The Pennsylvania State University
The Graduate School
Eberly College of Science

THREE NOVEL PROCEDURES TO CONTROL THE FALSE DISCOVERY RATE

A Dissertation in
Statistics
by
Daisy Philtron

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2014

The dissertation of Daisy Philtron was reviewed and approved* by the following:

Debashis Ghosh
Professor of Statistics and Public Health Sciences
Dissertation Adviser, Chair of Committee

Qunhua Li
Professor of Statistics

Murali Haran
Associate Professor of Statistics

Ross Hardison
Professor of Biochemistry and Molecular Biology

Aleksandra Slavkovic
Associate Professor of Statistics
Chair of Graduate Program

*Signatures are on file in the Graduate School.

Abstract

The field of multiple testing has seen a resurgence in the last twenty years after the seminal work of Benjamini and Hochberg (1995) that introduced the false discovery rate. With the proliferation of high-throughput data generation and very large-scale simultaneous testing problems in the arena of genetic analysis, the development of procedures to control the false discovery rate has taken on increased importance. In this dissertation we introduce three novel procedures with this specific goal. Each procedure is specifically tailored for a different situation in multiple testing.

The first procedure controls the false discovery rate when hypotheses are tested using next-generation sequencing data without experimental replication. In this situation the p-values used are discrete and as a result classical error control procedures are conservative. Existing approaches that are specifically for use with discrete p-values require the complete specification of each p-value's distribution function. When a small number of p-values have complicated distribution functions these approaches can be very slow. Our proposed procedure offers good error control properties, comparable power properties, and a computational advantage over existing procedures.

We further propose a procedure developed specifically to test the disjunction hypothesis, wich is appropriate when each gene or location studied is associated with multiple p-values of individual interest. This can occur when more than one aspect of an underlying process is measured. For example, cancer researchers may hope to detect genes that are both differentially expressed on a transcriptomic level and show evidence of copy number aberration. Currently used methods of p-value combination for this setting are overly conservative, resulting in very low power for detection. We introduce a method to test the disjunction hypothesis by using cumulative areas from the Voronoi diagram of two-dimensional vectors of p-values. Our method offers much improved power over existing methods, even in challenging situations, while maintaining appropriate error control.

Finally we introduce a non-parametric procedure to control the false discovery

rate while identifying reproducibility from the results of replicated high-throughput experiments. Experiments of this type are important because their results can identify sets of genes or binding sites for focused follow-up studies, however the variability from one experiment to another presents a well-known difficulty to researchers. We present a novel procedure to identify genes with consistent signals across replicated experiments. This procedure makes no model assumptions about reproducible genes and is free of tuning parameters. We show that it has good error control and power properties in a variety of different settings, as well as show some theoretical results.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The completion of this dissertation is a major accomplishment, however the journey to this point encompasses much more than this single document. Many individuals have contributed to my experience over the past five years at Penn State, and I would like to take time to acknowledge and thank them for their help, expertise, and encouragement.

First, thank you to my adviser Dr. Debashis Ghosh. Dr. Ghosh was always knowledgeable, quick to offer resources, and kind in his criticism of my burgeoning scholarship. Just as important, he was willing to listen to my crazy half-baked ideas, and always encouraging of my many endeavors.

Thank you also to my committee members for their excellent feedback after my comprehensive exam and for their time spent reviewing this dissertation. In particular, I appreciate the weeks of work that Dr. Qunhua Li spent working with Dr. Ghosh and me while the MaRR procedure was under development.

I would also like to acknowledge the excellent statistics professors at Penn State who went above and beyond. In particular, Dr. Naomi Altman has been very helpful for the development of Chapter 2 on discrete testing procedures. Also, Dr. David Hunter, Dr. Murali Haran, and Dr. Sesa Slavkovic have all been instrumental in my success at Penn State by making themselves available for all sorts of questions from coursework to personal choices.

Of course, my friends and classmates more than deserve an acknowledgment as well. We leaned on each other during our courses, our qualifying exams, and our research. I will greatly miss their company, sense of humor, and game nights. I sincerely could not have done this without them.

Thank you to my long-time mentor Millie Johnson for repeatedly telling me to stick with it and finish.

Thank you to my families. To my nuclear family: Mom, Dad, and Camye for developing my character and not getting mad when I was too busy for phone calls. To my Spanish family: Xiomara, Javier, and Leonor for the best winter trips anyone could imagine. Thank you to my husband: Dr. Jason Philtron, for his

support, patience, and willingness to go on adventures.

Finally, I want to acknowledge the unusual role that my bicycle has played in the completion of this document. I wrote this dissertation as my husband and I pedaled from Alaska to California, mostly in the tent at the end of each day of riding. It took 99 days and over 4,000 miles to complete the first draft, but we did it. Through it all, the bicycle did not get a single flat tire, and it safely carried my laptop and the many precious files needed to complete a dissertation.

Chapter 1
Introduction

Hypothesis testing is one of the core concepts of statistical practice. While it holds great power to make scientific discoveries, every test carries with it the potential for error. In the case of testing a single hypothesis, there are two types of error that a careful researcher must consider: Type I, and Type II. The term 'Type I' is used the describe the error that occurs when a true null hypothesis is falsely rejected. 'Type II' describes the error occurring when the alternative hypothesis is true but the null hypothesis is erroneously not rejected. Typically when we discuss error control, we are interested in the consideration of Type I errors. Classical hypothesis testing procedures control the probability of committing a Type I error while minimizing that of committing a Type II error.

When multiple hypotheses are tested simultaneously, adjustments must be made to ensure that Type I error rates are controlled for the entire set of hypotheses. Until recently, these sets of hypotheses were of relatively small size: less than a dozen. Control of the Family-Wise Error Rate (FWER) using a Bonferroni adjustment or related procedure, was generally adequate for the problems considered. With the rapid increase over the last twenty years of 'big data' and high-throughput genetic sequencing technology, the need for more powerful procedures has grown. Nearly 20 years ago in a landmark paper, Benjamini and Hochberg defined the False Discovery Rate (FDR) as an alternative to the FWER and revolutionized multiple testing.

The original FDR controlling procedure is the celebrated Benjamini-Hochberg (B-H) procedure. A large body of literature has been developed in the past twenty years proving properties of and describing modifications to this procedure. Modifications have been developed both boost power and extend the procedure's utility.

Additionally, alternative approaches have been developed and are in wide usage. Some of these approaches define and control error rates that are closely related to FDR, such as the positive false discovery rate (pFDR), the marginal false discovery rate (mFDR), and the local false discovery rate (lfdr).

In this dissertation we introduce three new procedures to control the false discovery rate in three specific settings. Chapter 2 describes the Adapted Generalized Benjamini-Hochberg procedure, which controls FDR when p-values are discrete and calculated using next-generation sequencing data. Chapter 3 proposes a procedure to control FDR using Voronoi diagrams when each hypothesis tested is associated with two p-values, each of which are of individual interest. Finally, Chapter 4 introduces the Maximum Rank Reproducibility (MaRR) procedure, a non-parametric approach to identifying reproducible genes from the results of replicated high-throughput experiments.

Each of these procedures is tailored for a different setting, and thus requires a separate literature review and background discussion. The common thread that unites them is the ultimate goal of false discovery rate control. For that reason, we formally define quantities related to Type I error control in Section 1.1, as well as briefly discuss related literature. We leave discussion of procedure–specific background to the relevant chapters.

1.1 Type I Error Rates

In this section we define and discuss Type I error rates in the context of simultaneous hypothesis testing. We also include a literature review of both modern and classical approaches to controlling these quantities.

Consider the possible outcomes for m simultaneous hypothesis tests presented in Table 1.1. In this table the rows represent the 'truth' with respect to the null hypothesis, and the columns represent the possible decisions from a hypothesis test. The numbers in each cell represent the number of tests whose outcomes agree with the corresponding column, and which pertain to the corresponding row. Using this notation, U is the number of truly null hypotheses that were correctly not rejected, V is the number of false rejections, T is the number of hypotheses that were not rejected when they should have been, and S is the number of correctly rejected hypotheses. R is the total number of rejections made.

Table 1.1: Decision outcomes for m hypothesis tests

	Fail to reject null	Reject null	Total
Null is true	U	V	m_0
Null is false	T	S	$m - m_0$
Total	$m - R$	R	m

Using this notation, we can now formally define error rates for simultaneous hypothesis testing.

1.1.1 The Family–Wise Error Rate

Perhaps the most well-known error rate for multiple testing problems, the family–wise error rate (FWER) has been in wide usage for decades. It is defined to be the probability of making *at least one* false rejection.

Definition 1. *Using the notation from Table 1.1, the family–wise error rate is defined as*

$$FWER = P(V > 0).$$

One of the most commonly used approaches to FWER control at a nominal level of α is application of a Bonferroni correction: null hypotheses are only rejected if their associated p-values are less than α/m. Clearly, for large m this value will be very small and the power of the tests will be compromised. Many related procedures were developed to boost power in this setting, including the Holm's sequentially-rejective Bonferroni method [2], a step-up procedure proposed by Hochberg (1988) [3], and Hommel's procedure [4]. A summary of these methods is provided by Shaffer (1995) [5].

The procedures mentioned above all use p-values as a basis for inference. When these p-values are discrete, however, these procedures are known to be conservative. For this reason, Tarone (1990) [6] developed an error control procedure specifically for discrete p-values. Subsequent authors have built on this procedure to further control the false discovery rate. These procedures are described in more detail in Chapter 2.

1.1.2 The False Discovery Rate

Before the advent of big data and widespread use of high-throughput genetic sequencing technology, m was typically less than a dozen. In this context, strong control of the FWER still allowed for an adequate level of power for most purposes. However, in many current biological applications m is in the thousands or tens of thousands, and controlling the FWER results in very low power for detection of true alternative signals. In the place of the FWER, we thus consider the expected proportion of false rejections: the false discovery rate (FDR) as introduced by Benjamini and Hochberg (1995) [7].

Definition 2. *Using the notation from Table 1.1, the false discovery rate is defined as*

$$FDR = E\left(\frac{V}{R}\right) = E\left(\frac{V}{R}\middle| R > 0\right) \cdot P(R > 0).$$

FDR increases the power to detect true alternative signals by allowing a controlled proportion of Type I errors. In this fashion, more errors are considered acceptable as more true signals are detected. The classical approach to FDR control is a procedure using ordered p-values that was proposed by Benjamini and Hochberg (1995) [7]. We describe this procedure in this section, as we use it throughout the subsequent chapters. Again using notation from Table 1.1, assume m hypotheses are tested, each of which is associated with a p-value $p_1, \ldots, p_m$. Assume m_0 of these p-values are associated with truly null hypotheses. To control FDR at a nominal level of $\alpha m_0/m$, the Benjamini-Hochberg (B–H) procedure is as follows:

B-H Procedure:
Order p-values as $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$. Define $\hat{k}$ as

$$\hat{k} = \max_{i=1,\ldots,m} \left\{ p_{(i)} \leq \frac{i}{m}\alpha \right\} \tag{1.1}$$

If $\hat{k}$ is well-defined, reject all hypothesis associated with $p_{(1)}, \ldots, p_{(\hat{k})}$, else reject nothing.

As originally proposed, the B–H procedure is suitable for continuous, independent p-values. In later work, it was shown by Benjamini and Yekutieli (2001) [8] to be valid for sets of p-values that satisfied certain positive dependency conditions. As with FWER, the B-H procedure is conservative when used on discrete p-values. Various modifications have been proposed to remedy this problem [9,10]. This dissertation further adds to this collection through proposal of the Adapted generalized Benjamini-Hochberg Procedure in Chapter 2.

FDR has become a widely accepted and used error rate in biology and other fields with large-scale multiple testing problems.

1.1.3 Positive False Discovery Rate

The positive false discovery rate (pFDR) was originally proposed by Storey (2002) [11] as an alternative to FDR. pFDR is defined to be the expected proportion of rejections that are false, given that at least one rejection is made.

Definition 3. *Using the notation from Table 1.1, the positive false discovery rate is defined as*

$$pFDR = E\left(\frac{V}{R}\middle| R > 0\right).$$

Storey argues that pFDR is a more useful quantity to control than FDR because of its close association with rejection regions. Researchers can define a rejection region, and then estimate the resulting pFDR if all tests with statistics contained in the region are rejected. This approach tends to provide the researcher with more information than application of the B-H procedure, which requires declaration of α, the error rate, and then determines the rejection region for p-values.

Closely related to pFDR is the use of q-values in place of p-values, as introduced by Storey and Tibshirani (2003) [12]. To understand a q-value first recall that the p-value for a single hypothesis test represents the probability of making a Type I error, and is calculated under the null hypothesis using the associated test statistic. The corresponding quantity in the false discovery rate paradigm is the q-value. The q-value for the i^{th} hypothesis is the minimum achievable pFDR if the critical region for p-values is defined as $(0, p_i)$. Thus, in the context of FDR control it may be desirable to use q-values directly. In fact, in some multiple testing settings such

as peak-callers for identification of protein binding sites, q-values are accepted and well-used.

1.1.4 Marginal False Discovery Rate

In both FDR and pFDR described above, we consider the expectation of a ratio. In contrast, the marginal false discovery rate (mFDR) as introduced by Genovese and Wasserman (2002) [13] is defined as a ratio of expectations.

Definition 4. *Using the notation from Table 1.1, the marginal false discovery rate is defined as*

$$mFDR = \frac{E[V]}{E[R]}.$$

mFDR has been shown to be asymptotically equivalent to FDR as the number of hypotheses tested increases. Specifically, $mFDR = FDR + O(m^{-1/2})$ [14]. In certain conditions, pFDR and mFDR are in fact equivalent. For this dissertation, we use estimated mFDR to control error rates in Chapter 4 when identifying reproducibility in replicate experiments.

1.1.5 Local False Discovery Rate

The local false discovery rate (lfdr) was first discussed by Efron (2004) [15]. In a broad sense, it is considered to be the Bayesian alternative to FDR control. Since the initial introduction in 2004, the topic has been visited by many authors since. In this body of literature, it is acknowledged that the bulk of p-values associated with truly null hypotheses may not follow the theoretically defined distribution of Uniform(0,1). This deviation has important implications for inference. Techniques assuming the theoretical null can result in very conservative or anti-conservative inference when this assumption is violated. To discuss lfdr, it is convenient to shift the focus from p-values in the unit interval to corresponding test statistics on real line. We thus consider the z-value associated with the hypothesis i, assuming that p_i is the corresponding p-value:

$$z_i = \Phi^{-1}(p_i) \tag{1.2}$$

Here $\Phi(\cdot)$ is the standard normal cumulative distribution function. If p_i follows the Uniform(0,1) distribution, then z_i follows the standard normal distribution. In practice, we assume that the z_i follow a mixture distribution:

$$z_i \sim f(z) = \delta f_0(z) + (1-\delta) f_1(z) \quad (i = 1, \ldots, m), \tag{1.3}$$

where f_0 is the null, or 'uninteresting' distribution and f_1 is the alternative or 'interesting' distribution. Under the theoretical null hypothesis, f_0 is the N(0,1) distribution, however in large-scale multiple testing problems the majority of values may behave differently. When this is the case, use of an empirically determined f_0 in place of $N(0,1)$ has important implications for the resulting inference. For example, if f_0 is estimated to be $N(0, 1.3)$, then inference based on the assumptions of $N(0,1)$ would result in elevated Type I error. Similarly, if the empirically estimated null is a narrower distribution such as $N(0, .75)$, procedures assuming the theoretical null will be conservative. In both cases, better inference is accomplished by estimating an empirical null distribution and using it as a basis for inference.

Efron (2004) [15] defined the *local false discovery rate* (fdr) as the posterior probability of a value z being from the null distribution, given the value of z.

Definition 5. *The local false discovery rate for a z-value z_i is defined as*

$$lfdr(z) = P\left(z_i \sim f_0 | z_i = z\right),$$

In his 2004 paper, Efron used the central peak of the histogram of observed p-values to estimate f_0. Other approaches to the estimate of δ and f_0 have been developed since then by Meinshausen and Rice (2006) [16], Jin and Cai (2007) [17], Strimmer (2008) [18], Muralidharan (2010) [19] and others. In this dissertation, we use the concept of the local false discovery rate in Chapter 3.

1.2 Dissertation objectives and layout

The goal of this dissertation is to advance the field of multiple testing by accomplishing three objectives related to false discovery rate control. Specifically, we introduce novel testing procedures for three specific situations:

G1 The simultaneous testing of hypotheses using discrete p-values arising from next-generation sequencing data when no replicates are available.

G2 The simultaneous testing of hypotheses when each gene or location studied is associated with two p-values, each of which is of individual interest.

G3 The identification of genes or locations which are consistently highly ranked in replicated high-throughput experiments.

We dedicate Chapters 2, 3, and 4 to accomplishing these objectives. In Chapter 2 we provide a literature review of discrete testing procedures before introducing the Adapted Generalized Benjamini-Hochberg procedure. In Chapter 3 we first define and discuss the disjunction hypothesis in addition to introducing the Voronoi diagram and its properties. Next we introduce a two-dimensional testing procedure that uses Voronoi diagrams to control the false discovery rate when testing the disjunction hypothesis. Chapter 4 defines the notion of reproducibility and introduces the Maximum Rank Reproducibility procedure, which is a novel and non-parametric approach to assessing reproducibility.

Chapter 2
The Adapted Generalized Benjamini-Hochberg Procedure for Next-Generation Sequencing Data

2.1 Introduction

Many procedures developed for hypothesis testing assume that the test statistic for each test follows a known, continuous distribution. The resulting p-value thus follows a Uniform(0,1) distribution. In certain settings, most notably next-generation sequencing data, these test statistics and p-values are discrete. Procedures developed for continuous p-values are conservative in these settings. For this reason, modifications of error control procedures have been developed to increase power for discrete p-values.

Specifically, Tarone (1990) [6] developed an adapted Bonferroni adjustment for discrete p-values to sharpen control of the family-wise error rate. Tarone's approach was later extended by Gilbert (2005) [10] to increase power while controlling the false discovery rate. Ferreira (2007) [9] also proposed a reformulation of the B-H procedure for discrete test statistics based on empirical distribution functions. Other related procedures have been proposed by Pounds and Cheng (2006) [20], Kulinskaya and Lewin (2009) [21], and Heyse (2011) [22]. These adaptations require specification of the full distribution of each p-value under the null

hypothesis. This can be challenging and computationally expensive when there are a very large number of tests, or when a small number of tests have p-values that can take on many possible values. This is the case with next-generation sequencing data, as the number of read counts attributed to each gene can vary from single digits to the tens of thousands in a single experiment.

The procedure described in this chapter addresses the discreteness problem through data-dependent simulations under the null hypothesis. This approach does not require specification of the p-values' distribution, offering a computational advantage in the presence of high-read genes. The approach is based on the generalized B-H procedure (Ghosh 2011 [23]), and offers a gain in power over continuous methods and displays similar power properties to other discrete methods.

This chapter proceeds as follows. Section 2.2 describes the generalized B-H procedure. Section 2.3 describes next-generation sequencing data and details the adapted generalized B-H procedure. Section 2.4 describes simulation studies performed to evaluate the performance of the proposed procedure. An application to data is described in Section 2.5. Further considerations and concluding remarks are included in Section 2.6.

2.2 The generalized Benjamini-Hochberg procedure

We first discuss the simple case of a family of testing procedures proposed by Ghosh (2011) [23]. Recall that the original B-H procedure compares ordered p-values $p_{(1)}, \ldots, p_{(m)}$ to a sequence of values $\alpha/m, \ldots, m\alpha/m$ to define $\hat{k}$:

$$\hat{k} = \max_{i=1,\ldots,m} \left\{ p_{(i)} \leq \frac{i}{m}\alpha \right\}$$

If $\hat{k}$ is well-defined, the hypotheses associated with $p_{(1)}, \ldots, p_{(\hat{k})}$ are rejected. The generalized Benjamini-Hochberg (genBH) procedure rephrases both the ordered p-values and the quantities $i\alpha/m$ in terms of spacings. This will later allow us to extend this procedure to the spacings of discrete p-values in Section 2.3.

To define p-value spacings, first consider the ordered p-values resulting from m simultaneous hypothesis tests: $p_{(1)}, \ldots, p_{(m)}$. Define their spacings as below.

Definition 6. *The i^{th} p-value spacing is defined as:*

$$\tilde{p}_i = p_{(i)} - p_{(i-1)}, \quad i = 1, \ldots, m+1,$$

where $p_{(0)} = 0$, and $p_{(m+1)} = 1$.

In the case of the genBH procedure, the original p-values are assumed independent and identically distributed as Uniform(0,1) under the null hypothesis. Thus their spacings are marginally distributed as Beta(1,m), and it is simple to calculate expectations such as $E[\tilde{p}_i] = (m+1)^{-1}$. Further, ordered p-values can be written in terms of spacings:

$$p_{(i)} = \sum_{j=1}^{i} \tilde{p}_j, \quad i = 1, \ldots, m. \tag{2.1}$$

Similarly, the quantities $i\alpha/m$ can be rephrased in terms of spacings:

$$\frac{i}{m}\alpha = i\frac{1}{m+1}\frac{m+1}{m}\alpha = i\frac{m+1}{m}E[\tilde{p}_i]\alpha \approx iE[\tilde{p}_i]\alpha \quad \text{for large } m. \tag{2.2}$$

By (2.1) and (2.2) the value $\hat{k}$ from the B-H procedure can be restated:

$$\tilde{k} = \max_{i=1,\ldots,n} \left\{ i : \frac{1}{i}\sum_{j=1}^{i} \tilde{p}_j \leq \alpha E[\tilde{p}_1] \right\} \tag{2.3}$$

The GenBH procedure uses this definition and proceeds as follows.

Generalized Benjamini-Hochberg Procedure:
Reject all hypotheses associated with $p_{(1)}, \ldots, p_{(\tilde{k})}$, where

$$\tilde{k} = \max_{i=1,\ldots,n} \left\{ i : \frac{1}{i}\sum_{j=1}^{i} \tilde{p}_j \leq \alpha E[\tilde{p}_1] \right\}. \tag{2.4}$$

If $\tilde{k}$ is not well-defined, reject nothing.

2.3 Next-Generation Sequencing data and the adapted generalized Benjamini-Hochberg procedure

In this section we describe the format of next-generation sequencing data, and introduce the adapted generalized Benjamini-Hochberg (AgenBH) procedure. The data format is important to the AgenBH procedure because we use the data to estimate expectations of p-value spacings under the null hypothesis.

2.3.1 Next-generation sequencing data

Next-generation sequencing data is produced by allowing cells to produce RNA, and then measuring the number of RNA-segments attributed to known segments of the genome (genes in particular). In this fashion, researchers can compare the behavior of genes in different settings such as differing tissue types, presence or absence of cancerous tissue, or control versus treatment.

The read counts for each gene may be very small, and in many cases even be zero. In the absence of replication, p-values resulting from testing for differences in read counts may be highly discrete in nature. In this chapter we consider experiments without replication whose data has the format detailed in Table 2.1. This table summarizes the number of read counts attributed to each gene under two different conditions: Control and Treatment. For each gene, we are interested in testing the null hypotheses that there is no difference in the number of read counts between these two conditions. The values T_i, R_{i1}, and R_{i2} respectively represent

Table 2.1: Next-generation sequencing data

	Control	Treatment	Total
gene 1	R_{11}	R_{12}	T_1
gene 2	R_{21}	R_{22}	T_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
gene n	R_{n1}	R_{n2}	T_n
total	$\sum_{i=1}^{n} R_{i1}$	$\sum_{i=1}^{n} R_{i2}$	$\sum_{i=1}^{n} T_i$

the number of read counts attributed to gene i overall, under the control, and under the treatment. These values are often low, necessitating the use of exact testing methods and yielding discrete p-values.

A common approach to testing for differential expression of a single gene i with this data type is to consider the 2×2 table composed of read counts for gene i in one row and the remaining read counts in the other row. An example table of this type is included as Table 2.2. In the presence of low read counts the assumptions needed to perform a Chi-square test are not met. Thus, a Fisher's exact test may be used to generate a discrete p-value to test this hypothesis. Under the null hypothesis, $R_{i1} \sim \text{binomial}(T_i, 0.5)$.

