5-2014

Family-Wise Error Rate Control in Quantitative Trait Loci (QTL) Mapping and Gene Ontology Graphs with Remarks on Family Selection

Garrett Saunders
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd
Part of the Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation
https://digitalcommons.usu.edu/etd/7021

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.
FAMILY-WISE ERROR RATE CONTROL IN QUANTITATIVE TRAIT LOCI (QTL) MAPPING AND GENE ONTOLOGY GRAPHS WITH REMARKS ON FAMILY SELECTION

by

Garrett Saunders

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences
(Statistics)

UTAH STATE UNIVERSITY
Logan, Utah

2014
ABSTRACT

Family-wise Error Rate Control in Quantitative Trait Loci (QTL) Mapping and Gene Ontology Graphs with Remarks on Family Selection

by

Garrett Saunders, Doctor of Philosophy

Utah State University, 2014

Major Professor: Dr. John R. Stevens
Department: Mathematics

The main aim of this dissertation is to meet real needs of practitioners in multiple hypothesis testing. The issue of multiplicity has become a significant concern in most fields of research as computational abilities have increased, allowing for the simultaneous testing of many (thousands or millions) statistical hypothesis tests. While many error rates have been defined to address this issue of multiplicity, this work considers only the most natural generalization of the Type I Error rate to multiple tests, the family-wise error rate (FWER). Much work has already been done to establish powerful yet general methods which control the FWER under arbitrary dependencies among tests. This work both introduces these methods and expands upon them as is detailed through its four main chapters. Chapter 1 contains general introductions and preliminaries important to the remainder of the work, particularly a previously published graphical weighted Bonferroni multiplicity adjustment. Chapter 2 then applies the principles introduced in Chapter 1 to achieve a substantial computational improvement to an existing FWER controlling multiplicity approach (the Focus Level method) for gene set testing in high throughput microarray and next generation sequencing studies using Gene Ontology graphs. This improvement to the Focus Level procedure, which we call the Short Focus Level procedure, is achieved by extending the
reach of graphical weighted Bonferroni testing to closed testing situations where restricted hypotheses are present. This is accomplished through Theorem 1 of Chapter 2. As a result of the improvement, the full top-down approach to the Focus Level procedure can now be performed, overcoming a significant disadvantage of the otherwise powerful approach to multiple testing. Chapter 3 presents a solution to a multiple testing difficulty within quantitative trait loci (QTL) mapping in natural populations for QTL LD (linkage disequilibrium) mapping models. Such models apply a two-hypothesis framework to the testing of thousands of genetic markers across the genome in search of QTL underlying a quantitative trait of interest. Inherent to the model is an unidentifiability issue where a parameter of interest is identifiable only under the alternative hypothesis. Through a second application of graphical weighted Bonferroni methods we show how the multiplicity can be accounted for while simultaneously accounting for the required logical structuring of the testing such that identifiability is preserved. Finally, Chapter 4 details some of the difficulties associated with the distributional assumptions for the test statistics of the two hypotheses of the LD-based QTL mapping framework. A novel bivariate testing strategy is proposed for these test statistics in order to overcome these distributional difficulties while preserving power in the multiplicity correction by reducing the number of tests performed. Chapter 5 concludes the work with a summary of the main contributions and future research goals aimed at continual improvement to the multiple testing issues inherent to both the fields of genetics and genomics.
PUBLIC ABSTRACT

Family-wise Error Rate Control in Quantitative Trait Loci (QTL) Mapping and Gene Ontology Graphs with Remarks on Family Selection

by

Garrett Saunders, Doctor of Philosophy
Utah State University, 2014

Major Professor: Dr. John R. Stevens
Department: Mathematics

One of the great aims of statistics, the science of collecting, analyzing, and interpreting data, is to protect against the probability of falsely rejecting an accepted claim, or hypothesis, given observed data stemming from some experiment. This is generally known as protecting against a Type I Error, or controlling the Type I Error rate. The extension of this protection against Type I Errors to the situation where thousands upon thousands of hypotheses are examined simultaneously is known as multiple hypothesis testing. This dissertation presents an improvement to an existing multiple hypothesis testing approach, the Focus Level method, specific to gene set testing (a branch of genomics) on Gene Ontology graphs. This improvement resolves a long standing computational difficulty of the Focus Level method, providing more than a 15,000-fold increase in computational efficiency. This dissertation also presents a solution to a multiple testing problem in genetics where a specific approach to mapping genes underlying quantitative traits of interest requires a multiplicity adjustment approach that both corrects for the number of tests while also ensuring logical consistency. The power advantage of the solution is demonstrated over the current standard approach to the problem. A side issue of this model framework led to the development of a new bivariate approach to quantitative trait marker detection, which is presented herein.
The overall contribution of this dissertation to the statistics literature is that it provides novel solutions that meet real needs of practitioners in genetics and genomics with the aim of ensuring both that truth is discovered and that discoveries are actually true.
ACKNOWLEDGMENTS

This work was funded by both a Utah Agricultural Experiment Station (UAES) project number UTA01062, associated with the W2112 multi-state project “Reproductive Performance in Domestic Ruminants” and a Utah State University VPR Research Catalyst Grant.
CONTENTS

ABSTRACT ......................................................... iii
PUBLIC ABSTRACT ................................................. v
ACKNOWLEDGMENTS ........................................ vii
LIST OF TABLES ........................................................ x
LIST OF FIGURES .................................................... xi

1 INTRODUCTION ....................................................... 1
  1.1 Error Rates ................................................. 3
  1.2 Principles for Family Selection ................................ 6
  1.3 Methods ..................................................... 9
    1.3.1 The Bonferroni Method .................................. 11
    1.3.2 The Holm Method ......................................... 12
    1.3.3 Closed Testing Procedures ............................... 15
    1.3.4 The Weighted Bonferroni Method ......................... 17
    1.3.5 The Weighted Holm Method ............................... 18
    1.3.6 Generalized Weighted Bonferroni Testing ............... 20

2 FWER CONTROL WITHIN GENE ONTOLOGY GRAPHS ............ 31
  2.1 Introduction ................................................ 31
    2.1.1 The Focus Level Method ................................ 35
    2.1.2 The Graphical Bonferroni Adjustment .................... 39
    2.1.3 Restricted Hypotheses Example .......................... 40
    2.1.4 Definitions and Preliminaries to Theorem 1 ............ 43
  2.2 Shortcuts for Restricted Hypotheses ......................... 46
    2.2.1 Theorem 1 ............................................ 46
    2.2.2 Discussion of Theorem 1 ................................ 47
  2.3 The Short Focus Level Procedure ................................ 48
    2.3.1 Power Analysis ......................................... 50
    2.3.2 Real Data Application .................................. 56
  2.4 Discussion .................................................. 59

3 A POWER IMPROVING MULTIPLICITY CORRECTION FOR LARGE-
SCALE SNP SELECTION IN LD-based QTL MAPPING ............ 61
  3.1 Abstract ................................................... 61
  3.2 Introduction ............................................... 62
  3.3 Materials and Methods ....................................... 65
    3.3.1 LD-based QTL Mapping Model ......................... 65
### 3.3.2 Two Hypothesis Tests

3.3.3 Graphical Bonferroni Adjustment

3.3.4 Rejection Scheme

3.3.5 The GBA as an IUT

### 3.4 Results

3.4.1 Power Simulation

3.4.2 Poplar Leaf Shape QTL Mapping Project

3.4.3 Mouse HDL Cholesterol QTL Mapping Project

### 3.5 Discussion

### 3.4 Results

4 QTL MAPPING: HYPOTHESES AND APPROACHES

4.1 Introduction

4.2 The QTL Hypotheses

4.3 Comparison of Univariate Approaches

4.3.1 Permutations

4.3.2 Simulation

4.3.3 Large-Sample Empirical

4.3.4 Simulation Results and Discussion

4.4 Null Kernel P-value Method

4.4.1 Location Testing for a Bivariate Normal

4.4.2 QTL Mapping Simulation Revisited

4.4.3 Mouse HDL Cholesterol QTL Mapping Revisited

4.5 Discussion

5 DISCUSSION

APPENDICES

A GBA Source Code and Help File

A.1 GBA Help File

A.2 GBA Source Code

CURRICULUM VITAE
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Random variables for the various possible counts that can occur when ( m ) hypotheses are tested simultaneously.</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Algorithm detailing the newly proposed Short Focus Level procedure.</td>
<td>51</td>
</tr>
<tr>
<td>2.2 Summary of results for Simulation 1. Power calculations were averaged over all levels of the effect size ( \lambda ) and both sizes of ( m ), the maximum leaf node gene set size, for each level of the sample size ( n ).</td>
<td>53</td>
</tr>
<tr>
<td>2.3 Allocation of simulated genes to the GO IDs of the GO graph in Figure 2.9.</td>
<td>54</td>
</tr>
<tr>
<td>2.4 Results of the power simulation for the GO Graph in Figure 2.9.</td>
<td>55</td>
</tr>
<tr>
<td>2.5 Experimental design for the \textit{in vivo} (IVV) and \textit{in vitro} (IVM) Oocyte Maturation RNA-Seq data. *Lane 3 contained quality problems and was removed from the analysis.</td>
<td>57</td>
</tr>
<tr>
<td>3.1 The theoretical joint probabilities of SNP (row) and QTL (column) genotypes.</td>
<td>66</td>
</tr>
<tr>
<td>3.2 The theoretical conditional probabilities of SNP genotype (columns) given QTL genotype (rows).</td>
<td>75</td>
</tr>
<tr>
<td>3.3 The results of the power simulation as depicted in Figure 3.5. Here, Bon is the Bonferroni adjustment and GBA (D) is the graphical Bonferroni adjustment when rejecting both nulls of interest (equivalent to IUT with the Holm adjustment), and GBA (L) is the graphical Bonferroni adjustment where rejecting ( H_0^L ) only is meaningful.</td>
<td>77</td>
</tr>
<tr>
<td>3.4 The significant results of the outbred mice HDL cholesterol QTL mapping project depicted in Figure 3.7. SNPs are ordered by significance level. Corresponding concurrence candidate gene and QTL from previous inbred crosses studies are shown.</td>
<td>82</td>
</tr>
<tr>
<td>4.1 The theoretical conditional probabilities of QTL genotype (columns) given marker genotype (rows).</td>
<td>90</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>As the number of hypotheses tested, ( m ), becomes large, the probability ( P_m ) of at least one test showing significant results when all null hypotheses are in fact true approaches one.</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>A simple noded diagram of the Bonferroni test for three hypotheses. Each node (or hypothesis) is tested with equal weight of ( \alpha/3 ).</td>
<td>22</td>
</tr>
<tr>
<td>1.3</td>
<td>A noded diagram of the weighted Bonferroni test for three hypotheses. Each node (or hypothesis) is tested with weight (assigned ( a \ priori )) of ( \alpha_1 = \alpha_3 = 0.005 ) and ( \alpha_2 = 0.04 ).</td>
<td>23</td>
</tr>
<tr>
<td>1.4</td>
<td>A noded diagram of the sequential Bonferroni test for the three hypotheses of Figure 1.2 with weighted directed edges. Each node (or hypothesis) is first tested with equal weight of ( \alpha/3 ) but if any hypothesis is rejected, its threshold ( \alpha_i ) is shared with the remaining hypotheses as specified by the weights along the arrows.</td>
<td>24</td>
</tr>
<tr>
<td>1.5</td>
<td>(a) The updated graph after ( H_1 ) has been rejected and removed from further consideration and its threshold has been shared equally between ( H_2 ) and ( H_3 ). Notice that all logical structures involving ( H_1 ) have been severed. (b) The final graph of ( H_3 ) obtained after ( H_2 ) is rejected at level ( \alpha/2 ), removed from consideration, and all of its threshold is passed on to ( H_3 ).</td>
<td>25</td>
</tr>
<tr>
<td>1.6</td>
<td>A first attempt at a testing diagram for testing the first family of hypotheses including ( H_1 ) and ( H_2 ), each at level ( \alpha/2 ), followed by the testing of the second family ( (H_3) ) at level ( \alpha ) in the case that both hypotheses in the first family ( (H_1 ) and ( H_2 )) are rejected.</td>
<td>28</td>
</tr>
<tr>
<td>1.7</td>
<td>If ( H_1 ) and ( H_2 ) are both rejected, the above diagram specifies that ( r_1 ) of ( \alpha ) will be passed along to ( H_3 ) and that ( r_2 ) of ( \alpha ) will be passed along to ( H_4 ). If at least one of ( H_1 ) or ( H_2 ) is not rejected, then neither ( H_3 ) or ( H_4 ) are tested.</td>
<td>30</td>
</tr>
<tr>
<td>2.1</td>
<td>Graphical demonstration of the ancestral relationships of the GO term GO:0007005 within the Biological Process GO graph. Arrows are directed from child to parent. The root BP GO term (GO:0008150) is the only node without ancestors in the graph.</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Three possible focus levels (filled nodes) for a simplified example GO graph.</td>
<td>37</td>
</tr>
</tbody>
</table>
2.3 Closures of the GO graphs from Figure 2.2 where the filled nodes represent the different choices of the focus level. .................................................. 38

2.4 (a) Graphical demonstration of the elementary hypotheses $H_1, \ldots, H_4$ and distinct intersection hypotheses. The null parameter space is shaded in gray for each hypothesis. Redundant intersection hypotheses are written in parentheses. (b) The closed test approach given the structure of the hypotheses. Testing begins with $H_{13}$, the full intersection hypothesis, and terminates at or before testing $H_2$ and $H_4$. .................................................. 42

2.5 Graphical Bonferroni adjustment approach for the partially nested elementary hypotheses $H_1, \ldots, H_4$ which performs the closed test described in Brannath and Bretz (2010) when Algorithm 1 is applied to the graph. .......................... 43

2.6 Flow chart demonstration of the equivalence of the graphical shortcut tailored from the methods of Bretz et al. (2009) to that of the full closed test procedure proposed in Brannath and Bretz (2010) within the context of the restricted hypotheses example of Section 2.1.3. At each step in the chart, the left graph represents the full closed test approach, while the right graph depicts the graphical shortcut. .................................................. 44

2.7 The suggested shortcut to the top-down portion of the Focus Level method exploits the natural consonance of the weighted Bonferroni tests applied to the GO graph to avoid closing the graph under all unions as in the original top-down approach. .................................................. 50

2.8 (a) The full closure of the example toy GO graph depicted in panel (b) that is currently utilized by the Focus Level method. (c) The graph $(\alpha, G)$ corresponding to the example toy GO graph depicted in panel (b) that is utilized by the proposed Short Focus Level procedure. .................................................. 52

2.9 Structure of the toy GO graph used in Simulation 2. Shaded nodes correspond to those GO IDs which were simulated to be significantly associated with the phenotype $Y$. .................................................. 55

2.10 Significant results from the gene set testing of porcine oocytes obtained from the Short Focus Level procedure using the full top down approach. .............. 58

3.1 A) Demonstration of the GBA testing scheme for a single marker. B) The updated graph after finding $H_5^L$ significant. .................................................. 69

3.2 Demonstration of the hierarchy of the GBA testing scheme for three markers. .......................... 70

3.3 Demonstration of the GBA testing scheme for three markers assuming that hypotheses $H_5^L$ and $H_6^L$ from the initial graph in Figure 3.2 are rejected. .......................... 71
3.4 A) The updated graph from Figure 3.3 assuming the hypothesis $H_0^{D1}$ of Figure 3.3 is rejected at the $\alpha/3$-level. B) Graph resulting from the rejection of the hypothesis $H_0^{D3}$ at the $\alpha/2$-level.

3.5 Power comparison between the graphical Bonferroni adjustment (GBA) and standard Bonferroni adjustment under different sample size, number of SNPs, and heritability (A: $H^2 = 0.1$, B: $H^2 = 0.4$).

3.6 The control of leaf shape according to the different genotypes (AA, Aa, aa) of the QTL identified by marker 1 on PC 4. This shows the increased sensitivity of the GBA approach as the effect of the aa QTL genotype is subtly different from that of genotypes AA and Aa, but nevertheless, corresponds to information which was undetected under the Bonferroni correction.

3.7 The negative log of the GBA-adjusted $p$-values for $H_0^D$ for each SNP in the mouse HDL cholesterol QTL mapping project. The red reference line corresponds to a 0.05 family-wise error rate.

4.1 A) Demonstration of the empirical cumulative density function (blue dots) for the Likelihood Ratio Test of $H_0^1$ against $H_A$ for synthetic data simulated under the null hypothesis of no QTL, $H_0$. B) The empirical cumulative density functions corresponding to the test of $D = 0$ for (i) synthetic marker and QTL data where both genotypes are known (large gray dots) which is consistent with a $\chi^2$ distribution, (ii) synthetic data simulated under $H_0^5$ (blue dots), and (iii) synthetic data simulated under $H_0^1$ (small gray dots). In both panels A and B, the cumulative density functions (underlaid in gray) for several $\chi^2$ distributions are also plotted for $\nu = 1, \ldots, 10$ for reference.

4.2 The resulting genome-wise significance thresholds for the LOD ($\log_{10}$ of the likelihood ratio, solid black line) according to each of the simulation, permutation, and theoretical approaches. The empirical approach did not identify any markers as supporting a linked QTL and was thus not included in the figure. Triangles demarcate the simulated QTL at 44.4 cM on the first chromosome and at 61.6 cM on the second chromosome.

4.3 (a) Graphical demonstration of the $-\log_{10}$ of the $P$-values corresponding to the testing of $H_0^1$ for each of the four univariate methods under consideration: permutations, simulations, empirical, and theoretical. A technical artifact is present in both the permutation and simulation graphs, where to allow the log transformation of zero (occurring for both the permutation and simulation methods), the value of $10^{-10}$ was first added to all raw $P$-values. (b) The $-\log_{10}$ of the $P$-values corresponding to the test of $H_0^2$.

4.4 Visualization of the Null Kernel method as applied to a sample of 1,000 $T$ and $U$ statistics simulated under the bivariate normal null distribution with zero mean, unit variances, and covariances of 0.3. The contours of the estimated null density $\hat{f}$ are superimposed for $1 - \hat{F}$ values of 0.05, 0.01, 0.001, and 0.0001.
4.5 Comparison of P-values ($-\log_{10}(p)$) obtained from either the Null Kernel method or Hotelling’s $T^2$ test. The four quadrants, I, II, III, and IV demonstrate the regions of agreement (I and III) and discord (II and IV) between the two methods.

4.6 The $-\log_{10}$ of the P-values from the Null Kernel and Hotelling’s $T^2$ methods for data simulated consistent with the null hypothesis. Both methods properly control the Type I Error with less than 5% of the data in the critical region. Specifically, the Null Kernel method shows a 4% Type I Error rate while Hotelling’s $T^2$ method shows a 3% rate.

4.7 Visualization of the level curves of the Null Kernel estimated (null) density $\hat{f}$ for the bivariate data corresponding to the test of $H_0^D$, $\chi^2_D$, and $H_0^L$, $\chi^2_L$.

4.8 The panels above compare the resulting adjusted P-values from each of the permutation, simulation, and theoretical approaches (Section 4.3) against the results of the Null Kernel method for the simulated QTL mapping data of Section 4.3. For ease of reference, the results for each method are highlighted individually in their own set of panels for each Chromosome (1-4), as labeled in the panel corresponding to Chromosome 4. The rug plots along the bottom of each panel show the locations of the synthetic markers.

4.9 The negative log of the Holm adjusted P-values for the Null Kernel approach applied to the 44,428 SNPs from the mice HDL QTL mapping study.

4.10 The joint plot of the observed test statistics for the mouse HDL QTL mapping data. Black points denote SNPs that were found significant after the Null Kernel P-values were adjusted for multiplicity using the Holm adjustment. For comparison, all points in the upper right quartile demarcated by the dashed lines were declared significant by the GBA method of Chapter 3 (where both hypotheses $H_0^D$ and $H_0^L$ were found significant).
CHAPTER 1
INTRODUCTION

The field of multiple hypothesis testing has a long, albeit relatively quiet history. As noted by Shaffer (1995), concerns about the multiplicity issue were voiced as early as 1843 by Cournot. He warned caution in the interpretation of results presented by researchers where, in his words as translated from the French,

...usually the attempts through which the experimenter passed don’t leave any traces; the public will only know the result that has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or cannot be attributed to chance (Cournot 1843; Shaffer 1995).

