given motif in an upstream sequence can be described as "single-site" in that the alternative hypothesis states the motif occurs exactly once in the upstream sequence. In this case, the probability of a given sequence occurring under the alternative becomes the sum of mutually exclusive events that each position is the

start site of the motif.

$$\begin{aligned}\Pr(\mathbf{s} \mid H_A) &= \sum_{j=1}^{L-w+1} \Pr(\mathbf{s} \cap \text{ the motif starts at position } j) \\ &= \sum_{j=1}^{L-w+1} (1-\beta)^{j-1} \cdot \Pr(\mathbf{s}_{[1:j-1]} \mid \theta_0) \quad \times \quad \beta \cdot \Pr(\mathbf{s}_{[j:j+w-1]} \mid \Theta) \\ &\quad \times \quad (1-\beta)^{L-w-j+1} \cdot \Pr(\mathbf{s}_{[j+w:L]} \mid \theta_0) \end{aligned} \tag{4.5}$$

With a site-independent background model, the likelihood ratio in collapses to a sum of ratios for every w-mer in the sequence

$$\begin{aligned} LR(\mathbf{s}) &= \frac{\sum_{j=1}^{L-w+1} \Pr(\mathbf{s}_{[1:j-1]} \mid \theta_0) \cdot \Pr(\mathbf{s}_{[j:j+w-1]} \mid \Theta) \cdot \Pr(\mathbf{s}_{[j+w:L]} \mid \theta_0) \cdot \beta \cdot (1-\beta)^{L-w}}{\Pr(\mathbf{s} \mid \theta_0) \cdot (1-\beta)^{L}} \\ &= \frac{\beta}{(1-\beta)^{w}} \cdot \sum_{j=1}^{L-w+1} \frac{\Pr(\mathbf{s}_{[j:j+w-1]} \mid \Theta)}{\Pr(\mathbf{s}_{[j:j+w-1]} \mid \theta_0)} \end{aligned} \tag{4.6}$$

For the null hypothesis that uses a 3rd-order Markov chain (MCN) for background instead of assuming independence, (4.6) does not exactly hold because of the conditional probabilities of the three positions just after each w-mer. In our setting, where true motifs are rare and $\mathcal{S}$ is large, one can assume there is minimal difference between the true LR and (4.6) and scores are computed accordingly.

4.2.3 Multi-site models

Based on the biological findings of repeated motifs occurring in the upstream regions of genes (Liu et al. 1995), it is potentially more powerful to use an alternative hypothesis that allows for multiple realizations to occur. Here we describe a "multi-site" model that allows for anywhere from 1 to $\frac{L}{w}$ motifs to occur in an independent manner under the

restriction of non-overlapping sites (Lawrence et al. 1993). In order to find the probability of observing the full sequence under this alternative hypothesis, a recursive formula is needed to cover the complete set of mutually exclusive motif occurrences. The recursive algorithm can be stated in terms of whether the last position is considered to either be from background or part of the known motif

$$\begin{aligned} \Pr(\mathbf{s}) &= \Pr(\mathbf{s}_{[1:L-w]}) \cdot \Pr(\mathbf{s}_{[L-w+1:L]} \mid \Theta) \cdot \beta \\ &+ \Pr(\mathbf{s}_{[1:L-1]}) \cdot \Pr(\mathbf{s}_{[L:L]} \mid \theta_0) \cdot (1-\beta) \end{aligned} \tag{4.7}$$

From this formulation, a recursive algorithm can be implemented in a forward or backward manner. To compute $\Pr(\mathbf{s} \mid H_A)$ from (4.3), the probability of every site being from background appears in the recursion and must be subtracted from $\Pr(\mathbf{s})$, yielding the likelihood ratio

$$LR(\mathbf{s}) = \frac{\Pr(\mathbf{s}) - (1-\beta)^L \cdot \Pr(\mathbf{s} \mid \theta_0)}{(1-\beta)^L \cdot \Pr(\mathbf{s} \mid \theta_0)} \tag{4.8}$$

In this model the likelihood ratio statistic no longer reduces to a sum of ratios of w-length sequence probabilities. For this reason, numerical underflow issues must be considered when computing probabilities for large L and motifs with large K-L distances from background. As noted in single-site model, when incorporating the MCN background model into the recursion scheme, sequences that occur immediately after a motif realization are assumed to depend on the last 3 positions of the observed motif.

4.2.4 Bayesian models

Rather than considering β to be a fixed constant based on a presumption of the overall frequency of motifs in the promoter regions of the human genome, we proposed putting a prior distribution on β. Since the conditional probability of $\mathbf{s}$ under the multi-site model is a sum of polynomials in β (and $1-\beta$), a Beta distribution has both proper support for β and yields the following marginal distribution for $\mathbf{s}$.

$$\begin{aligned}
\Pr(\mathbf{s} \mid H_A) &= \int \Pr(\mathbf{s} \mid \beta, H_A) \cdot \pi(\beta)\, d\beta \\
&= \int \sum_{i=1}^{\frac{L}{w}} a_i \cdot \beta^i \cdot (1-\beta)^{L-iw} \cdot \text{beta}(\gamma_1, \gamma_2) \cdot \beta^{\gamma_1 - 1} \cdot (1-\beta)^{\gamma_2 - 1}\, d\beta \\
&= \sum_{i=1}^{\frac{L}{w}} a_{i,L} \cdot \frac{\text{beta}(\gamma_1, \gamma_2)}{\text{beta}(i+\gamma_1, L-iw+\gamma_2)} \qquad (4.9)
\end{aligned}$$

where $a_{i,L}$ is the sum of the conditional probabilities where exactly i motifs are realized in $\mathbf{s}$. Each term is computed in a recursive manner similar to (4.7), but which further indexes across the number of upstream motif occurrences in addition to position

$$\begin{aligned}
a_{i,L} &= \Pr(\mathbf{s} \mid i \text{ motifs occur}) \\
&= \Pr(\mathbf{s}_{[1:L-w]} \mid i-1 \text{ motifs occur}) \cdot \Pr(\mathbf{s}_{[L-w+1:L]} \mid \Theta) \; + \\
&\quad \Pr(\mathbf{s}_{[1:L-1]} \mid i \text{ motifs occur}) \cdot \Pr(\mathbf{s}_{[L:L]} \mid \theta_0) \\
&= a_{i-1,L-w} \cdot \Pr(\mathbf{s}_{[L-w+1:L]} \mid \Theta) \; + a_{i,L-1} \cdot \Pr(\mathbf{s}_{[L:L]} \mid \theta_0) \qquad (4.10)
\end{aligned}$$

Because the recursion is now applied to both the length of the sequence, and also the number of possibly inserted motifs, the algorithm is computationally slower than the

models for fixed, β. The conditional distribution under the null reduces to

$$\Pr(\mathbf{s} \mid H_0) \;=\; \Pr(\mathbf{s} \mid \theta_0) \cdot \frac{\text{beta}(\gamma_1, \gamma_2)}{\text{beta}(\gamma_1, L + \gamma_2)} \tag{4.11}$$

and as in the fixed multi-site model, this is computed to a proportional constant within the recursive algorithm. Then (4.3) becomes a Bayes Factor for the two models from which the posterior probability of H_A can be determined.