Table 2.2: 2×2 table for a single gene g

	Control	Treatment	Total
gene g	R_{g1}	R_{g2}	T_g
all other genes	$\sum_{i \neq g} R_{i1}$	$\sum_{i \neq g} R_{i2}$	$\sum_{i \neq g} T_i$
total	$\sum_{i=1}^{n} R_{i1}$	$\sum_{i=1}^{n} R_{i2}$	$\sum_{i=1}^{n} T_i$

The read totals for each gene, T_g, can vary from zero to tens of thousands. For obvious reasons, genes with zero read counts must be excluded from analysis. In some cases, a gene showing evidence of highly differential expression will have a very large row total. In this case, the p-values for other genes with much lower read counts will be affected. Consider the 'toy' example set of read counts for four genes in Table 2.3. In this example, the p-values resulting from Fisher's exact test using the type of table in Table 2.2 are 0.007, 0.182, 0.001, and 0.207. Clearly, these p-values are affected by the differential expression of one highly expressed gene. To remedy this problem, an alternative approach is to conduct an exact binomial test using only the read counts from a single gene. In this fashion, the p-values from our toy dataset would be 1, 1, 0.07, and 0.000.

Table 2.3: A 'toy' example set of four genes.

	Control	Treatment	Total
gene 1	4	5	9
gene 2	3	2	5
gene 3	15	12	27
gene 4	360	52	412
total	382	71	453

2.3.2 The Adapted generalized Benjamini-Hochberg procedure

In this section we introduce the AgenBH procedure as an extension of the genBH procedure that is formulated specifically for next-generation sequencing data. Recall that the original genBH procedure used the expected values of spacings in order to define $\tilde{k}$. Most importantly, it took advantage of the fact that under the null hypothesis all spacings followed the same known marginal distribution. In the case of discrete p-values, this fact is no longer true. The marginal distribution of the spacings are both unknown and non-identical. Therefore, an adaptation of the genB-H procedure must allow for the $E[\tilde{p}_i]$ to be unknown and non-identical. Thus we rewrite $\tilde{k}$ from the genBH procedure below:

$$\tilde{k} = \max_{i=1,\ldots,n} \left\{ i : \frac{1}{i} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha E[\tilde{p}_1] \right\} \tag{2.5}$$

$$= \max_{i=1,\ldots,n} \left\{ i : \frac{1}{i} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha \frac{1}{i} \sum_{j=1}^{i} E[\tilde{p}_j] \right\}$$

$$= \max_{i=1,\ldots,n} \left\{ i : \sum_{j=1}^{i} \tilde{p}_j \leq \alpha \sum_{j=1}^{i} E[\tilde{p}_j] \right\}. \tag{2.6}$$

In the case of next-generation sequencing data, the $E[\tilde{p}_i]$ can be estimated under the null hypothesis from data-dependent simulations using the marginal total $T_1, \ldots, T_n$. To perform this estimation, we assume all null hypotheses are true: $R_{i1} \sim Binom(T_i, 0.5)$ for all i. For each of B iterations, we generate simulated read counts dependent on the marginal totals and according to this null distribution. These simulated read counts can then be used to calculate a set of p-values and therefore a set of p-value spacings. The B simulated sets of spacings are then used to estimate $E[\tilde{p}_i]$ for all i, assuming all null hypotheses to be true.

Assuming data follows the format of Table 2.1, the AgenBH proceeds as follows.

AgenBH Procedure:
Assume m discrete, ordered p-values $p_{(1)}, \ldots, p_{(m)}$ and their corresponding spacings $\tilde{p}_1, \ldots, \tilde{p}_{m+1}$.

1. For $b = 1, \ldots, B$, where B is the number of simulated sets of spacings under the null:

(a) Simulate $R_{i1}^{(b)} \sim$Binomial$(T_i, 0.5)$ for $i = 1, \ldots, m$.

(b) Using the simulated read counts, calculate $p_1^{(b)}, \ldots, p_m^{(b)}$ and $p_{(1)}^{(b)}, \ldots, p_{(m)}^{(b)}$.

(c) Calculate and save $\tilde{p}_1^{(b)}, \ldots, \tilde{p}_n^{(b)}$.

2. Estimate expectations as

$$\widehat{E[\tilde{p}_i]} = \frac{1}{B} \sum_{b=1}^{B} \tilde{p}_i^{(b)}, \quad i = 1, \ldots, m. \tag{2.7}$$

3. Define $\tilde{k}$ as

$$\tilde{k} = \max_{i=1,\ldots,m} \left\{ i : \sum_{j=1}^{i} \tilde{p}_j \leq \alpha \sum_{j=1}^{i} \widehat{E[\tilde{p}_j]} \right\}. \tag{2.8}$$

If $\tilde{k}$ is well-defined, reject all hypotheses associated with $p_{(1)}, \ldots, p_{(\tilde{k})}$, else reject nothing.

The value of B determines the accuracy of the estimates for $E[\tilde{p}_i]$, as well as the computational cost of the procedure. Higher values of B require more computing time, but also yield smaller standard errors for $\widehat{E[\tilde{p}_i]}$. We found that setting $B = 500$ was an acceptable compromise. In the simulations described below, $B = 500$ yielded standard errors for $E[\tilde{p}_i]$ less than 1%, while maintaining a slight computational advantage over existing methods for discrete p-values.

2.4 Simulation studies

In this section we describe three simulation studies performed to evaluate properties of the AgenBH procedure. For each study, the proportion of truly null genes, m_0/m, was fixed at 95%, 90%, or 80%. 100 data sets were simulated for each study under the following conditions.

- n=1,000
- $T_i \sim$lognormal(4,2) and rounded to the nearest integer, $i = 1, \ldots, m$.

- R_{i1} ~Binomial(T_i, p), where $p = 0.5$ for null genes, and
  $p \sim$ Uniform$((0, .4) \cup (.6, 1))$ for differentially expressed genes.

This simulation format and the distribution of T_i defined above follows those used by Dalsingh and Altman (2011) [24]. Any value of T_i exceeding 10,000 was truncated to 10,000. This truncation was performed because of the very high computational cost of specifying p-value distributions for competing discrete testing procedures. An advantage of the AgenBH procedure is that these distributions are unneeded. Two-sided p-values for each simulated gene were calculated using the binomial exact test. Each set of p-values was analyzed using the seven methods summarized in Table 2.4. The percentage of false rejections (achieved FDR) and the percentage of true alternative hypotheses detected (achieved power) were recorded for each data set and method.

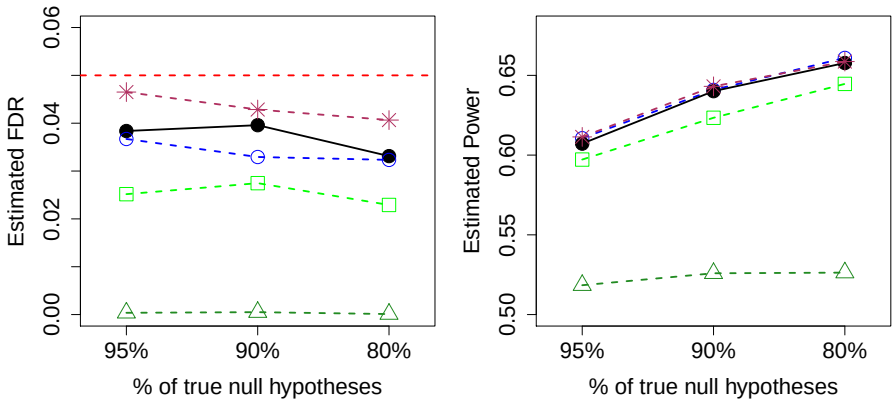
Table 2.4: Summary of methods applied to each set of p-values

Method	Reference
Classic B-H procedure	Benjamini and Hochberg (1995) [7]
Generalized B-H procedure	Ghosh (2011) [23]
AgenBH procedure, $B = 500$	Proposed
B-H procedure using mid-P values	Benjamini and Hochberg (1995) [7]
Step-up Heyse procedure	Heyse (2011) [22]
Step-down procedure using mid-P value	Benjamini and Liu (1999) [8]
Step-down Heller and Gur procedure	Heller and Gur (2011) [25]

2.4.1 Results

In this section we present the results of the simulation studies to compare performance of the AgenBH procedure to that of the procedures mentioned in Table 2.4. The results are summarized in Figure 2.1.

In these simulations, the AgenBH procedure maintained consistent and appropriate error control properties for all settings. Although other discrete methods illustrated slightly increased achieved FDR, the AgenBH procedure had comparable power properties. Finally, the AgenBH procedure was slightly faster computationally when all row total T_i were below 10,000, but it was approximately 10 times faster in the presence of even a very few row totals greater than 20,000.



(a) FDR (b) Power

Figure 2.1: Summary of achieved FDR (a) and power (b) for the AgenBH procedure, the B-H procedure, and other discrete competitors. The proposed procedure is represented by the black ●. Results from the classical B–H procedure are presented by green squares □. The blue ○ represents the B-H procedure with mid-P values, and the dark green △ represents the step-down procedure using mid-P values. The maroon ∗ represents both the step-up Heyse procedure and the step-down Heller and Gur procedure, which gave identical results.

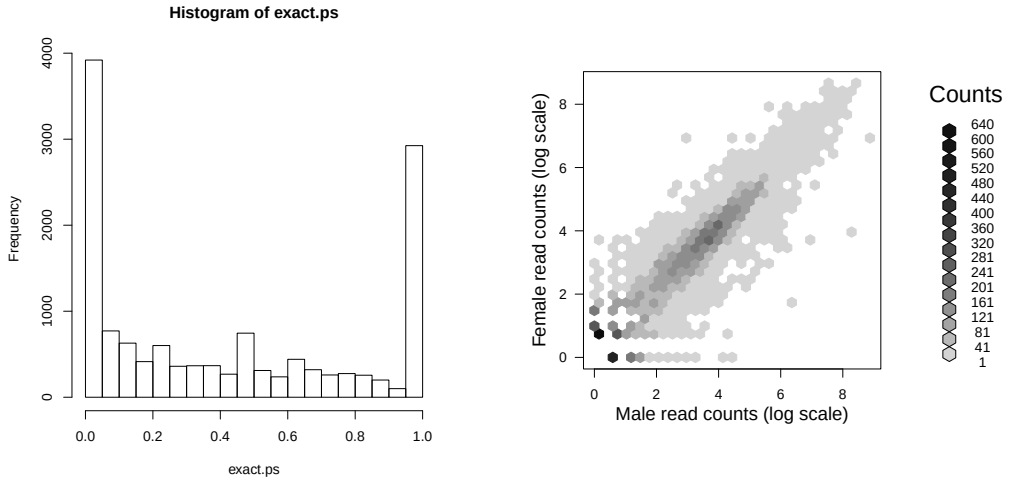
2.5 Application to data

In this section we illustrate use of the AgenBH procedure through application to an RNA-seq dataset. We use a publicly available dataset produced by Blekhman et al. (2010) [26] and composed of read counts from liver samples originating from male and female humans, chimpanzees, and rhesus monkeys. The data is available as raw reads from GEO dataset GSE17274. In this data analysis we consider only the human samples with the aim of finding differences in expression between a male sample and a female sample. Further, although there are three biological replicates for both male and female, we use only the first of each.

The data set illustrates the large variability in read counts that is typical in this type of experiment. There are 20,689 genes included in the two samples. Of these, 6,893 genes have no read counts for either sample, and 31 genes have more than 10,000 read counts. Before applying any analysis techniques, we remove these genes. We omit from analysis those genes without reads and genes with more than 10,000 reads. The read counts for the remaining 13,765 genes are highly skewed, with the first quartile, median, and third quartile being 7, 36 and 114

reads respectively. Because of the large number of genes with low read counts, we expect that methods specifically developed for discrete p-values to have an advantage over those that assume continuous p-values.

P-values were calculated using the binomial exact test. We favor the use of this test over Fisher's exact test to avoid the systematic inflation of significance due to the scenario described in Section 2.4 and Table 2.3. A histogram of the resulting p-values is included in Figure 2.2a.



(a) Discrete p-values calculated using binomial exact test

(b) Raw read counts for male and female liver samples

2.5.1 Results of data analysis

We applied all methods from Table 2.4 to the discrete p-values described above using an α value of 0.01. The results are summarized in Table 2.5. As expected, both the B-H and the genBH procedures found a fewer number of significant genes than the procedures designed for discrete values: AgenBH, and the step-up Heyse procedure. The B-H procedure with mid-P values found a comparable number of genes significant. We used DAVID to identify significant gene enrichment pathways, finding significant enrichment in oxidoreductase, endoplasmic reticulum, and lipid metabolic pathways. These findings confirm those of Blekhman et al. (2010) [26], who found significant differences between the sexes in the genes involved with lipid metabolism.

Although the step-up Heyse procedure found somewhat more significant genes, we expect this procedure and the AgenBH procedure to have comparable power properties based on simulation results. We also see that both step-down procedures found a lower number of genes significant.

Table 2.5: Number of genes found significant using each analysis method, with $\alpha = .01$.

Method	Number of genes
Classic B-H procedure	2014
Generalized B-H procedure	2014
AgenBH procedure, $B = 500$	2195
B-H procedure using mid-P values	2160
Step-up Heyse procedure	2320
Step-down procedure using mid-P value	931
Step-down Heller and Gur procedure	999

2.6 Concluding remarks

In this chapter we have discussed existing approaches for hypothesis testing using discrete p-values, described the simple case of the generalized Benjamini-Hochberg procedure, and introduced the AgenBH procedure. Our novel procedure is intuitively appealing for its simplicity, and practically appealing for its computational efficiency in the presence of a small number of very large read counts. It offers an improvement of power over the classical B–H procedure and displays similar power properties to other established procedures for discrete data. Computational efficiency could be increased by reduction of the simulation size, B, used to estimate the expected spacings, with the obvious trade-off of more variability in those estimates.

Chapter 3
Testing the disjunction hypothesis using Voronoi diagrams

3.1 Introduction

In current genetics and biological research, it is common for thousands of hypothesis tests to be performed simultaneously. When each gene or location of interest is associated with a single hypothesis and a single p-value, a variety of approaches are available to control type I error rates. We have previously described some of these approaches in Chapters 1 and 2. In some situations, however, each gene is associated with multiple hypotheses and thus multiple p-values. Multiple p-values per gene are frequently available in the arena of meta-analysis wherein information is combined across studies to test an overall hypothesis. Another setting in which multiple p-values are found, and which we are most interested in for this chapter, is when there are multiple measurements for each gene that are each individually of interest. For example, p-values may be available for each of two different aspects of a single underlying biological process.

With the shift from a single p-value per gene to a vector of p-values per gene, which we refer to here as a p-vector, clear specification of the null and alternative hypotheses is critical. If the goal is to pool information, as is common in meta-analysis, then conjunction or partial conjunction hypotheses are appropriate. The conjunction null hypothesis is that all p-values contained in a p-vector are from

a null distribution, and rejection is possible when at least one p-value shows evidence of being from a non-null distribution. Rejection of the partial conjunction hypothesis requires at least u of n p-values to show evidence of being from non-null distributions. There are scenarios, however, when the hypothesis associated with each p-value of the p-vector is of interest individually, and rejection should be possible only when there is evidence that all such hypotheses are non-null. In this case, the disjunction hypothesis is of primary interest. Distinctions between the conjunction, partial conjunction, and disjunction hypotheses are further described in Section 3.2 of this chapter.

Testing of the disjunction hypothesis is appropriate when multiple aspects of a single underlying biological process are measured. For example, there is interest in detection of genes related to cancer progression that are both differentially expressed on a transcriptomic level and show evidence of copy number aberrations in cancerous tissue [27–31]. Another motivating example is detection of periodic genes as explored by Lichtenberg et al. (2005) [32]. In this case, the disjunction hypothesis is considered using one p-value for periodicity and a second for regulation of expression for each gene. The most commonly used summary method for the disjunction hypothesis uses the maximum of all p-values for each test (Wilkinson 1951 [33]), and typically has very low power.

This chapter presents an approach for p-value combination appropriate for testing the disjunction hypothesis when there are two p-values associated with each gene or location. The approach considers p-vectors as locations on the unit square, where certain challenges arise that are absent in the case of single p-values. First, the strict ordering of p-values on the real line is lost. Second, relationships between p-vectors are complicated, and third, their components may be correlated. In light of these challenges, a method for large-scale simultaneous testing of the disjunction hypothesis must accomplish three objectives. It must account for the relative positioning of the p-vectors in the plane, allow for multiple ordering schemes and finally, control FDR under any correlation structure of the test statistics used to calculate the p-vectors' components.

The approach proposed here addresses these challenges through the use of Voronoi diagrams, flexible incorporation of ordering schemes, and empirical null distributions [15]. We begin this chapter in Section 3.2 by describing the disjunction hypothesis framework. Section 3.3 provides a thorough introduction to the

Voronoi diagram and some of its statistical applications. In Section 3.4 we introduce multiple ordering schemes for p-vectors. Section 3.5 describes a technique for summarizing p-vectors to a single value, and details how these values can be used to control FDR. We explore properties of the procedure through simulations in Section 3.6. A possible extension for higher dimensional p-vectors is discussed in Section 3.7. Finally, we apply the procedure to two genomic studies in Sections 3.8 and 3.9.

With the exception of Section 3.3, much of the content of this chapter was published in the Annals of Applied Statistics (Phillips and Ghosh 2014 [34]).

3.2 The disjunction of null hypotheses

In this section we discuss three different paradigms for testing a single overall hypothesis when each gene or location studied is associated with a vector of p-values. Clearly, in these situations the interpretation of significance depends on the specification of overall null and alternative hypotheses in relation to individual hypotheses for each component of the p-vector. Consider m p-vectors, each of length n, denoted

$$P_i = (p_{i1}, \ldots, p_{in}), \quad i = 1, \ldots, m. \tag{3.1}$$

In the context of a genomic study, i is the index of the individual gene, while n is the number of p-values associated with each gene. We employ notation used by Benjamini and Heller (2008) [35] to describe compound null and alternative hypotheses. Testing the global null hypothesis, also known as the conjunction of null hypotheses, is equivalent to testing that *at least one* of the p-values $p_{i1}, \ldots, p_{in}$ is considered significant.

$H_0^{1/n}$: all hypotheses associated with P_i are null
$H_A^{1/n}$: at least one hypothesis associated with P_i is non-null

P-value combination methods for testing the conjunction null include the well-known Fisher's and Stouffer's methods for combining p-values [36, 37]. A comparison of these and other methods is presented by Loughin (2004) [38]. Rejection of the conjunction null can result from the influence of a single highly significant p-value even when all other p-values show no evidence for the alternative hypothesis.

In this setting, the scientific conclusion from rejection is not as strong it would be if a level of increased consistency across p-values was enforced.

Benjamini and Heller (2008) [35] proposed techniques for addressing this weakness through testing of the partial conjunction hypothesis:

$$H_0^{u/n} : \text{at least } u \text{ of } n \text{ hypotheses associated with } P_i \text{ are null}$$
$$H_A^{u/n} : \text{at least } u \text{ of } n \text{ hypotheses associated with } P_i \text{ are non-null}$$

This hypothesis requires consistency of evidence across u studies that is not required in the conjunction framework, while still allowing for lack of significance for some associated p-values. It can be considered a compromise between the conjunction and disjunction hypotheses.

The disjunction hypothesis is also referred to as the disjunction of null hypotheses and can be expressed as follows.

$$H_0^{n/n} : \text{at least one hypothesis associated with } P_i \text{ is null}$$
$$H_A^{n/n} : \text{all hypotheses associated with } P_i \text{ are non-null}$$

This hypothesis is desirable when considering multiple p-values per test that are each of individual interest. The established p-value combination approach for testing the disjunction hypothesis is to simply select the maximum p-value of each p-vector [33]. Error control procedures can then by applied to these maximum values. This approach is generally conservative and exhibits low power. The procedure described in this chapter is suitable for testing the disjunction hypothesis and results in a gain of power over the maximum method of p-value combination.

3.3 The Voronoi Diagram

In this section we include a detailed introduction to the Voronoi diagram and a few of its statistical applications. Portions of this section were included as part of a review article for Wiley Computational Statistics (Phillips 2014 [39]).

A Voronoi diagram is a tessellation defined by a set of input points, often called 'seeds' or 'generators,' wherein each input is allocated a region of influence. For the hypothesis testing procedure introduced in this chapter, these seeds are in

fact p-vectors. Tessellations defined in this fashion have long been of interest in both statistics and a variety of applied fields. Because of their wide applicability, they have been known by various names. Terms in common usage include Voronoi tessellations or diagrams [40,41], Thiessen polygons [42], Dirichlet regions [43,44], S-mosaics [45], and Meijering cells [46]. Following modern convention [47], we use the term 'Voronoi diagram' in this dissertation.

3.3.1 Definition and properties

The Voronoi diagram has its simplest definition in the plane when generators are two-dimensional points and the Euclidean metric is used to define distance. The region of influence for a particular seed, commonly referred to as a 'Voronoi region', 'Voronoi cell' or simply 'cell', consists of all points closer to that seed than to any other. Diagrams of this type can be defined on the entire plane or can be restricted to a predetermined region containing the seeds. Consider a tessellation of a planar region $S \subseteq \mathbb{R}^2$. Let $\{\mathbf{x}_1,\ \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset S$, $1 \leq n \leq \infty$, be a set of two-dimensional points.

Definition 7. *The Voronoi diagram restricted to S and generated by $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is defined to be the collection of Voronoi cells:*

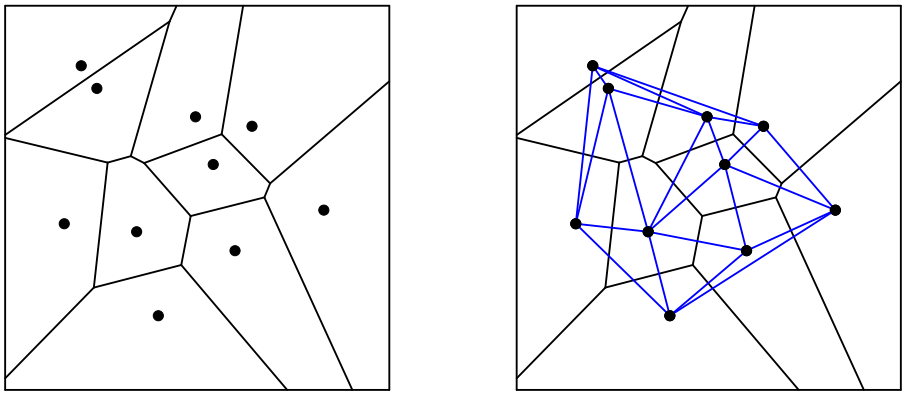
$$V_S(\mathbf{x}_i) = \{\mathbf{y} \in S : d(\mathbf{x}_i, \mathbf{y}) \leq d(\mathbf{x}_j, \mathbf{y}), i \neq j, j \in 1, \ldots, n\}$$

where $d(\mathbf{x}_i, \mathbf{y})$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{y}$.

In the specific context of two-dimensional p-vectors, we let the n seeds be p-vectors $P_1, \ldots, P_n$, and we set S to be the unit square. Figure 3.1a presents a simple planar Voronoi diagram restricted to the unit square and generated by a set of 10 p-vectors.

Key terms related to Voronoi diagrams include *Voronoi edge*, *Voronoi vertex*, and *Voronoi face*. A Voronoi edge is a line, half-line, or line segment that is part of the boundary of a Voronoi cell. A Voronoi vertex can be defined as either the end-point of a Voronoi edge, or alternatively as a point included in the boundaries of at least three Voronoi cells. Voronoi faces are boundaries of cells in higher dimensions, which may be of interest in future work but are not described in detail here. Without making any assumptions on the distribution of the seeds, many useful

properties of the resulting Voronoi diagram are known [47]. In particular, each simple planar Voronoi diagram is a unique tessellation for the seeds $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Furthermore, Voronoi cells are convex polygons. Two cells are said to be *neighbors* if they share a Voronoi edge. This definition of neighbor relates directly to the dual graph of the Voronoi diagram: the Delaunay triangulation. The properties of this triangulation are discussed in more detail below.



(a) Simple planar Voronoi diagram using ten generator points.

(b) Delaunay triangulation and its dual, the Voronoi diagram

Figure 3.1: The simple planar Voronoi diagram and its dual graph, the Delaunay triangulation.

3.3.1.1 Delaunay triangulation

The Delaunay triangulation is known as the dual graph to the Voronoi diagram due to a specific correspondence between the two constructions. To obtain the Delaunay triangulation generated by $\{\mathbf{x}_1,\ \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the generators of neighboring Voronoi cells are connected. A Delaunay triangulation and the corresponding Voronoi diagram is presented in Figure 3.1b. Under the assumption that no three generating points are collinear, the triangulation tessellates the region enclosed by the convex hull of generators into a collection of triangles.