Despite such early knowledge of the statistical difficulties in testing multiple hypotheses, methods for dealing with the multiplicity issue showed a slower start.

Tippet (1931) was among the first to write on the topic. His concern was that of controlling the probability that the largest of $m$ differences of means would equal or exceed some threshold $d$. He recognized that the probability statement for a single hypothesis test quickly lost meaning as several tests were considered separately, but simultaneously. To show this, Tippet let $P$ denote the probability that the difference of two means obtained from random sampling would be larger than some chosen threshold $d$ solely by chance. (In today’s vernacular the $P$ Tippet used is typically denoted by $\alpha$.) He then calculated the probability $P_m$ that the largest of $m$ differences of two means, each obtained from independent and identically distributed samples, would be greater than $d$ and hence derived the probabilistic statement (Tippet 1931)

\[ P_m = 1 - (1 - P)^m. \]
Fig. 1.1. As the number of hypotheses tested, m, becomes large, the probability $P_m$ of at least one test showing significant results when all null hypotheses are in fact true approaches one.

Tippet’s derivation demonstrates how rapidly $P_m$ (the probability that the largest of m differences of means will be larger than some threshold d) approaches 1 as m becomes large. Even when P is very small, say 0.05, and $m = 2$, the probability that the largest of the two differences is greater than d is $P_m = 0.0975$. As Figure 1.1 shows, ignoring the multiplicity issue leads to certainty in finding at least one significant result among m tests of significance as m becomes large. While in Tippet’s time m was typically less than 20, the problem has since compounded by advances in technology allowing researchers the ability to test $m > 1,000$ or even $m > 1,000,000$ hypotheses separately, but simultaneously.

As time progressed from Tippet’s first published work, the topic of multiple comparison procedures (MCPs) advanced little until, as Miller wrote, the “great spurt of interest and research in multiple comparisons” took place a few years later in the late forties and early fifties of the 20th century (Miller 1981). While works on MCPs were mostly scattered in journals during this era, MCPs later found their place in textbooks dedicated to just this topic with landmark works by Miller (1981), Hochberg and Tamhane (1987), Westfall and Young (1993), and Hsu (1996). The first world-wide conference dedicated to MCPs was held in Tel Aviv in 1996 with the eighth such conference at Southampton University, UK in 2013.

In an address to the 2010 Conference on MCPs, Yoav Benjamini stated that this “golden
era” of success for the field of multiple comparisons is largely due to “being able to address real current needs” (Benjamini 2010). He further argues, “that the vitality of [the] field in the future – as a research area – depends upon [the researcher’s] ability to continue and address the real needs of statistical analyses in current problems.” Hence, it should be the aim of every effort in multiple comparisons research to address “real needs” in such a way that it is clear to the practitioners both as to why multiple comparisons are needed and how they should be applied to the problems at hand.

1.1 Error Rates

Single statistical hypothesis testing is a fine balance between simultaneously protecting for errors against the null hypothesis (Type I Errors) and errors against the alternative (Type II Errors). Typically, a small probability of committing errors against the null hypothesis is assured by setting the probability of rejection under the null hypothesis at \( \alpha = 0.05 \). The aim of the statistician is then to find methods or suggest appropriate sample sizes such that not too many errors are correspondingly committed against the alternative. In other words, to keep the power high while protecting against Type I Errors. Miller described this balancing act well when he wrote, “The statistician is...inescapably strapped to a teeter-totter (or seesaw). As the error rates are forced down in one direction they must increase in the other” (Miller 1981). For reasons that will be apparent later on, it is important to note that while no formal mathematical proof can be given for the typical choice of \( \alpha = 0.05 \), perhaps Fisher stated the accepted tradition best when he wrote, “If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance; this is roughly equivalent to the corresponding limit \( P = .05 \)” (Fisher 1973).

Extending this protection of errors against the null hypothesis to several simultaneous hypothesis tests requires some thought and certainly has no unique solution. (See Section 1.2.9 of Dudoit and van der Laan (2008) for an explanation of many such possibilities.) In any case, the aim should be “to give satisfactory protection against wrong decisions” as wrote Holm (1979). One of the most common extensions is to simply protect against
Table 1.1. Random variables for the various possible counts that can occur when \(m\) hypotheses are tested simultaneously.

<table>
<thead>
<tr>
<th></th>
<th>Declared non-significant</th>
<th>Declared significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True null hypothesis</td>
<td>(U)</td>
<td>(V)</td>
<td>(m_0)</td>
</tr>
<tr>
<td>Non-true null hypothesis</td>
<td>(T)</td>
<td>(S)</td>
<td>(m - m_0)</td>
</tr>
<tr>
<td>(m - R)</td>
<td>(R)</td>
<td></td>
<td>(m)</td>
</tr>
</tbody>
</table>

the probability of committing any errors of the first kind. Correspondingly, the popular family-wise error rate (FWER) is defined as the probability of committing at least one Type I Error in the simultaneous testing of \(m\) null hypotheses.

Formally, let the unobservable random variables \(U\) and \(S\) denote the number of correct decisions of null and alternatively expressed hypotheses, respectively; the unobservable random variables \(V\) and \(T\) denote the number of errors against the null and alternatively expressed hypotheses, respectively; and the observable random variable \(R\) denote the total number of rejections. Then the following defines the FWER.

\[
P(V \geq 1) \quad \text{(FWER)}
\]

In contrast, the \(\alpha\)-level of the single hypothesis test discussed previously could be stated as

\[
E\left(\frac{V}{m}\right) \quad \text{(PCER)}
\]

where \(E\) denotes the expected value (or long run proportion) of \(V/m\). The per-comparison error rate (PCER) is the usual error rate that is controlled within single hypothesis testing by controlling only the probability of a Type I Error per hypothesis test. As demonstrated previously in Figure 1.1, controlling only the PCER for \(m\) simultaneous hypothesis tests will lead to certainty in experiencing at least one Type I Error as \(m\) becomes large. On the other hand, the bounding of the probability of any Type I Error, \(P(V \geq 1) \leq \alpha\), is more conservative than bounding only each comparison individually, \(E(V/m) = \alpha\) whenever
Importantly, the bound on the FWER holds for any size of $m$. It should be noted that controlling the FWER also controls the PCER, but at a lower level, so that the number of errors of the second kind will always be larger when controlling the FWER rather than the PCER. However, the loss of control on the probability of errors of the first kind (which are inevitable when only using the PCER for $m > 1$ tests) should discourage its use in multiple hypothesis testing (Hochberg and Tamhane 1987).

While this work focuses solely on methods which control the probability of any Type I Error (i.e., the FWER), an alternative error rate is so predominant in the literature that it deserves at least a brief explanation. A seminal paper by Benjamini and Hochberg (1995) proposed the FDR or False Discovery Rate as another option for error control in the multiple hypothesis testing scenario. It is defined by them as the expected number of false rejections given that at least one rejection has occurred. Note that within Table 1.1 the number of false rejections is given as $V$ and the number of total rejections is denoted as $R$, so that the formal definition of the FDR is

$$P(R > 0)E(V/R|R > 0).$$

The complexity of the statement is due to the difficulty that when $m = m_0$, all hypotheses are true so that every rejection is a false rejection causing $V = R$ and accordingly $V/R = 1$. In this scenario, the statement $P(R > 0)$ ensures that control over this error rate can still be achieved in the trivial case of no non-true null hypotheses as $E(V/R|R > 0) = 1$, which is itself uncontrolled.

The use of the FDR is appropriate for many situations, but is especially advocated for preliminary studies where it is acceptable for a certain proportion of discoveries to be false in exchange for an increase in overall discoveries as compared to FWER controlling methods (Liang and Nettleton 2010). This work deals exclusively with the FWER due to three main points. (i) The confirmatory rather than exploratory nature of the FWER (Hochberg and Tamhane 1987). (ii) The attractive property of the FWER which allows the researcher to safely consider any subset of the significant findings while still preserving control over
the FWER at the designated level. This is in direct contrast to the FDR, for which there is no assurance that a subset of the significant discoveries contains the same proportion of false discoveries as does the entire set of significant discoveries (Goeman and Solari 2014).

(iii) The logical flexibility of the FWER which allows for its application to directed graphs as in Chapter 2 (Goeman and Mansmann 2008).

1.2 Principles for Family Selection

The need for procedures which control a selected error rate for simultaneous multiple inference has been well established in the literature (Hochberg and Tamhane 1987; Hsu 1996; Miller 1981; Tippet 1931; Westfall and Young 1993), to mention just a few. Perhaps Diaconis (1985) stated it most succinctly when he wrote, “If enough statistics are computed, some of them will be sure to show structure.” Quality methods meeting the multiple comparison needs have been and are being produced. Still, at least one question remains difficult to answer, just what constitutes a family of hypotheses? Many authors have discussed this topic (Ahmed 1991; Hochberg and Tamhane 1987; Hsu 1996; Miller 1981; Saville 1990; Shaffer 1995; Westfall and Young 1993) with similar conclusions, that the selection of a family is a very subjective, yet important decision.

Critiques of multiple comparison procedures attack the sensitivity of the results to the often arbitrary selection of families (Ahmed 1991; Perry 1986; Saville 1990). They argue that the significance of the results can be greatly affected by how many hypotheses are included into the family. This phenomenon, which Saville defines as family inconsistency, is well demonstrated in Ahmed (1991) where results of the National Assessment of Educational Progress at the national level (780 comparisons) show no significant differences when multiplicity is accounted for but the results taken at a state level (39 comparisons) do show some significant differences under the same adjustment approach. Williams et al. (1999) studied this phenomenon in greater detail and recommended that controlling the FDR in place of the FWER was one way to avoid this difficulty as the FDR is more robust to changes in family size.

Holland and Cheung (2002) arrive at a similar conclusion as they proposed overcoming
the difficulties associated with selecting a family of hypotheses by considering only MCPs which are consistent in their findings across different sizes of the family. They define this property as *family-wise robustness* and discuss MCPs which control the FWER and satisfy various forms of this property. They point out that procedures which control the FDR are more family-wise robust than procedures which control the FWER. However, their conclusion unsatisfactorily circumvents the main issue of the subjective nature of selecting a family of hypotheses, especially for confirmatory studies where the use of the FWER, rather than the FDR, is advocated (Hochberg and Tamhane 1987).

In a more conservative approach, Bretz et al. (2009) discuss several ways of shifting significance between separate families of interest while controlling the FWER for all families simultaneously by using certain families as gatekeepers for others. Their work allows for importance rankings among families so that families of lesser importance are considered only after error rates have properly been controlled for the most important families. This often comes at a loss of power for the less important families in the hierarchy, but does well at controlling the overall error rate across all families. Still, the main issue of family selection remains unresolved.

Perhaps the best statement of the family selection problem was noted in the first textbook dedicated to multiple comparisons. Miller (1981) wrote that stating what constitutes a family of hypotheses “is the hardest part...because it is where statistics takes leave of mathematics and must be guided by subjective judgement.” He further emphasizes, “There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his own judgment for the problem at hand.” Miller’s original discussion provides the spectrum of possible choices for a family of hypotheses.

Two extremes of behavior are open to anyone involved in statistical inference. A nonmultiple comparisonist regards each separate statistical statement as a family, and does not give increased protection to any group of statements through group error rates. At the other extreme is the ultraconservative statistician who has just a single family consisting of every statistical statement he might make during his lifetime... There are a few statisticians who would adhere to the first principle, but the author has never met one of the latter variety. (Miller 1981)
Concerning the issue of family selection, Hochberg and Benjamini (1990) offer the following advice,

A conclusion is often based on a family of comparisons, for example, recommending one treatment over another in view of the outcome of several end points, or several side-effects. In a large experiment one may want to consider several such families separately.

Hochberg and Tamhane (1987) define the family as “any collection of inferences for which it is meaningful to take into account some combined measure of errors.” They go on to say, “the following are the two key reasons for regarding a set of inferences as a family [which they attribute to Cox (1965)]:

(i) To take into account the selection effect due to data-snooping.

(ii) To ensure the simultaneous correctness of a set of inferences so as to guarantee a correct overall decision.”

They reach the conclusion that a logical choice for the family is “all tests performed within a single experiment.” Westfall and Young (1993) admit this choice is reasonable, but also argue that this still does not resolve the issue as it is not always clear what tests belong to a given experiment. They go on to state, “there are no definitive answers to these questions because there can be no universal agreement on this controversy.” However, they do add that “proper reporting of the results of a multiplicity-adjusted analysis should include...a clear description of the family of tests considered, as well as some justification for why that particular family was chosen (ideally, it is defined before the experiment is run or data are collected).” Finally, they offer several suggestions as to when to consider a set of hypotheses a family:

- It is considered plausible that all null hypotheses tested may in fact be true.

- The analyst plans to make a serious claim whenever any result \( p < 0.05 \) is found.

- The analyst is prepared to perform much data manipulation to find a “significant” result. (For example, a doctoral candidate may require a significant result for a defensible thesis.)
• The experiment or survey is expensive and is unlikely to be repeated before serious actions are taken.

Efron (2008) claims two dangers of combining all hypotheses from an experiment into a single family—“over and under sensitivity within different subclasses of the experiment.” He further suggests “a helpful methodology...for diagnosing when separation may be necessary for a subset of the testing problems, as well as for carrying out separation in an efficient fashion.” Alternatively, Genovese et al. (2006) maintain a single family of hypotheses while improving power by weighting the p-values using \textit{a priori} information. This follows from Holm (1979) who first introduced weighting in context of FWER control for hypotheses which had greater “importance.”

Benjamini and Braun (2001) quote Tukey as having said that an “obligation of the statistician, as a methodological generalist, is to develop insights of value to the scientific enterprise.” They also state that “Tukey repeatedly emphasized that intelligent control of multiplicity depends crucially on the appropriate choice of the family of statements.” Further, again quoting Tukey, “one role for the statistician is to provide guidelines for use based on the nature of the problem and the aim of the analysis.” Hence, the consensus appears to remain as quoted previously from Benjamini (2010), “that the vitality of [multiple comparisons] in the future – as a research area – depends upon [the researcher’s] ability to continue and address the real needs of statistical analyses in current problems.” That is, the success of multiple comparisons depends upon researchers being able to address the needs of each experiment to ensure both that truth is discovered and that discoveries are actually true.

1.3 Methods

Within the vast field of family selection, the various error rates that can be applied, and the multiple comparison procedures (MCPs) which control them, there is a class of methods which consider only the resulting \(P\)-values from \(m\) tests of significance. Shaffer (1995) refers to these as “methods based on ordered \(P\)-values.” The \(m\) tests of significance
need not utilize the same statistics or even the same null distributions to obtain the raw \( P \)-values. It is necessary that the individual \( P \)-values measure the lowest probability threshold under the null hypothesis at which rejection of the null would be achieved. The objective of the \( P \)-value methods can be focused on either rejection schemes or adjusted \( P \)-values, with a general preference for the second as witnessed by the \texttt{p.adjust} function in R and \texttt{PROC MULTTEST} in SAS. The acronym MCP will therefore be used herein to denote a procedure which adjusts for multiplicity the \( P \)-values from \( m \) tests of significance in such a way that the adjusted \( P \)-values can be compared to the \( \alpha \)-threshold just as in the PCER approach but all the while controlling the FWER at level \( \alpha \).

While this work is focused on methods for ordered \( P \)-values which control the FWER, it will at best introduce only a small portion of such methods from all those available. However, it is distinct from those MCPs which concern simultaneous confidence intervals and the pairwise comparisons of means which Hsu (1996) treats in detail. The application of ordered \( P \)-value methods to two novel situations is the main effort of this work. Chapter 2 extends graphical weighted Bonferroni methods to restricted hypotheses, where the methods could not previously be applied. Utilizing this extension, the Short Focus Level, an improvement to the Focus Level method (Goeman and Mansmann 2008) is presented (applicable to gene set testing using Gene Ontology graphs) which is both computationally more efficient and, for certain scenarios, statistically more powerful. Chapter 3 presents a weighted version of the graphical Bonferroni adjustment (Bretz et al. 2009) which is tailored to deal with a specific testing structure found in the two step approach to QTL mapping. The power benefits of the new approach as compared to the standard Bonferroni adjustment are demonstrated both through simulations and real data. Chapter 4 presents a synthesized view of the hypotheses inherent to QTL testing, performs a simulation study to uncover the advantages and disadvantages for different selections of the null distribution for those hypotheses, and proposes a novel application of a bivariate Monte Carlo test for those hypotheses. Chapter 5 discusses all these results generally and presents possible future work within the field of multiple comparison procedures.
The following subsections are provided as a reference for the interested reader. They detail several methods from the existing MCP literature which are discussed throughout the following chapters, at times with little explanation. They can be skipped altogether or can be read as needed.

1.3.1 The Bonferroni Method

Let $H_1, \ldots, H_m$ be a family of hypotheses for which control of the FWER is desired. Many methods exist for FWER control, the simplest and most well known being the Bonferroni correction. Derived from the Bonferroni inequalities the resulting method states that the FWER will be controlled at level $\alpha$ if each individual test $H_i$ is controlled at level $\alpha_i$ (PCER) where

$$\alpha_i = \frac{\alpha}{m} \text{ for } i = 1, \ldots, m.$$ 

Proof. The Bonferroni inequalities, or sometimes known as Boole’s inequality, state that for any events $A_i$

$$P \left( \bigcup A_i \right) \leq \sum P(A_i),$$

where equality holds only when the $A_i$ are mutually independent. Rewriting the inequality for $A_i^c$

$$P \left( \bigcup A_i^c \right) \leq \sum P(A_i^c). \quad (1.1)$$

The Bonferroni correction is then obtained by defining $A_i$ to represent the event that the $i^{th}$ hypothesis $H_i$ is not rejected. Then $A_i^c$ represents the probability that $H_i$ is rejected. The probability of this is $\alpha_i$ if the null hypothesis $H_i$ holds and the level $\alpha_i$ is used as the criterion for rejection. From this, the probability that at least one of the hypotheses are rejected (assuming they are all true) is denoted by $P \left( \bigcup A_i^c \right)$. Setting the right-hand side of (1.1) equal to $\alpha$ (where $\alpha$ is the desired level of control over the FWER) provides

$$P \left( \bigcup A_i^c \right) \leq \sum P(A_i^c) = m \alpha_i = \alpha. \quad (1.2)$$
This implies that if $\alpha_i = \alpha/m$ the probability of at least one rejection ($P(\cup A_i^c)$) is at most $\alpha$. Hence, the FWER is controlled at $\alpha$ when $\alpha_i = \alpha/m$.

Perhaps the most important thing to note in the development of the Bonferroni correction is the distribution free approach. The method does not contain any assumptions about the distribution of the test statistics corresponding to each hypothesis. This, and the simplicity of the method result in a general purpose baseline method for multiple hypothesis error control. Simply dividing the desired level $\alpha$ by the number of tests performed ($m$) ensures FWER control. Not a bad place to start. The following methods improve upon this with minimal increases in complexity.

1.3.2 The Holm Method

Holm (1979) wrote an article detailing his sequentially rejective Bonferroni test which expanded on the Bonferroni method. Along with his method, he formalized the idea of free combinations of the null hypotheses. The idea is that any number ($m_0$) and any subset of the $m$ null hypotheses being tested could be correct, or actually true. In other words, there is no logical structure imposed on the hypotheses for which the truth or falsehood of some would imply the truth or falsehood of others. Holm claimed that only methods that control the error rate under free combinations need to be considered as these imply control under restricted combinations. He provided the following definition.

Definition (Holm 1979)

A multiple test procedure for testing hypotheses $H_1, \ldots, H_m$ is said to have multiple level of significance $\alpha$ (for free combinations) if for any non-empty index set $I \subseteq \{1, 2, 3, \ldots, m\}$ the supremum of the probability $P(\cup A_i^c)$ when $H_i$ are true for all $i \in I$ is less than or equal to $\alpha$ where $A_i^c$ denotes the event of rejecting $H_i$.