The posterior distribution of β from this model can by recognized from the kernel of the joint density as a weighted sum of Beta distributions

$$\begin{aligned} \Pr(\beta \mid \mathbf{s}) &\propto \Pr(\mathbf{s} \mid \beta) \cdot \pi(\beta) \\ &\propto \sum_{i=0}^{\frac{L}{w}} a_i \cdot \text{beta}(\gamma_1, \gamma_2) \cdot \beta^{\gamma_1 + i - 1} \cdot (1-\beta)^{\gamma_2 + L - iw - 1} \end{aligned}$$

so that

$$\Pr(\beta \mid \mathbf{s}) \;=\; \sum_{i=0}^{\frac{L}{w}} w_i \cdot \text{Beta}(\gamma_1 + i, \gamma_2 + L - iw) \tag{4.12}$$

where

$$w_i \;=\; a_i \cdot \frac{\text{beta}(\gamma_1, \gamma_2)}{\text{beta}(i + \gamma_1, L - iw + \gamma_2)} \cdot \Pr(\mathbf{s})^{-1} \tag{4.13}$$

The posterior mean of β given $\mathcal{S}$ can be solved directly by averaging across each sequence, and credible sets can be identified through numerical integration or an MCMC approach.

A simulation study is used to examine the different models and their respective sensitivity in different motifs. We further investigate the sensitivity of multi-site models to the choice of fixed values β or its hyperparameters $\{\gamma_1, \gamma_2\}$ that would correspond both to the small values of β that would generally be expected for a transcription factor, and also its uncertainty.

4.3 Simulation study of motif models

In order to evaluate the different models proposed for scoring the presence of known motifs in a set of sequences, simulated datasets were generated from the stated null and alternative hypotheses. Background sequences (of length $L = 5000$) were generated from the MCN model with transition probabilities estimated from all identified upstream sequences that were available from NCBI Build 36.1 of the human genome (http://genome.ucsc.edu/). For every human TF in the JASPAR database, realizations of the motif were generated from the ProductMultinomial distribution in (4.2) using the given PSWM as Θ. Non-overlapping motifs were inserted randomly into the background sequence for different realizations of β. Motifs were also inserted in a similar fashion to true upstream sequences randomly selected from NCBI Build 36.1 to confirm the results seen in simulated sequences would closely match those based on real structures of DNA.

To examine the performance of the models in Section 4.2, the scores for true background and true alternative sequences were compared using ROC curves of the true and false positive rates for the range possible predictors. The relative performance of models and motifs is quantified by the area under the curve (AUC). By examining performance in this manner, the results are invariant to monotone transformations and will be identical for likelihood ratios or Bayes factors (4.3) as compared to posterior probabilities (4.12) under any P_A. Figure 15 illustrates the relative performance of both the single- and multi-site models using ROC curves of representative motifs, and scatter plots of AUC.

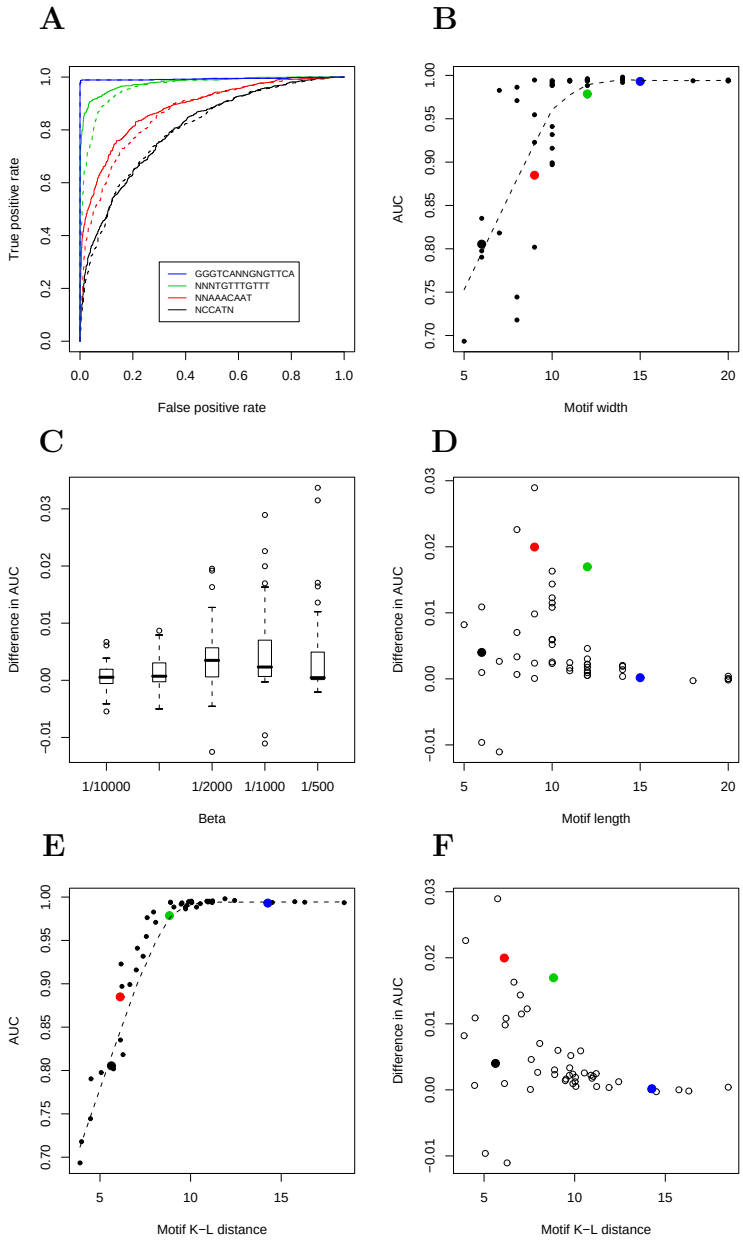

Figure 15: Simulated results for JASPAR motifs of varying length. (**A**) ROC curves for 4 representative motifs of length 6, 9, 12 and 15, scored by the single-site (dashed) and Bayesian multi-site model. (**B**) AUC is seen to increase with motif length, with the representative motifs from panel A shown in color. Multi-site models show the greatest improvement in performance when (**C**) β increases and (**D**) motifs are of moderate length between 6 and 12 base pairs. (**E**) and (**F**), panels B and D are are replotted against motif entropy, demonstrating a stronger correlation to AUC (rank correlation of 0.88), and improvement in the multi-site model for motifs with a moderate difference from background.

These results demonstrate ability to discrimination between true null and alternatives is correlated to motif length, but is shown to be more closely determined by the Kullback-Leibler distance from the background model.

$$H(\Theta, \theta_0) \;=\; \sum_{i=1}^{w} \sum_{j=1}^{4} \theta_{i,j} \cdot \log\left(\frac{\theta_{i,j}}{\theta_{0,j}}\right).$$

The improved performance of the multi-site models is demonstrated to be dependent on the realized β, and most notable for moderately-sized motifs. In shorter motifs, the relative abundance of close background sequences increase the false positive rate, while in longer motifs the chance of even a single occurrence being generate from background is so small that the distributions of scores fully discriminate.