Delaunay triangulations are used in applications independently of their dual Voronoi diagrams. For example, they are utilized in transmission of data over a network overlay [48], or between nodes in a wireless network [49]. They are also used as part of surface reconstruction efforts in computer graphics and mathemat-

ical modeling [50, 51]. In the laboratory setting, triangulation has been used to decompose clumps of nuclei for cell-based florescence imaging assays [52].

3.3.1.2 Poisson-Voronoi diagrams

A special case of the Voronoi Diagram that has been particularly well-studied is the Poisson-Voronoi diagram (PVD). A PVD is a Voronoi diagram whose generating points are the realization of a homogeneous Poisson point process. The definition and properties of such processes are well known [53, 54].

Many theoretical results about the characteristics of PVDs and their cells are known. For example, the PVD is a regular, normal, stationary, and isotropic tessellation [47]. In addition, moments and conditional moments of normalized cell area, number of edges, total edge length, and other characteristics of Poisson-Voronoi cells are known [47]. Calculation of the probability density functions for these quantities has not been achieved theoretically, however simulation studies have been performed to estimate them numerically using both Gamma and generalized Gamma distributions [55–57].

3.3.2 Statistical applications of Voronoi diagrams

The Voronoi diagram is a useful tool in spatial statistics because of its nonparametric nature and ability to partition regions in a manner determined entirely by data. For example, consider the enduring topic of cluster detection. Many powerful techniques such as the spatial scan statistic [58] and its variants exist for this problem. However, when clusters are irregularly shaped it may be difficult for these techniques to identify the cluster's domain. As an alternative approach, Voronoi diagrams constructed from spatial point data can be used to define potential domains as unions of adjacent Voronoi cells. For each candidate cluster domain, a likelihood is computed based on the sizes of the included Voronoi cells. The union of cells with maximum likelihood is chosen as the proposed cluster domain. This approach has been used to identify the boundaries of minefields from aerial data [59].

Another spatial problem that has benefited from Voronoi diagrams is the estimation of intensity functions of inhomogeneous Poisson point processes. One proposed approach is particularly simple: the estimated intensity for a given lo-

cation is defined as the inverse area of the Voronoi cell to which the location is assigned [60]. These estimates are approximately unbiased, nonparametric and spatially adaptive, however they can also be highly variable. A centroidal Voronoi estimator can be used to reduce this variability. Other approaches to model inhomogeneous spatial intensity functions also use Voronoi diagrams to discretize integrals [61], or as a method of quadrature [62, 63]. Similarly, Bayesian partitioning for estimation of risk surfaces [64] or non-stationary spatial processes [65] depends on Voronoi diagrams to discretize the considered regions. In these settings, the process being modeled is assumed stationary within each cell, but independence is assumed between cells. The use of Voronoi diagrams allows for discontinuities in correlation structure and simplifies computations.

Voronoi diagrams have also been used to extend procedures developed for one-dimensional spacings to higher-dimensional settings [66, 67]. Hyper-volumes of Voronoi cells serve as high-dimensional spacings, and statistical distances based on these Voronoi spacings have known asymptotic behavior [66]. Because this behavior matches that of spacings on the real line when the dimension is set to one, methods for spacings can be adapted to higher-dimensional settings through Voronoi diagrams. Such applications include comparison of high-dimensional probability densities, and the hypothesis testing procedure described in this chapter.

3.3.3 Generalizations of Voronoi diagrams

Important characteristics of a Voronoi diagram include the specification of a 'region of influence' for each generator, exhaustive covering of the region under consideration, and mutual exclusivity of cell interiors.

Various modifications to the simple planar diagram of Definition 7 are in use, resulting in generalized Voronoi diagrams that retain the characteristics mentioned above but differ in other properties. One modification is the change of distance from Euclidean, although other distances need not be formal mathematical metrics. Examples include exponential, logarithmic, and power distances. Weighted Voronoi diagrams belong to this category, and allow for generators to have differing 'importance' weights for allocation of influence regions. A modification not discussed in detail here is the inclusion of generators such as lines, line segments, or other objects embedded in the considered region.

3.3.4 Weighted Voronoi diagrams

The Voronoi diagrams described thus far assign all generators equal weight. In many examples, it is desirable to allow some generators to be considered more important, or higher weighted than others. For example, consider locations of human settlements as generator points. Construction of a Voronoi diagram based on these locations may want to incorporate relative population size and other factors through use of a weighted distance. Several such distances have been defined [47, 68] and are in use. This section briefly describes the definitions of additively, multiplicatively, and compoundly weighted Voronoi diagrams.

For $1 \leq n \leq \infty$, let $w_1, \ldots, w_n$ be positive multiplicative weights associated with the seeds $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and let $v_1, \ldots, v_n$ be the corresponding additive weights.

Definition 8. *The multiplicatively weighted distance is given by*

$$d_{MW}(\mathbf{y}, \mathbf{x}_i) = \frac{1}{w_i}||\mathbf{x}_i - \mathbf{y}||.$$

In $\mathbb{R}^2$, the Voronoi edges of multiplicatively weighted diagrams are the arcs of circles. Figure 3.2a illustrates a multiplicatively weighted diagram. An optimal algorithm for construction of these diagrams has been derived [69].

(a) Multiplicative (b) Additive (c) Compound

Figure 3.2: Three types of weighted Voronoi diagrams

Definition 9. *The additively weighted distance is given by*

$$d_{AW}(\mathbf{y}, \mathbf{x}_i) = ||\mathbf{x}_i - \mathbf{y}|| - v_i.$$

In $\mathbb{R}^2$, a Voronoi diagram generated using this distance is sometimes called a 'hyperbolic Dirichlet tessellation', as edges have the form of hyperbolic arcs or straight lines as illustrated in Figure 3.2b.

Definition 10. *The compoundly weighted distance is given by*

$$d_{CW}(\mathbf{y}, \mathbf{x}_i) = \frac{1}{w_i}||\mathbf{x}_i - \mathbf{y}|| - v_i.$$

This distance is a combination of multiplicative and additive weights, and is shown in Figure 3.2c.

These three types of weighted Voronoi diagrams are used in a variety settings. For example, multiplicative weighting has been used in applications such as determination of retail trade areas [70], in logistics districting and zone design [71], in proteomics [72], and for various purposes in GIS systems [73, 74]. Additive weighting has also been used in zoning problems [75], as well as for applications in molecular biology [76].

3.3.5 Voronoi diagrams in higher dimensions

The formulation of a simple planar Voronoi diagram can extend directly to m dimensions by letting $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be m-dimensional seeds, and letting $d(\mathbf{x}_i - \mathbf{y})$ be the Euclidean distance in m-space. For $m > 2$, Voronoi cells are m-dimensional polytopes with k-dimensional Voronoi faces ($0 < k < m$) marking the k-dimensional boundaries between adjacent Voronoi cells.

Some basic properties of higher-dimensional Voronoi diagrams are known, in particular for $m = 3$. Expectations of values such as the number of k-faces, and the volumes of polytopes for randomly chosen cells have been determined theoretically under the assumption that generator points are realizations of a Poisson point process [77]. Other properties such as the distribution of volumes have been determined through extensive simulation studies [56]. The development of feasible computing algorithms for higher-dimensional Voronoi diagrams is of ongoing interest.

3.3.6 Computational considerations

Conceptually, construction of Voronoi diagrams is quite straightforward. Computationally, their construction has been researched in the field of computational geometry for over three decades [78]. The families of approaches that have received the most attention can be roughly categorized as incremental construction algorithms, divide-and-conquer algorithms, sweep algorithms, and raise-and-compute algorithms.

Incremental construction algorithms compute a series of Voronoi diagrams, starting with a diagram generated by a small subset of seeds and finishing with the full diagram. As each seed is added, a new diagram is created until all generators have been included. This approach was described as early as 1980 [78], and has been shown to be the fastest approach family with an average time complexity of incremental insertion as low as $\mathcal{O}(n)$. The order in which seeds are added directly affects the algorithm's efficiency. Authors have used both bucketing [79] and random ordering [80, 81] to determine insertion order. Modifications of incremental approaches continue to be of interest, particularly in construction of higher dimensional Delaunay triangulations [82, 83], or of generalized Voronoi diagrams [84].

Divide-and-conquer algorithms divide the generators into two or more subsets of roughly equal size, perform tessellation on these subsets separately, and then merge the results to acquire the complete diagram. This approach, too, has been described and refined for decades [78, 85–87], and continues to be of interest both for simple diagrams [88], generalized diagrams [89], and for triangulation in higher dimensions [90]. In practice, average time complexity of divide-and-conquer algorithms have been shown [87] to be as low as $\mathcal{O}(n \log \log n)$.

In general, sweep algorithms reduce the complexity of a two-dimensional problem into a one-dimensional problem by computing along a one-dimensional line that sweeps through the region of interest until the entire computation is complete. While it cannot be applied directly to the problem of Voronoi diagrams, adaptations have made it possible. Such adaptations have been applied by various authors [91–93].

Computation of higher-order Voronoi diagrams is correspondingly more complex as m and n increase. A common approach to this problem is to exploit a known correspondence between the Voronoi diagram in m-dimensional space and the con-

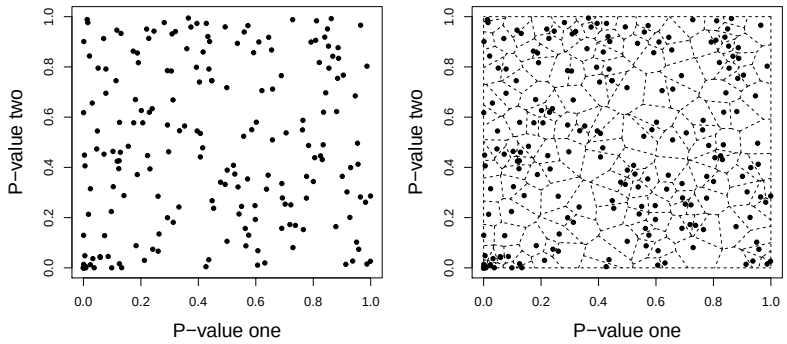
vex hull in $(m+1)$-dimensions. For this correspondence to hold, the m-dimensional generator points are raised to $(m+1)$-space by defining $x_{i,m+1} = \sum_{j=1}^{m} x_{i,j}^2$. These raised points are then used to calculate a convex hull. Methods for performing this calculation vary, and include both incremental and divide-and-conquer approaches. The efficiency of the convex hull calculations depends on the dimension. For even-dimensional space, incremental approaches are more efficient, while for odd-dimensional space divide-and-conquer algorithms are more efficient. Thus, to construct three-dimensional Voronoi diagrams, it is most efficient to raise the points to four-dimensional space and construct the convex hull using incremental techniques [47]. In fact, construction of three-dimensional Voronoi diagrams is optimally of $\mathcal{O}(n^2)$.

Programs such as R and Matlab are able to compute Voronoi diagrams and Delaunay triangulations through either built-in modules or freely available packages. At the core of many current implementations is qhull: a family of programs in C to compute Voronoi diagrams using an incremental approach combined with tools from convex hull calculations [94]. This code has been integrated into Matlab, where the function 'Voronoin' computes Voronoi diagrams for m-dimensional inputs. It has also been incorporated into the R package 'Geometry' [95]. Another R package, 'deldir' [96], computes and plots Voronoi diagrams in two dimensions using an incremental algorithm [78]. We use the deldir package in the implementation of the procedure described in this chapter. Both packages are available through CRAN, work efficiently, and can manage large numbers of points.

3.3.7 Voronoi Diagrams for hypothesis testing

In this section we motivate the use of Voronoi diagrams for hypothesis testing with p-vectors. Recall from Section 2.2 that Ghosh (2011) [23] rephrased the B-H procedure in terms of the one-dimensional spacings from ordered p-vectors. When we consider planar p-vectors, we must move to a different concept of 'spacing': the Voronoi diagram generated by p-vectors and restricted to the unit square. Ghosh (2011) summed consecutive spacings and used these values to define $\tilde{k}$ and declare significance. In the procedure proposed here, we instead sum the areas of Voronoi cells to define combined values for each gene. We can thus allow for multiple definitions of 'consecutive', and do so in Section 3.4.

An illustration of the Voronoi diagram generated by a sample set of 200 p-vectors is presented in Figure 3.3b. We follow this example in Section 3.4 before switching to a larger sample for subsequent sections.



(a) 200 p-vectors (b) Voronoi diagram

Figure 3.3: (a) 200 simulated p-vectors and (b) the corresponding Voronoi diagram.

3.4 Multiple Ordering Schemes

On the real line spacings are defined as the difference between consecutive ordered p-values. This definition is dependent upon an ordering of the p-values that is unique on the real line. However, this uniqueness is lost when p-vectors are considered as bivariate locations in the unit square. For this reason we present multiple ordering schemes for the plane, assuming that small values of p_{i1} and p_{i2} indicate evidence against the null. Thus the orderings begin at the origin, and each p-vector is ranked according to increasing values of D, its 'distance' from the origin. Each scheme defines D differently. Thus, for each definition $P_{(1)}$ is the p-vector with the smallest value of D, and $P_{(m)}$ with the largest. Here we describe the definition of D for each scheme.

1. *Euclidean Ordering* results in a movement from the origin in contours with the shape of circles. Define $D^{(E)}$ as the Euclidean distance from the origin.

$$D_i^{(E)} = \sqrt{p_{i1}^2 + p_{i2}^2}, \quad (i = 1, \ldots, m). \tag{3.2}$$

2. *Maximum Ordering* results in contours with the shape of squares. Define

$D^{(M)}$ as

$$D_i^{(M)} = \max\{p_{i1}, p_{i2}\}, \quad (i = 1, \ldots, m). \tag{3.3}$$

3. *Summation Ordering* is equivalent to beginning at the origin and moving out in contours of right isosceles triangles. In this case, $D^{(S)}$ is defined as

$$D_i^{(S)} = p_{i1} + p_{i2}, \quad (i = 1, \ldots, m). \tag{3.4}$$

4. *de Lichtenberg Ordering* is an ranking scheme proposed by de Lichtenberg et al. (2005) [32]. The scheme defines

$$D_i^{(L)} = p_{i1}p_{i2}\left(1 + \left(\frac{p_{i1}}{.001}\right)^2\right)\left(1 + \left(\frac{p_{i2}}{.001}\right)^2\right), \ (i = 1, \ldots, m). \tag{3.5}$$

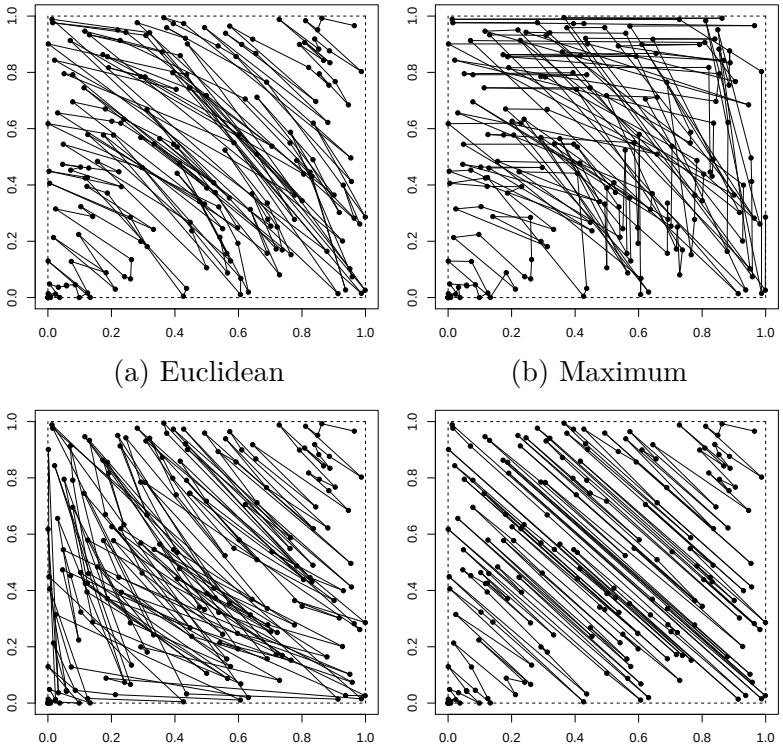
   Note that $D^{(L)}$ consists of four multiplicative factors. The first two weight $D^{(L)}$ according to the value of each individual component, and the last two penalize p-vectors that have only one very small component. For typical p-vectors the values for $D^{(L)}$ are very large as a result of division by .001 of both p_{i1} and p_{i2}. This magnitude is not a concern as the interest is only in their relative values for the purpose of ranking, and the values themselves are not of particular interest. The contour lines for this ordering scheme move from the origin in lines approximating an inverse function such as $y = 1/(x^3)$.

Figure 3.4 illustrates these four ordering schemes using the sample set of 200 p-vectors from Section 3.3.7. Table 3.1 presents a numerical example using five p-vectors.

It is noteworthy that three of the four ranking schemes described have concave contour lines: Euclidean, Maximum, and Summation. The remaining scheme, de Lichtenberg, has convex contour lines. As we will see, these characteristics have important implications for error control.

3.5 Summarizing p-vectors and declaring significance

The described ranking schemes can be combined with Voronoi cell areas to summarize each ranked p-vector as a single value in the interval (0,1). Define $A_{(i)}$ as the area of the Voronoi cell associated with the ordered p-vector $P_{(i)}$, and the



(a) Euclidean (b) Maximum

(c) de Lichtenberg (d) Summation

Figure 3.4: Illustration of all four ordering schemes with the sample set of 200 p-vectors. The solid lines join p-vectors that are considered 'consecutive' under each ordering.

cumulative sum of these ordered areas as

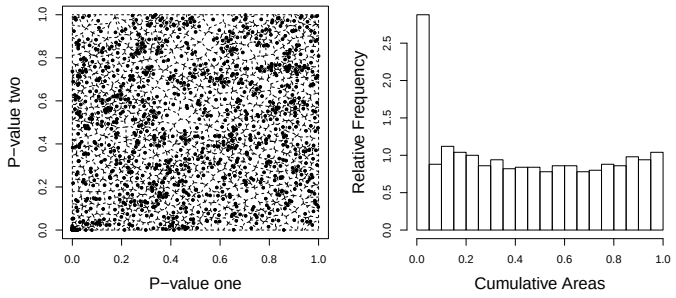
$$T_{(i)} = \sum_{j=1}^{i} A_{(j)}, \quad (i = 1, \ldots, m). \tag{3.6}$$

These cumulative sums serve as combined values in an analogous manner that cumulative spacings comprise p-values in one dimension. These $T_{(i)}$ reflect both the relative positioning of the p-vectors in space and their distance from the origin. They can be used to make decisions in the hypothesis testing framework. Figure 3.5 illustrates a sample set of 1000 p-vectors and a histogram of their cumulative areas. In this example the components of the p-vectors are independent, and 10% of p-vectors are associated with an alternative hypothesis.

The combined values for each hypothesis that result from the summation of cell areas can now be used in the hypothesis testing framework to declare significance

Table 3.1: Example of ordering results for five p-vectors. For each ranking scheme the distance, D is presented for each p-vector along with the resulting rank in parentheses.

P_i	$D_i^{(E)}$		$D_i^{(M)}$		$D_i^{(S)}$		$D_i^{(L)}$	
(.85, .51)	0.99	(3)	0.85	(3)	1.36	(4)	$8.1 \cdot 10^{10}$	(4)
(.91, .80)	1.21	(5)	0.91	(4)	1.71	(5)	$3.9 \cdot 10^{11}$	(5)
(.23, .97)	1.00	(4)	0.97	(5)	1.20	(3)	$1.1 \cdot 10^{10}$	(3)
(.62, .34)	0.71	(2)	0.62	(1)	0.96	(2)	$9.5 \cdot 10^{9}$	(2)
(.07, .63)	0.63	(1)	0.63	(2)	0.79	(1)	$8.9 \cdot 10^{7}$	(1)



(a) 1000 p-vectors (b) Combined values

Figure 3.5: An example of (a) 1000 simulated p-vectors with independent components and (b) a histogram of their cumulative areas calculating using the Euclidean ordering scheme. The sharp spike in the histogram corresponds to the p-vectors associated with alternative hypotheses.

and control error rates. In Sections 3.5.1 and 3.5.2 we describe two separate approaches: one is appropriate for p-vectors with independent components, and the other for p-vectors with positively correlated components. Because the situation is much simpler in the first case, that is where we begin.

3.5.1 Multiple hypothesis testing under independence

When the components of the p-vectors are assumed to be independent, standard multiple comparisons procedures such as the B-H procedure can be applied to the cumulative spacings with very good results. Simulation studies were performed to test the properties of FDR control and power of this approach. For each study 100

sets of test statistics were generated according to

$$(t_{i1}, t_{i2}) \sim MVN\left(\begin{pmatrix}\mu_i\\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right), \quad i = 1, \ldots, 2000;$$

10% of statistics were associated with an alternative hypothesis ($\mu_i = \mu_A$), while the remaining 90% were null ($\mu_i = 0$). P-vectors were formed from 2-sided p-values according to

$$P_i = (p_{i1}, p_{i2}) = (P(Z > |t_{i1}|), P(Z > |t_{i2}|)), \quad \text{where } Z \sim N(0, 1)$$

for $i = 1, \ldots, 2000$. The proposed method was applied to each data set, using the four described ordering schemes from Section 3.4. Additionally, the existing p-value combination technique based on the maximum was applied. After applying the B-H procedure to each set of summary values to declare significance, the achieved FDR and 1-non discovery rate (NDR) were recorded. Using notation from Table 1.1, 1-NDR is defined as $E[S/(S+T)]$. This quantity can be viewed as a measure of power.

Table 3.2: Power (1-NDR) results of simulation studies under independence

	Euclidean	Maximum	Summation	de Lichtenberg	Wilkinson
μ_A=2	0.200	0.189	0.216	0.205	0.005
μ_A=3	0.772	0.760	0.788	0.796	0.098
μ_A=4	0.976	0.975	0.977	0.979	0.744

Table 3.3: False Discovery Rate results of simulation studies under independence

	Euclidean	Maximum	Summation	de Lichtenberg	Wilkinson
μ_A=2	0.041	0.037	0.048	0.041	0.000
μ_A=3	0.042	0.038	0.049	0.056	0.000
μ_A=4	0.042	0.040	0.045	0.053	0.000

Tables 3.2 and 3.3 summarize the results for studies where $\mu_A = 2,\ 3,\ 4$ respectively. The results show that under all ordering schemes, the proposed combination

method results in greatly increased power. All concave schemes (Euclidean, Maximum, and Summation) control FDR at the desired level $\alpha = .05$, but the convex de Lichtenberg scheme does not. This difference becomes more pronounced when p-vectors with correlated components are considered. Additional simulations using correlated test statistics show that application of the B-H procedure to combined values is insufficient to control FDR when the correlation between test statistics surpasses 0.2. This loss of FDR control is a result of the increased concentration of p-vectors along the diagonal of the unit square under correlation, which changes the characteristics of the cumulative areas. In Section 3.5.2 we discuss approaches appropriate for multiple testing in these conditions.

3.5.2 Multiple hypothesis testing under dependence

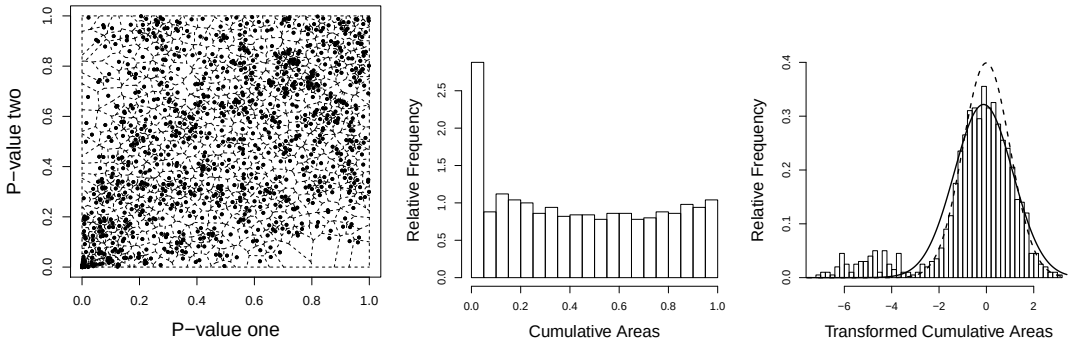
In certain settings the individual components of p-vectors may be correlated. For example, correlation between components may occur when different but related aspects of an underlying biological process are measured. Any technique used for testing the disjunction hypothesis in this setting should be robust to this structure. Using an empirical null approach [15, 97] for combined values in place of the B-H procedure results in FDR control for all positive correlation structures, although the trade-off is decreased power in the case of independent components.