This definition can be confusing, but is quite simple as the following translation shows.
Translation of Holm (1979)

A multiple test procedure for testing the hypotheses $H_1, \ldots, H_m$ controls the FWER for free combinations if the probability of committing at least one Type I Error is less than or equal to $\alpha$ for any configuration of true and false null hypotheses. This is called strong control of the FWER.

If this property is not satisfied, it may be that the method controls the FWER under the assumption that all nulls are true. This is called weak control of the FWER. Noting these definitions, it can be seen that the Bonferroni correction controls the FWER strongly.

To see this, let $I \subseteq \{1, 2, \ldots, m\}$ and note that only those null hypotheses $H_i$ with indices in $I$ are true. Hence, the only way to commit a Type I Error is by rejecting one of the hypotheses corresponding to the index set $I$. Let $k$ denote the number of hypotheses in $I$ and $A_i^c$ denote the event that $H_i$ is rejected. Since the probability of at least one of the hypothesis in $I$ getting rejected can be written as $P(\cup_{i \in I} A_i^c)$, it follows by Inequality (1.2) that $\alpha_i = \alpha/k$ applied to each test controls the FWER at $\alpha$ for any index set $I$. Since $\alpha/m \leq \alpha/k$ it holds that controlling each of the $m$ tests at level $\alpha_i = \alpha/m$ ensures that the true nulls will be controlled at $\alpha_i \leq \alpha/k$ no matter how many or which nulls are true.

Expanding on the above results, Holm’s method applies the Bonferroni method sequentially to the hypotheses by first ordering the resulting p-values from smallest to largest as $P(1), \ldots, P(m)$ according to the usual order statistics notation. The most significant hypothesis $H(1)$ corresponding to the result $P(1)$ is then tested with the usual Bonferroni correction of $\alpha_1 = \alpha/m$. If $H(1)$ is declared significant, then the method continues by checking $P(2)$ against $\alpha_2 = \alpha/(m-1)$. So long as rejections continue to occur, $P(i)$ is compared to $\alpha_i = \alpha/(m-i+1)$ until finally $P(m)$ is compared to $\alpha_m = \alpha$. If for any $i \in \{1, 2, \ldots, m\}$ the hypothesis $H(i)$ is not rejected, then the method stops and $H(i), \ldots, H(m)$ are retained, i.e. not rejected. This is summarized as follows.
Holm’s Method

Let \( P(\pi_1), \ldots, P(\pi_m) \) be the ordered p-values for the corresponding ordered hypotheses \( H(\pi_1), \ldots, H(\pi_m) \). Holm’s method rejects \( H(\pi_i) \) when for all \( j = 1, \ldots, i \)

\[
P(\pi_j) \leq \alpha_j = \alpha/(m - j + 1).
\]

Holm’s method controls the FWER strongly as his quick argument as follows shows. Let \( I \) denote the set of indices of true null hypotheses and let \( k \) denote the number of hypotheses in \( I \). Note that Inequality (1.2) shows that the probability of at least one Type I Error is controlled by the Bonferroni method so long as \( \alpha/k \) is applied to all \( k \) true nulls. Note that the probability that no Type I Error is committed is written \( P(P_i > \alpha/k \text{ for all } i \in I) \). Holm’s proof is then written formally (as in Holm (1979))

\[
P \left( P_i > \frac{\alpha}{k} \text{ for all } i \in I \right) = 1 - P \left( P_i \leq \frac{\alpha}{k} \text{ for some } i \in I \right)
\]

\[
\geq 1 - \sum_{i \in I} P \left( P_i \leq \frac{\alpha}{k} \right) \quad \text{(by (1.2))}
\]

\[
\geq 1 - k \frac{\alpha}{k} = 1 - \alpha.
\]

which demonstrates that the probability of none of the true nulls being rejected is at least \( 1 - \alpha \) so long as \( \alpha/k \) is used on all \( k \) true tests. Given the nature of the testing the number of true nulls remaining to be tested will always be less than or equal to the number of hypotheses remaining to be tested so that each true null will always be tested at least by \( \alpha/k \).

Just as with the Bonferroni method, Holm’s method is a distribution free approach to the multiple hypothesis testing issue. More importantly it is uniformly more powerful than the Bonferroni method as it will compare \( P(\pi_2), \ldots, P(\pi_m) \) to larger thresholds \( \alpha_i, i = 2, \ldots, m \) than will the Bonferroni method. For this reason it is clear that the Holm method should always be preferred over the Bonferroni method.
1.3.3 Closed Testing Procedures

This section presents the general theoretical work of Marcus et al. (1976) on devising stepwise multiple testing procedures which control the FWER. The paper was mainly motivated by the need for better multiple testing methods within the analysis of variance. However, their work ended up laying the foundation for much of the work in multiple hypothesis testing that came after this time. Before citing their closed testing procedure, we make note of the following definition.

Definition

A set of hypotheses $W$ is said to be closed under intersection if for any two hypotheses $H_i, H_j \in W$ their intersection hypothesis $w = H_i \cap H_j$ is also in $W$.

Applying this definition to a set of hypotheses is not immediately obvious. In fact, to really understand the work of Marcus et al. (1976) it is first necessary to understand some deeper theory about hypothesis testing. For a sufficient treatise on the matter see Chapter 8 of Casella and Berger (2002). For those familiar with the theory of hypothesis testing, the actual presentation of this section is better studied from the source Marcus et al. (1976). For those unfamiliar with such notation, or desiring a lower level approach to the same ideas, the current treatment will be useful.

Closed Testing Procedure

Let $W$ be a set of hypotheses that is closed under intersections and denote by $w$ an element of $W$. Reject the hypothesis $w_i$ if and only if all $w_j \subseteq w_i$ have been tested and rejected. Such a method controls the FWER at level $\alpha$ so long as each hypothesis is tested at level $\alpha$.

Example Proof.

The full proof of the closed testing procedure is left to Marcus et al. (1976). Here we provide an example that explains the rational behind the proof. Begin with the simple
hypotheses $H_1, H_2,$ and $H_3$. To form the set of hypothesis $W$ which is closed under intersections we must include the simple hypotheses $w_1 = H_1, w_2 = H_2,$ and $w_3 = H_3$ along with all their intersections:

$$w_4 : H_1 \cap H_2 \quad w_5 : H_1 \cap H_3 \quad w_6 : H_2 \cap H_3 \quad \text{and} \quad w_7 : H_1 \cap H_2 \cap H_3$$

where the numbering on the hypothesis $w$ could be arbitrarily chosen. It should be noted that this completes $W$ under intersections. For example $w_4 \cap w_5 = w_7$. Also, it is important to recognize the structure within $W$, i.e. $w_7 \subset w_6 \subset w_2$ or equally true is $w_7 \subset w_6 \subset w_3$, and so on, but $w_3 \not\subset w_2$ and $w_2 \not\subset w_3$. While any number of $H_1, H_2,$ and $H_3$ could be true null hypotheses, assume for the sake of this example that just $H_2$ and $H_3$ are true and that $H_1$ is false.

Denote by $A$ the event that any of the true nulls ($H_2$ or $H_3$ in this case) are rejected. Denote by $B$ the event that the intersection consisting of all true hypotheses ($w_6$ in this case) is rejected. By nature of the closed testing procedure, the only way for $A$ to happen is if $B$ has first occurred as $w_6 \subset w_3$ and $w_6 \subset w_2$. (Note that this implies $A \subset B$.) Recall the procedure, “Reject the hypothesis $w_i$ if and only if all $w_j \subset w_i$ have been tested and rejected.” The probability that both of these events occur then is, $P(A \cap B) = P(B)P(A|B)$ by the laws of conditional probability as $A$ is conditional on $B$. Since $w_6$ is tested at level $\alpha$ it follows that $P(B) = \alpha$ and since $P(A|B) \leq 1$ by the axioms of probability, it follows that $P(A \cap B) \leq \alpha$. However, since $A \cap B = A$ it follows that $P(A) = P(A \cap B) \leq \alpha$ and the FWER is controlled at level $\alpha$ as desired.

Some ideas closely related to the closed testing procedure stem from definitions supplied by Gabriel (1969). He initially established the idea of the closed testing procedure in Theorem 2 of his cited article for simultaneous test procedures much like the ANOVA F test that is performed before all pairwise comparisons are considered. The most important of Gabriel’s ideas as far as this work is concerned is coherence. The closed testing procedure described above is coherent in that if any hypothesis is rejected (say $w_1$) then all hypotheses
implied by it must also be rejected (in this case \( w_4, w_5, \) and \( w_7 \) would also have to be rejected).

### 1.3.4 The Weighted Bonferroni Method

Within the multiple hypothesis testing framework, there are times when structure can be introduced within the hypotheses. In other words, weights or importances can be assigned to each hypothesis as well as logical structures by which certain hypotheses are tested only if others are found first to be significant. Recall from Section 1.3.1 that the Bonferroni method strongly controls the FWER at level \( \alpha \) if each individual test \( H_i, i = 1, \ldots, m, \) is tested at level \( \alpha_i = \alpha/m. \) This was demonstrated by virtue of the Bonferroni inequality that provided that

\[
P(\text{at least one true } H_i \text{ is rejected}) \leq \sum_{i \in I} P(H_i \text{ is rejected}) = \sum_{i \in I} \alpha_i \leq \alpha, \quad (1.3)
\]

under the null hypotheses, where \( I \) denotes the index set of all true hypotheses. Inspection of Inequality (1.3) reveals that the Bonferroni method would still provide strong control of the FWER at level \( \alpha \) so long as the sum of the levels of the individual tests, \( \alpha_i, \) was again \( \alpha. \) To explain further, let \( k \) denote the number of indices in \( I \) (the number of true hypotheses) so that \( k \leq m. \) Then, the right hand side of Inequality (1.3) can be seen to be true as \( \sum_{i \in I} \alpha_i = k\alpha/m \leq \alpha. \) Under this approach, each individual test of \( H_i \) is given \( 1/m \) of \( \alpha \) so that the sum of \( m \) of these is once again \( \alpha, \) thereby ensuring that the sum of \( k \leq m \) of them is no more than \( \alpha. \) As the sum of the individual levels is all that is involved, the right hand side of Inequality (1.3) can be generalized for positive constants \( c_1, \ldots, c_m \) to obtain

\[
\sum_{i \in I} \alpha(c_i / \sum_{j \in I} c_j) \leq \alpha.
\]

This generalization allows departure from the original case of \( c_1 = \ldots = c_m = 1 \) while still controlling the FWER at level \( \alpha. \) The result is stated generally as follows.
Weighted Bonferroni

Let \( H_1, \ldots, H_m \) be a family of hypotheses for which control of the FWER is desired. Let \( c_1, \ldots, c_m \) be arbitrary positive constants assigned to the hypotheses \( H_1, \ldots, H_m \). The weighted Bonferroni method controls the FWER at level \( \alpha \) with the rule, reject \( H_i \) if

\[
P_i \leq \alpha_i = \alpha \left( \frac{c_i}{\sum_{j=1}^{m} c_j} \right).
\]

Note that this weighted method reduces to the original method when \( c_1 = \cdots = c_m = 1 \).

1.3.5 The Weighted Holm Method

Holm (1979), was one of the first to propose altering the proportion of \( \alpha \) used in a stepwise testing procedure, according to some \textit{a priori} knowledge that some hypotheses were "more important" than others. In the previous section the weighted Bonferroni procedure was presented to set the stage for his modified method, which is as follows.

Weighted Holm

Let \( H_1, \ldots, H_m \) be a family of hypotheses for which control of the FWER is desired. Let \( P_1, \ldots, P_m \) be the corresponding \( p \)-values and \( c_1, \ldots, c_m \) be positive constants assigned to each of the \( m \) hypotheses such that larger values of \( c_i \) imply greater importance for hypothesis \( H_i \). Define \( S_i = \frac{P_i}{c_i} \) and let \( S(1), \ldots, S(m) \) denote the ordered values of the \( S_i \). Denote by \( c(1), \ldots, c(m) \) the corresponding constants and by \( H(1), \ldots, H(m) \) the corresponding hypotheses. Then the weighted Holm procedure uses the rule, reject \( H(i) \) if

\[
S(j) \leq \alpha / \sum_{k=j}^{m} c(k), \text{ for all } j = 1, \ldots, i.
\]

As with the weighted Bonferroni method of the previous section, choosing \( c_1 = \cdots = c_m = 1 \) in the weighted Holm method results in the original (unweighted) method. This is seen as \( S_i = \frac{P_i}{c_i} = P_i \) in this case and \( \sum_{k=1}^{m} c(k) = m - j + 1 \) so that \( S(j) = P(j) \) is compared...
to $\alpha/(m-j+1)$. Similarly, setting all the weights equal to any constant, $c_1 = \cdots = c_m = c$, it follows that $S(j) = P(j)/c$, in other words, the ordering of the original $p$-values is unchanged when each $P_j$ is divided by the same constant. The resulting thresholds $\alpha c(j)/\sum_{k=j}^{m} c(k)$ reduce again to $\alpha c/c \sum_{k=j}^{m} 1 = \alpha/(m-j+1)$, the original unweighted Holm method. Thus, the magnitudes of the weights is not so important as their relative magnitudes.

Considering the method further, if we let $P(j)$ denote the $P_i$ that corresponds to $S(j)$ (not to be confused with $P(j)$) we can gain further insight into the weighted Holm method. In this case, the rejection formula can be rewritten as, reject $H(i)$ if

$$P[j] \leq \alpha(c(j)/\sum_{k=j}^{m} c(k)),$$

for all $j = 1, \ldots, i$.

Cast in this light, it is clear that $P(j)$ is compared to the proportion of $\alpha$ that $c(j)$ demands out of the total remaining weight, i.e. $\sum_{k=j}^{m} c(k)$, after $j = 1, \ldots, i - 1$ have already been rejected and “removed” from consideration. As Holm (1979) described it, “this implies an increase of power for alternatives to hypotheses with high values of $c_k$ at the cost of a decrease of power for alternatives to hypotheses with small values of $c_k$.”

To see that the weighted Holm method does control the FWER at level $\alpha$ in the strong sense, let $I$ be the set of all indices corresponding to the true hypotheses, and consider the case of free combinations among the hypotheses. (Recall that free combinations means that the truth of some hypotheses does not imply the truth of any other hypothesis. Further, controlling the FWER for free combinations implies control for restricted combinations (Holm 1979).) Let $C = \sum_{j \in I} c_j$, the total weight assigned to the true hypotheses. By virtue of the Bonferroni method, so long as all $P_i$ with $i \in I$ are tested by at most $\alpha(c_i/C)$, then the FWER is controlled at level $\alpha$. As it is $S_i = P_i/c_i$ that is actually tested in the weighted Holm method, it follows that so long as $S_i$ is tested by at most $\alpha/C$ for each $i \in I$, the FWER is controlled at level $\alpha$ as desired. To see this, note that the probability of at least one Type I Error (the definition of the FWER) can be written for the weighted Holm
method as

\[ P(S_i \leq \alpha/C \text{ for at least one } i \in I) = P(P_i \leq \alpha(c_i/C) \text{ for at least one } i \in I) \]

\[ \leq \sum_{i \in I} P(P_i \leq \alpha(c_i/C)) \]

\[ = \sum_{i \in I} \alpha(c_i/C) = \alpha. \]

It remains to be shown that the weighted Holm method will test each \( S_i \) belonging to a true \( H_i \) by at most \( \alpha/C \).

Consider that according to the weighted Holm method, the first \( S_i \) belonging to an \( i \in I \) will be tested by \( \alpha/\sum_{k=1}^{m} c_{(k)} \) where \( l \) is the position of \( S_i \) among all the ordered \( S_{(j)} \), \( j = 1, \ldots, m \). It is important to note that \( l \) could be any index from 1 to \( m \), but whatever the value of \( l \) the threshold against which \( S_{(l)} \) would be tested is given by the weighted Holm method as \( \alpha/\sum_{k=1}^{m} c_{(k)} \). Since \( S_{(l)} \) is the first weighted p-value in the ordered list of the \( S_{(j)} \) to correspond to a true null hypothesis, it follows that there are \( k - 1 \) remaining weighted p-values \( S_i \) (with \( i \in I \)) corresponding to true null hypotheses which have larger ordered indices than \( l \). Hence, \( \sum_{k=1}^{m} c_{(k)} \geq C \) so that \( \alpha/\sum_{k=1}^{m} c_{(k)} \leq \alpha/C \) (recall that \( C = \sum_{j \in I} c_j \)). If \( S_i > \alpha/C \) for each \( i \in I \), then all true hypotheses will be retained as the testing will stop by at most step \( l \) where the first weighted p-value corresponding to a true hypothesis is found. If at least one \( S_i \leq \alpha/C \), then the probability of rejecting at least one (the probability of at least one Type I Error) has already been demonstrated to be at most \( \alpha \). Hence, the weighted Holm method controls the FWER strongly at level \( \alpha \) as claimed.

1.3.6 Generalized Weighted Bonferroni Testing

Many Bonferroni type structured hypothesis testing methods have been proposed in the recent literature to control the FWER. While these include methods often referred to as gatekeeping procedures, fixed sequence tests, and fallback procedures, they are most clearly and succinctly summarized by the work of Bretz et al. (2009) which generalizes such methods into what they call a "graphical approach to sequentially rejective multiple test
procedures." For this reason, we discuss here only their work and not the work of those that led to their summarized approach. For the reader interested in reviewing the methods that led up to this approach see the references in Bretz et al. (2009).

The General Graphical Bonferroni Adjustment

Let $H_1, \ldots, H_m$ be a family of hypotheses for which control of the FWER at level $\alpha$ is desired and $M$ be the set of their indices, i.e. $M = \{1, \ldots, m\}$. Let $\alpha = (\alpha_1, \ldots, \alpha_m)$ be a vector of the thresholds at which each hypothesis $H_i$ will be tested with $\sum_{i=1}^{m} \alpha_i \leq \alpha$. Let $G = (g_{ij})$ denote an $m \times m$ transition matrix with entries $g_{ij}$ that are subject to the regularity conditions

$$0 \leq g_{ij} \leq 1, \quad g_{ii} = 0 \quad \text{and} \quad \sum_{k=1}^{m} g_{ik} \leq 1 \quad \text{for all} \quad i, j = 1, \ldots, m. \quad (1.5)$$

Then, for the observed $p$-values $p_1, \ldots, p_m$ the method, which strongly controls the FWER at level $\alpha$, is defined by the following algorithm.

0. Set $I = M$.

1. Let $j = \arg\min_{i \in I} p_i / \alpha_i$.

2. If $p_j \leq \alpha_j$, reject $H_j$; otherwise stop.

3. Update the graph:

$$I \rightarrow I \setminus \{j\}$$

$$\alpha_l \rightarrow \begin{cases} \alpha_l + \alpha_j g_{jl} , & l \in I \\ 0 & \text{otherwise} \end{cases}$$

$$g_{lk} \rightarrow \begin{cases} \frac{g_{lk} + g_{lj}g_{jk}}{1 - g_{lj}g_{jk}} , & l, k \in I, k \neq k \\ 0 & \text{otherwise} \end{cases}$$

4. If $|I| \geq 1$, go to step 1; otherwise stop.
Fig. 1.2. A simple noded diagram of the Bonferroni test for three hypotheses. Each node (or hypothesis) is tested with equal weight of \( \alpha/3 \).

To begin uncovering the algorithm proposed by Bretz et al. (2009) the basic Bonferroni (unweighted) method will be considered graphically (Figure 1.2). For simplicity, consider only \( m = 3 \) hypothesis tests with \( H_1, H_2, \) and \( H_3 \) and resulting \( p \)-values \( p_1 = 0.0032, \) \( p_2 = 0.022, \) and \( p_3 = 0.72. \) Then the unweighted Bonferroni test performed at level \( \alpha = 0.05 \) would test each hypothesis at level \( \alpha/3 \approx 0.0167 \) and could be depicted as shown in Figure 1.2, with significant results obtained only for \( H_1. \)

Considering the diagram of Figure 1.2, it can be seen how the weighted Bonferroni test could easily be applied in place of the Bonferroni test by reallocating the amount of \( \alpha \) that is partitioned to each node. Recall that the weighted Bonferroni method is applied by selecting positive constants \( c_1, c_2, \) and \( c_3 \) with larger (relative) values implying greater importance for the corresponding hypothesis. Once selected, these constants are used to obtain the individual thresholds \( \alpha_i = \alpha(c_i / \sum_{j=1}^{m} c_j). \) As stated previously, the magnitudes of the \( c_i \) are not important, only their relative magnitudes as each \( c_i \) is standardized by \( \sum_{j=1}^{m} c_j. \)

Considering how the \( \alpha_i \) are defined, it follows that \( 0 \leq \alpha_i < \alpha \) and that \( \sum_{i=1}^{m} \alpha_i = \alpha. \) Hence, the weighted Bonferroni method can be achieved simply by selecting the thresholds \( \alpha_i \) directly with the constraints that the \( \sum_{i=1}^{m} \alpha_i = \alpha \) and that each \( \alpha_i \) is between 0 and 1.