In using ROC curves and AUC, we note that the single-site model is not penalized by misspecifying β, since it is a scale factor for the LR (4.6). In the multi-site models, the higher-order polynomial terms in β dominate such that β can be seen as an approximate shift-effect on the log-ratios, and thus do not appreciably effect the ROC curves. Likewise, the rank correlation is highly preserved across choice of β ($r \geq 0.97$ for all TF motifs). Despite this conservation of rank order, misspecification of β will sharply affect the magnitude of posterior probabilities and thus the FDR of any particular choice for predicting motif presence (*e.g.,* $\Pr(H_A \mid \mathbf{s}) \geq 0.5$). For this reason, we favor the Bayesian prior for β proposed in (4.9). The robustness of scores was compared for Beta priors with modes at $\frac{1}{1000}$ and $\frac{1}{10000}$, (given by $\frac{\gamma_1 - 1}{\gamma_1 + \gamma_2 - 2}$), where the variance was allowed to range from a degenerate point mass to increase to a uniform prior. We have demonstrated via simulation that the posterior probabilities will be more robust under moderately informative priors.

4.4 TF and differential expression experiments

Having selected the Bayesian multi-site model as most appropriate for the potential multiple occurrences of motifs and an unknown β, we apply it to the upstream regions of known genes as a means of forming categories of genes that are potentially regulated by known TFs. In order to look for the potential co-regulation of these categories in gene expression data, we next derive a hypothesis testing framework that accommodates the probabilistic measures of gene membership as defined in (4.12). To achieve this goal, we extend the method SAFE from Chapters 2 and 3 to be a robust, resampling-based test for regressing differential expression against the probability of gene membership. We apply this method to two microarray datasets as examples of a exploratory and hypothesis-driven analysis.

4.4.1 Probabilistic functional categories

With the wide varieties of biological information being accumulated about genes across the entire human genome, it has been suggested that a more probabilistic approach to assigning gene function is needed (Fraser and Marcotte 2004). As our understanding of the role each gene plays in cellular biology is informed by an increasing number of disparate experiments, it is important to recognize variability in the data, and in particular the potential for false positive and false negative results when making inferences on a genomic level. When combining results from different experiments and technologies it is important to consider how such errors would affect the certainty of any joint conclusion. This perspective has been realized by Troyanskaya et al. (2003) in imple-

menting a Bayesian network to integrate several sources of inference on gene function in *Saccharomyces cerevisiae*, including gene expression data from microarray studies and protein-protein interactions from the yeast two-hybrid system. Functional categories derived from the multiple sources were shown to be in agreement with Gene Ontology annotation. We note that the collective assignment of cellular function to a set of genes can be equivalently stated as their membership to a functional category, and thus gene category testing in differential expression experiments can also be treated in a probabilistic manner.

In Chapter 2 a nomenclature was given for representing gene categories in the analysis of differential expression experiments. For an experimental design in which the expression of m genes are measured, $\mathbf{c}$ is a vector of indicator variables of length m such that $c_i = 1$ if the i-th gene belongs to the category, and $c_i = 0$ otherwise. To allow for uncertainty in a gene category, rather than using a vector of indicators where $c_i \in \{0, 1\}$ one can allow each coefficient to take any value in the interval $[0, 1]$ to represent the probability of inclusion to the category. When functional categories are defined in this manner, one can no longer describe an exact number of genes as being contained in the category. Rather, we define the size of a category based on its expectation. Using the set notation of a category, C, presented in Chapter 3, this quantity is given as

$$\begin{aligned} m_c &= E\left[\sum_{i=1}^{m} I\{i \in C\}\right] = \sum_{i=1}^{m} \Pr(i \in C) \\ &= \sum_{i=1}^{m} c_i \end{aligned} \tag{4.14}$$

4.4.2 Non-parametric regression techniques

When the membership of genes to a functional category is redefined to be a continuous measure, hypothesis tests can no longer be based on two-sample comparisons of the category to its complement (see Chapter 3 for the survey of methodologies that are described by this framework). With membership taking values in the $[0, 1]$ interval, finding an association between category membership and an increased amount of differential expression becomes a regression problem. In order to be generalizable to the diverse experimental designs that relate to differential expression, a permutation approach using distribution-free statistics is proposed.

To place the non-parametric regression problem in the SAFE framework, consider a local statistic, T, is chosen such that one expects a linear shift in the unknown distribution based on category membership $\Pr(T_i < t \mid c_i) = F(t - c_i\ b)$ the rank-based Wilcoxon estimate of the slope parameter b minimizes the dispersion function

$$D(b) = \sum_{i=1}^{m} \left(\text{Rank}(\epsilon_i(b)) - \frac{m+1}{2} \right) \epsilon_i(b) \tag{4.15}$$

where $\epsilon_i(b) = T_i - c_i \cdot b$. To test the null hypothesis that differential expression is unchanged by category membership, $H_0 : b = 0$, the Wilcoxon linear score statistic was proposed and characterized by Hajek and Sidak (1967), and is widely used in an asymptotic Z-test as follows

$$Z = \left((m+1)^2 \sum_{i=1}^{m} \left(c_i - \frac{m_C}{m} \right)^2 \right)^{-1/2} \quad \sum_{i=1}^{m} \left(c_i - \frac{m_C}{m} \right) \left(\text{Rank}(T_i) - \frac{m+1}{2} \right) \tag{4.16}$$

This classic statistic has been employed as an non-parametric approach to other regression problems in genetics including quantitative trait loci (QTL) (Haley and Knott 1992;

Kruglyak and Lander 1995; Zou et al. 2003). Alternative weights have also been proposed for non-parametric regression but are not explored here (Puri and Sen 1985).

As noted in discussions of SAFE with hard categories, the classical tests for this non-parametric regression statistic can not be applied to gene expression data, because of the correlation among genes. If we state the null hypothesis as having no increase in differential expression on average among the unknown true category members $\{i : i \in C\}$, hypothesis tests can be based on several resampling methods. Array permutation tests would be appropriate for experimental designs where an induced value of no association between the response and every gene is the null hypothesis one wants to test departures from. In this resampling scenario, the following global statistic will be rank invariant to (4.16), under a fixed soft category, $\mathbf{c}$, that is assumed by array permutation.

$$U = \sum_{i=1}^{m} c_i \cdot \text{Rank}(T_i) \tag{4.17}$$

and thus sufficient for obtaining an empirical p-value. As noted previously for hard categories, any FWER or FDR controlling procedures should avoid pooling before generating empirical p-values, because the variances of (4.17) across category will depend on the unknown correlations among gene members. Bootstrap-based tests of (4.17) can also be conducted in manner similar to Chapter 3.6. In the case of multiple classes of differentially expressed genes, the unconditional expected value of (4.17) can only be determined

if probability of gene membership is independent from differential expression

$$\begin{aligned} E_{H_0}[U] &= E\left[\sum_{i=1}^{m} c_i \cdot \text{Rank}(T_i)\right] \\ &= \sum_{i=1}^{m} E[c_i] \cdot E[\text{Rank}(T_i)] \\ &= \frac{m_C}{m} \cdot \sum_{i=1}^{m} E[\text{Rank}(T_i)] \\ &= \frac{m_c \cdot (m+1)}{2} \end{aligned} \tag{4.18}$$

based on the formulation in (3.25) for $\sum E[\text{Rank}(T_i)]$ under a K-class null. The bootstrap-based tests are again determined by the exclusion of $E_{H_0}[U]$ from standard resampling-based confidence intervals.