We described the use of an empirical null for determining statistical significance in Section 1.1.5. In the context of testing the disjunction hypothesis, we consider a transformation of the summarized cumulative areas $T_{(i)}$ as defined in (3.6):

$$Z_{(i)} = \Phi^{-1}(T_{(i)}), \quad (i = 1, \ldots, m) \tag{3.7}$$

where Φ is the cumulative distribution function for the standard normal random variable. Figure 3.6 illustrates the effect of the transformation from $T_{(i)}$ to $Z_{(i)}$ for a sample set of 1000 p-vectors. The transformation makes it easier to detect deviations from the null hypothesis, as true alternative hypotheses are presented as a second, smaller peak to the left of the null distribution instead of in a single spike for the original values. The bivariate test statistics used to calculate these p-vectors had a correlation of 0.7. The histogram of transformed values in Figure 3.6c shows evidence of a null distribution that differs from N(0,1), as the dependence structure of the p-vectors results in thicker tails than the theoretical null predicts.

For this reason, it is desirable to use an empirical null as a basis for our inference when the components of the p-vectors show evidence of correlation.



(a) $P_{(i)}$ (b) $T_{(i)}$ (c) $Z_{(i)}$

Figure 3.6: An example of (a) 1000 simulated p-vectors with dependent components, (b) a histogram of their cumulative areas calculating using the Euclidean ordering scheme, and (c) a histogram of the transformed cumulative areas with empirical (dashed) and theoretical (solid) null distributions.

We use these $Z_{(i)}$ to estimate an empirical null distribution and calculate local false discovery rates. In fact, we use a quantity closely related to lfdr to declare significance: the left-tail false discovery rate. For each value z the corresponding left tail FDR is defined as

$$lFDR(z) = P\left(z_i \sim f_0 | z_i \leq z\right). \tag{3.8}$$

Inference can be made based on estimated lfdr or lFDR values. We utilize an R package (mixFdr) developed by Muralidharan (2010) [19] and available through CRAN (http://cran.r-project.org/web/packages/mixfdr) which uses an empirical Bayes mixture method to fit an empirical null, estimate effect sizes, lfdr, and lFDR. The use of other packages or techniques is certainly possible. The function we used in simulation studies, mixFdr, includes two tuning parameters: J, the number of distributions to be estimated and P, a penalization parameter. A higher value of P encourages estimation of a larger null group and closer estimation of the central peak.

Careful calibration of J and P, and even experimentation with other techniques for empirical null estimation are desirable when a single data set is under consideration. The function mixFdr estimates left-tail false discovery rate for each $Z_{(i)}$,

and we declare significant all p-vectors with these estimates of left-tail FDR less than α. This approach results in appropriate error control that is robust to correlation in the components of p-vectors. Section 3.6 describes a simulation study performed to explore properties of power and FDR control when this approach is applied to the cumulative areas from ordered p-vectors

3.6 Simulation study with correlated components

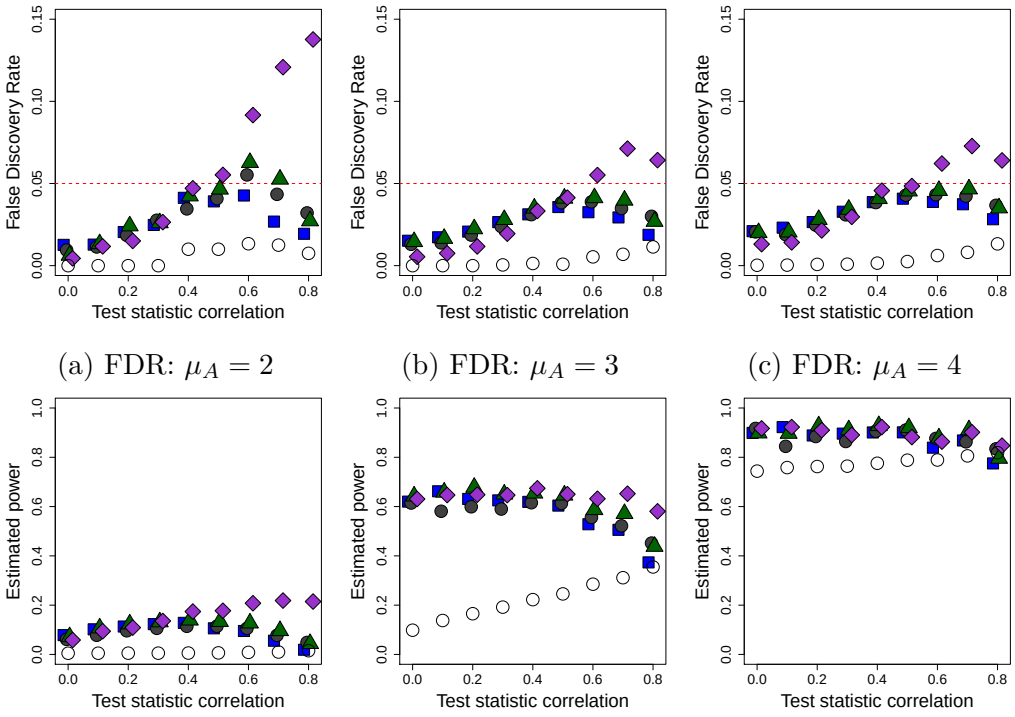
To illustrate properties of the procedure we ran three simulation studies: one each for strong, moderate, and weak alternative signals. We were interested in evaluating FDR control and power. For each simulated data set we set $m = 2000$, and generated test statistics by

$$(t_{i1}, t_{i2}) \sim MVN\left(\begin{pmatrix}\mu_i\\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),\ i = 1, \ldots, m$$

for strong, moderate, and weak alternative signals: $\mu_A = 4,\ 3,\ 2$ respectively. For null test statistics, $\mu_i = 0$. 10% of test statistics for each data set were generated from the alternative distribution. P-vectors were formed from 2-sided p-values.

For each simulation study ρ varied from 0 to .8 in increments of 0.1, and 100 data sets were simulated for each value of ρ. We performed the procedure using all four ordering schemes on each simulated data set, using mixFdr to estimate empirical null distributions and left-tail FDR for each data set, rejecting all hypotheses associated with p-vectors whose estimated left-tail FDR was less than 0.05. We set J=2 for all data sets, and after calibrating the fit of several example empirical nulls for weak, moderate, and strong signals, we set P=400, 800, and 1000 for the respective simulations. Additionally, the B-H procedure was applied to the maximum values from each p-vector to compare the proposed method to an existing approach. Figure 3.7 summarizes the results of all three simulation studies.

The proposed technique for combining p-values has improved power when compared to the existing procedure. This improvement is greatest for weak and moderate alternatives. These simulations further show that ordering scheme matters, although the differences between the three convex (Euclidean, maximum, summation) ordering schemes are small compared to the difference between them and



(a) FDR: $\mu_A = 2$ (b) FDR: $\mu_A = 3$ (c) FDR: $\mu_A = 4$

(d) 1-NDR: $\mu_A = 2$ (e) 1-NDR: $\mu_A = 3$ (f) 1-NDR: $\mu_A = 4$

Figure 3.7: Summarized results of simulation studies for test statistics with varying correlation structure. (a), (b), and (c) present FDR when alternative signals are 2, 3, and 4 standard deviations from the null. (d), (e), and (f) present 1-NDR for the same data sets. Solid squares, circles, triangles, and diamonds represent Euclidean, Maximum, Summation, and de Lichtenberg ordering schemes respectively. The open circles represent existing approach using the maximum of each component for inference.

the de Lichtenberg ordering. The de Lichtenberg ordering displays characteristics that differ considerably from the other three. Specifically, it shows a tendency to lose FDR control as the components of the p-vectors become more correlated. We present simulation results for this ordering for the sake of completeness, but we do not recommend using it for data analysis when testing the disjunction null. The nature of its contour lines suggest this ordering scheme is in fact more appropriate for testing the conjunction or partial conjunction hypothesis, as these lines resemble the contour lines for Fisher's or Stouffer's p-value combination techniques from Figure 1 of [98].

Table 3.4: Summary of distributions used to calculate null test statistics

% of total signals	μ_{i1}	μ_{i2}
10 %	0	3
10%	3	0
70%	3	3

3.6.1 Further simulation studies

In addition to the studies detailed in Sections 3.5.1 and 3.5.2, we also performed simulation studies to evaluate performance of the proposed approach in the presence of p-vectors constructed from test statistics with means $(0,\mu_A)$ or $(\mu_A, 0)$. These p-vectors should not be found significant in the disjunction setting, but should be under the conjunction framework. The purpose of performing this additional simulations is to verify that the procedure proposed in this chapter does control error rates according to the disjunction hypothesis in the presence of 'half-null' hypotheses.

For the first of these additional studies, we varied the correlation structure of all p-vectors while calculating 10% from test statistics with mean (0,3), and 10% from test statistics with mean (3,0). An additional 10% of test statistics had mean (3,3), and these are the signals that we hope to detect. The remaining 70% of p-vectors are calculating from test statistics with mean (0,0). To be clear, in this study we formed p-vectors considered null by calculating two-sided p-values from test statistics (t_{i1}, t_{i2}) generated according to:

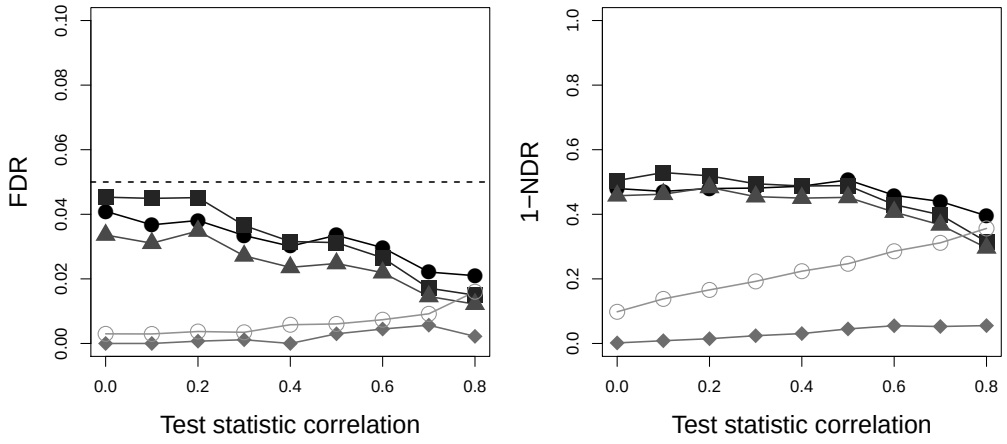
$$(t_{i1}, t_{i2}) \sim MVN\left(\begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),\ i = 1, \ldots, m \tag{3.9}$$

where ρ varies from 0 to 0.8, and the values of μ_{i1} and μ_{i2} are described in Table 3.4.

The p-vectors that we hope to detect were calculated from two-sided p-values using test statistics following the bivariate distribution below:

$$(t_{i1}, t_{i2}) \sim MVN\left(\begin{pmatrix}3\\ 3\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right),\ i = 1, \ldots, m \tag{3.10}$$

For each value of ρ from 0 to 0.8, we generated 100 data sets of size $m = 2000$. We calculated achieved FDR and 1-NDR using all four ordering schemes and the existing approach of Wilkinson. When applying the Voronoi p-value combination approach we fit an empirical null distribution and used the resulting estimation of left-tail FDR to declare significance. The results are summarized in Figure 3.8.



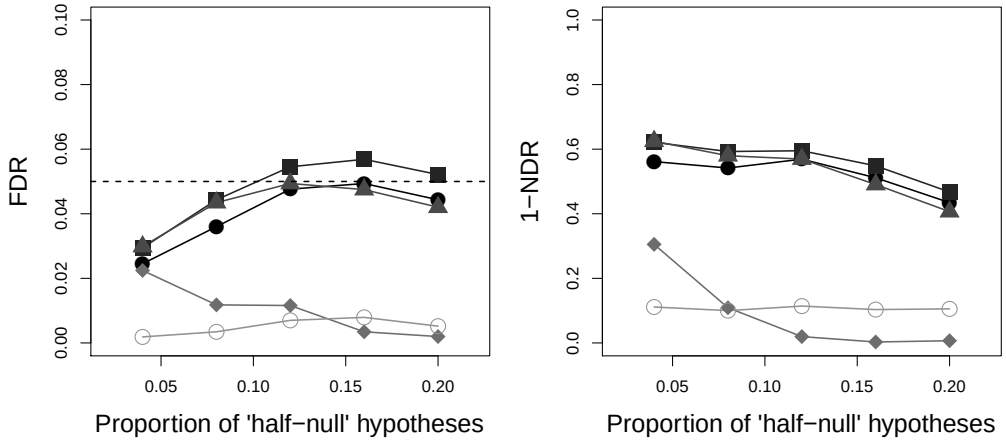
(a) FDR (b) 1-NDR

Figure 3.8: Summarized results of the first additional simulation study. Figure (a) presents estimated FDR while (b) presents estimated 1-NDR. Solid squares, circles, triangles, and diamonds represent Euclidean, Maximum, Summation, and de Lichtenberg orderings respectively. Open circles represent results from the existing approach using the maximum of each p-vector as a basis for inference. Note the poor performance of the de Lichtenberg ordering in comparison to the other three candidate orderings.

We can see from the error rates and power results that the proposed procedure performs well for all tested values of ρ in this situation, provided that one of the three recommended ordering schemes is used. The de Lichtenberg ordering does poorly in the presence of 'half-null' hypotheses.

The second additional simulation study evaluated performance when p-vectors had uncorrelated components with varying percentages of 'half-null' hypotheses. The p-vectors considered null were generated according to (3.9), with varying proportions of signals having means (0,4) and (4,0). We varied the total proportion of these 'half-null' p-vectors from 5% to 20%, with the remaining null test statistics having mean (0,0). In all cases, the p-vectors that we hope to declare significant are calculated from test statistics with mean (3,3). The purpose of these simu-

lation settings was to evaluate performance when the 'half-null' signals were in fact stronger in one dimension than the truly alternative signals. The simulation results are presented in Figure 3.9.



(a) FDR (b) 1-NDR

Figure 3.9: Summarized results of simulation studies using empirical null procedures with varying proportions of p-vectors calculated from test statistics with means (0,4) and (4,0). For each scenario, 10% of test statistics have mean (3,3). (a) presents estimated FDR while (b) presents estimated 1-NDR. Solid squares, circles, triangles, and diamonds represent Euclidean, Maximum, Summation, and de Lichtenberg orderings respectively. Open circles represent results from the existing approach using the maximum of each p-vector as a basis for inference. Again note the comparatively poor performance of the de Lichtenberg ordering.

Again, we see that the proposed procedure outperforms the existing Wilkinson procedure. The three concave-down orderings do very well, and the de Lichtenberg ordering performs poorly.

3.7 Extension to higher dimensions

The approach described in this chapter is suitable when there are two p-values associated with each hypothesis test, however in many situations three or more p-values will be available. In theory, the procedure can be extended to higher dimensions by replacing cumulative areas with cumulative volumes, hypervolumes, etc. In practice, however, the computation complexity for Voronoi cells increases quickly with dimension. Average time complexity is as low as $\mathcal{O}(n)$ in the plane,

but is at least $\mathcal{O}(n^2)$ in 3-space [77]. To avoid this disadvantage, we consider an alternative extension using the sets of all possible pairs of components. Consider a set of m 3-dimensional p-vectors.

$$P_i = (p_{i1}, p_{i2}, p_{i3}), \quad i = 1, \ldots, m. \tag{3.11}$$

Then define three sets of 2-dimensional p-vectors constructed via pairwise combination of components of P_i.

$$\{(p_{i1}, p_{i2})\}, \ \{(p_{i1}, p_{i3})\}, \ \{(p_{i2}, p_{i3})\}, \quad i = 1, \ldots, m. \tag{3.12}$$

For each of these sets of two-dimensional p-vectors the Voronoi diagram is computed and cell areas saved. Thus each p-vector, P_i is associated with three individual cell areas, $A_i^{1,2}$, $A_i^{1,3}$, and $A_i^{2,3}$, as well as an average area $\bar{A}_i = (A_i^{1,2} + A_i^{1,3} + A_i^{2,3})/3$. This average area can then be used in conjunction with an ordering scheme to create the summarized areas used for inference. Define $P_{(1)}$, $P_{(2)}, \ldots, P_{(m)}$ to be the p-vectors ranked according to a specified ordering scheme such as Euclidean distance from the origin, and $\bar{A}_{(1)}$, $\bar{A}_{(2)}, \ldots, \bar{A}_{(m)}$ to be the corresponding average areas. Then the cumulative average areas are defined as $\bar{T}_{(i)} = \sum_{j=1}^{i} \bar{A}_{(i)}$.

Multiple testing can then be performed on these summarized cumulative average areas using the methods described in Sections 3.5.1 and 3.5.2. Further investigation into the properties of this approach is necessary, as well as research on other possible extensions for higher dimensions. A preliminary simulation study using three-dimensional p-vectors with independent components was conducted with weak, moderate, and strong alternative test statistics. For each data set, test statistics were generated according to

$$(t_{i1}, t_{i2}, t_{i3}) \sim MVN\left(\begin{pmatrix}\mu_i\\ \mu_i\\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\right), \ i = 1, \ldots, m.$$

Three-dimensional P-vectors were formed from 2-sided p-values. The resulting p-vectors were ordered according to Euclidean distance from the origin. Hypothesis testing was performed using the B-H procedure on the summarized cumulative average areas. The existing technique of applying the B-H procedure to the set of

maximum p-values from each p-vector was also performed for comparison. Table 3.5 summarizes the findings of the simulations.

Table 3.5: Simulation results for proposed extension

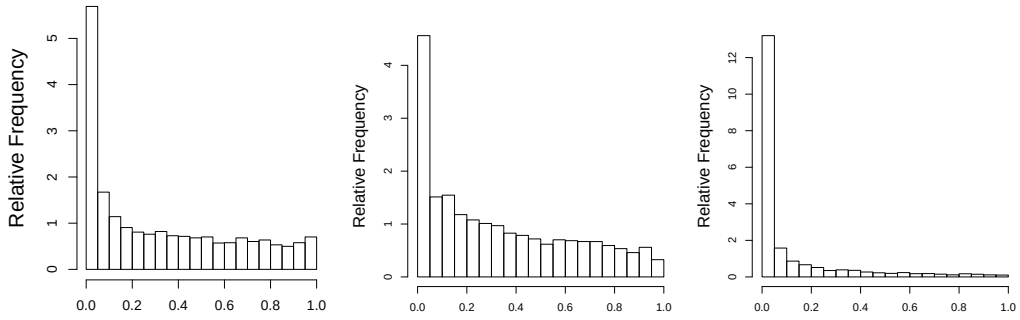
	Proposed Extension			Existing Approach		
	$\mu_A = 2$	$\mu_A = 3$	$\mu_A = 4$	$\mu_A = 2$	$\mu_A = 3$	$\mu_A = 4$
FDR	0.023	0.004	0.023	0.000	0.000	0.000
1-NDR	0.098	0.730	0.986	0.005	0.007	0.610

3.8 Application to Schizosaccaromyces Pombe data

In 2004 and 2005, three papers were published investigating the periodicity of genes in the fission yeast cell *Schizosaccharomyces pombe*. Specifically, Oliva et al. (2005) [1] produced three data sets including time points for three complete cell cycles using two different synchronization techniques. In their paper they identified 750 genes determined to be periodically expressed based on a ranking scheme and cut-off. We apply our approach to test for periodicity in a hypothesis testing framework using Fisher's exact G statistic to measure evidence of periodicity.

3.8.1 The data

Three microarray datasets from *Schizosaccharomyces pombe* from Oliva et al. (2005) were used: Elutriation a, Elutriation b, and Cdc25. The first two were produced using Elutriation synchronization, and the last using a Cdc25 block-release synchronization technique. We apply our technique on the two Elutriation sets, using Fisher's exact G statistic to calculate the p-vector for each gene. This test statistic requires evenly spaced time points, necessitating omission of any measurements that occur at uneven intervals. The Elutriation a data set includes 50 time points, however only 33 are at regular intervals of 8 minutes. For Elutriation b, many of the time points are technical repeats. We keep the first measurement in each case, leaving 31 evenly spaced time points for each gene taken at intervals of 10 minutes. The Cdc25 data set has a total of 51 evenly spaced time points,



(a) Elutriation a (b) Elutriation b (c) Cdc25

Figure 3.10: Histogram of p-values for (a) Elutriation a, (b) Elutriation b, and (c) Cdc25 block release. Note the strong evidence of periodicity in all three experiments, particularly Cdc25.

taken at intervals of 15 minutes. Only genes with complete measurements for all selected time points are considered.

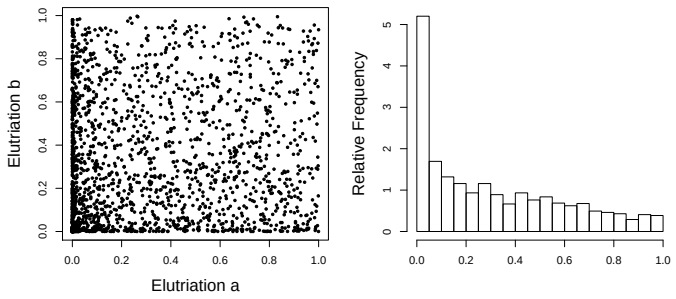
3.8.2 Results using existing procedures

The data show evidence of widespread periodicity. Considered separately, Elutriation a, Elutriation b, and Cdc25 have 22.8%, 26.4%, and 66% of p-values less than .05. Even controlling FDR using the B-H procedure on each set independently results in a very high rate of rejection. Table 3.6 presents a summary of the data and marginal analysis of all three data sets. Figure 3.10 presents histograms of the p-values when the data sets are considered independently.

Table 3.6: Summary of results from Oliva et al. (2005) [1] data sets considered separately

	Elutriation a	Elutriation b	Cdc25
Complete genes	3050	2394	3724
Evenly spaced time points	33	31	51
Genes (%) with p-values<.05,	868 (22.8%)	546 (26.4%)	2458 (66.0%)
Significant genes (%) using B-H	527 (17.3%)	155 (6.5%)	2252 (60.5%)

Consider the p-vectors formed using p-values generated by Elutriation a data and Elutriation b data. Note that these two Elutriation data sets were both gen-



(a) P-vectors (b) Cumulative areas

Figure 3.11: (a) P-vectors formed from Fisher's G statistic of Elutriation a and Elutriation b, (b) a histogram of cumulative cell areas formed using the Euclidean ordering scheme.

erated using the same synchronization technique, and the p-values generated by each repetition have roughly comparable marginal distributions. To test the disjunction hypothesis for Elutriation a and Elutriation b using an existing technique the maximum p-value for each gene is preserved. The B-H procedure is then applied to these maximum values. The resulting number of rejections is 15, which is surprisingly low. Figure 3.11a helps to explain this result. The p-vectors' components do not show evidence of correlation, thus considering only the maximum of each p-vector's components gives a distribution that is very different from either of the marginal distributions. Our proposed approach uses information from both p-values, and gives a different result.

3.8.3 Results using Voronoi P-value combination on Elutriation data

We apply our p-value combination method using the Euclidean, maximum, and summation ordering schemes to the p-vectors formed from the Elutriation a and b experiments. The p-vectors are plotted in Figure 3.11a. The components of the p-vectors do not show evidence of high correlation, and we apply the B-H procedure to the cumulative areas generated from each ordering scheme. This application results in 225, 213, and 249 rejections of the disjunction hypothesis using Euclidean, Maximum, and Summation orderings respectively.

Application of an empirical null approach to the combined areas yields a very different result. Because of the high amount of periodicity detected in the exper-

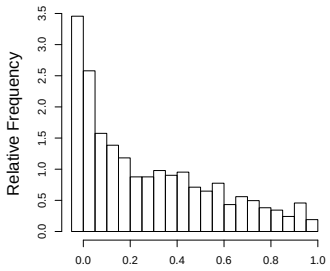


Figure 3.12: Cumulative average areas

iments, the empirical null is estimated to have a negative mean. This shift to the left of up to -0.87 results in rejection of far fewer genes: 15, 12, and 11 for the three concave ordering schemes. These genes could be considered significantly more periodic than the rest, although other genes also show evidence of periodicity.