At this point it should be emphasized that the allocation of \( \alpha \) to each hypothesis must be performed before testing begins and should depend on \textit{a priori} knowledge and not the resulting \( p \)-values. If \( \alpha \) was distributed after the fact, then one could easily reject both \( H_1 \) and \( H_2 \) of the above example by assigning say \( \alpha_1 = 0.0035, \alpha_2 = 0.025, \) and \( \alpha_3 = 0.0215. \) This choice indeed satisfies \( \sum_{i=1}^{3} \alpha_i \leq \alpha \) and \( 0 \leq \alpha_i \leq \alpha \) for all \( i = 1, 2, 3, \) but would not control the FWER at level \( \alpha \) due to the \textit{a posteriori} selection of the \( \alpha_i. \) However, if \textit{a priori} we felt that \( H_2 \) was the most important hypothesis and deserved \( \alpha_2 = 0.04 \) of the total level with \( \alpha_1 = \alpha_2 = 0.005 \) (as in Figure 1.3), then we would arrive at the same conclusion.
Fig. 1.3. A node diagram of the weighted Bonferroni test for three hypotheses. Each node (or hypothesis) is tested with weight (assigned a priori) of $\alpha_1 = \alpha_3 = 0.005$ and $\alpha_2 = 0.04$.

of rejecting both $H_1$ and $H_2$, but with control of the FWER at level $\alpha = 0.05$ by virtue of the weighted Bonferroni method.

While the thresholds $\alpha_i$ must be assigned a priori to maintain control of the FWER at level $\alpha$, the foregoing discussion does set forward an interesting idea. That is, if there was a way to shift the allocation of $\alpha$ to the places where it was needed most, then the power of the test could be increased. So long as the probability of at least one Type I Error is preserved at level $\alpha$ (the FWER) then any approach to reallocation of the $\alpha$ level would be acceptable.

Recall that the definition of the $\alpha$ level for a single hypothesis test is the probability of a Type I Error under the null hypothesis. In other words, $\alpha$ is intentionally kept small as a protection against a Type I Error under the assumption that the null hypothesis is true. However, once a decision is made about a hypothesis, of what use is the $\alpha$ level? The answer depends on the conclusion of the test. If the hypothesis is retained (not rejected) then the $\alpha$ level stands as the protection against the Type I Error and must remain, in a sense, fixed to that hypothesis test as we continue under the assumption that the null hypothesis is true. On the other hand, if the hypothesis is rejected, then we are willing to believe that the given hypothesis is alternatively expressed, meaning it is no longer probable to conclude that the null hypothesis is a proper assumption for that event and so the $\alpha$ level is therefore meaningless in that context.

Returning to the example in Figure 1.2 the question should be asked, what should be done with the $\alpha/3$ level that is left after rejecting $H_1$ if, according to the previous discussion, it is no longer of use? Here is the first consideration of introducing structure within the hypotheses. Before testing began, we did not know which (if any) hypotheses would show significant results. It would seem logical therefore to allow for the redistribution of any $\alpha_i$.
Fig. 1.4. A nodded diagram of the sequential Bonferroni test for the three hypotheses of Figure 1.2 with weighted directed edges. Each node (or hypothesis) is first tested with equal weight of $\alpha/3$ but if any hypothesis is rejected, its threshold $\alpha_i$ is shared with the remaining hypotheses as specified by the weights along the arrows.

to the other hypothesis tests if $H_i$ was found significant, providing more power towards the safe rejection of the other hypotheses. As we have no reason to give more of $\alpha_i$ to any one hypothesis over another, it seems logical in this case to share $\alpha_i/2$ with each remaining hypothesis. Consider how this is accomplished by use of the directed arrows in Figure 1.4.

As shown in Figure 1.4 each hypothesis is first tested according to the Bonferroni assigned thresholds of $\alpha/3$. With the $p$-values as previously stated, this allows for rejection of $H_1$ as $p_1 = 0.0032 < \alpha_1 = 0.05/3$ with retention of both $H_2$ and $H_3$ as $p_3 = 0.72$ and $p_2 = 0.022$ are both greater than $0.05/3$. Since $H_1$ was rejected, its threshold of $\alpha/3$ can now be passed along to both $H_2$ and $H_3$ with $1/2$ going to each. This provides a new graph as shown in Figure 1.5 as $H_1$ is now removed from consideration. Notice that the thresholds for testing $H_2$ and $H_3$ have now been increased to $\alpha/2$ as $\alpha/3 + 1/2(\alpha/3) = \alpha/2$. Also, the directed edges between $H_2$ and $H_3$ now show that if either $H_2$ or $H_3$ is rejected, then all of the $\alpha$ level for that hypothesis is shifted to the testing of the other hypothesis.

With the new $\alpha/2$ level, $H_2$ can now be declared significant as $p_2 = 0.022 < 0.05/2$. This then implies that $H_3$ is the only hypothesis remaining and is to be tested at level $\alpha$ as $\alpha/2 + 1(\alpha/2) = \alpha$. Clearly $H_3$ is non significant, and thus is retained and the testing is complete. The astute reader will notice at this point that $H_1$ was tested at level $\alpha/3$, followed by $H_2$ being tested at $\alpha/2$, and finally $H_3$ was tested at level $\alpha$. This is precisely the unweighted Holm method because $p_1 < p_2 < p_3$. What is even more surprising is that
Fig. 1.5. (a) The updated graph after $H_1$ has been rejected and removed from further consideration and its threshold has been shared equally between $H_2$ and $H_3$. Notice that all logical structures involving $H_1$ have been severed. (b) The final graph of $H_3$ obtained after $H_2$ is rejected at level $\alpha/2$, removed from consideration, and all of its threshold is passed on to $H_3$.

we have applied the graphical Bonferroni adjustment just as described in Section 1.3.6, The General Graphical Bonferroni Adjustment.

To summarize how the graphical Bonferroni adjustment (GBA) has just been applied, consider that first all of the regularity conditions were met prior to any testing. The assigned thresholds $\alpha_i$ summed to $\alpha$. Also, the amount of the $\alpha/3$ threshold from $H_1$ to be passed on to $H_2$ was $1/2$ as was the amount to be passed to $H_3$ so that to total proportion of $\alpha/3$ being passed on to the other hypotheses in the case that $H_1$ was rejected was $1/2 + 1/2 = 1$. The same is true for $H_2$ as well as $H_3$. Thus the requirement that $\sum_{k=1}^{m} g_{ik} \leq 1$ is satisfied as is $0 \leq g_{ij} \leq 1$ whenever $i \neq j$ and was zero whenever $i = j$ as no hypothesis returned any proportion of its $\alpha_i$ level to itself.

With the regularity conditions met, Step 0 is applied so that $I = M = \{1, 2, 3\}$. Then, Step 1 assigns $j = 1$ as $p_1/\alpha_1 < p_2/\alpha_2 < p_3/\alpha_3$. Step 2 rejects $H_1$ as $p_1 \leq \alpha_1$. Step 3 updates the graph as shown in the left side of Figure 1.5 by first removing $\{j\}$ from $I$ so that currently $I = \{2, 3\}$. The thresholds for $H_2$ and $H_3$ (all hypotheses with indices still in $I$) are then updated by the rule

$$\alpha_2 = \alpha_2 + \alpha_1 (1/2) \quad \alpha_3 = \alpha_3 + \alpha_1 (1/2).$$
Finally, the outgoing edges for nodes $H_2$ and $H_3$ are restandardized to both sum to 1 and allow for the removal of $H_1$ by the rules (see Figure 1.5)

\[
g_{22} = 0, \quad g_{23} = \frac{g_{23} + g_{21}g_{13}}{1 - g_{21}g_{12}}, \quad g_{32} = \frac{g_{32} + g_{31}g_{12}}{1 - g_{31}g_{13}}, \quad g_{33} = 0
\]

so that

\[
g_{22} = 0, \quad g_{23} = \frac{1/2 + (1/2)(1/2)}{1 - (1/2)(1/2)} = 1, \quad g_{32} = \frac{1/2 + (1/2)(1/2)}{1 - (1/2)(1/2)} = 1, \quad g_{33} = 0.
\]

Step 4 finds that $|I| = 2$ and so the process is reiterated to obtain the reduced graph of Figure 1.5.

At this point, while still new to the reader, the GBA is likely only mysterious in the reassignment of the $g_{ik}$ in Step 3 of the algorithm. First, we emphasize that the notation $g_{ik}$ is used to describe the proportion of $\alpha_l$ that is reallocated to $\alpha_k$ in the case that $H_j$ is rejected. While the number of different indices is at first confusing, it is important to note that rejecting $H_j$ causes that many $g_{ik}$ must be updated due to the removal of all $g_{lj}$ and $g_{jk}$ from the graph when $H_j$ is removed, i.e., rejected. Assuming $H_j$ is rejected, the Bretz algorithm then provides the rule

\[
g_{ik} \rightarrow \begin{cases} 
\frac{g_{ik} + g_{ij}g_{jk}}{1 - g_{ij}g_{jk}}, & l, k \in I, k \neq k \\
0 & \text{otherwise.}
\end{cases}
\]

To update $g_{ik}$ (the proportion of $\alpha_l$ being passed to $\alpha_k$ now that $H_j$ is removed) any proportion of $\alpha_l$ that was being passed from $H_l$ to $H_k$ through $H_j$ should be combined with any proportion that is being passed directly from $H_l$ to $H_k$. However, the effort is complicated by the proportion of $\alpha_l$ that was being returned to $H_l$ via first the outgoing $g_{lj}$ and then the incoming $g_{jl}$ (which have now both been removed). In a sense, $g_{jl}$ of whatever was being sent to $H_j$ by $g_{lj}$ was being returned to $H_l$, so that $g_{lj}g_{jl}$ describes the proportion previously cycled between $H_l$ and $H_j$. Similarly, $g_{lj}g_{jk}$ represents the proportion previously being passed from $H_l$ through $H_j$ to $H_k$. Hence, to update $g_{ik}$ we add to the direct line, $g_{ik}$.
any proportion that was being passed from $H_i$ to $H_k$ indirectly through $H_j$ to obtain the numerator $g_{ik} + g_{ij}g_{jk}$. Then, to ensure that 100% of whatever proportion was previously being sent out from $H_i$ is still being sent out, we standardize the newly obtained proportion for $g_{ik}$ by dividing by $1 - g_{ij}g_{jk}$, one minus the proportion being cycled between $H_i$ and $H_j$.

With the application of the algorithm to the graphical method in place, the next point of interest is how different choices of initial $\alpha$ allocation and weight selection affect the method. It was already demonstrated that equal division of $\alpha$ to each hypothesis $(\alpha_1 = \cdots = \alpha_m = \alpha/m)$ with equal outgoing weight to all other hypotheses in the case of a rejection ($g_{ij} = 1/(m - 1)$ for all $i, j \in \{1, \ldots, m\}$ where $i \neq j$ and $g_{ii} = 0$) resulted in the Holm method. Preserving the equal outgoing weight to all other hypotheses while initially allocating the $\alpha$ level unequally between hypotheses results in the weighted Holm method. The important note here is that each hypothesis is logically connected to all other hypotheses with equal weight sharing ($g_{ij}$) in either method. Keeping all logical connections while varying the magnitudes of the $g_{ij}$ generalizes the weighted Holm method. Keeping only select logical connections opens the doors to many possible testing strategies, few of which are actually named, but which include the *Fixed Sequence Test* and *Fallback Procedure*. For details see Bretz *et al.* (2009).

One final note on the GBA is pertinent to the current treatise. All testing presented to this point concerned a single family of hypotheses. It may be the case that several families of hypotheses exist and it is desired to introduce logical structures between these families, perhaps testing certain families only if some (or all) hypotheses from another family are first found to be significant. The GBA allows for this possibility through the use of what is called *epsilon edges*, denoted by $\epsilon$. Bretz *et al.* (2009) establish calculation rules for $\epsilon$ with positive real numbers $x$ by $x + \epsilon = x$, $x\epsilon = 0$, $\epsilon^0 = 1$, and for all non-negative integers $k, l$

$$
\frac{\epsilon^k}{\epsilon^l} = \begin{cases} 
0 & \text{if } k > l \\
1 & \text{if } k = l \\
\infty & \text{if } k < l.
\end{cases}
$$

(1.6)
With these rules, the GBA can be directly applied, and ensure that no positive amount of \( \alpha \) is passed along an \( \epsilon \) edge unless the proper hypotheses are first rejected. Returning to the previous example, assume that \( H_1 \) and \( H_2 \) belong to one family and that \( H_3 \) belongs to another family. Perhaps the first family with \( H_1 \) and \( H_2 \) is of primary concern, and that if both of these hypotheses are rejected, \( H_3 \) will be tested with the full level \( \alpha \) passed along from \( H_1 \) and \( H_2 \). If one or both of \( H_1 \) and \( H_2 \) are not rejected, then \( H_3 \) will not be tested. Before the presentation of the epsilon edge, an attempt at such an approach may have looked something like “Attempt 1” in Figure 1.6.

\[
\begin{array}{c}
\text{Attempt 1:} \\
H_1 \xrightarrow{1} H_2 \xrightarrow{\alpha/2} 0 \\
\end{array}
\]

\[
\begin{array}{c}
\text{Attempt 2:} \\
H_1 \xrightarrow{1} H_2 \xrightarrow{\alpha/2} 0 \\
\end{array}
\]

\[
\begin{array}{c}
\text{Attempt 3:} \\
H_1 \xrightarrow{1 - \epsilon} H_2 \xrightarrow{\epsilon} 0 \\
\end{array}
\]

**Fig. 1.6.** A first attempt at a testing diagram for testing the first family of hypotheses including \( H_1 \) and \( H_2 \), each at level \( \alpha/2 \), followed by the testing of the second family (\( H_3 \)) at level \( \alpha \) in the case that both hypotheses in the first family (\( H_1 \) and \( H_2 \)) are rejected.

The difficulty with this first attempt is that we have performed a Holm method test on the first family with no way to graphically represent the passing on of the \( \alpha/2 \) levels to \( H_3 \) if significance is found in both \( H_1 \) and \( H_2 \). Yet, from previous discussion, it follows that \( \alpha \) would be free to be redistributed to \( H_3 \) in this case. Returning to the algorithm of the GBA, notice that if \( H_1 \) were to be rejected at level \( \alpha/2 \), then the graph provides that all of its \( \alpha/2 \) would be passed on to \( H_2 \), which is appropriate for the current testing strategy. The updating algorithm would then give \( H_2 \) a level of \( \alpha \) with no outgoing weights to pass on to \( H_3 \). Hence, we need in this case an edge connecting \( H_2 \) to \( H_3 \) which only becomes
“active” after $H_1$ and $H_2$ are both rejected. Consider that this could be accomplished by inserting an $\epsilon$ edge between $H_2$ and $H_3$ as shown in “Attempt 2” of Figure 1.6.

The difficulty with this second attempt is not immediately obvious. According to the calculation rules established previously, $1 + \epsilon = 1$ so that the outgoing weight allocation from $H_2$ still satisfies the requirement of being less than or equal to one. However, consider what happens to the $g_{ik}$ in Step 3 of the GBA if the graph of Attempt 2 is used. Let’s assume that $H_1$ is rejected in Step 2 so that Step 3 first sets $I = \{2, 3\}$ and then performs the following updates.

$$
\alpha_2 = \alpha_2 + \alpha_1(1) \quad \alpha_3 = \alpha_3 + \alpha_1(0)
$$

$$
g_{22} = 0, \quad g_{23} = \frac{g_{23} + g_{21}g_{13}}{1 - g_{21}g_{12}}, \quad g_{32} = \frac{g_{32} + g_{31}g_{12}}{1 - g_{31}g_{13}}, \quad g_{33} = 0
$$

From these then we have $\alpha_2 = \alpha$, $\alpha_3 = 0$,

$$
g_{22} = 0, \quad g_{23} = \frac{\epsilon + (1)(0)}{1 - (1)(1)}, \quad g_{32} = \frac{0 + (0)(0)}{1 - (0)(0)}, \quad \text{and } g_{33} = 0.
$$

This results in a logical fallacy for the calculation of $g_{23}$ as $1/0$ is undefined. However, notice what happens in the calculations if we use the logical connections shown in Attempt 3. In this case we have the same calculations for $g_{22}$, $g_{32}$, and $g_{33}$, but a different result for $g_{23}$ as

$$
g_{23} = \frac{\epsilon + (1 - \epsilon)(0)}{1 - (1 - \epsilon)(1)} = \frac{\epsilon}{\epsilon} = 1.
$$

Thus, only Attempt 3 properly joins $H_2$ to $H_3$, passing all of $\alpha$ to $H_3$ in the case that $H_2$ is subsequently rejected. Hence, as this example proposes, it follows that an extra regularity condition must be established when using $\epsilon$ edges, even though the established calculation rules would suggest otherwise.

**Extra Regularity Condition for $\epsilon$ Edges**

When using $\epsilon$ edges in the Bretz method, all outgoing edges $g_{ij}$ must satisfy the previous constraints of The General Graphical Bonferroni Adjustment of Section 1.3.6 under the
Fig. 1.7. If $H_1$ and $H_2$ are both rejected, the above diagram specifies that $r_1$ of $\alpha$ will be passed along to $H_3$ and that $r_2$ of $\alpha$ will be passed along to $H_4$. If at least one of $H_1$ or $H_2$ is not rejected, then neither $H_3$ or $H_4$ are tested.

regular calculation rules for positive real numbers. In other words, $\epsilon$ must be treated in these calculations as a positive real number and not as specified in the calculation rules of Equation (1.6).

From this extra rule on the regularity conditions, we see that if we wish to add an $\epsilon$-edge as in Attempt 2 of Figure 1.6, then we are forced to change the weight on the outgoing edge from $H_2$ to $H_1$ from 1 to $1 - \epsilon$ in order that all outgoing edges from $H_2$ sum to 1. This requirement was overlooked when the calculation rules of Equation (1.6) were implemented.

Finally, if it is desired to have several outgoing edges (from a single node) incorporate $\epsilon$ edges, then multiplying each $\epsilon$ edge by weights of $r_1, \ldots, r_n$ such that $\sum_{i=1}^n r_i = 1$ will accomplish the desired result. For example, assume there was a second hypothesis $H_4$ in our example included in the second family (which previously contained only $H_3$). Perhaps instead of passing all of $\alpha$ to $H_3$ in the case that both $H_1$ and $H_2$ are rejected (as was previously done) suppose that it is desired to share $\alpha$ between $H_3$ and $H_4$ according to the proportion $r_1 + r_2 = 1$. Then Figure 1.7 shows how this could be accomplished.

The methods of this section have only briefly been established, and their full versatility remains for the reader to explore. However, the rules and approaches necessary for the practitioner to establish their own FWER controlling testing format have been meticulously established. It remains only to apply them. Further examples demonstrating the versatility of the method can be found in Bretz et al. (2009) as well as Chapters 2 and 3 of this work.
CHAPTER 2
FWER CONTROL WITHIN GENE ONTOLOGY GRAPHS

2.1 Introduction

Microarray technology and next generation sequencing have played an important role in discovering important associations between gene expression patterns and phenotype (Malone and Oliver 2011). An excellent source for an introduction to the microarray and next generation sequencing technologies can be found in the review by Jaluria et al. (2007). Such gene expression technologies have been instrumental in discoveries ranging from the retarding of aging in mice brought about by caloric restrictions in diet (Lee et al. 1999) to the identification of various types of diffuse large B-cell lymphoma in humans (Alizadeh et al. 2000); from characterizing the transcriptomes of in vitro manipulated porcine embryos (Isom et al. 2013) to uncovering the underlying genes and pathways involved in Alzheimer’s disease (Miller et al. 2013). While both microarray and next generation sequencing technologies allow researchers to study the differential expression of genes across conditions or treatments, each has their advantages and disadvantages (Malone and Oliver 2011). However, in either case, the resulting increase in genetic knowledge has allowed researchers to group genes with common function into gene sets and test these gene sets for differential expression (Efron and Tibshirani 2007; Goeman et al. 2004).