We further note that by allowing $\{c_i\}$ to take continuous values on the interval $[0, 1]$, soft categories also provide a natural solution to the problem of multiple representations of a gene on an array that was alluded to in Chapter 2. While gene enrichment tests and the basic SAFE framework allow such genes to have more influence on the category, one could easily down-weight the corresponding probesets ($c_i = \frac{1}{k} \cdot \Pr(H_A \mid \mathbf{s})$ if a gene occurs k times) such that each gene contributes equally to determining a category's behavior. A value between 1 and $\frac{1}{k}$ could also be chosen to reflect the added certainty in multiply-spotted genes.

Data example 1: Lung Carcinoma study

For an example of an exploratory analysis of TF gene categories and differential expression, we examined a subset of the lung carcinoma dataset from Bhattacharjee et al. (2001), that has been previously used in a SAFE analysis of GO and Pfam categories

(Barry et al. 2005). Data pre-processing is described in detail in Chapter 2 and resulted in considering a filtered set of 7299 expressed genes. This analysis focuses on the subset of normal ($n_1 = 16$) and carcinoid ($n_2 = 20$) samples. PSWM were obtained from combining all 49 human motifs available from the JASPAR database along with 93 motifs from TRANSFAC that had a minimal length of $w \geq 6$ and column totals $N \geq 50$. Posterior probabilities were calculated for every upstream sequence available using the Bayesian multi-site model with hyperparameters ($\gamma_1 = 2$, $\gamma_2 = 1000$). First, the robustness of SAFE to the model prior, P_A, was examined by taking values across five orders of magnitude from $P_A = 5e-1$ to $5e-6$. Next, significant categories are presented for a reasonable model prior as illustration of the utility and biological interpretation of findings.

To generate soft gene categories for the lung carcinoma dataset, REFSEQ IDs were obtained from the annotation for the hgu95av2 Affymetrix array from Bioconductor (http://www.bioconductor.org). When mapping the posterior probabilities to the set of expressed probesets, two considerations must be made regarding duplication. When probesets are linked to multiple REFSEQ IDs with upstream sequence information, the average score is taken; conversely, when multiple probesets map to the same gene, the probabilities are down-weighted in the manner described in section 4.4.2. In this weighting scheme each gene will carry equal weight in the category, rather than each probeset.

In Figure 16A, we note that gene category size is sharply affected by the choice of P_A with the median expected category size decreasing from 1561 genes to 0.06 genes across the five orders of magnitude over which P_A varied. Despite the sensitivity of m_C to the model prior, the SAFE results are shown to be more robust. The order of significance

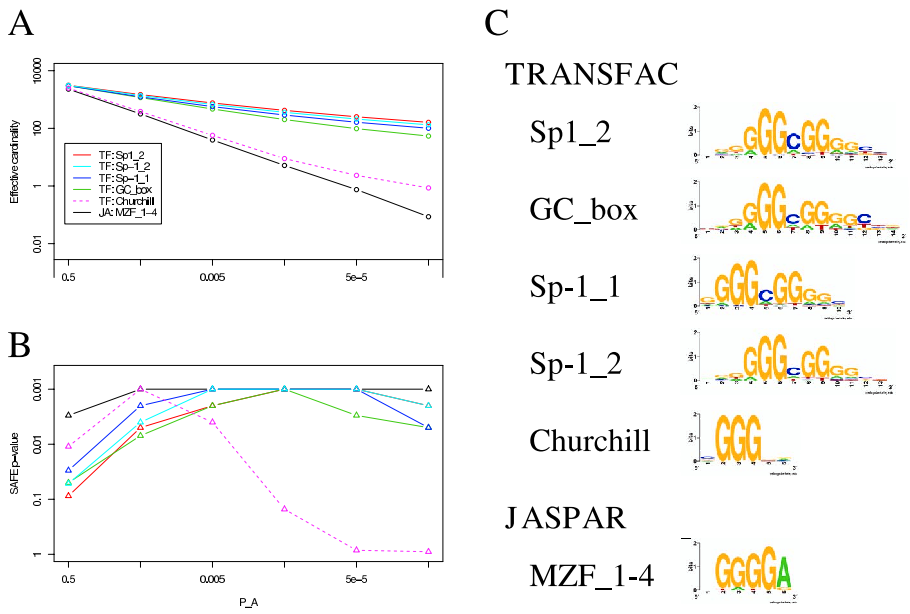

Figure 16: (**A**) The expected category size, (**B**) and SAFE permutation-based p-value is plotted against choice of model prior for five TF that remained the most significant throughout, and also one category ("Churchill") where significance level was not robust. (**C**) Sequence logos demonstrate the close relationship among four of the TRANSFAC motifs, and the lower K-L distance from background in the less stable category.

of p-values was well preserved across the range of P_A, as demonstrated by several of the most significant categories plotted in Figure 16B. The TRANSFAC TF "Churchill" is also shown as a rare instance where the SAFE results varied sharply across model prior. The sequence logos in Figure 16C demonstrates that this motif is shorter and less distinguishable from background in terms of K-L distance as the other motifs. Despite the poor performance of the "Churchill" motif, overall this illustrates the robustness to choice in model prior that results from using a regression-type global statistic for soft

categories.

The most significant categories in this analysis appear to be several related motifs for Sp-like TFs, reaching the minimum attainable empirical p-value for 1000 permutations under a moderate model prior $P_A = 0.0005$, and collectively had a Yekutieli-Benjamini FDR estimate < 0.05. We feel that this model prior is representative of what one would expect for the number of genes being regulated by a single TF. In examining the sequence logos (Figure 16C), these motifs are highly conserved around a consensus sequence of GGGGCGGGG. These findings demonstrate the capability of SAFE in finding significant TF gene categories, but that in an exploratory analysis one must examine the sequence logos in more detail to discern whether they constitute separate biological findings.

4.4.3 Data example 2: a leukemia and Down-syndrome study

As an second example of a SAFE analysis of TF regulated genes and differential expression, we analyzed a microarray dataset of leukemia patients with and without Down syndrome (DS) from Bourquin et al. (2006). Sample information and raw Affymetrix data was available for the comparison of 24 DS and 39 non-DS patients. Data was preprocessed as described (Bourquin et al. 2006), and a gene-specific SAM analysis (Tusher et al. 2001) was performed to reproduce the most significant gene-specific effects that were used as predictors by the authors.

In addition to the gene-specific analysis, the authors also conducted a category test (GSEA) for several sets of genes. The category members had been identified as homologs to genes shown to be regulated by the GATA1 TF in a murine model (Welch et al. 2004).

The authors identified a marginally significant increase in differential expression in one of the categories. This dataset provides an opportunity to use soft categories of GATA1 in a more of a hypothesis-driven SAFE analysis.

From the supplemental information to Bourquin et al. (2006), we constructed our own version of hard categories of probesets for each of the three gene sets, and also for their union. These were compared in a SAFE analysis to soft categories based on several GATA TF motifs available from the TRANSFAC database. As a second step, an exploratory analysis of TRANSFAC motifs was also performed to look for more significant results, and also potential interactions with GATA motifs. All SAFE analyses were performed with a one-sided Student's t-statistic relating to increased expression in Down Syndrome patients. This corresponds to the direction of hypotheses conducted by the original authors (albeit with a different measure).