The two considerations of the combined values reflect two different scientific questions. By using the B-H procedure on the combined areas, the genes found are those that show significant periodic expression in both elutriation experiments. The genes found using the empirical null procedure are those genes that are significantly periodic in both experiments relative to the majority of genes. This distinction explains the difference in numbers of genes found significant.

3.8.4 Extension of procedure to include Cdc25

The extension to three dimensions described in Section 3.7 can be applied to the 3-dimensional p-vectors formed from Oliva et al. [1] data. We order the three-dimensional p-vectors according to their Euclidean distance from the origin, and calculate the three Voronoi cell areas associated with each p-vector. From these we calculate each p-vectors average cell area, and then cumulative average areas. Figure 3.12 presents a histogram of these values. Note that many of these cumulative areas are quite small as result of the high number of very small p-values from the Cdc25 experiment. Application of the B-H procedure to the cumulative average areas formed using the Euclidean ordering scheme results in rejection of 165 disjunction hypotheses. These 165 genes are those that show significant evidence of periodic expression in all three of the experiments performed by [1]. The existing procedure using the maximum values yields a mere 12 rejections for these experiments. Using an empirical null approach on the transformed cumulative av-

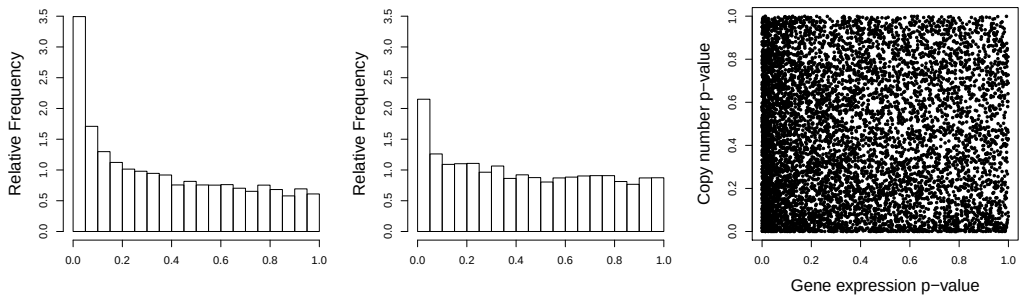
erage areas yields results similar to those discussed in Section 3.8.3. Because of the evidence of widespread periodicity throughout the experiment, only 8 genes show behavior that is significantly more periodic in comparison to the majority of genes when all three experiments are considered.

3.9 An Application related to Prostate Cancer

Identification of genes implicated in cancer progression is a research topic of great interest. Several studies have shown interest in identifying genes that show both alterations in copy number and evidence of differential expression in cancerous tumors [27–31]. We applied our method to data produced by Kim et al. (2007) [27] in a study on prostate cancer progression. Data on copy number and gene expression was gathered for 7,534 genes using prostate cell populations from low-grade and high-grade samples of cancerous tissue. For details on data acquisition and cleaning see the Kim et al. (2007) [27] paper.

We calculated t-statistics for genetic expression and copy number aberrations comparing tissue types. For each of 7534 genes we compute a two-dimensional p-vector from the resulting 2-sided p-values based on the t-statistics. Figures 3.13a and 3.13b present histograms of the expression and copy number p-values, while Figure 3.13c presents a representation of the resulting p-vectors. Upon close inspection, it is revealed that the smallest copy number p-values are much smaller than the smallest gene expression p-values. Thus, application of the B-H procedure to copy number p-values yields 62 significant genes, while application to the expression p-values fails to yield any. It is unsurprising, then, that application of the B-H procedure to the set of all maximum p-values for each gene also produces no significant results.

Using Voronoi p-value combination followed by the B-H procedure on the summarized values at $\alpha = .05$ gives 12, 14, and 25 rejections for Euclidean, maximum, and summation orderings. Guided by the results of simulation, we consider the rejections made using the Summation ordering. Of these 25, four were mapped to official gene names, and all four were listed in the COSMIC database of cancer genes [99]. These four genes are CABLES2, PAK1IP1, CAMKV, and TSHZ1. The COSMIC results suggest that there are mutations in these four genes that are found in a variety of cancers, thus strengthening the evidence of these genes being



(a) Gene Expression (b) Copy Numbers (c) Resulting P-vectors

Figure 3.13: Histogram of p-values for (a) expression, and (b) copy number. (c) presents the resulting p-vectors in the unit square.

Table 3.7: Summary of Gene Functional Classification from DAVID

Number of Genes	Enrichment Score	Keywords
16	0.73	Peptidase; Serine; Endopeptidase; Kringle
11	0.58	Transmembrane; Membrane; Extracellular; Cytoplasmic
3	0.50	Transport: protein, intracellular, vesicle; Golgi apparatus
3	0.36	Catabolic process; Proteosome; Proteolysis
12	0.27	Lumen: nuclear, intracellular, organelle, membrane-enclosed; Phosphoroprotein; nucleolus; ATP binding
3	0.22	GTP-binding; Nucleotide binding: guanyl, purine; Ribonucleotide binding: guanyl, purine
20	0.12	Transmembrane; Membrane; Glycoprotein;

putative oncogenes in prostate cancer.

To use DAVID [100,101] for further investigation of our results a larger gene list was necessary. For this purpose, we performed the B-H procedure on the combined p-values at $\alpha = .20$. Under the summation ordering, this yielded 306 rejections, 102 of which could be mapped to recognized genes by DAVID. The functional annotation tool found significant enrichment (adjusted p-value of 0.019) in the Fibrinolysis pathway. Fibrinolysis has been associated with prostate cancer for decades [102]. Tumor classifications for different malignancies have been proposed based on the behavior of this pathway [103]. Results for functional classification of the 102 genes are summarized in Table 3.7.

3.10 Discussion

In this chapter we have presented a novel approach to p-value combination for testing the disjunction hypothesis when two p-values are considered for each test. The approach uses an extension of one-dimensional spacings, Voronoi cell areas, in combination with concave ordering schemes to define cumulative areas suitable for multiple testing techniques. When the majority of p-vectors have independent components, techniques such as the B-H procedure can be directly applied. If the components are correlated, empirical null techniques are more suitable. Simulation studies showed that the approach has appropriate error control properties and results in a gain of power over the existing method. This increased power is of particular interest for detection of genes related to biological processes or implicated in cancer progression.

Four candidate ordering schemes were described, and simulations were used to test their performance in several settings. The concave up ordering proposed by de Lichtenberg et al. [32] failed to control FDR in the paradigm of the disjunction hypothesis. As discussed in Section 3.5.1, we suspect that concavity of an ordering's contour lines is vital to its FDR control characteristics. Specifically, as contours become increasingly concave down, the procedure is more conservative. The reverse applies when considering concave up schemes. For this reason, we recommend using the summation ordering in practice, as it represents the boundary case between concave up and down. This offers the least conservative, and thus most powerful, procedure that retains appropriate FDR control.

This approach can be extended in several meaningful directions. The conjunction or partial conjunction hypotheses could be tested by defining suitable ordering schemes such as the minimum, or product. Extension to higher dimensions is also of utmost interest, particularly considering the scale of current biological and genomic experiments. In Section 3.7 we described a potential extension to three or more dimensions, but further investigation of this and other techniques is necessary.

Chapter 4
The MaRR Procedure: Assessing reproducibility in replicate experiments

4.1 Introduction

The use of high-throughput technologies is now an essential part of modern biological research. This technology can be used to identify differentially or periodically expressed genes, and to produce ChIP-seq data for identification of protein binding sites. Sets of genes selected from high-throughput experiments are important for focused follow-up studies, however a well-known difficulty met by researchers is the variability of results even among experiments that are technical or biological replicates. For this reason, statistical methods for assessing agreement between experiments has been a recent research topic for statisticians. Genes or binding sites that show consistency across replicate experiments are often called *reproducible*, and those that do not are termed *irreproducible*.

Spearman's pairwise rank correlation can be used to assess reproducibility of gene sets, however its properties are dependent in some part upon how stringent the requirements are for inclusion of genes in the calculation. More stringent requirements produce higher values of rank correlation than more lenient requirements, even for the same experiments. Further, Spearman's rank correlation does not provide for error control. An alternative approach was proposed by [104], which avoids parametric assumptions by 'bucketing' genetic signals from both studies.

The determination of these 'buckets', however, is not straight-forward and may be difficult in practice. A comprehensive approach to assessing and describing reproducibility, including error control, was proposed by [105]. This approach uses a copula mixture model on ranked data to estimate effect sizes, correlation, variance, proportion of reproducible signals, and irreproducible discovery rate (idr) for all genes considered. This approach uses an expectation-maximization (EM) algorithm to determine parameter estimates and is model-based, so inferences may be sensitive to model misspecification or choice of starting parameters. A related procedure to identify reproducibility in the presence of missing data was described by [106], and similarly uses a copula-mixture model and EM algorithm.

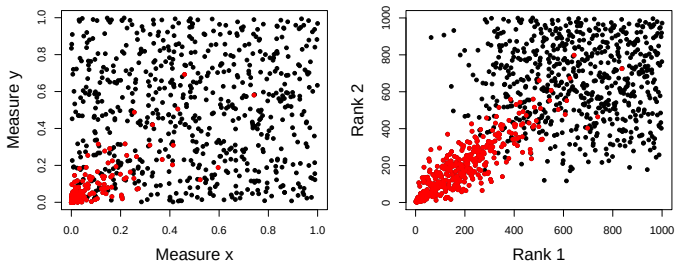
An approach that identifies reproducible genes in a non-parametric fashion and is free of tuning parameters represents a valuable development in this field. The procedure introduced in this chapter is non-parametric with respect to the underlying distribution of reproducible signals, requires no tuning parameters, and shows to have desirable properties in terms of discriminative power and error control. To derive the procedure, we make certain assumptions about the set of irreproducible genes in a sample, and use a maximum rank statistic as a basis for inference. The exact distribution of the maximum rank statistic can be derived for irreproducible genes in certain settings. By comparing theoretical to observed survival functions, the proposed procedure identifies where the change from reproducible to irreproducible signals begins with respect to the maximum rank statistic. Estimated marginal false discovery rates for each gene are then calculated based on the distribution of irreproducible maximum rank statistics.

This chapter proceeds as follows. Section 4.2 introduces the data format and defines the maximum rank statistic. In Section 4.2.1 estimators are derived in an ideal setting and shown to be asymptotically consistent. More realistic settings and estimation of false discovery rates are described in Section 4.2.2, where we also include a summary of the procedure. The finite-sample properties of the proposed procedures are evaluated using simulation studies in Section 4.3. Section 4.4 describes analyses on published genetic data sets. Finally, we conclude with some discussion in Section 4.5.

4.2 Data description and procedure formulation

We first introduce notation necessary to describe the proposed procedure. We assume that each gene or location studied is associated with a continuous measure from each of two experiments, for example a fold expression score, test statistic, p-value, or q-value. Let x_i be the measure from the first experiment for gene i, and y_i be the corresponding measure from the second experiment. With n genes we thus have two sets of measures: $x_1 \ldots, x_n$ from the first experiment and $y_1, \ldots, y_n$ for the second. We further assume no missing data is present.

These measurements are converted to rank statistics. Each gene g is thus associated with two ranks: (R_g^x, R_g^y), where R_g^x is the rank of x_g among $x_1, \ldots, x_n$, and similarly for R_g^y. Because the original measures are assumed to be continuous, we assume no ties are present. Figure 4.1 provides an example data set of p-values and rank statistics for $n = 1000$ genes, of which 350 are reproducible. Figure 4.1a displays bivariate p-values from two replicate experiments in the unit square, and Figure 4.1b shows their corresponding bivariate ranks. In these figures, the red points indicate pairs of p-values and ranks for reproducible genes, black points indicate the same for irreproducible genes. Genes whose measures indicate



(a) P-values (b) Rank pairs

Figure 4.1: p-values (a), and rank pairs (b) for 1000 genes. 350 of these pairs are from reproducible genes, as indicated in red.

the most interest to the researcher, for example those showing strong evidence of differential expression, are said to be “highly ranked", which means that the numerical values of their ranks are small. That is, the most highly ranked gene in a set has rank 1. Genes with reproducible measurements should be consistently highly ranked for both replicate experiments, and are expected to have positive correlation in their ranks. Genes with irreproducible measurements are assumed

to have independent ranks.

The procedure proposed in this chapter uses the maximum rank for each gene to determine which genes are reproducible. This statistic is defined below:

Definition 11.

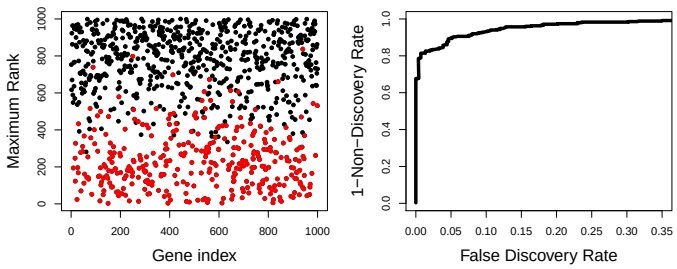
$$M_g = \max\left\{R_g^x, R_g^y\right\}, \quad g = 1, \ldots, n.$$

Table 4.1 provides a sample data set of $n = 4$ genes to illustrate the calculation of maximum rank statistics. The use of this statistic is compelling for its simplicity and also for its discriminative properties in the classification of reproducible genes. Genes that are consistently highly ranked will have a relatively low value for their maximum rank statistic, while inconsistent or low ranked genes will have higher values. For this reason, choosing a threshold based on the maximum rank can effectively separate reproducible from irreproducible signals. Figure 4.2 continues the example from Figure 4.1 by presenting maximum rank statistics and the corresponding receiver operating characteristic (ROC) curve generated when they are used to classify genes. This ROC curve illustrates the very good discriminative power of M_g, providing some evidence to base our procedure on this statistic. For this reason, we call it the 'Maximum Rank Reproducibility' (MaRR) procedure.

Table 4.1: Sample data, ranks, and maximum rank statistics from four genes, assuming larger values of x_g, y_g indicate more interest to the researcher.

index (g)	x_g	y_g	(R_g^x, R_g^y)	M_g
1	1.0	1.3	(3,2)	3
2	-0.2	0.0	(4,3)	4
3	1.2	-1.0	(2,4)	4
4	2.4	2.2	(1,1)	1

The joint distribution of all n maximum rank statistics has a complicated covariance structure, as no more than two of these statistics can take on any single value. We can, however, calculate the exact marginal distribution functions for irreproducible genes when certain conditions are in place and the proportion of reproducible genes in the sample is assumed to be π_1. We define these conditions in Section 4.2.1, and present the marginal distribution functions as Proposition 1 and Corollaries 1 and 2.



(a) M_g (b) ROC curve

Figure 4.2: Maximum rank statistics (a) and corresponding receiver operating characteristic (ROC) curve using M_g as the basis for declaring reproducibility (b). These figures continue the example from Figure 4.1. As before, red points indicate reproducibility and black indicate irreproducibility.

The MaRR procedure compares observed and theoretical marginal distribution functions to estimate π_1. To derive this estimator and show that it is consistent, we re-scale the maximum rank statistics to the unit interval by considering M_g/n. When $\pi_1 = 0$, all n genes are irreproducible and the marginal probability mass function for M_h/n can be calculated exactly.

Proposition 1. *Assume that all n genes are irreproducible. Then the marginal probability mass function for the normalized maximum rank statistic M_h/n is*

$$f_{n,0}(i/n) = \begin{cases} \frac{2i-1}{n^2} & 0 < i/n \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

Proof. Let there be n genes with independent bivariate ranks (R_g^x, R_g^y) and maximum rank statistics $M_g,\ g = 1, \ldots, n$. Define the random variable:

$$W_n(i) = \begin{cases} 2 & \text{if } i \text{ is twice a maximum rank} \\ 1 & \text{if } i \text{ is a unique maximum} \\ 0 & \text{if } i \text{ is not a maximum rank} \end{cases} \tag{4.1}$$

We calculate $E[W_n(i)]$ because $P(M_g = i) = E[W_n(i)]/n$.
Consider the probability mass function of $W_n(i)$.

$$P(W_n(i) = 2) = P\left(\exists g : (R_g^x, R_g^y) \in \{(l, i) : l < i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(i, l) : l < i\}\right)$$

$$
\begin{aligned}
&= \frac{i-1}{n} \cdot \frac{i-1}{n-1} \\
&= \frac{(i-1)^2}{n(n-1)} \qquad (4.2)
\end{aligned}
$$

$$
\begin{aligned}
P(W_n(i)=1) &= P\left(\exists g : (R_g^x, R_g^y) = (i,i)\right) \\
&\quad + P\left(\exists g : (R_g^x, R_g^y) \in \{(l,i) : l < i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(m,i) : m > i\}\right) \\
&\quad + P\left(\exists g : (R_g^x, R_g^y) \in \{(l,i) : l > i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(m,i) : m < i\}\right) \\
&= \frac{1}{n} + \frac{i-1}{n} \cdot \frac{n-i}{n-1} + \frac{i-1}{n} \cdot \frac{n-i}{n-1} \\
&= \frac{-n-1+2(n+1)i-2i^2}{n(n-1)} \qquad (4.3)
\end{aligned}
$$

$$
\begin{aligned}
P(Wn(i)=0) &= 1 - P(W_n(i)=2) = P(W_n(i)=1) \\
&= \frac{n^2+i^2-2ni}{n(n-1)} \qquad (4.4)
\end{aligned}
$$

Then the expectation of $W_n(i)$ can be determined under the null.

$$
\begin{aligned}
E(W_n(i)) &= 0 \cdot P(W_n(i)=0) + 1 \cdot P(W_n(i)=1) + 2 \cdot P(W_n(i)=2) \\
&= \frac{-n-1+2(n+1)i-2i^2}{n(n-1)} + \frac{2(i-1)^2}{n(n-1)} \\
&= \frac{2i-1}{n} \qquad (4.5)
\end{aligned}
$$

Thus the marginal probability mass function is known.

$$
P(M_g = i) = \frac{1}{n} E(W_n(i)) = \begin{cases} \frac{2i-1}{n^2} & \text{for } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \qquad (4.6)
$$

The marginal pmf for M_h/n is thus

$$
f_{n,0}(i/n) = P(M_h = i) = \begin{cases} \frac{2i-1}{n^2} & \text{for } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \qquad (4.7)
$$

□

4.2.1 Derivation of estimate $\hat{\pi}_1$ under ideal setting

In this section we derive a procedure to estimate π_1 in an ideal setting by making strong assumptions about the behavior of the marginal ranks R_g^x and R_g^y. We later relax these assumptions, discuss the properties of $\hat{\pi}_1$ in realistic settings, and derive estimates for FDR in Section 4.2.2. We call the setting consistent with the strong assumptions below 'ideal', and settings consistent with relaxed assumptions 'realistic'. For clarity of notation, we use the index h to indicate a gene that is assumed to have irreproducible measures.

Assumptions under the ideal setting

(I1) Reproducible signals are always ranked higher than irreproducible signals, i.e. $R_g^x < R_h^x$ and $R_g^y < R_h^y$ if gene g is reproducible and gene h is irreproducible.

(I2) The correlation between the ranks of reproducible signals is non-negative.

(I3) The two ranks per irreproducible gene are *independent*.

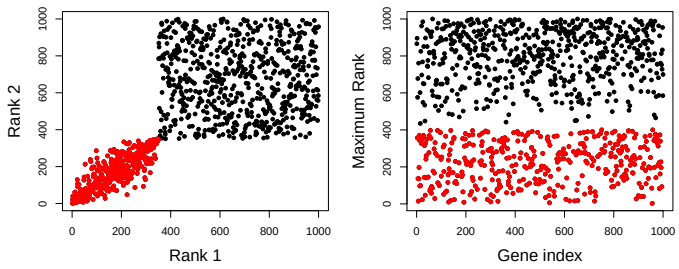
As a result of assumption (I1), $M_g < M_h$ for all reproducible genes g and irreproducible genes h. Letting π_1 be the proportion of reproducible genes, this implies that all genes g such that $M_g/n \leq \pi_1$ are reproducible, and all genes h such that $M_h/n > \pi_1$ are irreproducible. Rank pairs and maximum rank statistics for a sample data set generated under the ideal assumptions with $\pi_1 = 0.4$ are provided in Figures 4.3a and 4.3b. We can now derive the relevant distribution functions for M_h/n when $\pi_1 > 0$. For notational simplicity, define:

$$j_{\pi_1} = \max_{i=1,\ldots,n} \{i : i/n \leq \pi_1\} = \lfloor n\pi_1 \rfloor. \tag{4.8}$$

In the ideal setting described above, reproducible genes must have maximum ranks no more than j_{π_1}, thus possible values for M_h are $j_{\pi_1+1}, \ldots, n$. Adaptation of $f_{n,0}$ in Proposition 1 gives the marginal mass function for M_h/n dependent on π_1:

Corollary 1.

$$f_{n,\pi_1}(i/n) = \begin{cases} \frac{2(i-j_{\pi_1})-1}{(n-j_{\pi_1})^2} & \pi_1 < i/n \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$



(a) Rank pairs (b) Maximum ranks

Figure 4.3: Data from 1000 genes generated under the assumptions for the ideal setting. 400 of these genes (red) are assumed to be reproducible, and the remaining 600 genes (black) are irreproducible.

Proof. Assume (I1), (I2), and (I3). Further assume $\pi_1 \in (0,1)$ is fixed, and let $j_{\pi_1} = \lfloor n\pi_1 \rfloor$. Then the marginal pmf of M_h/n for an irreproducible gene h is $f_{n,\pi_1}(i/n) = f_{n-j_{\pi_1},0}(i/n - j_{\pi_1})$:

$$f_{n,\pi_1}(i/n) = \begin{cases} \frac{2(i-j_{\pi_1})-1}{(n-j_{\pi_1})^2} & \pi_1 < i/n \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

□

As a further result, the marginal cumulative distribution, F_{n,π_1} and survival, S_{n,π_1}, functions can be calculated.