Gene set testing allows for the quantification of the significance of activity level differences between treatment groups for specific biological processes of interest. For example, a recent study on human longevity compared the gene expression profiles corresponding to 1,808 different biological processes for nonagenarians and a control group to identify 73 biological processes associated with longevity (Passtoors et al. 2012). When there are relatively few gene sets (biological processes) of a priori interest (1,808 in Passtoors et al. 2012),
the impact of the multiplicity correction for the tests of differential expression of
the gene sets can be greatly lessened as compared to individually testing all member genes
(45,164 in Passtoors et al. (2012)), improving the power of the test. Even when no a priori
gene set of interest can be specified, it can still be highly beneficial to test all known gene
sets from a biological process database for differential expression, as the number of gene
sets is still typically magnitudes smaller than the corresponding number of individual genes
(Goeman and Mansmann 2008; Goeman et al. 2004).

One rich source of gene set knowledge is found in the Gene Ontology database (Ashburner
et al. 2000). The Gene Ontology (GO) provides a controlled vocabulary that is not
specific to any particular species. This vocabulary is divided into three general ontologies,
Molecular Function (MF), Cellular Component (CC), and Biological Process (BP). Indi-
vidual GO Terms form the basis of these vocabularies and are structured through parent
child relationships with more general terms as parents and more specific terms as children.
Each GO Term typically contains a definition of its biological process (molecular function
or cellular component) and other annotation as well as a mapping of all known gene prod-
ucts involved in its specified process (function or component). For example, consider the
biological process GO Term GO:0007005, mitochondria organization, which is defined as “A
process that is carried out at the cellular level which results in the assembly, arrangement
of constituent parts, or disassembly of a mitochondrion” (Ashburner et al. 2000). There are
currently 4,794 gene products annotated to GO:0007005. Further, GO:0007005 is a direct
child of the BP GO term GO:0006996, organelle organization, and by inheritance, an off-
spring of 6 other BP GO terms including the root biological process term, GO:0008150 (see
Figure 2.1). Similarly, GO:0007005 is the parent of 19 other BP GO terms. The structure
of the GO ontologies is such that any parent term contains all mapped gene products (and
thereby all genes which map to those gene products) of its children terms as well as any
other genes mapped directly to it. For example, there are currently 34,866 gene products
annotated to the parent term (GO:0006996) of GO:0007005. The root Biological Process
term, GO:0008150, contains all gene products mapped to any other GO Term in the Bi-
Table: GO Term Node Table

<table>
<thead>
<tr>
<th>Node</th>
<th>Term ID</th>
<th>Term Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0008150</td>
<td>biological process</td>
</tr>
<tr>
<td>2</td>
<td>GO:0009987</td>
<td>cellular process</td>
</tr>
<tr>
<td>3</td>
<td>GO:0071840</td>
<td>cellular component organization or biogenesis</td>
</tr>
<tr>
<td>4</td>
<td>GO:0016043</td>
<td>cellular component organization</td>
</tr>
<tr>
<td>5</td>
<td>GO:0006996</td>
<td>organelle organization</td>
</tr>
<tr>
<td>6</td>
<td>GO:0007005</td>
<td>mitochondrion organization</td>
</tr>
</tbody>
</table>

Fig. 2.1. Graphical demonstration of the ancestral relationships of the GO term GO:0007005 within the Biological Process GO graph. Arrows are directed from child to parent. The root BP GO term (GO:0008150) is the only node without ancestors in the graph.

Many methods of gene set testing have been proposed in the literature as reviewed in Goeman and Buhlmann (2007). These can essentially be divided into two classes of gene set testing, often referred to as competitive tests and self contained tests. The competitive tests compare the expression profiles of the genes in the set to those not in the set. The self contained tests focus only on those genes within the set and compares them to some fixed standard. While the first are more popular (Efron and Tibshirani 2007; Khatri and Draghici 2005), the second have been shown to be more powerful (Fridley et al. 2010; Goeman and Buhlmann 2007). Further, the null hypothesis associated with the self contained tests,

\[ H_0^{self}: \text{no genes in the gene set are differentially expressed,} \]

has been shown to be the more logical generalization of single gene testing (with other advantages that will be explained later on) as compared to the competitive test null hypothesis

\[ H_0^{comp}: \text{the genes in the gene set are at most as often differentially expressed as the genes in the complement of the gene set.} \]

While gene set testing methods are varied in their approach, they are alike in that they test each GO term, i.e. gene set, individually. Thus, when more than one GO term is

ological Process ontology plus any others mapped solely to it so that currently there are 563,081 mapping gene products.
tested simultaneously (typically hundreds or thousands are tested simultaneously) some sort of multiplicity adjustment is necessary to preserve control over either the family-wise error rate (FWER) or the false discovery rate (FDR) or a derivative of these error rates. The FDR is typically the error rate of choice in exploratory studies where follow up confirmatory studies are then conducted (Nettleton et al. 2008). On the other hand, the FWER is typically the suggested error rate for confirmatory studies (Hochberg and Tamhane 1987). We also suggest that the FWER is highly appropriate for exploratory gene set studies as, in our experience, it is seldom more results that are desired, but the most promising real significances that are sought. The FWER offers the best error rate control for such conclusions (Hochberg and Tamhane 1987).

Goeman and Mansmann (2008) proposed the powerful Focus Level method of multiplicity adjustment for self contained gene set testing which takes into account the structure of the GO graph while controlling (strongly) for the FWER. This approach is more powerful than standard FWER controlling methods such as the Bonferroni and uniformly more powerful Bonferroni-Holm (Holm 1979) procedures for multiple testing with GO graphs (Goeman and Mansmann 2008). The Focus Level method allows the researcher to select the level of the GO graph in which they are most interested. This is called the focus level. The procedure then applies a top-down and bottom-up approach from the specified focus level. First, the terms in the focus level are tested using the multiplicity approach of Holm (1979). Then, in the bottom-up approach, any term above the focus level is declared significant when any of its offspring in the focus level have been declared significant. This inheritance of P-values is accomplished through the assumption that a parent term must be differentially expressed if any of its children terms are differentially expressed, a logical assumption for the GO graph structure. In the top-down procedure, significance of the children of the focus level terms is decided through an application of the closed testing procedure of Marcus et al. (1976) (see Section 1.3.3 for an introduction).

While the Focus Level method is a powerful approach to adjusting for multiplicity, it quickly becomes computationally infeasible when the selected focus level contains a large
This computational limitation makes it essentially impossible to perform the full top-down approach, a rather significant disadvantage (Liang and Nettleton 2010). Using the full top-down approach provides researchers the default focus level of the root node (GO:0008150 in the context of the BP GO ontology) whenever they have no a priori interest in a given focus level, a common scenario, see for example Liang and Nettleton (2010). This also allows adjusted \( P \)-values to be considered apart from their context in the GO graph which is advantageous to reporting on single significant gene sets of interest. Discussions of the significant findings of the Focus Level method are currently restricted to their context within the GO graph (Goeman and Mansmann 2008).

This work proposes an improvement to the top-down portion of the Focus Level method of Goeman and Mansmann (2008) which we call the Short Focus Level as it performs a shortcut of the full Focus Level method. This is accomplished using a novel application of the general graphical Bonferroni adjustment for multiple testing as proposed by Bretz et al. (2009), which is a generalization of closed testing based on weighted Bonferroni tests (Hommel et al. 2007). (See Sections 1.3.3, 1.3.4, and 1.3.6 for an introduction to each of these topics.) The Short Focus Level procedure shows a significant improvement in computational speed (as much as \( \sim 15,000 \) times faster) while maintaining similar power to that of the original Focus Level procedure and even showing a gain in power over the original Focus Level procedure for certain scenarios while experiencing a loss in power for others. Most importantly, the computational improvements are such that the full top-down method can now be performed on a standard operating system within just a few minutes. The R code (R Core Team 2013) for the improved Focus Level procedure currently consists of two functions, `shortFindFocus` and `shortFocusLevel`, which will be included in the forthcoming `mvGST` package (Mecham 2014; Stevens and Mecham 2014).

2.1.1 The Focus Level Method

basic assumptions underly the method.

A1. A non-differentially expressed parent gene set implies the children gene sets are also non-differentially expressed.

A2. If the children gene sets form a partition of the parent gene set and are all non-differentially expressed, then the parent gene set is also non-differentially expressed.

These assumptions require that the null hypothesis for each gene set is that no genes in the gene set are differentially expressed. The alternative in each case being that at least one gene in the set is differentially expressed. Thus, only self contained gene set testing methods (which utilize this hypothesis framework) can be used to test the gene sets of the GO graph if the Focus Level method of multiplicity adjustment is used. This excludes gene set enrichment methods such as those proposed in Khatri and Drăghici (2005) but supports very well the Global Test of Goeman et al. (2004), \( P \)-value combination methods such as Fisher’s and Stouffer’s methods (Fridley et al. 2010; Stevens and Isom 2012), as well as Global Ancova (Hummel et al. 2008), PLAGE (Tomfohr et al. 2005), and SAM-GS (Dinu et al. 2007).

As prescribed by Goeman and Mansmann (2008) there are two requirements in the selection of the focus level.

FL1. No offspring of a focus level term be contained in the focus level.

FL2. All remaining terms are either ancestors or offspring of the focus level terms.

Figure 2.2 demonstrates on a simplified toy GO graph how the focus level (filled nodes) could be chosen. The full bottom-up approach (panel (a) of Figure 2.2) selects all GO Terms corresponding to terminal nodes as the focus level, in this example, nodes \( C \), \( D \), and \( E \). The full top-down approach (panel (c) of Figure 2.2) selects the root node, \( A \) in this case, as the focus level. Finally, in a typical GO graph there are many (hundreds or thousands) of options for focus levels contained somewhere in the middle of the GO graph. In the simplified example graphs of Figure 2.2 the most logical intermediate focus level is
Fig. 2.2. Three possible focus levels (filled nodes) for a simplified example GO graph. 

Demonstrated with nodes B and F (panel (b)). It would also be possible to use nodes C, D and F as the focus level but such choices in actual GO graphs do not provide a consistent level of specificity in the graph and would not be as logical a choice. Choosing nodes C, D, E, and F as the focus level would not be allowed as E is a child of F, violating the requirement that the focus level must not contain any offspring of another focus level term (E is an offspring term of F). Choosing only node B as the focus level would also not be allowed as node F is neither an ancestor or offspring term of B, violating the second requirement.

The top-down portion of the Focus Level procedure of Goeman and Mansmann (2008), which applies the closed testing approach of Marcus et al. (1976), requires closing the GO graph under all unions from the focus level down. This is done by treating each focus level term, along with all of its offspring terms, as separate graphs which are each closed under all possible unions. As these separate closed graphs will share common elements, the full closed graph $\bar{G}$ is obtained by unioning each of the separately closed graphs into a single graph which is also unioned to all ancestor terms of the focus level.

To demonstrate, consider the closures of each of the example GO graphs from Figure 2.2 as shown in Figure 2.3. In each case, the nodes above the focus level remain unchanged, while the creation of several sets not present in the original example GO graph (depicted with rounded rectangles) are required in order to close the graph under all possible unions from the focus level down. Since the closing of the graph is only required from the selected focus level down, it is clear from Figure 2.3 that the more offspring terms the focus level contains, the greater the number of sets that must be created to close the graph. Closing the
Fig. 2.3. Closures of the GO graphs from Figure 2.2 where the filled nodes represent the different choices of the focus level.

The graph can quickly become computationally infeasible in practice. Importantly, performing the full top-down approach as in panel (c) of Figure 2.3 is rarely possible in real applications due to the computational burden.

To partially amend the computational difficulties of the Focus Level method, Goeman and Mansmann (2008) implement a more efficient method of computing the closed graph using what they term “atom sets.” These atom sets form a core collection of gene sets which form a basis for all gene sets in the graph. All other gene sets in the graph (as well as its closure) can be created through unions of the atom sets. This ensures the size of the closed graph is $2^k - 1$, where $k$ is the number of atom sets, which is often smaller (and never larger) than the size of the original closed graph. Further, Goeman and Mansmann (2008) recommend selecting the focus level so that no more than 9-12 atom sets are required to recreate the offspring of any single focus level term. They also suggest computing only the smallest few adjusted $P$-values to save computation time in place of computing all adjusted $P$-values.

This work offers an alternative solution to improve on the computational speed of the top-down portion of the Focus Level method through an application of the general graphical Bonferroni adjustment of Bretz et al. (2009). This allows for a short-cut of length $m$ in place of the currently applied full closed testing approach of Marcus et al. (1976). In the following section (an abbreviated version of Section 1.3.6), we summarize the general
graphical Bonferroni adjustment approach and show how we tailor the method for a powerful application to the top-down portion of the Focus Level method.

2.1.2 The Graphical Bonferroni Adjustment

Bretz et al. (2009) proposed a powerful and versatile graphical generalization of weighted Bonferroni based closed testing (Marcus et al. 1976) which provides strong control of the family-wise error rate (FWER) at a specified level $\alpha$. Their approach represents all $m$ hypotheses of interest, $H_1, \ldots, H_m$ as nodes in a directed graph. Each node can be thought of here as a gene set, with a corresponding hypothesis $H_i$ testing for differential expression. Node $i$, representing hypothesis $H_i$, is allocated a local threshold $\alpha_i$ for all $i = 1, \ldots, m$. Nodes are joined by edges with weights $g_{ij}$ dictating the proportion of the local threshold $\alpha_i$ that is allocated to all connected hypotheses (nodes) $H_j$ in the case that hypothesis $H_i$ is rejected. The structure of the graph as well as the size of the local thresholds $\alpha_i$ and edge weights $g_{ij}$ is dependent on the objectives of the study. The versatility of the method is in the generality of the regularity conditions and updating algorithm for the directed graph.

The regularity conditions require the following (Bretz et al. 2009):

R1. The local thresholds $\alpha_1, \ldots, \alpha_m$ satisfy $\sum_{i=1}^{m} \alpha_i \leq \alpha$.

R2. The edge weights satisfy $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$, and $\sum_{k=1}^{m} g_{ik} \leq 1$ for all $i, j = 1, \ldots, m$.

The updating algorithm defines a sequentially rejective test procedure and is given as follows (Bretz et al. 2009). Note that $p_i$ represents the observed $P$-value for the test of hypothesis $H_i$.

Algorithm 1

0. Set $I = \{1, \ldots, m\}$.

1. Let $j = \text{argmin}_{i \in I} p_i / \alpha_i$.

2. If $p_j \leq \alpha_j$, reject $H_j$; otherwise stop.
3. Update the graph:

\[ I \rightarrow I \setminus \{j\} \]
\[
\alpha_l \rightarrow \begin{cases} 
\alpha_l + \alpha_j g_{jl}, & l \in I \\
0 & \text{otherwise}
\end{cases}
\]
\[
g_{lk} \rightarrow \begin{cases} 
\frac{g_{lk} + g_{lj} g_{jk}}{1 - g_{lj} g_{jl}}, & l, k \in I, l \neq k \\
0 & \text{otherwise}
\end{cases}
\]

4. If \(|I| \geq 1\), go to step 1; otherwise stop.

The proof that Algorithm 1 defines a sequentially rejective closed testing procedure which strongly controls the FWER at level \(\alpha\) is found in the Appendix of Bretz et al. (2009), and depends directly on Theorem 1 from Hommel et al. (2007). Both Brannath and Bretz (2010) and Goeman and Solari (2010) claim that Theorem 1 from Hommel et al. (2007) cannot be directly applied to the hypotheses of the GO graph as the hypotheses are nested, creating logical restrictions. In their own words, Brannath and Bretz (2010) claim that “the shortcut procedure of Hommel et al. (2007) cannot be applied to restricted hypotheses.” Similarly, Goeman and Solari (2010) state, “these methods [Bretz et al. (2009)] cannot make use of logical relationships between hypotheses and, as such, do not incorporate graph-based methods which exploit such relationships, such as [the Focus Level procedure] of Goeman and Mansmann (2008).” However, Section 2.1.3 below presents a restricted hypotheses example where the methods of Bretz et al. (2009) can be applied. Section 2.1.4 sets forward some important notation and vocabulary and then demonstrates that while these claims are technically true, the methods of Bretz et al. (2009) can be applied to the Focus Level method if one of the assumptions underlying Theorem 1 of Hommel et al. (2007) is slightly relaxed. We prove this with Theorem 1 in Section 2.2.

2.1.3 Restricted Hypotheses Example

Let \(H_1, \ldots, H_m\) denote \(m\) hypotheses of interest and call these the elementary hypotheses. Let \(I\) denote a non-empty index set such that \(I \subseteq \{1, \ldots, m\}\) and denote an intersection
hypothesis by $H_I$ where $H_I = \cap_{k \in I} H_i$. The closed test procedure (Marcus et al. 1976) utilizes the intersection closed set of hypotheses $\mathcal{H} = \{H_I : I \subseteq \{1, \ldots, m\}, I \neq \emptyset\}$. In the case that the hypotheses are unrestricted, $|\mathcal{H}| = 2^m - 1$ and Algorithm 1 of Bretz et al. (2009) is proven to hold. On the other hand, the hypotheses are restricted if for index sets $I$ and $J$ it is true that $I \neq J$ and $H_I = H_J$ so that $|\mathcal{H}| < 2^m - 1$. In this case, Algorithm 1 cannot currently be applied (Brannath and Bretz 2010; Goeman and Solari 2010).

As the hypotheses corresponding to any GO graph are always restricted, the methods of Bretz et al. (2009) cannot be applied to the GO graph under the current framework. However, the following closed test example from Brannath and Bretz (2010) can be extended to demonstrate how Algorithm 1 can be applied to the case of restricted hypotheses. This example sets the stage for Section 2.2, where we relax the assumptions of Theorem 1 of Hommel et al. (2007) to formally establish how the methods of Bretz et al. (2009) can indeed be applied to restricted hypotheses, and hence, the GO graph.

Consider the partially nested elementary hypotheses $H_1$, $H_2$, $H_3$, and $H_4$ defined as follows for the parameters $\theta_1$ and $\theta_2$ where $\delta_1, \delta_2 > 0$.

\[ H_1 : \theta_1 \leq -\delta_1, \quad H_2 : \theta_1 \leq 0, \quad H_3 : \theta_2 \leq -\delta_2, \quad H_4 : \theta_2 \leq 0 \tag{2.1} \]

The full closure family of hypotheses $\mathcal{H}$ of these four elementary hypotheses would contain $2^4 - 1 = 15$ distinct intersection hypotheses if they were unrestricted. However, the restrictions stemming from the partial nesting of $H_1$ with $H_2$ ($H_1 \subset H_2$) and $H_3$ with $H_4$ ($H_3 \subset H_4$) reduce the final closure to just eight distinct intersection hypotheses. For example, $H_{12} = H_1 \cap H_2 = H_1$ and $H_{34} = H_3 \cap H_4 = H_3$. Computing all intersections and retaining only the distinct intersection hypotheses shows

\[ \mathcal{H} = \{H_1, H_2, H_3, H_4, H_{13}, H_{14}, H_{23}, H_{24}\}. \tag{2.2} \]

Each of the null parameter spaces corresponding to the hypotheses in $\mathcal{H}$ are graphically depicted in panel (a) of Figure 2.4.
Fig. 2.4. (a) Graphical demonstration of the elementary hypotheses $H_1, \ldots, H_4$ and distinct intersection hypotheses. The null parameter space is shaded in gray for each hypothesis. Redundant intersection hypotheses are written in parentheses. (b) The closed test approach given the structure of the hypotheses. Testing begins with $H_{13}$, the full intersection hypothesis, and terminates at or before testing $H_2$ and $H_4$.