The SAFE results are presented in Table 5. We first note that none of the permutation-based p-values are as significant as those reported by the Kolmogorov-Smirnoff type analysis done by the original authors (Bourquin et al. 2006). However, it is unknown if this difference stems from having dissimilar probeset annotations, the different mechanisms applied for multiply-spotted genes, or the fact that Kolmogorov-Smirnoff statistics, unlike the Wilcoxon rank sum, are sensitive to differences that do not necessarily relate to increased amounts of differential expression Damian and Gorfine (2004). Despite this difference in results, we observe that the more powerful bootstrap-based versions of SAFE yielded Z-scores and quantile-based p-values that are substantially more significant. The bootstrap also provides enough power to reject several categories under conservative FDR and FWER controlling procedures.

Table 5: SAFE results for murine homologs, GATA TF and other motifs from the TRANSFAC database.

Category	**Size**	**Perm.** p	**Boot.** Z	**Boot.** p
Geneset C: Erythroid genes	24	0.047	4.818	0.0001
Geneset A: Down-regulated	49	0.037	3.773	0.0003
Geneset B: Up-regulated	34	0.393	0.847	0.1968
A $\cup$ B $\cup$ C	103	0.021	4.882	0.0001
GATA-3	27.2	0.011	5.664	0.0001
GATA-1	20.1	0.161	1.836	0.0444
GATA-6	20.9	0.167	1.549	0.0704
GATA	33.1	0.226	1.309	0.0982
GATA-2	20.1	0.264	1.102	0.1395
p53	11.4	0.029	9.609	0.0001
Max	16.1	0.006	5.911	0.0001
E47	18.1	0.016	5.174	0.0001
cap	15.9	0.041	3.540	0.0010

A more interesting discovery in this analysis is that a soft category based on the "GATA-3" motif ($w = 9$ and $N = 63$), was more significant than any of the hard categories based on the homologous gene sets. This is noteworthy because while the gene sets represent a manually curation of laboratory findings, the soft category is derived entirely by computational algorithms searching for patterns in DNA sequences. Taken together, these results support the hypotheses of GATA involvement in leukemia and Down syndrome; however, we further examined of other TRANSFAC motifs, several of which were seen to produce equally or more significant results, including the TFs "p53" ($w = 10$ and $N = 98$), "MAX" ($w = 14$ and $N = 100$), "E47" motifs ($w = 16$ and $N = 100$), and a total of 12 TFs passing a Benjamini-Hochberg FDR controlling procedure for $\alpha = 0.01$. These data suggest that the differences in the patient populations may be more biologically complex than is conjectured, and serve to illustrate that caution is needed in presuming a distinct relationship between GATA TFs and Down Syndrome based on the gene category results found in Bourquin et al. (2006).

4.5 Extensions of TF scores and gene expression

4.5.1 Consideration of TF modules

In addition to the discovery of novel TFs in may different organisms, increasing attention is being given to understanding their complex interactions in controlling transcription. Model-based approaches have been developed for discovering *de novo* modules from sets of implicated sequences, but the computational strategies become more complex as one moves from yeast into higher-order eukaryotes and humans (Gupta and Liu 2005). The

difficulties stem from needing to consider longer upstream sequences, with motifs occurring as far as a few kilobases away; in addition, binding sites can be of shorter length and with a motif having less consensus; and finally, more frequent and presumably non-regulating repeats are observed in background sequence. In order to score upstream sequences across the human genome, we propose adapting the single-motif models already presented into a position-independent score of multiple motifs. In this way, testing for differential expression in jointly occurring motifs could be evidence for cis-regulation across the conditions of the experiment.

We extend the multi-site model in (4.9) to consider an alternative hypothesis that realizations of two (or more) motifs occur in a sequence $\mathbf{s}$.

$$\begin{aligned}\Pr(\mathbf{s}) &= \Pr(\mathbf{s}_{[1:L-w_1]}) \cdot \Pr(\mathbf{s}_{[L-w_1+1:L]} \mid \Theta_1) \cdot \beta_1 \\ &+ \Pr(\mathbf{s}_{[1:L-w_2]}) \cdot \Pr(\mathbf{s}_{[L-w_2+1:L]} \mid \Theta_2) \cdot \beta_2 \\ &+ \Pr(\mathbf{s}_{[1:L-1]}) \cdot \Pr(\mathbf{s}_{[L:L]} \mid \theta_0) \cdot (1 - \beta_1 - \beta_2) \end{aligned} \tag{4.19}$$

where Θ_1 and Θ_2 are derived from the PSWMs of the respective motifs. In order to compute the conditional probability of $\mathbf{s}$ under the alternative hypothesis, the events of background alone, or only one motif occurring must be subtracted as follows

$$\Pr(\mathbf{s} \mid H_A) = \Pr(\mathbf{s}) - \Pr(\mathbf{s} \mid \Theta_1, \theta_0) - \Pr(\mathbf{s} \mid \Theta_2, \theta_0) + \Pr(\mathbf{s} \mid \theta_0) \tag{4.20}$$

where the appropriate motif abundance ratios, β_1 and β_2, are used for calculating the probabilities $\{\Pr(\mathbf{s} \mid \Theta_i, \theta_0)\}$. In computing likelihood ratios and posterior probabilities, it could be valid to test against a null hypothesis of only a background model, or also

the complement set of single-motif events

$$\Pr(\mathbf{s} \mid H_0) = \Pr(\mathbf{s} \mid \theta_0) \qquad \text{or} \tag{4.21}$$

$$\Pr(\mathbf{s} \mid H_0) = \Pr(\mathbf{s} \mid \Theta_1, \theta_0) + \Pr(\mathbf{s} \mid \Theta_2, \theta_0) - \Pr(\mathbf{s} \mid \theta_0) \tag{4.22}$$

To illustrate the potential utility of these models, we return to the leukemia and Down Syndrome dataset where several motifs were found to be significant. The joint occurrences of the most significant motifs were scored as a preliminary experiment that avoids the computational effort required to consider all $\binom{93}{2}$ possible pairs among the TRANSFAC motifs. Models with fixed abundance ratios of $\beta_1 = \beta_2 = \frac{1}{1000}$ and the null in (4.21) were run, but future consideration is warranted as to when each null model would be appropriate. The following table gives the pairwise permutation and bootstrap SAFE results for the considered TRANSFAC motifs.

These results demonstrate the ability of SAFE to identify potential interaction of multiple TFs in regulating gene expression. In particular p53 and GATA-3 showed an interesting result, both in having a larger number of joint occurrences than expected and also a nominally significant amount of DE (Table 6). This result may be of biological interest such that further investigation is warranted. Although it would not be necessary for cis-regulation, if the predicted sites of the two motifs occur in a non-random manner in the upstream sequences, this could be seen as additional evidence of a biological event.

It should be noted that one may be more interested in identifying motif interactions where the marginal results are not significant. However, considering the $\frac{L \cdot (L-1)}{2}$ pairs, or the even greater number of higher order effects, becomes a computation challenge for the long upstream sequences and whole-genome scans. For this reason we did not

Table 6: SAFE results for select pairs of TRANSFAC TFs from the single-motif analysis of the leukemia Down-syndrome dataset from section 4.4.3.