Corollary 2. *Let $\pi_1 \in (0,1)$, $x \in (0,1)$, and $i_x = \lfloor nx \rfloor$. Then the marginal cumulative distribution and survival functions are:*

$$F_{n,\pi_1}(x) = \begin{cases} 0 & x < \pi_1 \\ \frac{(i_x - j_{\pi_1})^2}{(n-j_{\pi_1})^2} & \pi_1 \leq x \leq 1 \end{cases} \qquad S_{n,\pi_1}(x) = \begin{cases} 1 & x < \pi_1 \\ 1 - \frac{(i_x - j_{\pi_1})^2}{(n-j_{\pi_1})^2} & \pi_1 \leq x \leq 1 \end{cases}$$

Proof. Let $\pi_1 \in (0,1)$, $x \in (\pi_1, 1)$, and $i_x = \lfloor nx \rfloor$. Then we can derive the marginal cumulative distribution function of M_h/n in the ideal setting:

$$\begin{aligned} F_{n,\pi_0}(x) &= P(M_h/n \leq x) \\ &= P(\pi_1 < M_h/n \leq x) = \sum_{i=j_{\pi_1}}^{k_x} P(M_h = i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=j_{\pi_1}+1}^{k_x} \frac{2(i-j_{\pi_1})-1}{(n-j_{\pi_1})^2} = \frac{1}{(n-j_{\pi_1})^2}\left(2\sum_{i=j_{\pi_1}+1}^{k_x} i - \sum_{i=j_{\pi_1}+1}^{k_x} 1\right) \\
&= \frac{1}{(n-j_{\pi_1})^2}\left(2\frac{(k_x-j_{\pi_1})(k_x-j_{\pi_1}+1)}{2} - (k_x-j_{\pi_1})\right) \\
&= \frac{1}{(n-j_{\pi_1})^2}(k_x-j_{\pi_1})(k_x-j_{\pi_1}+1-1) \\
&= \frac{(k_x-j_{\pi_1})^2}{(n-j_{\pi_1})^2}
\end{aligned}
$$

□

We also derive the limiting marginal distribution of M_h/n:

Theorem 1. *Let $\pi_1 \in (0,1)$ be fixed, and assume (I1), (I2), and (I3). Then as $n \to \infty$ the marginal limiting distributions of the random variable M_h/n are as below:*

$$
F_{n,\pi_1}(x) \to F_{\pi_1}(x) = \begin{cases} 0 & x < \pi_1 \\ \frac{(x-\pi_1)^2}{(1-\pi_1)^2} & \pi_1 \le x \le 1 \\ 1 & 1 < x \end{cases},
$$

$$
S_{n,\pi_1}(x) \to S_{\pi_1}(x) = \begin{cases} 1 & x < \pi_1 \\ 1 - \frac{(x-\pi_1)^2}{(1-\pi_1)^2} & \pi_1 \le x \le 1 \\ 0 & 1 < x \end{cases}
$$

Proof. Let $\pi_1 \in (0,1)$, and $x \in (0,1)$ be fixed, and assume (I1), (I2), and (I3). For a fixed n, let $k_x = \lfloor nx \rfloor$, and $j_{\pi_1} = \lfloor n\pi_1 \rfloor$. Then as n tends to infinity:

$$
\lim_{n\to\infty} j_{\pi_1}/n \to \pi_1, \text{ and } \lim_{n\to\infty} k_x \to x.
$$

Thus the limiting cumulative distribution function can be derived.

$$
\begin{aligned}
F_{n,\pi_1}(x) &= \frac{(k_x-j_{\pi_1})^2}{(n-j_{\pi_1})^2} = \frac{(k_x-j_{\pi_1})^2/n^2}{(n-j_{\pi_1})^2/n^2} \\
&= \frac{(k_x/n - j_{\pi_1}/n)^2}{(n/n - j_{\pi_1}/n)^2} \\
&\to \frac{(x-\pi_1)^2}{(1-\pi_1)^2}
\end{aligned}
$$

□

The MaRR procedure estimates π_1 by comparing observed and theoretical distributions of maximum rank statistics. A classical approach to this comparison would focus on the cumulative distribution functions. The Cramer-von Mises statistic [107, 108] is a goodness-of-fit statistic that achieves this purpose, and is well-known in change-point literature. In the setting described here, however, the exact distribution of M_g/n is only known for $M_g/n > \pi_1$. Using the cumulative distribution function would require assumptions or knowledge about the distribution of maximum rank statistics for reproducible genes. We avoid this issue by using the survival function, allowing consideration of only statistics associated with irreproducible genes. Define the empirical survival function as:

$$\hat{S}_n(x) = \frac{1}{n}\sum_{g=1}^{n} I(M_g/n \geq x), \quad x \in (0,1). \tag{4.9}$$

We expect the $\hat{\pi}_1$ for which $S_{\hat{\pi}_1}$ is closest to $\hat{S}_n$ to be a consistent estimate for π_1. To define "closest", we use the weighted squared difference between the two functions for $\lambda \in (0,1)$:

Definition 12.

$$SS(\lambda) = \frac{1}{1-\lambda}\int_{\lambda}^{1} \left(\hat{S}_n(x) - (1-\lambda)S_\lambda(x)\right)^2 dx$$

The definition of $SS(\lambda)$ is that of a weighted integral. The factor $(1-\lambda)^{-1}$ is included because as λ varies, the range of the integral varies. Inside the integral, the factor $(1-\lambda)$ is necessary because the M_g follow a mixture distribution: $M_g \sim \lambda G + (1-\lambda)F_\lambda$, where G is the unknown distribution of reproducible genes, and F_λ is defined in Corollary 2. Therefore the theoretical S_λ must be normalized by $(1-\lambda)$. $SS(\lambda)$ can be calculated for any value of $\lambda \in (0,1)$, and we expect it to be small for values of λ close to the true π_1, and larger for values far from π_1. Thus we define the estimate $\hat{\pi}_1$ and declare it to be asymptotically consistent below:

Theorem 2. *Let the ideal assumptions hold, and define the estimate $\hat{\pi}_1$ of π_1 as*

$$\hat{\pi}_1 = \underset{\lambda\in(0,1)}{\arg\inf}\{SS(\lambda)\}.$$

Then as $n \to \infty$,

$$\hat{\pi}_1 \xrightarrow{p} \pi_1$$

Proof. Assume π_1 is fixed, $\hat{\pi}_1$, $\hat{S}_n(x)$, $S_n(x)$, and $SS(\lambda)$ are as defined previously. Then the proof outline to show $\hat{\pi}_1 \leftarrow \pi_1$ is below:

1. Show that the random variables M_h are *absolutely regular*.
2. Use result from [109] and apply the Glivenko-Cantelli theorem to show that $SS(\pi_1) \to 0$ as $n \to \infty$.
3. Show $SS(\lambda) \nrightarrow 0$ for $\lambda \neq \pi_1$.
4. Use (2) and (3) to prove consistency.

1. Following [109], a sequence of random variables $X_1, \ldots, X_N$ is *absolutely regular* if the dependence coefficients $\beta(n),\ n = 1, 2 \ldots$ go to 0, where

$$\beta(n) = \sup_{A \in \sigma(X_1, \ldots, X_k, X_{n+1+1}, \ldots)} \left| P(X) - P_0^k(X) P_{n+k+1}^{\infty}(A) \right|.$$

Here, P is based on the full joint distribution, and $P_1^{\infty}, P_{-\infty}^0$ are joint distributions for $X_{-\infty}, \ldots, X_0$ and $X_1 \ldots, X_{\infty}$ respectively and independently. Because the M_h have possible values $1, \ldots, N$, we must let n and N go to ∞ together. Thus set $N = n^2$ and consider:

$$\begin{aligned}
\beta(n) &= \sup_{A \in \sigma(M_1, \ldots, M_k, M_{n+1}, \ldots, M_N)} \left| P(A) - P_0^k(A) P_{n+k+1}^N(A) \right| \\
&\leq |1 - P((A_1, \ldots, M_k, M_{n+1}, \ldots, M_N) = (1, 2, \ldots, N-n))| \\
&= \left| 1 - \sum_{i=1}^{N-n} P(M_h = i) \right| = \left| 1 - \frac{(N-n)^2}{N^2} \right| \\
&= \left| 1 - (1 + \frac{n^2 - 2n}{N^2}) \right| \\
&= \frac{n^2 - 2n}{N^2} \\
&\to 0
\end{aligned}$$

Thus the M_h are absolutely regular.

2. By Nobel and Dembo (1992) [109], the Glivenko-Cantelli theorem holds for M_h/n using the marginal distribution. Using the result from Theorem 1:

$$\begin{aligned}
&\sup_{x\in(\pi_1,1)} |\hat{S}_n(x) - (1-\pi_1)S_{1-\pi_1}(x)| \overset{a.s.}{\to} 0 \\
\Rightarrow &\int_{\pi_1}^{1} (\hat{S}_n(x) - (1-\pi_1)S_{1-\pi_1}(x))^2\, dx \overset{a.s.}{\to} 0 \\
\Rightarrow &SS(\pi_1) \to 0.
\end{aligned}$$

3. For an ideal setting where the proportion of reproducible signals is $\pi_1^* \neq \pi_1$,

$$\hat{S}_n(x) \to S_{\pi_1^*}(x)$$

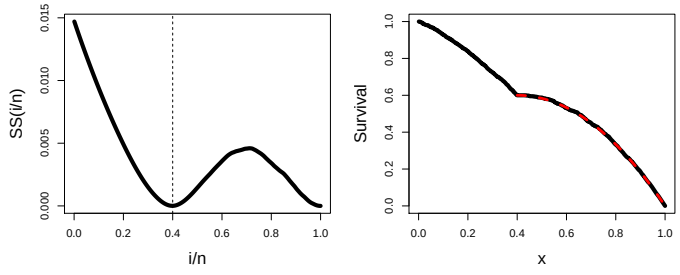
and $\exists x \in (\pi_1, 1)$ such that $S_{\pi_1}(x) \neq S_{\pi_1^*}(x)$. Because $S_{\pi_1}(x)$, $S_{\pi_1^*}(x)$ are continuous,

$$\int_{\pi_1}^{1} (\hat{S}_n(x) - \pi_0 S_{\pi_0}(x))^2\, dx \nrightarrow 0 \;\Rightarrow SS(\pi_1) \neq 0$$

4. Thus by (1), (2), (3) above,

$$\hat{\pi}_1 = \underset{\lambda\in(0,1)}{\arg\inf}\{SS(\lambda)\} \to \pi_1$$

□



(a) $SS(i/n)$ (b) Survival functions

Figure 4.4: Continuing the example from Figure 4.3, (a) shows the values of $SS(i/n)$ for $i = 0, 1, \ldots, n$, and (b) the empirical survival function (solid black) overlaid with the theoretical survival function (dashed red)for this data. Here, the true π_1 is 0.40, and the resulting estimate is $\hat{\pi}_1 = 0.40$.

Figure 4.4 illustrates the calculation of $\hat{\pi}_1$ for the sample ranks presented in Figure 4.3a. For a fixed n, the possible values for $\hat{\pi}_1$ can be restricted to $0, 1/n, \ldots, (n-1)/n$. As seen in Figure 4.4, $SS(i/n)$ is very small near the correct π_1, and for i/n close to 1. It is small for $i/n \in (0.9, 1)$ because this part of the survival function is very similar regardless of the true value of π_1. For this reason, in practice it is necessary to consider only values of $\hat{\pi}_1 \in (0, .9)$. Once $\hat{\pi}_1$ has been determined, it is assumed that gene g is reproducible if $M_g/n \leq \hat{\pi}_1$. In the ideal setting, it also means that gene h is irreproducible if $M_h/n > \hat{\pi}_1$. This perfect split is unrealistic, however, and we must define an estimate of the false discovery rate for each gene with $M_g/n > \hat{\pi}_1$. In the next section we relax the ideal assumptions, and derive a false discovery rate estimate.

4.2.2 Estimation of false discovery rates in realistic settings

In this section we make assumptions that are reasonably met by nearly all data sets, and use them in conjunction with $\hat{\pi}_1$ to estimate marginal false discovery rates of rejection regions for M_g. For a realistic setting, we continue to assume (I2) and (I3) from the idealistic setting, but relax (I1):

Assumptions for a realistic setting

(R1) Reproducible signals *tend* to be ranked higher than irreproducible signals. Thus R_g^x tends to be smaller than R_h^x and R_g^y tends to be smaller than R_h^y if gene g is reproducible and gene h is irreproducible.

(I2) The correlation between the ranks of reproducible signals is non-negative.

(I3) The two ranks per irreproducible gene are *independent*.

The primary difference between these assumptions and those of the ideal case is the lack of a clear split between reproducible and irreproducible signals in terms of the value of M_g. The estimator $\hat{\pi}_1$ derived in Section 4.2.1 is consistent in the ideal case, but systematically underestimates π_1 in the more realistic setting. Through simulations similar to those described in Sections 4.3.1 and 4.3.2, we have determined that $\hat{\pi}_1$ is a good estimate of when reproducible signals *begin* to transition to irreproducible signals. For convenience, we now move away from

the unit interval and work with the originally scaled $M_g = 1, \ldots, n$. We define a discrete and re-scaled version of $\hat{\pi}_1$:

$$\hat{k} = \underset{i=0,1,\ldots,\lfloor .9n \rfloor}{\arg\min} \{SS(i/n)\} \tag{4.10}$$

This value $\hat{k}$ marks approximately the minimum value of M_h for irreproducible genes. In other words, if $M_g \leq \hat{k}$ then it is almost certainly associated with a reproducible gene. Genes with $M_g > \hat{k}$ are potentially irreproducible.

Note that we recommend using $\lfloor .9n \rfloor$ as the maximum possible value for $\hat{k}$. We choose this value to ensure $\hat{k}$ is estimated as the first local minimum in the $SS(i/n)$ curve, as this curve tends to zero as i approaches n. For certain datasets with small effect size, $\lfloor .9n \rfloor$ may need to be reduced to ensure accuracy. Sample datasets and $SS(i/n)$ curves are presented in Figure 4.5, which is further described later in this section.

Recall that this procedure uses the value of M_g to determine which genes are reproducible. It accomplishes this task by choosing a critical value, $\hat{N}$, and declaring all genes associated with $M_g \leq \hat{N}$ as reproducible. This approach is akin to defining a rejection rejection as $(0, \hat{N})$, and rejecting the null hypothesis of irreproducibility for all signals with M_g in this region [110]. We use the term 'false discovery' to describe the type 1 error committed when an irreproducible gene is declared reproducible. In this procedure, we estimate a marginal false discovery rate [13] based on a rejection region. This quantity is closely related to the classical false discovery rate (FDR) as defined by Benjamini and Hochberg (1995) [7], and we defined and briefly discussed it in Section 1.1.2. We will again define it here using slightly different notation according to Table 4.2. We change notation for this section to avoid confusing between the R of total rejections and common notation for ranks.

Table 4.2: Decision outcomes for m hypothesis tests

	Fail to reject null	Reject null	Total
Null is true	U	V	m_0
Null is false	T	S	$m - m_0$
Total	$m - Q$	Q	m

Consider Table 4.2 detailing possible decision outcomes for m simultaneous hypothesis tests. This table is identical to Table 1.1, however we have exchanged the notation for total number of rejections R and replaced it with Q to avoid confusion with notation for ranks. Using this new notation, the marginal false discovery rate (mFDR) [13] is defined below:

$$mFDR = \frac{E[V]}{E[Q]} \tag{4.11}$$

To describe our approach to mFDR estimation, we use the following notation:

$$Q(i) = \sum_{g=1}^{n} I(M_g \leq i) = \#\text{ genes declared reproducible for critical region } (0, i) \tag{4.12}$$

$$V_k(i) = \#\text{ irreproducible genes declared reproducible with } k < M_g \leq i \tag{4.13}$$

Using this notation, the estimated mFDR for using i as the threshold value for declaring reproducibility is below:

$$\widehat{mFDR}(i) = \frac{E(V_{\hat{k}}(i)))}{Q(i)} \tag{4.14}$$

The denominator of this expression is determined directly from data, however the numerator must be calculated using the distribution of M_h calculated in Section 4.2.1, and dependent on $\hat{\pi}_1 = \hat{k}/n$. With the value of $\hat{k}$ determined, all genes with $M_g \leq \hat{k}$ are declared reproducible. We further assume there are $n - \hat{k}$ irreproducible genes, h, with $M_h > \hat{k}$. In a realistic setting, there is no clear division between gene types and some reproducible genes will have $M_g > \hat{k}$. Thus there will be $n - Q(\hat{k})$ genes with $M_g > \hat{k}$. We therefore include the factor $(n-\hat{k})/(n-Q(\hat{k}))$ in calculations to express the proportion of signals with $M_g > \hat{k}$ that are irreproducible. The calculation of the numerator is thus detailed below:

$$\begin{aligned} E(V_{\hat{k}}(i)) &= (n - \hat{k}) \cdot P_{n,\hat{k}/n}(M_h \leq i) \cdot \frac{n - \hat{k}}{n - Q(\hat{k})} \\ &= (n - \hat{k}) \cdot \frac{(i - \hat{k})^2}{(n - \hat{k})^2} \cdot \frac{n - \hat{k}}{n - Q(\hat{k})} \end{aligned}$$

$$= \frac{(i-\hat{k})^2}{n - Q(\hat{k})},\ i = \hat{k}+1, \ldots, n \tag{4.15}$$

We can then define the estimated mFDR associated with any rejection region $(0, i)$.

$$\widehat{mFDR}(i) = \frac{E(V_{\hat{k}}(i))}{Q(i)} = \frac{(i-\hat{k})^2}{Q(i)(n - Q(\hat{k}))},\quad i = \hat{k}+1, \ldots, n \tag{4.16}$$

Thus the false discovery rate is controlled at a nominal level α if the threshold value $\hat{N}$ is chosen to be

$$\hat{N} = \max_{\hat{k}<i\leq n} \{i : \widehat{mFDR}(i) \leq \alpha\}, \tag{4.17}$$

and all genes with maximum rank statistics less than or equal to $\hat{N}$ are declared reproducible. We apply this procedure to three sample data sets and present the results visually in Figure 4.5. For each data set, we plot the rank pairs, the $SS(i/n)$ curves used to determine $\hat{k}$, and the actual maximum rank statistics. Achieved false discovery rates are included.

We now summarize the Maximum Rank Reproducibility procedure for set of n genes each with two measurements generated from replicated experiments.

MaRR Procedure:

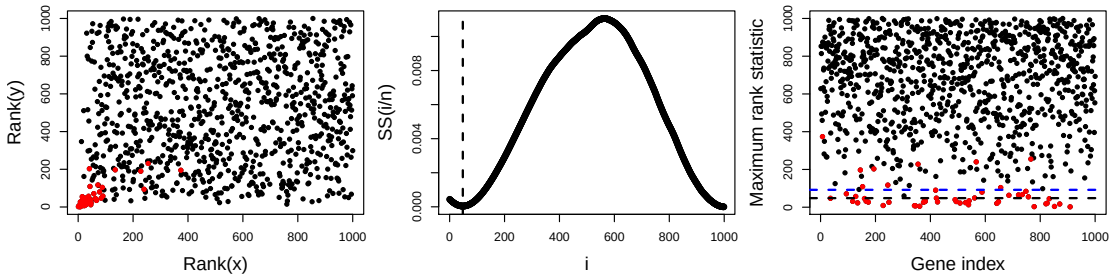
To control FDR at a nominal level of α, Define:

$$\hat{k} = \underset{i=0,1,\ldots,\lfloor .9n \rfloor}{\arg\min} \left\{ SS(i/n) = (1-i/n)^{-1} \int_{i/n}^{1} \left(\hat{S}_n(x) - (1-i/n) S_{i/n}(x) \right)^2 dx \right\},$$

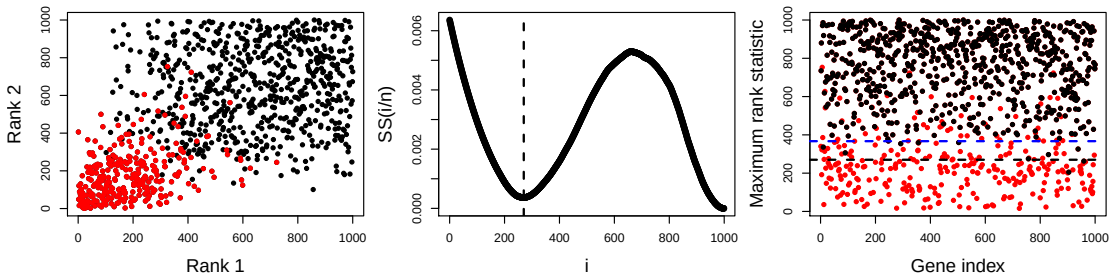
where $\hat{S}_n(x)$ is the empirical survival function, and $S_{i/n}(x)$ is the limiting survival function defined in Theorem 1. Define $\hat{N}$ as:

$$\hat{N} = \max_{\hat{k}<i\leq n} \left\{ i : \widehat{mFDR}(i) = \frac{(i-\hat{k})^2}{Q(i)(n - Q(\hat{k}))} \leq \alpha \right\}$$

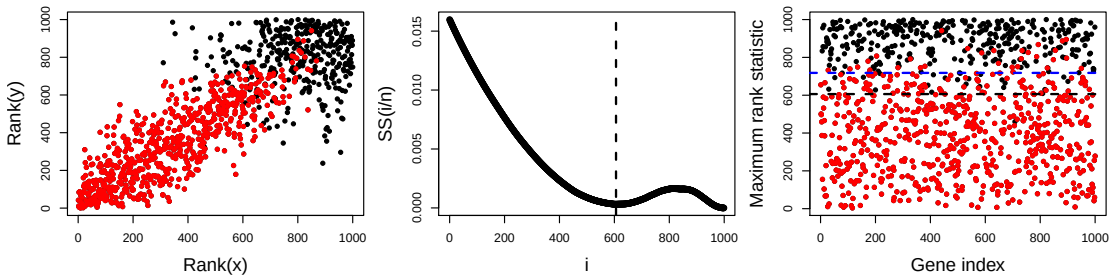
where $Q(i)$ is the number of genes with $M_g \leq i$. Reject all genes g associated with maximum ranks M_g less than or equal to $\hat{N}$.



(a) 50 reproducible genes, $\hat{k} = 46$, $\hat{N} = 87$, and achieved FDR is 0.028



(b) 300 reproducible genes, $\hat{k} = 270$, $\hat{N} = 367$, and achieved FDR is 0.038



(c) 650 reproducible genes, $\hat{k} = 606$, $\hat{N} = 718$, and achieved FDR is 0.043

Figure 4.5: Illustration of procedure for three settings, each with $n = 1000$: (a): $\pi_1 = 0.05$, (b): $\pi_1 = 0.30$, (c): $\pi_1 = 0.65$. The left column presents the rank pairs for each data set, with red dots indicating reproducible genes. The middle column shows the $SS(i/n)$ curves used to determine $\hat{k}$. The right column shows the actual maximum rank statistics, with horizontal lines indicating estimated values of $\hat{k}$ and $\hat{N}$.

4.3 Simulation studies

In this section we describe two simulation studies performed to assess the performance of the MaRR procedure. For both studies we compare our results to those of the copula mixture approach of Li et al. (2011) [105]. Li et al. compared their approach to existing p-value combination techniques such as Fisher's and Stouf-

fer's, finding in all cases that their approach offered a clear advantage. Because these comparisons have already been performed, in these simulation studies we compare the MaRR approach only to the copula mixture model. Both approaches perform analysis on the rank scale, negating the need to calculate p-values and allowing the simulations to proceed directly from simulated Z-statistics.

In both studies we set $\alpha = 0.05$, and vary the proportion of reproducible signals, π_1. In the first study we simulate data following Li et al., and in the second we use simulation settings that violate assumptions employed by the copula mixture model approach. We evaluate FDR control in all settings and for both procedures.

4.3.1 Settings for Simulation A

In the first study we consider three settings as detailed in Table 4.3, varying both the proportion of reproducible genes (π_1) and the correlation between these signals (ρ). We assume that large values of test statistics will be highly ranked, corresponding to calculation of right-tail one-sided p-values. The test statistics for reproducible signals are generated as follows:

$$\begin{pmatrix} Z_{g,1} \\ Z_{g,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{4.18}$$

Irreproducible signals are generated from the standard bivariate normal distribution:

$$\begin{pmatrix} Z_{h,1} \\ Z_{h,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \tag{4.19}$$

For each setting, we simulated 100 data sets of size $n = 2000$. We applied the MaRR procedure with $\alpha = 0.05$ to each data set, recording the resulting achieved false discovery rates. The copula mixture procedure requires specification of initial parameter values for μ_1, ρ, $\sigma = 1$, and π_1 to perform an expectation maximization algorithm. We thus performed the procedure ten times on each data set, drawing initial parameters from uniform distributions with the domains detailed in Table 4.3. The results from the parameters that gave the highest likelihood were recorded. The achieved false discovery rate was recorded by declaring as reproducible all signals with estimated irreproducible discovery rate less than $\alpha = 0.05$. The results are presented and discussed in Section 4.3.4.

Table 4.3: Summary of parameter values for simulation A, and initial parameter input ranges for the copula-based approach.

			Domains of Uniform distributions for tuning parameters		
Setting	π_1	ρ	$\pi_{1,initial}$	$\rho_{initial}$	$\mu_{initial}$
A_1	0.05	0.84	(0.0,0.1)	(0.40,0.84)	(1.5,3.5)
A_2	0.30	0.40	(0.2,0.4)	(0.40,0.84)	(1.5,3.5)
A_3	0.80	0.84	(0.7,0.9)	(0.40,0.84)	(1.5,3.5)

4.3.2 Settings for Simulation B

In the second simulation, we again assume that each pair of reproducible signals follows a bivariate normal distribution with positive correlation but we now assume there are two different groups of reproducible signals. This simulation setting is in fact similar the S4 setting used by Li et al., when the presence of a second group of reproducible signals with different correlation structures violates assumptions made by the copula-mixture model procedure. We perform this simulation to test the performance of our procedure in a situation that may be difficult for existing methods. Each pair of reproducible test statistics is thus drawn from:

$$\begin{pmatrix} Z_{g,1} \\ Z_{g,2} \end{pmatrix} \Bigg| K_g = i \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix} \right), \quad K_g \sim Bernoulli(0.8) + 1, \qquad (4.20)$$

where $\mu_1,\ \mu_2,\ \rho_1$, and ρ_2 are detailed in Table 4.4 for each of three different settings. Irreproducible signals are generated following Equation 4.19. As with simulation A, for each setting we generate 100 data sets of size $n = 2000$ and record results from both our proposed procedure and the copula mixture model approach. For this simulation, we selected the highest likelihood from 25 initial starting parameters for the copula mixture model approach. Domains of the uniform distributions from which we drew these parameters are detailed in Table 4.4.