Brannath and Bretz (2010) apply a closed test approach to $\mathcal{H}$ beginning with the raw $p$-values $p_1, p_2, p_3,$ and $p_4$ obtained from testing the original elementary hypotheses $H_1, H_2, H_3,$ and $H_4,$ each with $\alpha$-level tests, respectively. To define the closed test approach, they compute the closed test $p$-values $p_{H_i}$ for each hypotheses $H_i$ in $\mathcal{H}$ by the following rules. First, $p_{H_1} = p_1$ and $p_{H_3} = p_3$. Second, $p_{H_2} = \max\{p_1, p_2\}$ and $p_{H_4} = \max\{p_3, p_4\}$. Finally, $p_{H_{ij}} = \min\{1, 2p_{H_i}, 2p_{H_j}\}, i = 1, 2$ and $j = 3, 4$. The closed test procedure (Marcus et al. 1976) is then applied to $\mathcal{H}$ as depicted in panel (b) of Figure 2.4 using the closed test $p$-values $p_{H_i}$ as explained in the following paragraph.

The closed test procedure only tests a hypothesis $H_i \in \mathcal{H}$ if all hypotheses implying $H_i$ are first rejected. For example, $H_1$ can only be tested by the closed test procedure if $H_{13}$ and $H_{14}$ are first rejected, see panel (b) of Figure 2.4. In other words, the hypothesis corresponding to a child node is only tested if its parent node hypothesis is first rejected. Brannath and Bretz (2010) state that, “this closed test procedure controls the family-wise error rate strongly at level $\alpha$ and reflects the logical constraints among the elementary
Fig. 2.5. Graphical Bonferroni adjustment approach for the partially nested elementary hypotheses $H_1, \ldots, H_4$ which performs the closed test described in Brannath and Bretz (2010) when Algorithm 1 is applied to the graph.

hypotheses." We show that this closed test approach for these restricted hypotheses can be performed using the directed graph of Figure 2.5 and Algorithm 1 from Bretz et al. (2009).

Consider the sequential rejection procedure resulting from the application of Algorithm 1 (Bretz et al. 2009) to the directed graph shown in Figure 2.5. Initial local thresholds of $\alpha/2$ are assigned to $H_1$ and $H_3$ and local thresholds of zero assigned to $H_2$ and $H_4$ as depicted in Figure 2.5. The weighted edges provide for the reallocation of the local thresholds in the case of rejection of either $H_1$ or $H_3$. If neither $H_1$ nor $H_3$ can be rejected at the $\alpha/2$-level, then the testing is stopped with no rejections. This corresponds to the first step of the closed test procedure described previously, as proposed in Brannath and Bretz (2010). As can be seen in panel (b) of Figure 2.4, the closed test requires the rejection of the intersection hypothesis $H_{13}$ before any other rejection can occur. This requires that the previously defined closed test $p$-value $p_{H_{13}} = \min\{2p_{H_1}, 2p_{H_3}\}$ satisfy $p_{H_{13}} < \alpha$. Since $p_{H_1}$ and $p_{H_3}$ were defined to be $p_1$ and $p_3$ respectively for this particular example, it follows that $p_{H_{13}} < \alpha$ implies $2\min\{p_1, p_3\} < \alpha$, witnessing that the methods agree on their starting analysis using only the values of $p_1$ and $p_3$. The flow chart in Figure 2.6 further demonstrates that the two approaches agree for all possible test scenarios and hence, that the shortcut of Bretz et al. (2009) can successfully be applied to this example of restricted hypotheses.

2.1.4 Definitions and Preliminaries to Theorem 1

A deeper inspection of Figure 2.6 will reveal the reason why the shortcut from Bretz
Fig. 2.6. Flow chart demonstration of the equivalence of the graphical shortcut tailored from the methods of Bretz et al. (2009) to that of the full closed test procedure proposed in Brannath and Bretz (2010) within the context of the restricted hypotheses example of Section 2.1.3. At each step in the chart, the left graph represents the full closed test approach, while the right graph depicts the graphical shortcut.
et al. (2009) can be applied to the example of restricted hypotheses of the previous section. To explain how, we must first define two terms, *consonance* and *natural consonance*.

The traditional definition of *consonance* (Gabriel 1969) relies on the idea of maximal hypotheses. It states that consonance is the property of certain closed tests where rejection of an intersection hypothesis $H_1 \in \mathcal{H}$ implies rejection of a maximal hypothesis $H \in \mathcal{H}$. Here, a maximal hypothesis $H \in \mathcal{H}$ is such that there is no $H' \in \mathcal{H}$ with $H' \supset H$. (When the closed test corresponding to the hypotheses in $\mathcal{H}$ is depicted graphically, as in panel (b) of Figure 2.4, it can be seen that maximal hypotheses correspond to the leaf nodes of the graph. Further, in context of the GO graph, maximal hypotheses correspond to the leaf nodes of the graph, while the minimal hypothesis corresponds to the root node of the graph.) From the example of the previous section, it can be seen that only $H_2$ and $H_4$ are maximal. Thus, the closed test of the example is not consonant in the traditional sense as rejection of the intersection hypothesis $H_{13}$ does not imply the rejection of either of the maximal hypotheses $H_2$ or $H_4$.

*Natural consonance* is a similar, but slightly more relaxed property than *consonance*, and differs in that it implies the rejection of only an elementary hypothesis (not necessarily a leaf node in the closure graph) whenever any other hypothesis $H_1 \in \mathcal{H}$ is first rejected. This relaxed definition is more recent and is due to Brannath and Bretz (2010). Importantly, it is easier for a closed test to satisfy the property of *natural consonance* than that of *consonance*. The claims of both Brannath and Bretz (2010) and Goeman and Solari (2010) that Algorithm 1 (Bretz et al. 2009) is not applicable to restricted hypotheses rest on the subtle difficulty of how consonance is defined. Note (v) following Theorem 2 of Brannath and Bretz (2010) claims that “consonance with respect to the elementary hypotheses [natural consonance] always implies the existence of a nested shortcut of size $m$,” where $m$ is the number of elementary hypotheses. The *natural consonance* of the closed test allows for the shortcut from Bretz et al. (2009) to be applied to the restricted hypothesis example of the previous section, as explained in the following paragraph.
Examining the flow chart of Figure 2.6 will reveal that the closed test procedure proposed by Brannath and Bretz (2010) has this property of consonance with respect to the elementary hypotheses \( H_1, H_2, H_3, \) and \( H_4, \) i.e., the closed test for this example is naturally consonant. This follows from the fact that rejection of the intersection hypothesis \( H_{13} \) implies rejection of either of the hypotheses \( H_1 \) or \( H_3 \) which are two of the original four elementary hypotheses. Note as before that rejection of \( H_{13} \) requires that either \( 2p_1 < \alpha \) or \( 2p_3 < \alpha \) by the definition of \( p_{H_{13}}. \) If say \( 2p_1 < \alpha, \) then \( H_{13} \) is rejected. Further, since \( 2p_1 < \alpha, \) \( H_{14} \) is also rejected as \( p_{H_{14}} = \min\{1, 2p_{H_1}, 2p_{H_4}\} = \min\{2p_1, 2p_{H_4}\} < \alpha. \) Most importantly, \( 2p_1 < \alpha \) provides for \( H_1 \) to be rejected, as the closed test \( p \)-value \( p_{H_1} \) requires only \( p_1 < \alpha \) which is certainly satisfied if \( 2p_1 < \alpha. \) Hence, in this case, the rejection of the intersection hypothesis \( H_{13} \) implied rejection of the elementary hypothesis \( H_1. \) A similar scenario holds for the elementary hypothesis \( H_3 \) if \( 2p_3 < \alpha \) instead of (or as well as) \( 2p_1 < \alpha. \) Finally, rejection of \( H_{24} \) similarly implies rejection of either \( H_2 \) or \( H_4. \) Thus, the closed test procedure for these restricted hypotheses admits the shortcut of Bretz et al. (2009) because of the consonance of the closed test with respect to the elementary hypotheses, i.e. the closed test is naturally consonant.

2.2 Shortcuts for Restricted Hypotheses

We now extend Theorem 1 of Hommel et al. (2007) to restricted hypotheses, and thereby verify the appropriateness of the graphical shortcut of Bretz et al. (2009) for restricted hypotheses. To this end, let \( m \) elementary hypotheses \( H_1, \ldots, H_m \) of interest be given and denote by \( \mathcal{H} \) their closure under intersection. For the purposes of Theorem 1, \( \mathcal{H} \) can be either restricted or unrestricted. Let \( \alpha_i(I) \) denote the local significance levels for an intersection hypothesis \( H_I \in \mathcal{H} \) where \( \sum_{i \in I} \alpha_i \leq \alpha \) for all non-empty \( I \subseteq \{1, \ldots, m\}. \)

2.2.1 Theorem 1

(Extension of Theorem 1 from Hommel et al. (2007) to restricted hypotheses.) If for \( \emptyset \neq I, J \subseteq \{1, \ldots, m\} \) with \( \emptyset \neq H_I \subset H_J \) it holds that \( \alpha_i(I) \leq \alpha_i(J), \) then the closed test for \( \mathcal{H} \) based on local Bonferroni tests is naturally consonant and a shortcut equivalent to
the following procedure is possible (adapted from Bretz et al. (2009)).

0. Set $M = \{1, \ldots, m\}$.

1. Set $I$ equal to the smallest subset of $M$ such that $H_I = H_M$.

2. Reject $H_j$ if there exists $j \in I$ such that $p_j \leq \alpha_j(I)$. If no such $j$ exists, then stop.

3. Set $M \rightarrow M \setminus j$.

4. If $|M| \geq 1$ return to Step 1. Otherwise, stop.

Proof. First, note that in the case of unrestricted hypotheses, natural consonance and consonance are identical (Brannath and Bretz 2010) so that the proof is already demonstrated in Theorem 1 of Hommel et al. (2007). Consider then the case of restricted hypotheses in the sense that for $\emptyset \neq I, J \subseteq \{1, \ldots, m\}$ with $I \neq J$ it is true that $\emptyset \neq H_I = H_J$ so that $|\mathcal{H}| < 2^m - 1$. Then, for $I, J$ with $\emptyset \neq H_I \subset H_J$ it follows from $\alpha_i(I) \leq \alpha_i(J)$ that $p_j \leq \alpha_j(I)$ implies $p_j \leq \alpha_j(J)$. Thus, rejection of $H_I$ implies rejection of some elementary hypothesis $H_j$, witnessing that the closed test for $\mathcal{H}$ is indeed naturally consonant.

### 2.2.2 Discussion of Theorem 1

Some comments are in order regarding Theorem 1 in Section 2.2.1. First, while an intersection hypothesis $H_I$ may not be unique in $\mathcal{H}$, it must not be empty for the nested shortcut of length $m$ to exist. Second, the only difference between the proof here and the proof for unrestricted hypotheses (Hommel et al. 2007) is in the definition of consonance. Here we follow the suggestion in Brannath and Bretz (2010) and allow natural consonance, which can be seen as a loosening of the requirements of consonance to include all elementary hypotheses instead of just all maximal hypotheses. The important distinction is that for unrestricted hypotheses, all elementary hypotheses are maximal. The same is not necessarily true for restricted hypotheses. Third, as in the example of Section 2.1.3, restricted hypotheses are often the result of nested elementary hypotheses. This is certainly the case for the hypotheses attached to the gene sets of the GO graphs. Fourth, the main importance of
the extended Theorem 1 in Section 2.2.1 rests with its assurance that a naturally consonant closed test based on weighted Bonferroni tests exists so long as the monotonicity condition \( \alpha_i(I) \leq \alpha_i(J) \) is satisfied for all \( \emptyset \neq H_i \subset H_j \) in \( \mathcal{H} \). Fifth, Theorem 1 does not specify that any graph with local thresholds of \( \alpha = (\alpha_1, \ldots, \alpha_m) \) and edge weights \( G = \{g\}_{ij} \), denoted by \((\alpha, G)\), can combine with Algorithm 1 and lead to a consonant closed test. It simply specifies the conditions under which a consonant closed test based on local Bonferroni tests can be formed.

One important rule on the graph \((\alpha, G)\) when the hypotheses are restricted is that the local threshold \( \alpha_i \) for an elementary hypothesis \( H_i \) must remain zero until all elementary hypotheses \( H_j \) with \( H_j \subset H_i \) are first rejected. This property can be seen to hold for the graph of Figure 2.5. However, if the graph in Figure 2.5 allowed for any of \( H_1 \)'s threshold to be passed to \( H_4 \) or similarly, if \( H_3 \) allowed for anything to be passed to \( H_2 \), this property would no longer hold. So, while Theorem 1 assures that a consonant closed test exists when local Bonferroni tests are used for the testing of each \( H \in \mathcal{H} \), not just any graph \((\alpha, G)\) will result in that consonant closed test. In the following section we demonstrate how a graph \((\alpha, G)\) can be applied to the GO graph such that a consonant closed test based on weighted Bonferroni tests is achieved through the application of Algorithm 1.

That Algorithm 1, when applied to a graph \((\alpha, G)\), preserves the monotonic property that \( \alpha_j(I) \leq \alpha_j(J) \) for \( I, J \) such that \( H_I \subset H_J \) can be seen by noting that Algorithm 1 only provides for the local thresholds \( \alpha_i \) to remain the same size or increase. Never does it allow for them to become smaller. Further, at any point in the iterative process, the local thresholds \( \alpha_i \) define the weighted Bonferroni test thresholds \( \alpha_j(I) \) for the intersection hypothesis \( I \) corresponding to the intersection of the elementary hypotheses with non-zero thresholds (see for example Figure 2.6). Hence, as \( H_J \) will be tested only after \( H_I \) is first rejected whenever \( H_I \subset H_J \), it follows that Algorithm 1 will provide \( \alpha_j(I) \leq \alpha_j(J) \).

2.3 The Short Focus Level Procedure

We obtain the Short Focus Level procedure by modifying the top-down portion of the Focus Level method. This is done by tailoring the general graphical shortcut (Bretz et al.
2009) to a GO graph as follows. Label the $m$ hypotheses corresponding to the test of significance for each GO term (gene set) as $H_1, \ldots, H_m$ starting with the root node and proceeding in an organized manner through each level of the GO graph, ending with the terminal nodes. (The precise ordering is not important.) Let $F \subset M = \{1, \ldots, m\}$ denote the index set of the nodes corresponding to the pre-selected focus level of the GO graph. For all $m_F$ nodes in the focus level, assign local significance levels of $\alpha_i = \alpha/m_F$ to each hypothesis $H_i$ with $i \in F$. Assign initial local significance levels of 0 to all children nodes of the focus level. Note that nodes above the focus level will still be tested using the bottom-up approach of the Focus Level method and are not considered when applying the top-down portion of the method.

Using the structure of the GO graph, assign to each edge from parent node $i$ to child node $j$ a weight of $g_{ij} = 1/m_i$, where $m_i$ denotes the number of children nodes of node $i$. After all edge weights have been assigned for the edges defined by the GO graph, all terminal nodes are individually joined with $m_F$ new edges to each of the $m_F$ focus level nodes. These new edges are given weights of $1/m_F$. (In the case that a terminal node is also a focus level node, then edges are made only to all other focus level nodes with weight $1/(m_F - 1)$.)

At this point, a modified form of Algorithm 1 of Bretz et al. (2009) is applied to the resulting directed graph to obtain the final set of significant hypotheses. The modifications ensure that no child node is tested before all parent nodes are first found significant, maintaining the strong control of the FWER under the restricted hypotheses of the GO graph as well as maintaining Property FL2 of the basic assumptions underlying the Focus Level method (Section 2.1.1). Figure 2.7 demonstrates the application of the described graphical Bonferroni adjustment to the top-down portion of the Example GO graphs of Figure 2.2. Comparing Figure 2.3 to Figure 2.7 provides a heuristic understanding of how the new top-down approach is computationally faster than the original closure approach because no new nodes need to be created.
Fig. 2.7. The suggested shortcut to the top-down portion of the Focus Level method exploits the natural consonance of the weighted Bonferroni tests applied to the GO graph to avoid closing the graph under all unions as in the original top-down approach.

An algorithm which implements the Short Focus Level procedure is detailed in Table 2.1. Here, $H$ denotes the index set of testable hypotheses (nodes) and $w = \{w_i\}_{i \in H}$ the corresponding set of weights such that $\alpha/w_i$ provides the local thresholds $\alpha_i$ for each hypothesis $H_i$ indexed by $i \in H$. As described previously, $F \subset \{1, \ldots, m\}$ denotes the index set of all pre-selected focus level nodes. The notation $C_i$ denotes the index set of children nodes of the parent hypothesis $H_i$. Similarly, the notations $P_i$ and $A_i$ denote the parents and all ancestors, respectively, of the node corresponding to the hypothesis $H_i$. Finally, we use $R$ and $S$ to denote the index sets of the current and cumulative rejected hypotheses, respectively.

2.3.1 Power Analysis

A natural question at this point concerns the advantages and disadvantages of changing the top-down portion of the Focus Level procedure from the original closed test approach as in Goeman and Mansmann (2008) to the graphical shortcut of Bretz et al. (2009) as proposed for the Short Focus Level. If the local tests for each intersection hypothesis were originally performed with weighted Bonferroni tests, then the difference between the methods would be that the first performed the full closure test requiring the testing of somewhere on the order of $2^m - 1$ intersection hypotheses, while the second, which applies a shortcut, would test no more than $m$ hypotheses with no reduction in the power of the tests. When using the Global Test for each intersection hypothesis as suggested by Goeman and Mansmann (2008), the answer to the differences in computation time and power is not as clear.
Table 2.1. Algorithm detailing the newly proposed Short Focus Level procedure.

0. Set $H = F$ and $w_i = m_F$ for each $i \in H$.
1. Add $i \in H$ to $R$ if $p_i < \alpha / w_i$.

If $R \neq \emptyset$, perform Steps 2-4. Otherwise, STOP.

2. Update $H$ and $w$:
   i. Set $H = H \setminus R$.
   ii. for $i \in R$,
      add $C_i$ to $H$ and set $w_j = w_j + w_i \cdot |C_i|$ for all $j \in C_i$.
   iii. for all $i \in H$ with $(P_i \cap H)$ non-empty,
       remove $i$ from $H$ and add $w_i \cdot |P_i \cap H|$ to $w_j$ for all $j \in (P_i \cap H)$.
3. Add $R$ and $\bigcup_{i \in R} A_i$ to $S$.
4. Set $R = \emptyset$, return to Step 1.

The final set of rejected hypotheses will be contained in $S$.

The following simulations demonstrate that neither method is uniformly more powerful than the other, with each having the advantage for certain scenarios. However, as these simulations demonstrate, the newly proposed Short Focus Level procedure is uniformly (and exponentially) computationally faster than the Focus Level method which will hopefully better enable its use by practitioners.

Simulation 1

The following simulation based on the toy GO graph depicted in Figure 2.8 panel (b) demonstrates the advantages and disadvantages of moving to the newly proposed graphical shortcut of Bretz et al. (2009) in the top-down portion of the Focus Level procedure. The simulation was performed with the phenotype $Y$ as a dichotomous class variable (say, treatment and control) and the data $X$ representing an RNA-Seq counts matrix with rows as genes ($m$) and columns as samples ($n$). The number of samples belonging to the treatment group was simulated according to a binomial($n$, 0.5) distribution, where $n$ is the total number of samples, with the added rule that at least two samples were in each group. This
Fig. 2.8. (a) The full closure of the example toy GO graph depicted in panel (b) that is currently utilized by the Focus Level method. (c) The graph \((\alpha, G)\) corresponding to the example toy GO graph depicted in panel (b) that is utilized by the proposed Short Focus Level procedure.

allowed for unbalanced data, with the tendency towards fairly balanced designs. Separate simulations for sample sizes of \(n = 5, 20,\) and 100 were performed.

The structure of gene assignments to the sets A, B, C, D, E, and F of Figure 2.5, as well as the total number of genes assigned, was allowed to vary in each simulation according to certain parameters. Genes were first assigned to the leaf node gene sets C, D, and E. This was accomplished by randomly selecting both the number of distinct sets in each of these sets (anywhere from 1 to a maximum specified size of either 10 or 40) as well as the number of genes shared by all possible combinations of the leaf node gene sets. Common genes between all or many gene sets was discouraged with small probabilities of occurrence, while common genes between a few gene sets was allowed to occur more frequently. Following the assignments of genes to leaf nodes, parent nodes were randomly assigned new genes (anywhere from 1 to the maximum specified size) as well as all genes contained by their children nodes. The result was a nested graph with at least some overlap common to many gene sets, as is the case within GO Graphs.