TF pair	**Size**	**Perm.** p	**Boot.** Z	**Boot.** p
p53 + Max	18.6	0.383	0.461	0.308
p53 + GATA-3	186.9	0.026	2.618	0.008
p53 + E47	37.5	0.110	1.865	0.041
Max + GATA-3	61.8	0.137	1.253	0.104
Max + E47	17.1	0.278	0.625	0.269

implement an exhaustive search, but note that the process can be highly parallelized across either the number of motif pairs or the number of upstream sequences. We suggest this framework would be very useful for any exploratory analysis that identifies multiple significant motifs.

4.5.2 An iterative approach to updating PSWMs

Throughout this chapter, we have focused on treating PSWMs as fixed and known while developing a hypothesis testing framework for gene expression data. Here we propose and give a preliminary example of an algorithm for updating Θ based on the joint information of a gene's upstream sequence and measures of differential expression from a

DNA microarray experiment.

The TF binding motifs from the JASPAR and TRANSFAC databases are identified from different types of experimentation, from *in vitro* assays that immunoprecipitate bound oligomers (Funk et al. 1992) to computational algorithms for finding patterns in longer implicated sequences from a literature search or experiment (Matys et al. 2003). With such different sources of historical data, it would be difficult to design a fully Bayesian approach that would both allow for different variability in priors, and would also be of a conjugate form such that posterior probabilities can be computed in a timely manner when considering whole genomes. As a different approach, we propose an iterative algorithm that updates the position-specific weight matrix from initial values of the parameters $\Theta^{(0)}$ by sampling from either the genome-wide upstream sequences, or a subset selected by external data (*e.g.* differential expression).

In the first stage for updating Θ, a motif model from section 4.2 gives the probability of having at least one true motif based on the sequence information alone, $\Pr(H_A \mid \mathbf{s}, \Theta^0, \theta_0)$. Bayes theorem can then used to get the probability of the alternative hypothesis conditional on both sequence information and a local statistics as a measure of differential expression, t.

$$\Pr(H_A \mid t, \mathbf{s}, \Theta^{(i)}, \theta_0) \quad = \quad \frac{\Pr(t \mid H_A) \cdot \Pr(H_A \mid \mathbf{s}, \Theta^{(i)}, \theta_0)}{\Pr(t \mid H_A) \cdot \Pr(H_A \mid \mathbf{s}, \Theta^{(i)}, \theta_0) + \Pr(t \mid H_0) \cdot \Pr(H_0 \mid \mathbf{s}, \Theta^{(i)}, \theta_0)} \tag{4.23}$$

This formulation requires specifying the distribution of local statistics under both the null and alternative hypothesis of TF motif presence. In many basic experimental designs, the distributions can be taken as mixtures of a central and non-central distributions of

simple test statistics, where the mixing parameters relate to the sensitivity and specificity of motif presence resulting in regulation and differential expression.

Once the probability of motifs is updated, sequences can be sampled based on the posterior probability, along with the particular start sites using the following joint distribution Let the indicator $A_i = 1$ if a true motif starts at the i-th position of $\mathbf{s}$, then the joint distribution of start sites can be decomposed into the following marginal conditional probabilities

$$\begin{aligned} Pr(A_1 \dots A_{L-w+1} \mid \mathbf{s}) &= Pr(A_{L-w+1} \mid \mathbf{s}) \times Pr(A_{L-w} \mid \mathbf{s}, A_{L-w+1}) \times \dots \\ &\times \quad Pr(A_1 \mid \mathbf{s}, A_2 \dots A_{L-w+1}) \end{aligned} \tag{4.24}$$

where the probability of the last position being in a true motif is

$$\begin{aligned} Pr(A_{L-w+1} = 1 \mid \mathbf{s}) &= \frac{Pr(\mathbf{s} \cap A_{L-w+1} = 1)}{Pr(\mathbf{s})} \\ &= \frac{Pr(\mathbf{s}_{[1:L-w]}) \cdot \beta \cdot Pr(\mathbf{s}_{[L-w+1:L]} \mid \Theta)}{Pr(\mathbf{s})} \end{aligned} \tag{4.25}$$

and the conditional probabilities of upstream start sites take the form

$$Pr(A_i = 1 \mid \mathbf{s}, A_{i+1} \dots A_{L-w+1}) \quad =$$

$$\begin{cases} 0 \quad \text{if } A_{i+1} = 1 \cup A_{i+2} = 1 \dots \cup A_{i+w-1} = 1 \\ \\ \frac{Pr(\mathbf{s}_{[1:i-1]}) \cdot \beta \cdot Pr(\mathbf{s}_{[i:i+w-1]} \mid \Theta)}{Pr(\mathbf{s}_{[1:i+w-1]})} \quad \text{otherwise} \end{cases} \tag{4.26}$$

In this way, start sites can be sampled in a backward or forward manner from an implicated sequence using probabilities that have already be defined in (4.7) and thus would require no further computation once $Pr(H_A \mid \mathbf{s})$ is obtained. Algorithmically,

the set of start sites in an implicated sequence are determined by first considering if $A_{L-w+1} = 1$; if so, one jumps to $A_{L-2 \cdot w+1}$ as the next nearest possible start site, otherwise A_{L-w} is the next position to be considered. An update to Θ is then obtained from the frequency counts in the resampled PSWM. Lastly, we note that a Dirichlet prior can be put on Θ such that the updated multinomial probabilities become the weighed sum of frequency counts and hyperparameters (Liu et al. 2002)

To demonstrate the potential ability of this algorithm to refine motif PSWM estimates, a small microarray dataset from Yoon et al. (2002) is used as evidence of TF regulation. In the study, homologous recombination was used to knock out either one or both copies of the p53 gene in a human cell line. A linear association is assumed between p53 copy number and activity so that the gene-specific measure of differential expression is a one-sided t-statistic from a simple linear regression model.

Gene expression data was preprocessed as described and mapped to REFSEQ IDs using Bioconductor annotation package. The p53 gene ("NM_000546") was confirmed to be highly differentially expressed as previously published ($t = 9.54$, $p = 0.0003$). Although severely limited by the number of unique resamples, a bootstrap-based SAFE analysis showed a p53 motif from TRANSFAC is marginally significant ($p = 0.015$), suggesting that information on differential expression might improve the estimated motif.

As a initial run of the iterative procedure, 200 updates were generated from the starting values of the p53 PSWM from TRANSFAC (Figure 17). Sequence scores were obtained from the Bayesian multi-site model with $\gamma_1 = 2$ and $\gamma_2 = 1000$ and a model prior of $P_A = 0.01$. To decrease the computational time, a single update of β was obtained from 4.12 and then posterior probabilities in each update are obtained from the multi-

site model conditional on β. Preliminary results suggest that the estimate of β remains largely unchanged across iterations.

Posterior probabilities of regulation were obtained based on the observed t-statistic, assuming it has a central distribution under the null, and a non-centrality parameter of $\delta = 2$ under the alternative. 50 implicated sequences were then sampled with replacement, and an unspecified number of start sites were obtained from (4.24). Figure 17 shows a general convergence of the algorithm to a new motif. Further, the K-L distance to the background increased during the iterations, representing that a more distinct motif was identified. Because the motif changed substantially from the starting PSWM, consideration should be given as to whether the specific binding site is being identified. Since this approach is dependent on the differential expression data, it will be of benefit to verify that for any microarray experiment direct regulation by the TF factor occurs in the most significantly DE genes.