4.3.3 Settings for Simulation C

In this third simulation study, we investigate the FDR controlling properties of the MaRR procedure when assumption (I3) is violated. (I3) states that the ranks of irreproducible signals are independent, and upon this assumption hinges the

Table 4.4: Summary of parameter values for simulation B, and initial parameter input ranges for the copula-based approach.

	Parameters			Domains of Uniform distributions for tuning parameters		
Setting	π_1	$\mu_1,\ \rho_1$	$\mu_2,\ \rho_2$	$\pi_{1,initial}$	$\rho_{initial}$	$\mu_{initial}$
B_1	0.20	3, 0.5	2, 0.3	(0.0,0.5)	(0.20,0.60)	(1,3)
B_2	0.50	3, 0.5	2, 0.3	(0.3,0.7)	(0.20,0.60)	(1,3)
B_3	0.60	3, 0.3	2, 0.7	(0.3,0.9)	(0.20,0.90)	(1,3)

estimation of $\hat{k}$. To explore this situation, we let 50% of signals be reproducible and generated according to

$$\begin{pmatrix} Z_{g,1} \\ Z_{g,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}, \begin{pmatrix} 1 & .84 \\ .84 & 1 \end{pmatrix} \right). \tag{4.21}$$

The irreproducible signals are generated following

$$\begin{pmatrix} Z_{h,1} \\ Z_{h,2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix} \right), \tag{4.22}$$

and we allow ρ_0 to vary from 0 to 0.20 following the settings detailed in Table 4.5. For each setting we generated 100 data sets each of size $n = 2000$, applied the MaRR procedure, and recorded the achieved false discovery rate.

Table 4.5: Summary of parameter values for simulation C

Setting	$\mu_1,\ \rho_1$	ρ_0
C_1	2.5, 0.84	0.00
C_2	2.5, 0.84	0.05
C_3	2.5, 0.84	0.10
C_4	2.5, 0.84	0.15
C_5	2.5, 0.84	0.20

4.3.4 Simulation results

We discuss the results of the simulation studies, and use them to evaluate the performance of the MaRR procedure in different settings. The results from simu-

lations A and B are presented in Figures 4.6 and 4.7 respectively. In all settings but A_1, the MaRR procedure is conservative in its estimation of $m\widehat{FDR}$. Setting A_1 contained only 5% reproducible signals, and with this low level of reproducibility the MaRR procedure was slightly anti-conservative. This tendency is explained by the estimation of $m\widehat{FDR}$: $m\widehat{FDR}$ is assumed to be 0 for rejection regions contained in $(0, \hat{\pi}_1)$. When even a single false rejection occurs in this interval, the achieved FDR is thus inflated greatly because the total number of rejections will be small for small π_1. In practice, researchers very concerned with strict FDR control could set $\hat{k}$=0 when $\hat{\pi}_1$ is suspected to be very small.

Noticeably, the MaRR procedure produces reasonably well-calibrated FDR in each setting without any need for input parameters. The copula mixture model approach does better in some settings, particularly A_1, but worse in the settings with either lower correlation between reproducible signals (A_2), or when there are more than one group of reproducible signals (B_1, B_2).

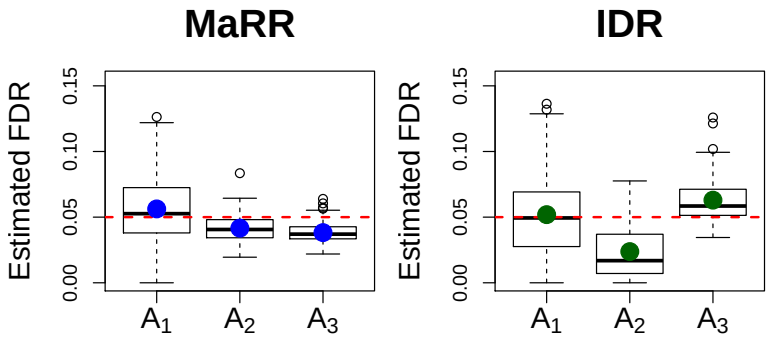


Figure 4.6: FDR results for simulation A based on 100 simulated data sets in each setting. Green circles represent mean FDR values using the Li et al. procedure, and blue circles represent mean FDR values using proposed procedure.

These simulations show two primary facts. First, the MaRR procedure performs well in a variety of situations, with care needed only when the proportion of reproducible signals is low. Second, the MaRR procedure is a valuable tool because it requires no initial input values or calibration. Results from the MaRR procedure such as $\hat{\pi}_1$ and the Spearman's rank correlation of signals determined to be reproducible could be used as initial parameter values in other procedures such as the copula-mixture model approach.

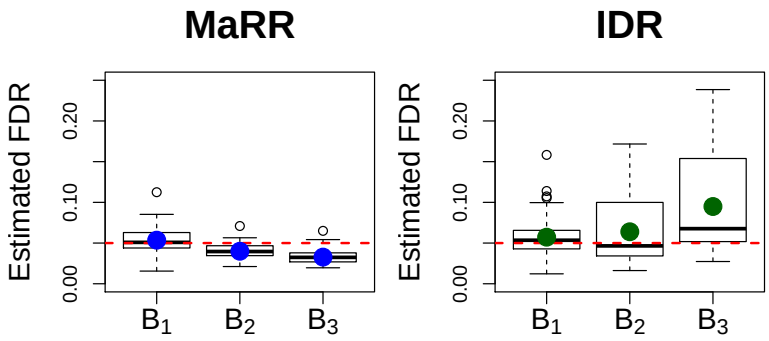


Figure 4.7: FDR results for simulation B. Green circles represent mean FDR values using the Li et al. procedure, and blue circles represent mean FDR values using proposed procedure.

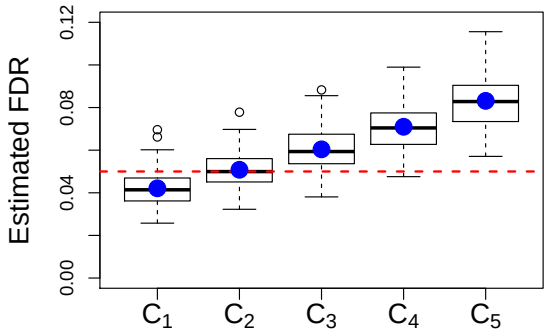


Figure 4.8: FDR results for simulation C,. Blue circles represent mean FDR values. Notice the linear behavior of estimated FDR as a function of ρ_0.

The results from Simulation C are presented in Figure 4.8. These results show that when assumption (I3) is violated the MaRR procedure is anti-conservative and loses FDR control. Further, it appears that the level of anti-conservativeness is a linear function of the correlation between irreproducible ranks. Although the results of this simulation study are initially discouraging, they also present a compelling opportunity for improvement. Through modeling of correlation between genes declared irreproducible, adjustments could be made to accurately control FDR. This will certainly be considered in future work.

4.4 Data application

We apply our method to two different problems: (1) assessment of three different peak callers in ChIP-seq experiments; and (2) analysis of cell cycle data involving the fission yeast schizosaccharomyces pombe.

4.4.1 Peak caller assessment

One current use of high-throughput technologies is the identification of protein binding sites along a genome via ChIP-seq experiments. In these experiments the sections of genome bound by the protein of interest are enriched through immunoprecipitation, and these enriched regions are sequenced to generate tag counts. Measurement of the relative importance of binding sites is performed via computational transformation of these tag counts. The algorithm used for this transformation is called a peak-caller. Many peak-callers are in current use, including some that generate as outputs p-values or q-values for each candidate region. Typically, regions are identified for further study when these values are in pre-specified interval.

In this section of the manuscript, we use the output of nine different peak-callers applied to biological replicate ChIP-seq experiments of a transcription factor CTCF. For details of data generation and processing, refer to [105]. The goal of this analysis to identify the level of reproducibility of protein binding sites for the various peak-callers through application of the MaRR procedure.

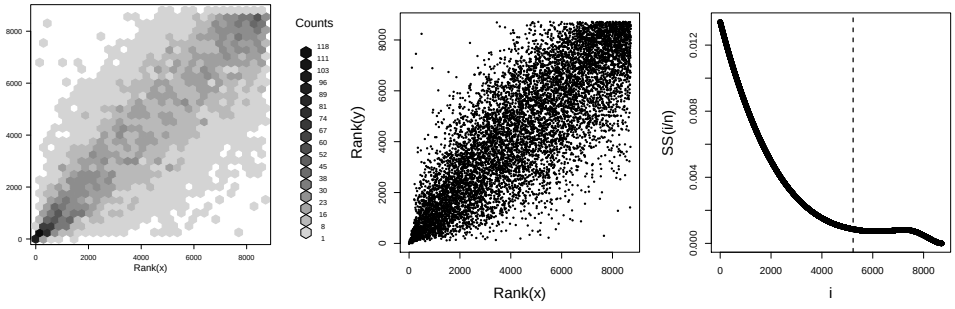
Li et al. [105] applied their copula-based method to the following nine peak callers: Peakseq [111], MACS [112], SPP [113], Fseq [114], Hotspot [115], Erange [116], Cisgenome [117], Quest [118], and Sissrs [119]. Of these nine, six were found to have both reproducible and irreproducible components, and the remaining three were concluded to have only one component with a uniform correlation structure according to a likelihood ratio test for the number of components. We applied the MaRR approach to each of these peak-callers, with results summarized in Table 4.6. The two methods are in general agreement for six of the nine, but the methods reach different conclusions for the callers believed to have only one component. We describe two of these three in more detail, along our results for one peak caller for which both methods agree.

Table 4.6: Summary of results for the copula-model and MaRR approaches for each of nine peak callers. We calculated a rank correlation using the MaRR procedure by computing Spearman's rank correlation coefficient for the signals declared reproducible.

		IDR		MaRR		
Peak Caller	n	π_1	ρ	$\hat{\pi}_1$	$Q(\hat{N})/n$	ρ
Peakseq	35994	0.69	0.89	0.735	0.777	0.87
MACS	32853	0.84	0.89	0.730	0.773	0.84
SPP	33497	0.77	0.88	0.722	0.764	0.83
Fseq	27362	0.74	0.82	0.721	0.758	0.76
Hotspot	23810	0.69	0.88	0.733	0.776	0.86
Erange	8716	0.72	0.81	0.736	0.765	0.78
Cisgenome	11372	0.85	0.65	0.369	0.352	0.46
Quest	7651	0.72	0.67	0.240	0.300	0.60
Sissrs	6526	1	0.24	0.024	0.027	0.24

4.4.1.1 Agreement: Erange

The Erange peak caller shows a high proportion of reproducible genes with moderate correlation between these ranks. See Figure 4.9 for illustration of these bivariate ranks. The MaRR procedure estimated $\hat{\pi}_1$ as 0.736 for this set of genes, and ultimately declared 6668 (76.5%) to be reproducible. Spearman's rank correlation for the ranks of only these genes is 0.78. This result agrees well with that using the copula mixture model, which estimated π_1 as 0.72 and ρ as 0.81.



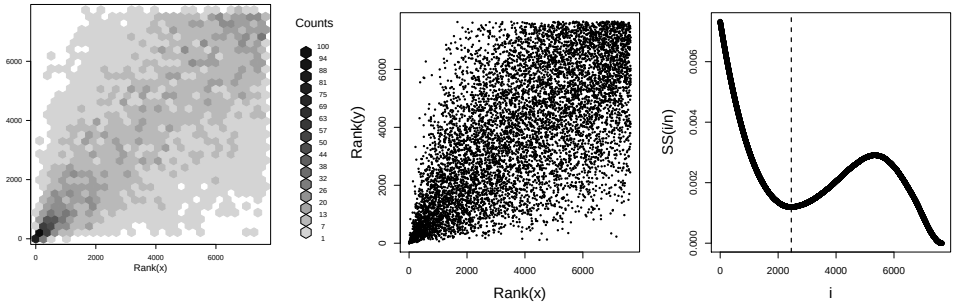
(a) Hexbin visualization (b) Rank Pairs (c) $SS(i)$

Figure 4.9: Analysis of Erange peak-caller: hexbin representation of frequencies (a), graphical representation of bivariate ranks (b), and the values of $SS(i)$ for all possible i (c).

4.4.1.2 Discrepancy: Quest and Sissrs

There were three peak callers for which the MaRR procedure gave a different result from the copula-based procedure: Quest, Sissrs, and Cisgenome. For these three the copula-based procedure indicated that all genes were part of one group of signals. In the case of Sissrs, this group had weakly correlated signals, and for Quest and Cisgenome the conclusion was that a stringent cut-off had been applied and thus very little of the noise component was kept in the data set. The MaRR procedure, however, indicated for all three that there is evidence of at least two groups of signals that transition from reproducible to irreproducible. In this section we discuss two of these. Our analyses for Cisgenome and Quest gave similar results, thus we include discussion here only of Sissrs and Quest.

First we consider the Quest peak caller. Graphical representations of the bivariate ranks and the $SS(i/n)$ curve are provided in Figure 4.10. The MaRR procedure estimates $\hat{\pi}_1$ to be 0.32, and ultimately finds 2292 genes, or 30%, to be reproducible. The Spearman's rank correlation for these 2292 genes is 0.60, indicating a reasonable level of agreement between ranks for reproducible genes.

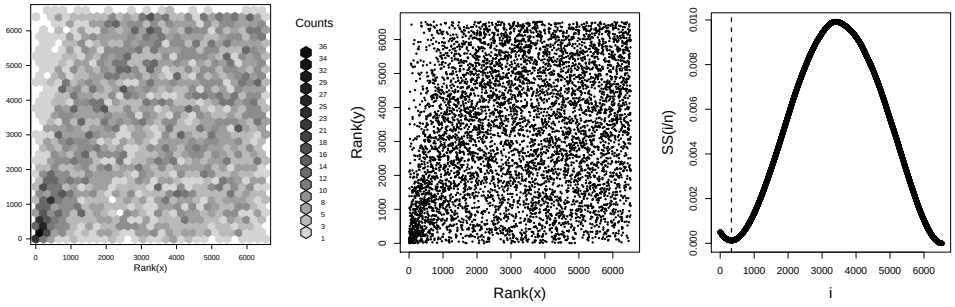


(a) Hexbin visualization (b) Rank Pairs (c) $SS(i)$

Figure 4.10: Analysis of Quest peak-caller: hexbin representation of frequencies (a), graphical representation of bivariate ranks (b), and the values of $SS(i)$ for all possible i (c).

We next consider the Sissrs peak caller. Figure 4.11 shows the bivariate ranks of all genes for this peak caller, which are spread in a manner consistent with a high proportion of irreproducible genes. In this case, the MaRR procedure estimates $\hat{\pi}_1$ to be 0.052, but ultimately declares only 177, or 2.7% of genes, to be reproducible. The Spearman's rank correlation for these 177 genes is 0.24. Upon close inspection,

this result is similar to the conclusion made by Li et al., who surmised that all genes were from a single group with low correlation between replications.



(a) Hexbin visualization (b) Rank Pairs (c) $SS(i)$

Figure 4.11: Analysis of Sissrs peak-caller: hexbin representation of frequencies (a), graphical representation of bivariate ranks (b), and the values of $SS(i)$ for all possible i (c).

4.4.2 Periodicity of yeast cell genes

In the early 2000's numerous studies were performed to analyze the periodicity of gene expression for the yeast organism schizosaccharomyces pombe. [120] provides a good summary of these analyses. Here we analyze data produced by [1]. To generate this data, yeast cells were synchronized and expression levels measured at equal time points after synchronization. To assess evidence of periodicity, a Fisher's G-statistic was calculated for each gene from each experiment. Three data sets were generated, two of which were produced using the same synchronization technique and the third using a different synchronization technique. We analyze the two data sets both produced using elutriation synchronization to assess the reproducibility between these two experiments, and compare our results to existing p-value combination approaches. The lack of reproducibility between these experiments has been discussed in [120], despite the large number of very small p-values in both experiments. This is an interesting data set to apply the MaRR approach to because of this discrepancy.

Figure 4.12 presents histograms of p-values calculated using Fisher's G-statistic for each experiment, and a plot of the bivariate p-values in the unit square. We include for analysis only genes with complete measurements at all time points in

both experiments, a total of 1865. Notice the large proportion of very small p-values for each data set, potentially indicating a large proportion of genes with significantly periodic expression.

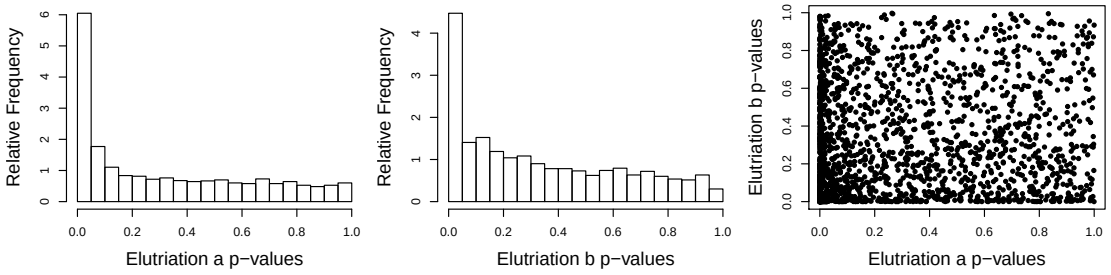
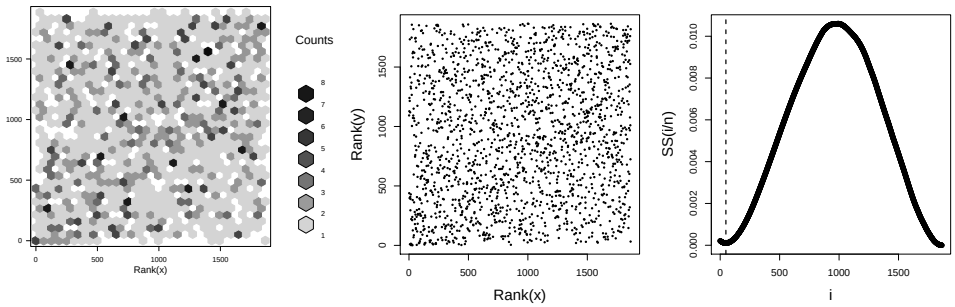


Figure 4.12: Oliva et al. data

To test for periodic expression in both experiments simultaneously, various p-value combination techniques can be applied. Fisher's combined p-value test [121] and Stouffer's p-value combination approach [37] give 499 and 438 rejections respectively, after controlling FDR using the B-H procedure. This high number of rejections results from the underlying hypothesis tested by these procedures: the gene is periodically expressed in *at least one experiment*. To test a stricter hypotheses, that the gene is periodic in *both* experiments, application of the B-H procedure to the maximum p-value for each gene [33] gives 15 rejections. These same hypotheses can be tested using the Voronoi P-value combination approach [34], resulting in 249 rejections. These rejections are indicative of the large number of genes with low p-values in both experiments.

The question of reproducibility between the experiments is separate from the hypothesis of significant periodicity: instead of using actual values of p-values, we are interested in their relative values, and how comparable they are between studies. Figure 4.13 shows the bivariate ranks produced from ranking the p-values from elutriation a and elutriation b. This figure clearly shows a very low level of agreement between the ranks of these two data sets. Application of the MaRR procedure shows that $\hat{\pi}_1$ is correspondingly low: 5%. In fact, only 7 genes are found to reproducible for $\alpha = .05$. This result confirms that there is very little reproducibility between replicate experiments for periodic genes, despite the large marginal evidence of periodicity.



(a) Hexbin visualization (b) Rank Pairs (c) $SS(i)$

Figure 4.13: Analysis of Oliva et al. elutriation a and b data: hexbin representation of frequencies (a), graphical representation of bivariate ranks (b), and the values of $SS(i)$ for all possible i (c).

4.5 Discussion

In this chapter we have introduced a nonparametric approach to detect reproducible genes from replicate experiments. The MaRR approach is appropriate for many different measurement types, does not rely on tuning parameters, and makes very few assumptions about the distribution of reproducible signals. For these reasons, it offers an advantage over existing procedures that may be sensitive to misspecified tuning parameters, or to choice of threshold values.

Although the MaRR procedure was shown to be robust to different settings for reproducible signals, there are also situations in which the MaRR procedure was unable to control FDR at the specified level. These situations included the violation of the assumption that irreproducible ranks are independent, and the presence of a very low proportion of reproducible signals. In future work we hope to explore adjustments to the MaRR procedure that will solve these issues.

This procedure can be extended to higher dimensions in a straightforward fashion through derivation of the maximum rank distribution in the ideal case when more than two replicated experiments are considered. Similarly, we could derive the distribution of the second highest maximum rank statistic, third highest, etc. Further research into these scenarios is related to the problem of the partial conjunction hypothesis for reproducibility [35]. Future research on using observed maximum rank statistics will boost power for the MaRR procedure by modeling the distribution of maximum ranks from reproducible signals.

In practice, researchers may wish to combine results from both the MaRR and copula-based procedures. The estimate of $\hat{\pi}_1$ from the former could serve as a tuned initial value for the latter, leading to better estimation of effect size, a less conservative estimate of the false discovery rate, and more accurate modeling of the covariance structure for reproducible signals. Another noteworthy combination of methods could be the use of Spearman's rank correlation to the set of genes identified as reproducible by the MaRR procedure. This will give a good indication of the strength of correlation for a set of genes predetermined to meet the criteria of reproducibility.

Chapter 5
Concluding remarks

In this dissertation we have introduced and described novel procedures to control the false discovery rate in three specific settings. For each setting we have motivated the need for type I error control, reviewed existing approaches, and discussed relevant topics such as Voronoi diagrams. We have described all three procedures and shown their efficacy through both simulations and application to real data sets. The procedures introduced here advance the field of multiple testing by giving researchers three powerful new tools to use in future genetic studies.

This chapter proceeds in three sections. First, we will summarize the content of the dissertation in Section 5.1. Next, we list the specific contributions that have been made to the field in Section 5.2. Finally, we discuss future directions of this work in Section 5.3.

5.1 Dissertation Summary

In this section we summarize the topics discussed in this dissertation.

In Chapter 1 we defined quantities related to type I error control in the context of multiple testing: the family-wise error rate, false discovery rate, positive false discovery rate, marginal false discovery rate, and local false discovery rate. We included discussion of existing procedures developed for control of these quantities and described in detail the celebrated Benjamini-Hochberg procedure.

In Chapter 2 we discussed difficulties that arise when the p-values considered for hypothesis testing are discrete and introduced a procedure that controls the false discovery rate for the specific setting of unreplicated next-generation sequencing data. The Adapted generalized Benjamini-Hochberg procedure uses the marginal

read counts for each gene in a data-dependent simulation-based estimation of expected p-value spacings under the null hypothesis. These estimates serve as part of re-phrasing of the Benjamini-Hochberg procedure to declare significance. Through simulations, we showed that the AgenBH procedure effectively controls the false discovery rate, has improved power properties over the traditional B-H procedure, and offers a slight computational advantage over existing discrete methods.

Chapter 3 introduced a novel procedure to test the disjunction hypothesis using Voronoi diagrams. In this chapter we first motivated the need for such a procedure using two separate genetic examples. Next, we discussed the disjunction framework and provided a detailed introduction to the definition, use, and construction of the Voronoi diagram. We then introduced our p-value combination approach and illustrated its properties in four simulation studies. The approach defines two-dimensional spacings using a Voronoi diagram, and creates summarized values using cumulative cell areas. These summarized values are then suitable for application of error-control procedures such the B–H procedure or empirical null approaches. Finally, we applied the procedure to two datasets: the first had the aim to detect periodically expressed yeast genes while the second aimed to identify genes implicated in prostate cancer progression. We found that this procedure offers a clear advantage over existing approaches in terms of power properties. It also has the desirable property of being robust to any correlation structure of the test statistics used to generate two-dimensional p-vectors. There are numerous extensions for this work which will be of interest in the future.

The third novel procedure, which we introduced in Chapter 4, is suitable for the identification of reproducible genes in replicate experiments. In this chapter we first motivated the need for a non-parametric approach for the assessment of reproducibility before introducing the Maximum Rank Reproducibility procedure. The MaRR procedure uses a maximum rank statistic to estimate the proportion of reproducible genes by comparing observed and theoretical survival functions. We also detailed how to estimate marginal false discovery rates. We investigated power and error control properties in two sets of simulation studies. In these studies we compared the performance of the MaRR procedure to a model-based approach. In this chapter we applied the MaRR procedure to the evaluation of nine peak-calling algorithms. We also revisited the yeast data analyzed in Chapter 3. We found that our procedure performs well in a variety of settings, including those that violate

assumptions made by the model-based approach.