The data counts matrix \(X\) was simulated using an actual RNA-Seq data set as a sampling distribution for the per-gene means in the control group. Specifically, the counts \(k_{ij}\) for all samples \(j\) assigned to the control group were generated from a \(\text{NB}(\mu_i, \mu_i + \mu_i^2/d)\)
Table 2.2. Summary of results for Simulation 1. Power calculations were averaged over all levels of the effect size $\lambda$ and both sizes of $m$, the maximum leaf node gene set size, for each level of the sample size $n$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Method</th>
<th>Node</th>
<th>Mean Computation Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$A$</td>
<td>$B$</td>
</tr>
<tr>
<td>5</td>
<td>FL</td>
<td>0.447</td>
<td>0.428</td>
</tr>
<tr>
<td></td>
<td>SFL</td>
<td>0.447</td>
<td>0.366</td>
</tr>
<tr>
<td>20</td>
<td>FL</td>
<td>0.574</td>
<td>0.567</td>
</tr>
<tr>
<td></td>
<td>SFL</td>
<td>0.574</td>
<td>0.552</td>
</tr>
<tr>
<td>100</td>
<td>FL</td>
<td>0.642</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td>SFL</td>
<td>0.642</td>
<td>0.623</td>
</tr>
</tbody>
</table>

FL: Focus Level
SFL: Short Focus Level

distribution, where the means $\mu_i$ were randomly sampled from the per-gene means of the control group from the actual RNA-Seq data set. The scaling parameter $d$ was set at 10 for all simulations. Leaf node gene sets (any of nodes $C$, $D$, or $E$ in Figure 2.5) were then selected at random to be significant. Each gene mapping to the selected significant leaf nodes was assigned a treatment mean of $\tilde{\mu}_i = 2^\beta \mu_i$ where $\mu_i$ denotes the mean sampled from the actual RNA-Seq data for gene $i$ and $\beta_i$ was an effect size obtained from a Poisson($\lambda$) distribution with the parameter $\lambda$ set to one of 0, 1, 2, or 3. Thus, not all genes in the significant gene sets necessarily had non-zero effect sizes. The actual counts $k_{ij}$ for all samples $j$ assigned to the treatment group were obtained from a $\text{NB}(\tilde{\mu}_i + \frac{\mu_i^2}{d})$ distribution where, as with the control group, $d = 10$ was constant across all simulations. (See Fridley et al. (2010) for a similar simulation approach where single gene sets were the object of interest as opposed to an entire GO graphs as in this simulation.)

The averaged results of Simulation 1 are presented in Table 2.2. This example shows greater power for the current implementation of the Focus Level procedure where the Global-test (Goeman et al. 2004) is used to test all intersection hypotheses and all elementary hypotheses. The greatest power differences of the two methods appear for small sample sizes, $n = 5$ in this simulation, and for nodes with relatively few child nodes. The power of the two methods is comparable otherwise. Importantly, the computation time for the
Table 2.3. Allocation of simulated genes to the GO IDs of the GO graph in Figure 2.9.

<table>
<thead>
<tr>
<th>GO ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>1-100</td>
<td>1-40</td>
<td>21-60</td>
<td>61-100</td>
<td>1-10</td>
<td>11:20</td>
<td>21:40</td>
</tr>
<tr>
<td>GO ID</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Genes</td>
<td>41-50</td>
<td>51-60</td>
<td>61-80</td>
<td>71-90</td>
<td>81-100</td>
<td>72-79</td>
<td>82-89</td>
</tr>
</tbody>
</table>

Short Focus Level procedure is significantly faster, even for this extremely small toy GO graph whose closure contains just 14 nodes. Interestingly, the Focus Level procedure as it is currently implemented seems to operate best, computationally speaking, when the sample size is moderate, \( n = 20 \) in this simulation.

Simulation 2

A second simulation study using the toy GO graph of Figure 2.9 was also used to compare power and computation time of the original Focus Level method to the Short Focus Level. The closure of the toy GO graph in Figure 2.9 is more complex than that of the previous simulation, containing 574 nodes as compared to the 14 of Figure 2.8, panel (a). This simulation considered the continuous phenotype \( Y \sim N(0, 1) \) and its correlation with simulated gene expression values \( X \). For this simulation \( m = 100 \) genes were partitioned to the 14 GO IDs of Figure 2.9 as specified in Table 2.3. Expression values \( X_{ij} \) for each sample \( i = 1, \ldots, n \) and gene \( j = 1, \ldots, m \) were generated as \( N(0,1) \) variates. GO IDs 6, 7, and 13 were designated as significant by adding \( rY, r \in [0,1] \), to the expression values of the corresponding genes (i.e., the columns of \( X \) corresponding to genes in GO IDs 6, 7, and 13). Thus, by inheritance, GO IDs 1, 2, 3, 4, 10, and 11 were also significantly associated with the phenotype \( Y \). Values of \( r \) close to 1 provided a strong signal and greater power for detection while \( r \) near zero resulted in a very weak signal and correspondingly very low power for detection. Goeman’s Globaltest (Goeman et al. 2004) was used to test each GO ID for association with the phenotypic variable \( Y \). Given that Simulation 1 suggested that the current Focus Level procedure performs best at a moderate sample size, \( n = 20 \) was used for this simulation.
Fig. 2.9. Structure of the toy GO graph used in Simulation 2. Shaded nodes correspond to those GO IDs which were simulated to be significantly associated with the phenotype Y.

Table 2.4. Results of the power simulation for the GO Graph in Figure 2.9.

<table>
<thead>
<tr>
<th>GO:01</th>
<th>GO:02</th>
<th>GO:03</th>
<th>GO:04</th>
<th>GO:06</th>
<th>GO:07</th>
<th>GO:10</th>
<th>GO:11</th>
<th>GO:13</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>0.995</td>
<td>0.968</td>
<td>0.890</td>
<td>0.462</td>
<td>0.512</td>
<td>0.872</td>
<td>0.380</td>
<td>0.399</td>
<td>0.344</td>
</tr>
<tr>
<td>SFL</td>
<td>0.995</td>
<td>0.988</td>
<td>0.952</td>
<td>0.543</td>
<td>0.837</td>
<td>0.949</td>
<td>0.489</td>
<td>0.476</td>
<td>0.445</td>
</tr>
</tbody>
</table>

FL: Focus Level
SFL: Short Focus Level

Power and computation time were averaged over 1,000 simulations. Results are presented in Table 2.4 for the most interesting case of $r = 0.5$. They show the Short Focus Level method having greater power at every GO ID. The computational speed advantage of the improvement is also manifest, showing nearly a 15,000 fold increase in speed over the current Focus Level procedure. This second simulation emphasizes the fact that neither approach to the Focus Level procedure is uniformly more powerful than the other. While it is clear that each has the advantage in certain scenarios, at least theoretically, more work needs to be completed to determine exactly where each is most appropriate. Practically speaking however, the computational advantage and similar statistical power (on average) of the Short Focus Level should solicit its use except perhaps for choices of the focus level near the leaf nodes of the graph where the current Focus Level method is computationally tractable.
2.3.2 Real Data Application

A drawback to the otherwise powerful Focus Level method is the computational burden which prohibits the full top-down approach from being applied to real data sets (Goeman and Mansmann 2008). When no *a priori* focus level exists, as is often the case (Liang and Nettleton 2010), the root node of the GO graph is a logical default choice, but requires the full top-down approach. Under the newly proposed Short Focus Level method, this is now a computational possibility. The following application to RNA-Seq counts data from porcine oocytes demonstrates the performance of the full top-down approach of the Short Focus Level procedure to real data. The Biological Process (BP) root node was selected as the focus level for this study due to there being no other focus level of greater *a priori* interest.

It is well known that *in vivo* (naturally) matured oocytes show far greater developmental competence than do those matured *in vitro* (Cox et al. 2013). Yet, the underlying genetics are still not well understood. To uncover the genetic differences of *in vitro* matured oocytes as compared to those matured naturally (*in vivo*), transcript counts for 4 *in vivo* and 4 *in vitro* matured porcine oocytes were obtained using the Illumina RNA-Seq platform (Wang et al. 2009). Lanes were populated as shown in Table 2.5. These data from the lab of Dr. Clay Isom of the Utah State University Department of Animal, Dairy, and Veterinary Sciences are reported on here with permission.

Individual *P*-values testing the differential expression of 12,625 genes were calculated using the DESeq package of Bioconductor (Anders and Huber 2010; Gentleman et al. 2004) with pig mother, as identified in Table 2.5, included as a covariate. Specifically, these *P*-values were obtained under the null hypotheses that the per-gene expression strength of the *in vivo* matured oocytes (IVV) is equal to that of the *in vitro* matured oocytes (IVV) when accounting for any pig mother effect. This was done through the DESeq package (Anders and Huber 2010) which compares a full model (regressing the RNA-Seq counts on both the oocyte type and pig mother by a generalized linear model) to a reduced model (regressing only on the pig mother) to determine significance for a given gene.
Table 2.5. Experimental design for the *in vivo* (IVV) and *in vitro* (IMV) Oocyte Maturation RNA-Seq data. *Lane 3 contained quality problems and was removed from the analysis.

<table>
<thead>
<tr>
<th>Oocyte No.</th>
<th>Embryo Type</th>
<th>Pig (Mother)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IVV</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>IVV</td>
<td>2</td>
</tr>
<tr>
<td>3*</td>
<td>IVV</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>IVM</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>IVM</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>IVM</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>IVM</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>IVV</td>
<td>3</td>
</tr>
</tbody>
</table>

A gene set analysis using the GO BP ontology was then performed to characterize differentially expressed gene products between the two types of oocytes (IVV and IMV). P-values for each of 5,687 BP GO Terms containing at least 5 of the 12,625 Entrez IDs from the single gene (DESeq) analysis were calculated using Stouffer’s Method (Stouffer *et al.* 1949). Stouffer’s method transforms each of the P-values (from the single gene analysis) corresponding to an individual gene in the gene set to a standard normal Z-score. A single P-value for the gene set is then obtained from the mean of the Z-scores by computing the appropriate tail probability (from a standard normal distribution) beyond the mean Z-score. Stouffer’s P-value combination method was applied here as it is more powerful for the consensus alternative than say Fisher’s P-value combination test (Fisher 1973) or Goeman’s global test (Goeman *et al.* 2004), see discussions in Stevens and Isom (2012).

Finally, multiplicity adjusted gene set P-values for each BP term were calculated using the Short Focus Level procedure, with the root BP GO term (GO:0008510) as the focus level. This adjustment (the full top-down approach) took just 3 minutes and 23 seconds of processing time on an Intel Pentium M 1.86GHz processor with 1 GB of RAM. The current Focus Level method is computationally intractable for these data.

Figure 2.10 reports the significant subgraph (Goeman and Mansmann 2008) obtained from the Short Focus Level method containing 113 of the original 5,687 BP terms. Since the full top-down approach was performed, these GO terms, which are differentially expressed
Fig. 2.10. Significant results from the gene set testing of porcine oocytes obtained from the Short Focus Level procedure using the full top down approach.

between the two types of oocytes (IVV and IVM), can be discussed either individually or within their context of this significant subgraph. Advantaged by the FWER control of the Short Focus Level procedure, any subset of the significant results can also be reported on (while the others ignored) with the assurance that the FWER remains controlled at the specified level for the selected sets. Possible interpretation discussions of the results include the significant differential activity (between in vivo and in vitro maturated oocytes) of biological processes “response to bacterium” (node 74 in Figure 2.10), “double-strand break repair” (node 110), and “ribonucleoside metabolic process” (node 93), among others.
2.4 Discussion

As pointed out in Goeman and Mansmann (2008), the GO graphs are structured and “it is wasteful not to make use of that structure” in correcting for multiplicity. Further, they stress the importance of not making any assumptions on the joint distribution of the test statistics corresponding to each of the gene sets in the GO graph while correcting for multiplicity. The Focus Level procedure both avoids any such assumptions and capitalizes on the inherent structure of the GO graph to adjust for the multiple tests performed, resulting in a powerful approach. Another advantage of the Focus Level method is the possibility of incorporating biological knowledge into the adjustment approach through the selection of the focus level, where the method has the greatest power.

This work improves upon the Focus Level procedure of Goeman and Mansmann (2008) by altering the top-down portion of the method to utilize the graphical shortcut of Bretz et al. (2009) in place of the full closed testing approach of Marcus et al. (1976) as originally suggested by Goeman and Mansmann (2008). This was made possible by extending the result from Hommel et al. (2007) to restricted hypotheses (Theorem 1) as the hypotheses corresponding to the GO graph are always restricted.

The main advantage of the Short Focus Level procedure proposed in this work is the exponential decrease in computational burden. This provides for the most logical default choice of the root node of the GO graph as the focus level when no other a priori choice can be specified. Another advantage of the improvement is in the ability to consider the adjusted $P$-values apart from their context within the significant subgraph of the full GO graph under the full top-down approach. When the focus level is selected to be anything other than the root node, individual hypotheses must be considered in context of their position within the significant subgraph. However, this is not altogether a disadvantage as “the interpretation of an individual adjusted $P$-value should depend on the location in the graph where it occurs” (Goeman and Mansmann 2008).

It is our hope that this shortcut for the Focus Level procedure, the Short Focus Level, will result in more wide-spread use of the method. Still, future work remains to be done.
The simulations performed within this work demonstrate that each approach appears to be more powerful under different circumstances. Hence, further theoretical work is needed to determine the conditions under which each method is most powerful.
CHAPTER 3
A POWER IMPROVING MULTIPLICITY CORRECTION FOR LARGE-SCALE SNP SELECTION IN LD-based QTL MAPPING

3.1 Abstract

Controlling for the multiplicity effect is an essential part of determining statistical significance in single-marker genome-wide association scans (GWAS) on Single Nucleotide Polymorphisms (SNPs). Bonferroni adjustments are a commonly used approach due to their simplicity, but are conservative and have low power for large-scale tests. The permutation test, which is a powerful and popular tool, is computationally expensive and may provide misleading results in the presence of family structure. This work proposes a computationally efficient and powerful multiple testing correction for Linkage Disequilibrium (LD) based Quantitative Trait Loci (QTL) mapping on the basis of graphical weighted-Bonferroni methods which have been shown to synthesize weighted Bonferroni-based closed testing procedures into a powerful and versatile graphical approach. The proposed multiplicity adjustment, termed the Graphical Bonferroni Adjustment (GBA), capitalizes on the different priorities for the two hypothesis tests involved in LD-based QTL mapping to gain an increase in power and maintain computational efficiency and conceptual simplicity. The proposed GBA maintains strong control of the family-wise error rate (FWER). The performance of the GBA as compared to the standard Bonferroni correction was examined both theoretically and by simulation. We observe an increase in power for higher heritabilities and larger sample sizes. We also applied the proposed method to a real outbred mouse HDL cholesterol QTL mapping project where we detected more significant QTLs than were detected in the literature, while still ensuring strong control of the FWER.
3.2 Introduction

Linkage Disequilibrium (LD) analysis plays a fundamental role in Quantitative Trait Loci (QTL) mapping, as a tool for uncovering both biological trait and disease regulating genes. Many biological traits and diseases are influenced by variation at multiple loci and hence it is possible to determine the rough genomic position of the causative variations through associations between SNPs and phenotype (Ardlie et al. 2002; Dawson et al. 2002; Fu et al. 2013; Goldstein and Weale 2001; Martin et al. 2000; Morris and Kaplan 2002; Reich et al. 2001; Sachidanandam et al. 2001; Service et al. 1999; Skipper et al. 2004; Terwilliger and Weiss 1998; Wang et al. 2011; Weiss and Clark 2002). Although modeling the epistatic effects of multiple SNPs is of great interest, the single-locus QTL mapping remains a powerful tool as associations can generally only be found over small distances (Mooney 2005). Moreover, as tens-of-thousands of SNPs for genome-wide association studies (GWAS) are under demand (Sachidanandam et al. 2001), the single SNP analysis can at least provide a necessary initial screening selection to detect a subset of promising candidates for further exploration (Doerge 2002; Li et al. 2011).

Despite the great progress which has already been made within LD-based QTL mapping, powerful and computationally efficient methods for large-scale simultaneous testing of individual SNPs with strong control of the family-wise error rate (FWER) are still lacking (Johnson et al. 2010; Nyholt 2004). FWER is the most accepted error rate used to determine significance levels for large-scale testing problems where the goal is to provide conclusive results. In some studies, researchers often control the False Discovery Rate (FDR) to obtain a large pool of potentially significant SNPs and then select only the most significant subset for validation due to cost restrictions. However, this rule can lead to unwanted results as the FDR is controlled only for all selected SNPs, and provides no promise of control for an arbitrarily selected subset of the significant SNPs (Goeman and Solari 2014). Thus, we recommend controlling the FWER (in place of the FDR) in exploratory scenarios where only the most promising results will be considered.

The Bonferroni correction, as one of the most widely used statistical procedures, is
often employed to control the FWER when multiple tests are conducted. However, the
Bonferroni correction is not favorable in large-scale testing because it substantially reduces
the statistical power, hence decreasing the chances of detecting SNPs with real effects
(Nakagawa 2004). While permutation procedures have been widely employed to adjust for
correlated tests, they are computationally expensive (Doerge 2002; Gao et al. 2008; Han
et al. 2009). In LD-based QTL studies, the high likelihood of dependencies among SNPs
demands a new multiplicity adjustment approach that maintains simplicity under arbitrary
dependencies but is more powerful than the standard Bonferroni correction.

In the LD-based QTL model proposed by Fu et al. (2013), detecting a significant QTL
that is associated with a certain phenotype requires two hypothesis tests for each SNP,
one testing for the existence of a QTL (i.e., whether or not the QTL is associated with
the phenotype), and the other testing for the strength of the LD between the SNP and the
existing QTL (i.e., whether or not the QTL is successfully detected by the model). Although
the existence of a significant QTL is the ultimate goal, the degree of LD between the QTL
and SNP is also critical in guaranteeing the basic assumptions of the model. Under the
assumptions in Fu et al. (2013), the unobservable QTL can be mapped by its association
with the observable SNP through the conditional probability of the genotype of the QTL
given the genotype of the SNP. Therefore, of greatest interest are QTLs that are both
significantly existing and strongly linked with a SNP.

Although the LD-based QTL model has been successful in locating significant QTLs
(Das and Wu 2008; Fu et al. 2013; Lou et al. 2003), the Bonferroni multiplicity correction
approach used previously ignored two important issues. First, if the QTL existence test
fails to reject the hypothesis of ‘no QTL,’ then the LD between the SNP and QTL is not
identifiable in the model. Second, while it is of greatest interest to identify those SNPs
for which both hypothesis tests are significant (existing and linked QTL), there is also a
secondary interest in those SNPs for which only the first hypothesis test is rejected.

In this article, we propose a power improving multiplicity correction approach specially
designed for the two hypothesis framework of the LD-based QTL mapping of Fu et al. (2013)
on the basis of graphical weighted-Bonferroni methods (Bretz et al. 2009). Such methods have been shown to synthesize weighted Bonferroni-based closed testing procedures such as the "weighted Bonferroni-Holm procedure, fixed sequence tests, gatekeeping procedures, and the fallback procedure into a powerful and versatile graphical approach" (Bretz et al. 2009), which we tailor here for LD-based QTL mapping. By introducing a logical structuring for the two tests involved for each SNP, i.e., higher order for QTL existence test (primary) than the LD testing (secondary), the LD test will never be investigated if the primary test concludes that the QTL does not exist. This avoids the previously mentioned unidentifiability issue which occurs for the linkage test under the hypothesis of 'no QTL.' None of the current LD-based QTL mapping methodologies directly overcome this challenge when performing these two tests (Das and Wu 2008; Fu et al. 2013; Lou et al. 2003). Further, the proposed multiplicity adjustment approach preserves control of the FWER for both the large-scale number of SNPs and the two hypothesis tests performed for each SNP while allowing for an increase of power (over the standard Bonferroni correction) towards the discovery of SNPs showing evidence of a linked QTL.