4.6 Discussion

In this chapter, we have developed a new approach to hypothesis tests of functional categories and gene expression data. By allowing a probabilistic measure for category membership, SAFE can be extended to biological situations where function is less well understood, or resulting in a more continuous measure.

To incorporate transcription factor analysis into SAFE, we derived probabilistic measures of TF regulation that are based on the presence of known motifs in the surrounding sequences of genes. Models used for motif discovery can be readily adapted to this

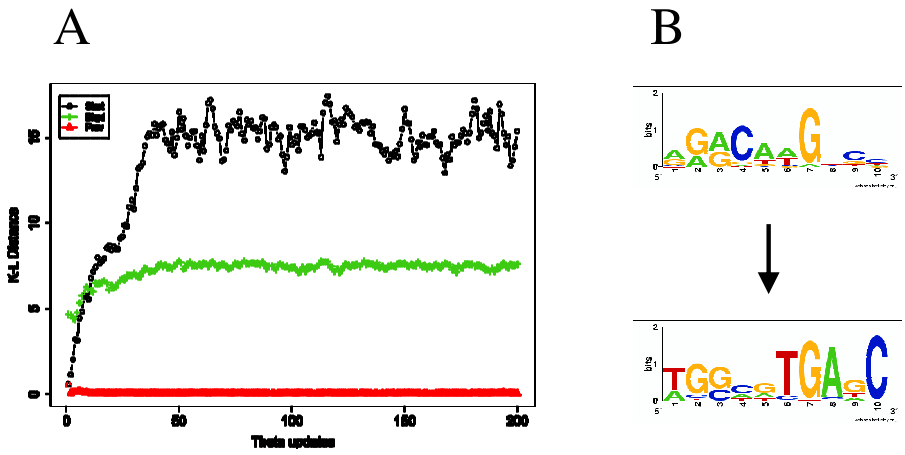

Figure 17: (**A**) Kullback-Leibler distance of the updated PSWM to the starting motif, background, and the previous iteration. (**B**) Sequence logos demonstrate the difference in the starting and final motif.

problem, and we note that future developments of models that better relate to the underlying biology could be implemented and lead to better estimates of gene regulation. For instance, parameters for the position of motif start sites could be incorporated if a consistent pattern of location were to be observed. This may also be important when addressing the multiple motifs thought to occur cis-regulated modules.

In addition, the concept of "soft categories" can allow SAFE to be extended to other types of functional annotation, such as chromosomal location where a continuous measure like physical or linkage map distance may provide more power to detect causal events of differential expression that are based on a particular locus (*e.g.,* an unknown amount

of loss of heterozygocity). This would also provide a framework for considering multiple and/or conflicting sources of annotation that define gene function. The implementation of SAFE to these scenarios will be highly determined by the data structures of the particular problem and quality of information.

Lastly, we note that the proposed use of differential expression data to improve motif estimation offers a novel mechanism that could seen by biologists as an interesting way of combining the two sources of information into both motif discovery and better understanding transcriptional regulation. Further work is necessary to establish the appropriate manner of parameterizing the model and sampling updated motif start sites. First, simulations can create random measures of differential expression for the true null and alternative gene sequences that were generated in section 4.3. This would allow us to understand the capability of the algorithm to converge on local or global maximum in the likelihood surface of the ProductMultinomial. The effect of details like the Dirichlet prior and sampling scheme for start sites can also be better established in the controlled setting of simulation. In real sequence and expression data, it will be important to learn if this approach can refine true motifs rather than converging on other known departures from background in the sequences that possibly reflect other less interesting features of DNA structure (*e.g.,* single-nucleotide repeats or the TATA box for RNA polymerase binding). Also, because one might expect that in studies involving cell culture and biological samples only a small fraction of the differential expressed genes result from direct TF activation, it may be important to gather information from multiple experiments and designs to be able to distinguish the genes that are directly regulated.

REFERENCES

Agudo, D., Gómez-Esquer, F., Martínez-Arribas, F., Núñez-Villar, M. J., Pollán, M., and Schneider, J. (2004). Predictive makers and cancer prevention Nup88 mRNA overexpression is associated with high aggressiveness of breast cancer. *Int J Cancer*, 109(5):717–720.

Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20:578–580.

Allison, D. B., Cui, X. Q., Page, G. P., and et al. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and G., S. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25–29.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36.

Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54.

Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949.

Beißbarth, T. and Speed, T. P. (2004). Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

Ben-shaul, Y., Bergman, H., and Soreq, H. (2005). Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21:1129–1137.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, 57:289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate under dependency. *Ann Stat*, 29:1165–1188.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74:457–468.

Berriz, G. F., King, O. D., Bryant, B., Sander, C., and Roth, F. P. (2003). Characterizing gene sets with FuncAssociate. *Bioinformatics*, 19:2502–2504.

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

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31:365–370.

Boorsma, A., Foat, B. C., Vis, D., Klis, F., and Bussemaker, H. J. (2005). T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Research*, 33:W592–W595 Suppl. S JUL 1 2005.

Bourquin, J. P., Subramanian, A., Langebrake, C., Reinhardte, D., Bernard, O., Ballerini, P., Baruchel, A., Cave, H., Dastugue, N., Hasle, H., Kaspers, G. L., Lessard, M., Michaux, L., Vyas, P., wering, E. V., Zwaan, C. M., Golub, T. R., and Orkin, S. H. (2006). Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 103:3339–3344.

Braga-neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20:374–380.

Breslin, T., Eden, P., and Krogh, M. (2004). Comparing functional annotation analyses with catmap. *Bmc Bioinformatics*, 5:193.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares Jr, M., and Haussler, D. (2001). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267.

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Gen*, 21:33–37.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury, Australia, second edition.

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H. Y., He, Y. D. D., veer, L. J. V., Bartelink, H., de rijn, M. V., Brown, P. O., and de vijver, M. J. V. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102:3738–3743.

Chen, Y., Dougherty, E. R., and Bittner, M. (1997). Ratio-based decisions and the quantitative analysis of DNA microarray images. *Biomedical Optics*, 2:364–374.

Chu, T. M., Weir, B. S., and Wolfinger, R. D. (2004). Comparison of li-wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays. *Bioinformatics*, 20(4):500–506.

Cox, D. R. (1972). Regression models in life tables (with discussion). *J R Statist Soc B*, 34:187–220.

Cui, X. G., Hwang, J. T. G., Qiu, J., and et al. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6:59–75.

Damian, D. and Gorfine, M. (2004). Statistical concerns about the GSEA procedure. *Nature Genetics*, 36:663–663.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7.

Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, 81(2):98–04.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.

Dudoit, S., Yang, Y. H., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26.

Efron, B. (1981). Censored-data and the bootstrap. *Journal Of The American Statistical Association*, 76:312–319.

Efron, B. (1987). Better bootstrap confidence-intervals. *Journal Of The American Statistical Association*, 82:171–185.

Efron, B. and Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Chapman and Hall/CRC, New York, New York, second edition.

Eisen, M. B., Spellman, P. T., O., B. P., and Botstein, D. (2001). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868.

Fraser, A. G. and Marcotte, E. M. (2004). A probabilistic view of gene function. *Nat Genet*, 36(6):559–564.