5.2 Contributions

This dissertation made multiple contributions to the larger field of statistics and to the specific field of multiple testing. We list them here in the order in which they appeared.

1. Development and presentation of the Adapted generalized Benjamini-Hochberg procedure for control of the false discovery rate when discrete p-values are calculated using next-generation sequencing data.

2. Presentation of a concise summary of the Voronoi diagram, including definitions, computational concerns, and existing statistical applications.

3. Development and presentation of a powerful new procedure to test the disjunction hypothesis using Voronoi diagrams. This is the first time that the Voronoi diagram has been used as an extension of one-dimensional spacings in the context of multiple hypothesis testing. It is also the first new approach to testing the disjunction hypothesis to be proposed in over 50 years.

4. Development and presentation of the Maximum Rank Reproducibility procedure for identification of reproducibility in high-throughput replicate experiments. This novel procedure is a valuable development in this area because it is free of model assumptions about reproducible genes and requires no tuning parameters.

5.3 Future Directions

In this final section we highlight extensions and related problems that represent future research opportunities arising from the topics discussed in this dissertation.

The Voronoi approach to p-value combination was shown to be a valuable tool for testing the disjunction hypothesis using two-dimensional p-vectors. A wealth of future directions are possible stemming from the material included in Chapter 3:

1. Extension of the procedure to higher dimensions through higher-dimensional Voronoi diagrams.

2. Extension to higher dimensions through the approaches related to that described in Section 3.7.

3. Investigation of FDR controlling properties as a function of the ordering scheme used to define cumulative sums of cell areas.

4. Extension of the procedure to the conjunction or partial conjunction hypotheses through revised ordering schemes.

The MaRR procedure also offers a number of future directions for research:

1. Extension to higher dimensions through derivation of the distribution function for maximum rank statistics calculated using more than two marginal ranks.

2. Extension to higher dimensions and to the partial conjunction hypothesis framework through derivation of distribution functions for other statistics calculated using marginal ranks.

3. Sharpening of error control through modeling of observed maximum rank statistics, both reproducible and irreproducible.

Bibliography

[1] Oliva, A., A. Rosebrock, F. Ferrezuelo, S. Pyne, H. Chen, S. Skiena, B. Futcher, and J. Leatherwood (2005) "The cell cycle–regulated genes of Schizosaccharomyces pombe," *PLoS Biol.*, **3**(7), p. e225.

[2] Holm, S. (1979) "A simple sequentially rejective Bonferroni test procedure," *Scand. J. Stat.*, **6**, pp. 65–70.

[3] Hochberg, Y. (1988) "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, **75**, pp. 800–803.

[4] Hommel, G. (1989) "A comparison of two modified Bonferroni procedures," *Biometrika*, **76**, pp. 624–625.

[5] Shaffer, J. P. (1995) "Multiple hypothesis testing," *Annu. Rev. Psychol.*, **46**, pp. 561–584.

[6] Tarone, R. (1990) "A modified Bonferroni method for discrete data," *Biometrics*, pp. 515–522.

[7] Benjamini, Y. and Y. Hochberg (1995) "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**(1), pp. 289–300.

[8] Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi, and I. Golani (2001) "Controlling the false discovery rate in behavior genetics research," *Behavioural brain research*, **125**(1), pp. 279–284.

[9] Ferreira, J. (2007) "The Benjamini-Hochberg method in the case of discrete test statistics," *The International Journal of Biostatistics*, **3**(1), p. 11.

[10] Gilbert, P. (2005) "A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(1), pp. 143–158.

[11] Storey, J. (2002) "A direct approach to false discovery rates," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**(3), pp. 479–498.

[12] Storey, J. D. and R. Tibshirani (2003) "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, **100**(16), pp. 9440–9445.

[13] Genovese, C. and L. Wasserman (2002) "Operating characteristics and extensions of the false discovery rate procedure," *J. Roy. Stat. Soc. B*, **64**, pp. 499–517.

[14] ——— (2004) "A stochastic process approach to false discovery control," *Ann. Statist.*, **32**(3), pp. 1035–1061.

[15] Efron, B. (2004) "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *J. Amer. Statist. Assoc.*, **99**(465), p. 96.

[16] Meinshausen, N. and J. Rice (2006) "Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses," *Ann. Stat.*, **34**(1), pp. 373–393.

[17] Jin, J. and T. Tony Cai (2007) "Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons," *J. Amer. Statist. Assoc.*, **102**(478), pp. 495–506.

[18] Strimmer, K. (2008) "A unified approach to false discovery rate estimation," *BMC bioinformatics*, **9**(1), p. 303.

[19] Muralidharan, O. (2010) "An empirical Bayes mixture method for effect size and false discovery rate estimates," *Ann. Appl. Stat.*, **4**(1).

[20] Pounds, S. and C. Cheng (2006) "Robust estimation of the false discovery rate," *Bioinformatics*, **22**(16), pp. 1979–1987.

[21] Kulinskaya, E. and A. Lewin (2009) "On fuzzy familywise error rate and false discovery rate procedures for discrete distributions," *Biometrika*, **96**(1), pp. 201–211.

[22] Heyse, J. (2011) "A false discovery rate procedure for categorical data," *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, **4**, p. 43.

[23] Ghosh, D. (2011) "Generalized Benjamini-Hochberg procedures using spacings," *J. Indian Soc. Agricultural Statist.*

[24] Dialsingh, I., "False Discovery Rates when the test statistic is discrete," .

[25] Heller, R. and H. Gur (2011) "False discovery rate controlling procedures for discrete tests," *arXiv preprint arXiv:1112.4627*.

[26] Blekhman, R., J. Marioni, P. Zumbo, and Y. Gilad (2010) "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, **20**, pp. 189–189.

[27] Kim, J. H., S. M. Dhanasekaran, R. Mehra, S. A. Tomlins, W. Gu, J. Yu, C. Kumar-Sinha, X. Cao, A. Dash, L. Wang, et al. (2007) "Integrative analysis of genomic aberrations associated with prostate cancer progression," *Cancer Res.*, **67**(17), pp. 8229–8239.

[28] Tsafrir, D., M. Bacolod, Z. Selvanayagam, I. Tsafrir, J. Shia, Z. Zeng, H. Liu, C. Krier, R. F. Stengel, F. Barany, et al. (2006) "Relationship of gene expression and chromosomal abnormalities in colorectal cancer," *Cancer Res.*, **66**(4), pp. 2129–2137.

[29] Fritz, B., F. Schubert, G. Wrobel, C. Schwaenen, S. Wessendorf, M. Nessling, C. Korz, R. J. Rieker, K. Montgomery, R. Kucherlapati, et al. (2002) "Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma," *Cancer Res.*, **62**(11), pp. 2993–2998.

[30] Pollack, J. R., T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O. Brown (2002) "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proc. Natl. Acad. Sci. USA*, **99**(20), pp. 12963–12968.

[31] Tonon, G., K.-K. Wong, G. Maulik, C. Brennan, B. Feng, Y. Zhang, D. B. Khatry, A. Protopopov, M. J. You, A. J. Aguirre, et al. (2005) "High-resolution genomic profiles of human lung cancer," *Proc. Natl. Acad. Sci. USA*, **102**(27), pp. 9625–9630.

[32] de Lichtenberg, U., L. Jensen, A. Fausbøll, T. Jensen, P. Bork, and S. Brunak (2005) "Comparison of computational methods for the identification of cell cycle-regulated genes," *Bioinformatics*, **21**(7), p. 1164.

[33] Wilkinson, B. (1951) "A statistical consideration in psychological research." *Psychol. Bull.*, **48**(3), p. 156.

[34] Phillips, D. and D. Ghosh (2014) "Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics," *Ann. Appl. Statist.*, **8**(2), pp. 801–823.

[35] Benjamini, Y. and R. Heller (2008) "Screening for partial conjunction hypotheses," *Biometrics*, **64**(4), pp. 1215–1222.

[36] Fisher, S. R. A. (1932) *Stat. Meth. Res. Worker*, Edinburgh, London.

[37] Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr (1949) "The American soldier: Adjustment during army life, Vol. I," *Stud. Soc. Psychol. World War II. Princeton: Princeton University Press*.

[38] Loughin, T. M. (2004) "A systematic comparison of methods for combining p-values from independent tests," *Comput. Statist. Data Anal.*, **47**(3), pp. 467–485.

[39] Phillips, D. (2014) "Tessellation," *WIREs Coput Stat*, **6**, pp. 202–209.

[40] Delaunay, B. (1932) "Neue Darstellung der geometrischen krystallographie," *Zeitschrift fúr Krystallographie*, **84**, pp. 109–149.

[41] Voronoi, G. (1908) "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs." *Journal fur die reine und angewandte Mathematik*, **134**, pp. 198–287.

[42] Thierssen, A. (1911) "Precipitation averages for large areas," *Monthly Weather Review*, **39**, pp. 1082–1084.

[43] Delaunay, B. (1929) "Sur la partition régulière del'espace a 4 dimensions. Première parti e," *Bulletin de L'Academie des Sciences de S'URSS, Classe des Sciences Mathèmatiques et Naturelles*, **1**, pp. 79–110.

[44] ——— (1929) "Sur la partition régulière del'espace a 4 dimensions. Deuxième partie," *Bulletin de L'Academie des Sciences de S'URSS, Classe des Sciences Mathèmatiques et Naturelles*, **1**, pp. 147–169.

[45] Pielou, E. (1977) *Mathematical Ecology*, Wiley-Interscience, New York.

[46] Meijering, J. (1953) "Interface area, edge length and number of vertices in crystal aggregates with random nucleation," *Philips Research Reports*, **8**, pp. 270–290.

[47] Okabe, A., B. Boots, K. Sugihara, and S. Chiu (2000) *Spatial Tessellations: concepts and applications of Voronoi diagrams*, second ed., John Wiley AND Sons.

[48] Liebeherr, J., M. Nahas, and W. Si (2002) "Application-layer multicasting with delaunay triangulation overlays," *IEEE Journal on Selected Areas in Communications*, **20**(8), pp. 1472–1488.

[49] Li, X.-Y., G. Calinescu, P.-J. Wan, and Y. Wang (2003) "Localized delaunay triangulation with application in ad hoc wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, **14**(10), pp. 1035–1047.

[50] Amenta, N., S. Choi, T. K. Dey, and N. Leekha (2000) "A simple algorithm for homeomorphic surface reconstruction," in *Proceedings of the sixteenth annual symposium on Computational geometry*, ACM, pp. 213–222.

[51] Wan, M., Y. Wang, and D. Wang (2011) "Variational surface reconstruction based on delaunay triangulation and graph cut," *International journal for numerical methods in engineering*, **85**(2), pp. 206–229.

[52] Wen, Q., H. Chang, and B. Parvin (2009) "A Delaunay triangulation approach for segmenting clumps of nuclei," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009*, IEEE, pp. 9–12.

[53] Diggle, P. (2013) *Statistical analysis of spatial and spatio-temporal point patterns*, 3rd ed., Taylor and Francis.

[54] Stoyan, D., W. S. Kendall, J. Mecke, and L. Ruschendorf (1995) *Stochastic geometry and its applications*, vol. 2, Wiley Chichester.

[55] Hinde, A. and R. Miles (1980) "Monte Carlo estimates of the distributions of the random polygons of the Voronoi tessellation with respect to a Poisson process," *Journal of Statistical Computation and Simulation*, **10**(3-4), pp. 205–223.

[56] Tanemura, M. (2003) "Statistical distributions of Poisson Voronoi cells in two and three dimensions," *Forma*, **18**(4), pp. 221–247.

[57] Ferenc, J.-S. and Z. Néda (2007) "On the size distribution of Poisson Voronoi cells," *Physica A: Statistical Mechanics and its Applications*, **385**(2), pp. 518–526.

[58] Kulldorff, M. (1997) "A spatial scan statistic," *Commun Stat A Theory Methods*, **16**, pp. 1481–1496.

[59] Allard, D. and C. Fraley (1997) "Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation," *Journal of the American Statistical Association*, **92**(440), pp. 1485–1493.

[60] Barr, C. D. and F. P. Schoenberg (2010) "On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process," *Biometrika*, **97**(4), pp. 977–984.

[61] Wager, C., B. Coull, and N. Lange (2004) "Modeling spatial intensity for replicated inhomogeneous point patterns in brain imaging," *J Roy Stat Soc B Met*, **66**, pp. 429–466.

[62] Berman, M. and R. Turner (1992) "Approximating point process likelihoods with GLIM," *Commun Stat A Theory Methods*, **410**, pp. 31–38.

[63] Baddeley, A. and R. Turner (2000) "Practical maximum pseudolikelihood for spatial oint patterns," *Aust N Z J Stat*, **42**, pp. 283–322.

[64] Denison, D. and C. Homes (2001) "Bayesian partitioning for estimating disease risk," *Biometrics*, **57**, pp. 143–149.

[65] Kim, H.-M., B. K. Mallick, and C. Holmes (2005) "Analyzing nonstationary spatial data using piecewise Gaussian processes," *Journal of the American Statistical Association*, **100**(470), pp. 653–668.

[66] Jiménez, R. and J. Yukich (2002) "Asymptotics for statistical distances based on Voronoi tessellations," *Journal of Theoretical Probability*, **15**(2), pp. 503–541.

[67] ——— (2005) "Statistical distances based on Euclidean graphs," in *Recent Advances in Applied Probability*, pp. 223–239.

[68] Deza, M. M. and E. Deza (2009) *Encyclopedia of distances*, chap. 20, Springer.

[69] Aurenhammer, F. and H. Edelsbrunner (1984) "An optimal algorithm for constructing the weighted Voronoi diagram in the plane," *Pattern Recognition*, **17**(2), pp. 251–257.

[70] Boots, B. and R. South (1998) "Modeling retail trade areas using higher-order, multiplicatively weighted Voronoi diagrams," *Journal of Retailing*, **73**(4), pp. 519–536.

[71] Galvão, L. C., A. G. Novaes, J. Souza de Cursi, and J. a. C. Souza (2006) "A multiplicatively-weighted Voronoi diagram approach to logistics districting," *Computers & operations research*, **33**(1), pp. 93–114.

[72] Orel, R. and M. Randić (2012) "On characterizing proteomics maps by using weighted Voronoi maps," *Journal of Mathematical Chemistry*, **50**(10), pp. 2689–2702.

[73] Dong, P. (2008) "Generating and updating multiplicatively weighted Voronoi diagrams for point, line and polygon features in GIS," *Computers and Geosciences*, **34**(4), pp. 411–421.

[74] Mu, L. (2004) "Polygon Characterization With the Multiplicatively Weighted Voronoi DiagramâĹŮ," *The Professional Geographer*, **56**(2), pp. 223–239.

[75] Moreno-Regidor, P., J. García López de Lacalle, and M.-A. Manso-Callejo (2012) "Zone design of specific sizes using adaptive additively weighted Voronoi diagrams," *International Journal of Geographical Information Science*, **26**(10), pp. 1811–1829.

[76] Will, H.-M. (1998) "Fast and efficient computation of additively weighted Voronoi cells for applications in molecular biology," in *Algorithm TheoryâĂŤSWAT'98*, Springer, pp. 310–321.

[77] Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu (2009) *Spatial tessellations: concepts and applications of Voronoi diagrams*, vol. 501, Wiley.

[78] Lee, D.-T. and B. J. Schachter (1980) "Two algorithms for constructing a Delaunay triangulation," *International Journal of Computer & Information Sciences*, **9**(3), pp. 219–242.

[79] Su, P. and R. L. Scot Drysdale (1997) "A comparison of sequential Delaunay triangulation algorithms," *Computational Geometry*, **7**(5), pp. 361–385.

[80] Guibas, L. J., D. E. Knuth, and M. Sharir (1992) "Randomized incremental construction of Delaunay and Voronoi diagrams," *Algorithmica*, **7**(1-6), pp. 381–413.

[81] De Berg, M., K. Dobrindt, and O. Schwarzkopf (1995) "On lazy randomized incremental construction," *Discrete & Computational Geometry*, **14**(1), pp. 261–286.

[82] Liu, J.-F., J.-H. Yan, and S. Lo (2013) "A new insertion sequence for incremental Delaunay triangulation," *Acta Mechanica Sinica*, pp. 1–11.

[83] Isenburg, M., Y. Liu, J. Shewchuk, and J. Snoeyink (2006) "Streaming computation of Delaunay triangulations," in *ACM Transactions on Graphics (TOG)*, vol. 25, pp. 1049–1056.

[84] Held, M. and S. Huber (2009) "Topology-oriented incremental computation of Voronoi diagrams of circular arcs and straight-line segments," *Computer-Aided Design*, **41**(5), pp. 327–338.

[85] Shamos, M. I. and D. Hoey (1975) "Closest-point problems," in *Foundations of Computer Science, 1975., 16th Annual Symposium on*, IEEE, pp. 151–162.

[86] Guibas, L. and J. Stolfi (1985) "Primitives for the manipulation of general subdivisions and the computation of Voronoi," *ACM Transactions on Graphics (TOG)*, **4**(2), pp. 74–123.

[87] Dwyer, R. A. (1987) "A faster divide-and-conquer algorithm for constructing Delaunay triangulations," *Algorithmica*, **2**(1-4), pp. 137–151.

[88] Yang, S.-W., Y. Choi, and C.-K. Jung (2011) "A divide-and-conquer Delaunay triangulation algorithm with a vertex array and flip operations in two-dimensional space," *International Journal of Precision Engineering and Manufacturing*, **12**(3), pp. 435–442.

[89] Aichholzer, O., W. Aigner, F. Aurenhammer, T. Hackl, B. Jüttler, E. Pilgerstorfer, and M. Rabl (2010) "Divide-and-conquer for Voronoi diagrams revisited," *Computational Geometry*, **43**(8), pp. 688–699.

[90] Chen, M.-B., T.-R. Chuang, and J.-J. Wu (2006) "Parallel divide-and-conquer scheme for 2D Delaunay triangulation," *Concurrency and Computation: Practice and Experience*, **18**(12), pp. 1595–1612.

[91] Fortune, S. (1987) "A sweepline algorithm for Voronoi diagrams," *Algorithmica*, **2**(1-4), pp. 153–174.

[92] Zalik, B. (2005) "An efficient sweep-line Delaunay triangulation algorithm," *Computer-Aided Design*, **37**(10), pp. 1027–1038.

[93] Biniaz, A. and G. Dastghaibyfard (2012) "A faster circle-sweep Delaunay triangulation algorithm," *Advances in Engineering Software*, **43**(1), pp. 1–13.

[94] Barber, C. B., D. P. Dobkin, and H. Huhdanpaa (1996) "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software (TOMS)*, **22**(4), pp. 469–483.

[95] Barber, C. B., K. Habel, R. Grasman, R. B. Gramacy, A. Stahel, and D. C. Sterratt (2013) *Geometry: Mesh generation and surface tesselation*, r package version 0.3-3.
URL `http://CRAN.R-project.org/package=geometry`

[96] Turner, R. (2013) *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.*, r package version 0.0-22.
URL `http://CRAN.R-project.org/package=deldir`

[97] Efron, B. (2007) "Correlation and large-scale simultaneous significance testing," *J. Amer. Statist. Assoc.*, **102**(477).

[98] Owen, A. B. (2009) "Karl Pearson's meta-analysis revisited," ***Ann. Statist.***, pp. 3867–3892.

[99] Forbes, S. A., N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, et al. (2011) "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic Acids Res.*, **39**(suppl 1), pp. D945–D950.

[100] Da Wei Huang, B. T. S., R. A. Lempicki, et al. (2008) "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protoc.*, **4**(1), pp. 44–57.

[101] Da Wei Huang, B. T., Sherman, R. A. Lempicki, et al. (2009) "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, **37**(1), pp. 1–13.

[102] Tagnon, H. J., W. F. Whitmore, and N. R. Shulman (1952) "Fibrinolysis in metastatic cancer of the prostate," *Cancer*, **5**(1), pp. 9–12.

[103] Zacharski, L. R., M. Z. Wojtukiewicz, V. Costantini, D. L. Ornstein, V. A. Memoli, et al. (1992) "Pathways of coagulation/fibrinolysis activation in malignancy." in *Seminars in thrombosis and hemostasis*, vol. 18, p. 104.

[104] Shabtai, D., G. Glaever, and C. Nislow (2012) "An algorithm for chemical genomic profiling that minimizes batch effects: bucket evaluations," *BMC Bioinformatics*, **13**(1), p. 245.

[105] Li, Q., J. Brown, H. Huang, and P. Bickel (2011) "Measuring reproducibility of high-throughput experiments," *Ann. Appl. Statist.*, **5**(3), pp. 1752–1779.

[106] Wei, Y. and H. Ji (2013) "A Survival Copula Mixture Model for Comparing Two Genomic Rank Lists," *arXiv preprint arXiv:1311.7122*.

[107] Cramer, H. (1928) "On the composition of elementary errors: II, statistical applications," *Skand. Akt.*, **11**, pp. 141–180.

[108] von Mises, R. (1931) "Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik," *Leipzig: Deuticke*.

[109] Nobel, A. and A. Dembo (1993) "A note on uniform laws of averages for dependent processes," *Statistics and Probability Letters*.

[110] Storey, J. D., J. E. Taylor, and D. Siegmund (2004) "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66**(1), pp. 187–205.

[111] Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein (2009) "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nature Biotechnology*, **27**, pp. 66–75.

[112] Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu (2008) "Model-based analysis of ChIP-seq (MACS)," *Genome Biology*, **9**, p. R137.

[113] Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park (2008) "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature Biotechnology*, **26**, pp. 1351–1359.

[114] Boyle, A. P., J. Guinney, G. E. Crawford, and T. S. Furey (2008) "F-Seq: A feature density estimator for high-throughput sequence tags," *Bioinformatics*, **24**, pp. 2537–2538.

[115] Thurman, R., M. Hawrylycz, S. Kuehn, E. Haugen, and S. Stamatoyannopoulos (2011) "Hotspot: A scan statistic for identifying enriched regions of short-read sequence tags." *Unpublished manuscript, Univ. Washington*.

[116] Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008) "Mapping and quantifying mammalian transcriptomes by RNA-seq," *Nature Methods*.

[117] Ji, H., H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong (2008) "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*.

[118] Valouev, A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow (2008) "Genome-wide analysis of transcription factor binding sites based on ChIP-seq data," *Nature Methods*.

[119] Jothi, R., S. Cuddapah, A. Barski, K. Cui, and K. Zhao (2008) "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data." *Nucleic Acids Res.*

[120] Marguerat, S., T. Jensen, U. de Lichtenberg, B. Wilhelm, L. Jensen, and J. Bähler (2006) "The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast," *Yeast*, **23**(4), pp. 261–277.

[121] Fisher, R. (1929) "Tests of significance in harmonic analysis," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **125**(796), pp. 54–59.

Vita

Daisy Philtron

Education

- Ph.D. Statistics, Pennsylvania State University, 2014
  Thesis title: *Three novel procedures to control the false discovery rate*
  Adviser: Dr. Debashis Ghosh.
- M.S. Mathematics, Western Washington University, 2009
  Thesis Project: *A proof of the central limit theorem using Fourier transforms*
  Adviser: Dr. Arpad Benyi.
- B.S. Mathematics, Western Washington University, 2008.

Publications

- Phillips, D., and Ghosh, D. (2014). Testing the disjunction hypothesis using Voronoi diagrams, with applications to genetics. *Annals of Applied Statistics*, **8**(2), 801–823.
- Phillips, D. (2014). Tessellation. *WIREs Computational Statistics*, **6**, 202–209.
- Hall, N., Mercer, L., Phillips, D., Shaw, J., and Anderson, A. D. (2012). Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. *Genetics Research.* **94**, 151-161.

Honors and Awards

- 2013, Gertrude M. Cox Scholarship, American Statistical Association
- 2013, Harold F. Martin Graduate Assistant Outstanding Teaching Award, The Pennsylvania State University
- 2013, Eberly College of Science Endowed Innovative Teaching Award in Statistics, The Pennsylvania State University
- 2009, Distinguished Graduate Fellowship, Pennsylvania State University
- 2009, Richard Green Award for Academic Excellence,
  Department of Mathematics, Western Washington University
- 2008, Presidential Scholar, College of Science and Technology,
  Western Washington University
- 2008, Outstanding Graduate of Mathematics, Department of Mathematics,
  Western Washington University