Funk, W. D., Pak, D. T., Karas, R. H., W.E., W., and Shay, J. W. (1992). A transcriptionally active DNA-binding site for human p53 protein complexes. *Mol Cell Biol*, 12(6):2866–2871.

Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999). Ploidy regulation of gene expression. *Science*, 285(5425):251–254.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1):1–77.

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

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Gupta, M. and Liu, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102:7079–7084.

Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests*. Academic Press, New York, first edition.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative traints in line crosses using flanking markers. *Heredity*, 69:315–324.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods*. Wiley, New York, New York, second edition.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandivanian Journal of Statistics*, 6:65–70.

Honore, B., Ostergaard, M., and Vorum, H. (2004). Functional genomics studied by proteomics. *Bioessays*, 26(8):901–915.

Hosack, D. A., Dennis Jr., G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol*, 4(10):R70.

Hu, J. and Wright, F. A. (2005). Detecting differential gene expression in oligonucleotide arrays using a mean-variance model. In submission.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Bioinformatics*, 4(2):249–264.

Jenson, S. T., S., L. X., Zhou, X., and Liu, J. S. (2004). Computational discovery of gene regulatory binding motifs: A Bayesian perspective. *Statistical Science*, 19(1):188–204.

Kanehisa, M. (1997). A database for post-genome analysis. *Trends Genet*, 13(9):375–376.

Kau, T. R., Way, J. C., and Silver, P. A. (2004). Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nat Rev Cancer*, 4(2):106–117.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201.

Kim, C. C. and Falkow, S. (2003). Significance analysis of lexical bias in microarray data. *BMC Bioinformatics*, 4(1):12.

Kim, S.-Y. and Volsky, D. J. (2005). Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res*, 13(4):703–716.

Kohane, I. S., Kho, A. T., and Butte, A. J. (2003). *Microarrays for an integrative genomics*. The MIT Press, Cambridge, Massachusetts, first edition.

Kruglyak, L. and Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139:1421–1428.

Landgrebe, J., Wurst, W., and Welzl, G. (2002). Permutation-validated principal components analysis of microarray data. *Genome Biol*, 3(4):RESEARCH0019.

Lathe, R. (1986). Nomenclature for incompletely specified bases in nucleic-acid sequences - recommendations 1984. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 83:1–5.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins-structure Function And Genetics*, 7:41–51.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14:1085–1094.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index omputation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31–36.

Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M. A. (2003). NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*, 31(1):82–86.

Liu, J. S., Gupta, M., Liu, X. S., and Lawrence, C. L. (2002). *Case Studies in Bayesian Statistics*, chapter Statistical models for motif discovery. Springer-Verlag, New York.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J Amer Statist Assoc*, 90:1156–1170.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-margoulis, O. V., Kloos, D. U., Land, S., Lewicki-potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). Transfac (R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31:374–378.

Michl, P., Barth, C., Buchholz, M., Lerch, M. M., Rolke, M., Holzmann, K. H., Menke, A., Fensterer, H., Giehl, K., Lohr, M., Leder, G., Iwamura, T., Adler, G., and Gress, T. M. (2003). Claudin-4 expression decreases invasiveness and metastatic potential of pancreatic cancer. *Cancer Res*, 63(19):6265–6271.

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

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol*, 8(1):37–52.

Nichols, L. S., Ashfaq, R., and Iacobuzio-Donahue, C. A. (2004). Claudin 4 protein expression in primary and metastatic pancreatic cancer support for use as a therapeutic target. *Amer J Clin Pathol*, 121(2):226–230.

Nishioka, M., Kohno, T., Tani, M., Yanaihara, N., Tomizawa, Y., Otsuka, A., Sasaki, S., Kobayashi, K., Niki, T., Maeshima, A., Sekido, Y., Minna, J. D., Sone, S., and Yokota, J. (2002). MYO18B, a candidate tumor suppressor gene at chromosome 22q12.1, deleted, mutated and methylated in human lung cancer. *Proc Natl Acad Sci U S A*, 99:12269–12274.

Pavlidis, P., Qin, J., Arango, V., Mann, J. J., and Sibille, E. (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, 29:1213–1222.

Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8:250–254.

Polansky, A. M. and Schucany, W. R. (1997). Kernel smoothing to improve bootstrap confidence intervals. *J R Statist Soc B*, 59(4):821–838.

Puri, M. L. and Sen, P. K. (1985). *Nonparametric methods in general linear models*. Wiley, New York, New York, first edition.

Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat Appl Genet Mol Bio*, 3(1):16.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed

genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and B., L. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32:D91–D94.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, 270(5235):467–470.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18:6097–6100.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, 28:405–420.

Staden, R. (1990). Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*, 5(1):89–96.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat*, 31:2013–2035.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proc Natl Acad Sci U S A*, 100:9440–9445.

Tani, M., Ito, J., Nishioka, M., Kohno, T., Tachibana, K., Shiraishi, M., Takenoshita, S., and Yokota, J. (2004). Correlation between histone acetylation and expression of the MYO18B gene in human lung cancer cells. *Genes Chromosomes Cancer*, 40(2):146–151.

Thomas, G. B. J. and Finney, R. L. (1992). *Maxima, Minima, and Saddle Points*. Calculus and Analytic Geometry. Addison-Wesley, Reading, MA, eighth edition.

Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S., and Lawrence, C. E. (2004). Decoding human regulatory circuits. *Genome Research*, 14:1967–1974.

Troendle, J. F., Korn, E. L., and Mcshane, L. M. (2004). An example of slow convergence of the bootstrap in high dimensions. *American Statistician*, 58:25–29.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A*, 100:8348–8353.

Troyanskaya, O. G., Garber, M., Brown, P. O., Botstein, D., and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461.

Tsai, C. A., Hsueh, H. M., and Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics*, 59:1071–1081.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121.

Virtaneva, K. I., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de la Chapelle, A., and Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci U S A*, 98(3):1124–1129.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics*, 26:225–228.

Welch, J. J., Watts, J. A., Vakoc, C. R., Yao, Y., Wang, H., Hardison, R. C., Blobel, G. A., Chodosh, L. A., and Weiss, M. J. (2004). Global regulation of erythroid gene expression by transcription factor gata-1. *Blood*, 104:3136–3147.

Westfall, P. H. and Young, S. S. (1989). P-value adjustment for multiple tests in multivariate binomial models. *J Amer Statist Assoc*, 84:780–786.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing : examples and methods for P-value adjustment*. Wiley, New York, New York, first edition.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol*, 8(6):625–637.

Yates, F. (1984). Tests of significance for 2 x 2 contingency tables. *J R Statist Soc A*, 147:426–463.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Statist Plann Inference*, 82:171–196.

Yoon, H., Liyanarachchi, S., Wright, F. A., Davuluri, R., Lockman, J. C., de la Chapelle, A., and Pellegata, N. S. (2002). Gene expression profiling of isogenic cells with different TP53 gene dosage reveals numerous genes that are affected by TP53 dosage and iden-

tifies CSPG2 as a direct target of p53. *Proc Natl Acad Sci U S A*, 99(24):15632–15637.

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

Zhang, B., Schmoyer, D., Kirov1, S., and Snoddy, J. (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16.

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

Zhou, X., Kao, M. C., and Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A*, 99(20):12783–12788.

Zou, F., Yandell, B. S., and Fine, J. P. (2003). Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics*, 165(3):1599–1605.