In the following section we present the LD-based QTL model of Fu et al. (2013) and the two tests involved. Next, we describe in detail how we design the logical structuring to perform the multiplicity correction for the LD-based QTL model. The significance of the power advantage of the proposed method over the previous Bonferroni correction (which could not account for the unidentifiability issue) is established first theoretically and then demonstrated through both simulations and actual QTL mapping data for HDL cholesterol in outbred mice. Since sample size, heritability, and number of SNPs all determine the power of the method, we illustrate the power through simulations for heritability of 0.1 and 0.4, sample size small (100), medium (300), and large (500), and number of SNPs changing from 1, 10, 50, 100, 500, to 1,000. We conclude with a discussion of the results. The R code (and help file) for the developed approach is available at the author's website: www.stat.usu.edu/gsaunders, and is also included in Appendix A of this dissertation.
3.3 Materials and Methods

3.3.1 LD-based QTL Mapping Model

Fu et al. (2013) suggest a mixture model to map the rough location of the QTL regulating a certain biological trait or disease. Under this model, QTL mapping is accomplished by statistically modeling the genotypic variation through not only the association between phenotype and the putative QTL, but also the association between the QTL and SNP. Since the SNP genotype is observable, the probabilities of a putative QTL genotype can be inferred from the conditional probability of QTL genotype \(A\) given the SNP genotype \(M\), as long as there exists LD between the SNP and putative QTL (Wu et al. 2007).

The mixture model of Fu et al. (2013) assumes each individual’s phenotype \(Y_i\), \(i = 1, \ldots, n\), is a random variate from density \(f_l(Y_i|\theta_l)\), where \(l \in \{1, 2, 3\}\) denotes three distinct QTL genotypes. Each QTL genotype is assumed to induce a separate distribution of phenotypes. Typically, normal distributions are assumed for each \(f_l(Y_i|\theta_l)\) with \(\theta_l = (\mu_l, \sigma)\). From these assumptions, the corresponding likelihood is expressed as (Fu et al. 2013)

\[
L(\omega, \mu, \sigma|Y, M) = \prod_{i=1}^{n} \sum_{l=1}^{3} \omega_{li} f_l(Y_i|\mu_l, \sigma),
\]

where \(\omega_{li}\) is the conditional probability of individual \(i\) having QTL genotype \(l\) given their SNP genotypes, \(\mu_l\) is the phenotypic mean for QTL genotype \(l\), \(\sigma\) is the common standard deviation for all genotypes, and \(f_l(Y_i|\mu_l, \sigma)\) is the probability density of observations for individual \(i\) at QTL genotype \(l\) (Fu et al. 2010, 2013; Wang and Wu 2004; Wu et al. 2007).

The probability of the SNP’s major allele \((M)\) is denoted by \(p\), and correspondingly \(1 - p\) for the minor allele \((m)\). Similarly, the probability of the QTL’s major allele \((A)\) is denoted by \(q\), and correspondingly \(1 - q\) for the minor allele \((a)\). Together, the SNP and QTL form four haplotypes \((MA, Ma, mA, \text{ and } ma)\) with corresponding frequencies \(p_{11} = pq + D, p_{10} = p(1 - q) - D, p_{01} = (1 - p)q - D, \text{ and } p_{00} = (1 - p)(1 - q) + D\), respectively. Here, \(D\) is the linkage disequilibrium between SNP and QTL. The conditional probabilities \(\omega_{li}\) of the QTL’s various genotypes \((AA, Aa, \text{ and } aa)\) can be calculated upon
Table 3.1. The theoretical joint probabilities of SNP (row) and QTL (column) genotypes.

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>$p_{11}^2$</td>
<td>$2p_{11}p_{10}$</td>
<td>$p_{10}^2$</td>
</tr>
<tr>
<td>Mm</td>
<td>$2p_{11}p_{01}$</td>
<td>$2(p_{11}p_{00} + p_{10}p_{01})$</td>
<td>$2p_{10}p_{00}$</td>
</tr>
<tr>
<td>mm</td>
<td>$p_{01}^2$</td>
<td>$2p_{01}p_{00}$</td>
<td>$p_{00}^2$</td>
</tr>
</tbody>
</table>

the observed SNP genotypes ($MM$, $Mm$, and $mm$) from the joint probabilities in Table 3.1 (Fu et al. 2013; Wang and Wu 2004). Hence, $\omega_{ii}$ is a function of $p$, $q$, and $D$. The EM algorithm is then applied to the likelihood in (3.1) to obtain maximum likelihood estimates for all parameters (Fu et al. 2013; Wang and Wu 2004).

3.3.2 Two Hypothesis Tests

Through the likelihood in (3.1), the hypotheses

\[ H_0^L: \mu_1 = \mu_2 = \mu_3 = \mu \quad \text{vs} \quad H_1^L: \text{one of the equalities above does not hold} \]  

(3.2)

can be used to test if the QTL is significantly associated with phenotype. We call the test of $H_0^L$ against $H_1^L$ the QTL existence test. Since all the unknown parameters in (3.1) were estimated by maximum likelihood estimates (MLEs), a log likelihood ratio statistic can be used to test the hypotheses in (3.2) (Fu et al. 2013). The resulting test statistic ($\chi_5^2$) is asymptotically distributed as a $\chi_5^2$ under $H_0^L$ for large enough samples (Wilks 1938). (Discussions about the validity of this asymptotic distribution can be found in Chapter 4 of this work.)

On the other hand, linkage disequilibrium, denoted by $D$, between the SNP and QTL can be tested by means of the hypotheses

\[ H_0^D: D = 0 \quad \text{vs} \quad H_1^D: D \neq 0. \]  

(3.3)
Once the existence of a QTL is established ($H^L_1$), the test statistic used to judge whether or not the QTL is significantly associated with the SNP is (Brown 1975):

$$
\chi^2_D^* = \frac{n \hat{D}^2}{\hat{\rho}(1 - \hat{\rho})\hat{q}(1 - \hat{q})}
$$

(3.4)

$$
\equiv n \hat{r}^2.
$$

(3.5)

Here, $\hat{r}^2$ is the square of the correlation coefficient between the SNP and QTL that has been used in most of the related literature (Hedrick 1987; Pritchard and Przeworski 2001). Under $H^D_0$, $\chi^2_D$ is asymptotically distributed as $\chi^2_1$ (Fu et al. 2013), from which the tail probability (p-value) of the observed level of association can be determined. (However, see discussions about this distributional assumption in Chapter 4).

While discovering evidence for the existence of a QTL is of interest, of greatest interest is whether or not an existing QTL is linked to a SNP. Hence, those SNPs for which the QTL existence test (3.2) is significant are of interest as they manifest some evidence for a linked QTL, but of greatest interest are those SNPs for which both the existence and linkage tests, (3.2) and (3.3), are rejected as these manifest the greatest promise of supporting a linked QTL. However, there is an important identifiability issue within this two hypothesis framework as the parameter $D$ is not identifiable under the null hypothesis $H^L_0$. That is, the parameter $\omega_{l|i}$ falls out of the model when the means are equal, as the $f_l(Y_{i|\mu_l, \sigma})$, $l = 1, 2, 3$ in the likelihood (3.1) are identical in this case, resulting in $\sum_{i=1}^{3} \omega_{l|i} f_l(Y_{i|\mu_l, \sigma}) = f(Y_{i|\mu, \sigma})$. Hence, the likelihood reduces to $L(\mu, \sigma|Y, M) = \prod_{i=1}^{n} f(Y_{i|\mu, \sigma})$, so that $D$, which is contained within $\omega_{l|i}$, cannot be computed. Hence, either the testing approach or the multiplicity adjustment must account for this to ensure identifiability is preserved.

Inspired by the graphical Bonferroni adjustment (Bretz et al. 2009), we design the multiple testing adjustment to control for this identifiability issue. To do this, we select the QTL existence test (3.2) to be of primary importance and the LD test (3.3) to be of secondary importance. If the primary test (3.2) is not rejected, the secondary test will not be investigated. As a result, our proposed multiplicity correction approach increases the
power over the previously applied Bonferroni correction, while preserving strong control of the FWER and avoiding the unidentifiability issue inherent to the Bonferroni approach in this two hypothesis setting.

3.3.3 Graphical Bonferroni Adjustment

The graphical weighted-Bonferroni method of Bretz et al. (2009) is a versatile and easily communicated general adjustment method for multiple testing. Provided as a generalized framework, it must be specially tailored to each testing situation. Generally speaking, it is most powerful for situations where hypotheses can be partitioned into levels of importance such that the most important hypotheses are tested first and the lower level hypotheses are tested only if the higher level hypotheses show significant results.

Under the graphical weighted-Bonferroni adjustment method, all hypotheses of interest are depicted as nodes in a directed acyclic graph. (A detailed introduction to the graphical approach is provided in Section 1.3.6 of this work.) Local significance thresholds for each node (hypothesis) dictate the local level at which each hypothesis is tested. Weighted edges between all nodes map the logical structuring of the designated testing approach. When a hypothesis is rejected, the weighted edges dictate the proportion of the locally assigned significance threshold that is passed from the rejected node to all connected nodes. Thus, the graph induces an iterative testing approach that is shown to result in a closed-test that admits a short-cut (Bretz et al. 2009). Further, Algorithm 1 of Bretz et al. (2009) provides a simple updating technique that performs the short-cut. Strong control of the FWER at level $\alpha$ is proven to occur so long as three regularity conditions are met: 1) the sum of the local significance thresholds is no more than $\alpha$, 2) the sum of outgoing edge weights from each node are no larger than unity, and 3) no node has an edge connecting to itself (Bretz et al. 2009).

3.3.4 Rejection Scheme

Since the ultimate goal is to discover linked QTL, the first interest is in testing $H_0^L$ in (3.2) to see if the phenotype shows evidence of association with a latent QTL. Depending
Fig. 3.1. A) Demonstration of the GBA testing scheme for a single marker. B) The updated graph after finding $H_0^L$ significant.

on the results of the test of (3.2), the testing for the given SNP will either end, or interest will be turned to testing $H_0^D$ in (3.3) to see if the SNP shows evidence of association with the QTL. Figure 3.1A demonstrates how all of $\alpha$ is used to test the first hypothesis, $H_0^L$, and none of $\alpha$ is initially given to the testing of $H_0^D$. That is, node $H_0^L$ has local significance threshold $\alpha$, and $H_0^D$ has local significance threshold 0. Assuming $H_0^L$ is claimed significant, the node belonging to $H_0^L$ would be removed and all of $\alpha$ passed on to $H_0^D$ as signified by the edge weight of 1 along the path from $H_0^L$ to $H_0^D$. At this point, $H_0^D$ is tested at level $\alpha$, its new local significance threshold given the rejection of $H_0^L$ as shown in Figure 3.1B.

Note that adjusted $p$-values could be similarly obtained for each node. The adjusted $p$-value for $H_0^L$ would be the same as the unadjusted value. The adjusted $p$-value for $H_0^D$ would be either the larger of its unadjusted value and $H_0^L$'s unadjusted value (if $H_0^L$ was significant at level $\alpha$) or 1 (if $H_0^L$ was not significant at level $\alpha$). The structuring ensures that a child node (such as $H_0^D$ in Figure 3.1) cannot have a smaller adjusted $p$-value than its parent node ($H_0^L$ in Figure 3.1).

As a result, for the single SNP analysis, either both hypotheses will be tested at level $\alpha$, or the testing will stop after $H_0^L$ without considering $H_0^D$. Alternatively, the standard Bonferroni correction would test both hypotheses at $\alpha/2$. Hence, the Bonferroni adjustment has less power, due to its smaller thresholds. Compared to the traditional Bonferroni, the only potential disadvantage of the GBA method is that it skips testing $H_0^D$ if $H_0^L$ is not significant. However, this potential disadvantage becomes an advantage for the LD-based QTL model because $D$ is not identifiable under the null of $H_0^L$. Thus, the only situation in which the Bonferroni method would have a possible advantage over the GBA method is
not applicable here.

Our proposed GBA method further achieves an advantage in the case of multiple SNPs through sharing of the \( \alpha \)-level between SNPs. Say there are \( m \) SNPs to be tested for both \( H_0^L \) and \( H_0^D \). Let \( H_0^{L_i} \) and \( H_0^{D_i} \) denote these two hypotheses respectively for the \( i \)th SNP, \( i = 1, \ldots, m \). Figure 3.2 demonstrates the case of multiple SNPs, taking \( m = 3 \) as an example for simplicity. In addition to the schemes demonstrated in Figure 3.1, Figure 3.2 shows two more rules. First, it includes the extra edge weights from each \( H_0^{D_i} \) node to all non-parent \( H_0^{L_i} \) nodes, i.e. to all \( H_0^{L_j} \) with \( j \neq i \). This allows for additional \( \alpha \)-sharing between SNPs when both hypotheses (i.e. \( H_0^{L_i} \) and \( H_0^{D_i} \)) are rejected for any given SNP \( i \). Second, the \( \alpha \)-level is split with a Bonferroni type allocation between the \( m \) top-level hypotheses while none of \( \alpha \) is initially provided to the \( m \) lower-level hypotheses. Upon rejection of a higher-level hypothesis, the lower-level child hypothesis receives all of the \( \alpha/m \)-level of the parent (edge weight of 1). If the lower-level child hypothesis is then also found to be significant, its \( \alpha \) threshold is then shared between all remaining higher-level hypotheses (edge weights of 1/2).

The power advantage of our proposed GBA over the Bonferroni method is evident from the larger thresholds. Where the Bonferroni method would test each hypothesis at the \( \alpha/6 \)-level, the GBA tests each hypothesis by thresholds that are no smaller than \( \alpha/3 \). To demonstrate, assume that \( H_0^{L_1} \) and \( H_0^{L_3} \) from Figure 3.2 are rejected at the \( \alpha/3 \) level, but that \( H_0^{L_2} \) is not. Then nodes corresponding to the rejected hypotheses \( H_0^{L_1} \) and \( H_0^{L_3} \) are removed and all \( \alpha \) thresholds and edge weights are updated as shown in Figure 3.3. Notice in Figure 3.3 the reconnecting of edge weights which previously attached to \( H_0^{L_1} \) and \( H_0^{L_3} \).
Fig. 3.3. Demonstration of the GBA testing scheme for three markers assuming that hypotheses $H_0^{L1}$ and $H_0^{L3}$ from the initial graph in Figure 3.2 are rejected from $H_0^{D2}$, $H_0^{D1}$, and $H_0^{D3}$. This demonstrates how edges determine not only the weight that will be passed, but also define the inheritance of edge weights.

Assume now that $H_0^{D1}$ of Figure 3.3 can be rejected at the $\alpha/3$-level. The graph updating (Figure 3.4A) becomes more complicated with this rejection because the rejected hypothesis is both sending out and taking in edge weight from the same hypotheses (nodes). Specifically, $H_0^{D2}$ is set to send half of its threshold to $H_0^{D3}$ and the other half to $H_0^{D1}$. Of the half that the now rejected $H_0^{D1}$ would have received from $H_0^{D2}$, half is designated to $H_0^{L2}$ and the other half designated to go to $H_0^{D3}$. This assignment causes the updated $H_0^{D2}$ to send a total weight of $3/4$ to $H_0^{D3}$. However, recalling the logical structure of the hypotheses, it can be seen that $H_0^{D2}$ will not be considered for testing unless $H_0^{L2}$ is first rejected. Hence, the $1/4$ that $H_0^{D2}$ would pass on to $H_0^{D3}$ through $H_0^{D2}$ at this point is not logically possible as this would require testing $H_0^{D2}$ before testing $H_0^{L2}$. This logical restriction allows us to move the $1/4$ out from $H_0^{D2}$ by means of the only other path available, so that $H_0^{D3}$ receives a total weight of $1$ from $H_0^{D2}$, as shown in Figure 3.4A.

The node corresponding to $H_0^{D3}$ in Figure 3.3 was sending half of its threshold to $H_0^{D1}$ and the other half to $H_0^{L2}$. With the removal of $H_0^{D1}$, now assumed to be significant, $H_0^{D3}$ will now be doubly joined to $H_0^{L2}$ and to itself by inheriting the outgoing paths from $H_0^{D1}$ to both $H_0^{L2}$ and $H_0^{D3}$. This junction of $H_0^{D3}$ to itself would specify that the $1/2$ that was going from $H_0^{D3}$ to $H_0^{D1}$ times the $1/2$ that was going from $H_0^{D1}$ to $H_0^{D3}$ would result in $H_0^{D3}$ returning $1/4$ to itself. Since it is not possible for $H_0^{D3}$ to pass $1/4$ back to itself, it passes to $H_0^{L2}$ the original $1/2$ it was already sending to $H_0^{L2}$, plus the $1/4$ inherited by...
Fig. 3.4. A) The updated graph from Figure 3.3 assuming the hypothesis $H_0^{D1}$ of Figure 3.3 is rejected at the $\alpha/3$-level. B) Graph resulting from the rejection of the hypothesis $H_0^{D3}$ at the $\alpha/2$-level.

$H_0^{L2}$ from $H_0^{D3}$ via $H_0^{D1}$ plus the $1/4$ that $H_0^{D3}$ re-inherited from $H_0^{D1}$. The result is to have $H_0^{D3}$ send all of its threshold to $H_0^{L2}$. This can also be viewed more simply by the fact that upon removal of $H_0^{D1}$ from the graph (due to its rejection), $H_0^{D3}$ is left with only one outgoing edge to $H_0^{L2}$, hence all of its threshold must be passed to $H_0^{L2}$.

The final graph resulting from the rejection of $H_0^{D1}$ in Figure 3.3 is depicted in Figure 3.4A. At this point it could be possible that $H_0^{L2}$ is rejected, but to demonstrate a more interesting scenario, assume that $H_0^{D3}$ only can be rejected at the $\alpha/2$-level. The resulting graph with $H_0^{D3}$ removed is depicted in Figure 3.4B. Interestingly, both $H_0^{L2}$ and (if significant) $H_0^{D2}$ can now be tested at the full level $\alpha$.

3.3.5 The GBA as an IUT

It was suggested by a reviewer that the GBA approach might be accomplished through a conceptually simpler (but computationally equivalent) approach using the idea of an Intersection Union Test (IUT) (Berger 1982). This is accomplished under the assumption that of interest is only the case that both hypotheses $H_0^L$ and $H_0^D$ are rejected simultaneously for a given SNP. As this is certainly the most interesting scenario, the method is worth considering. The IUT portion of the approach is performed by taking the maximum of the $P$-values corresponding to the tests of $H_0^L$ and $H_0^D$ for a given SNP. Once this maximum is obtained for all SNPs, a Bonferroni-Holm correction (Holm 1979) is applied to the maximum $P$-values to obtain a final list of significant SNPs. This approach is identical to the GBA approach when only the rejection of both hypotheses is of interest. In the case that the
decision on $H_0^L$ is of interest aside from the decision on $H_0^D$, then the GBA is more powerful.

The IUT approach is essentially the approach of the original multiplicity correction in Fu et al. (2013) but where only a Bonferroni correction was applied. Technically, the original work in Fu et al. (2013) performed a Bonferroni correction across all SNPs separately for the testing of $H_0^L$ over each SNP, and then for the testing of $H_0^D$ over each SNP. The maximum of the corrected values was then selected as bearing on the decision as to whether or not a given SNP was significantly linked to a QTL. Thus, the ordering of the multiplicity correction and IUT were reversed from that suggested by the reviewer. However, the results were the same due to the equal adjustment of the Bonferroni approach across all SNPs. However, for the extension to the Holm adjustment, the IUT must first be applied, and the resulting maximum $P$-values then adjusted for multiplicity, as suggested by the reviewer. The following argument demonstrates the equality of the IUT with a Holm correction for multiplicity to the GBA approach (when only the rejection of both hypotheses is of interest).

Under the IUT approach, there is a single $P$-value for each SNP which represents the maximum of the tests of $H_0^D$ and $H_0^L$ for that SNP. Denote this maximum value by $P_j^M$ for SNP $j = 1, \ldots, m$ so that $P_j^M = \max\{p_j^L, p_j^D\}$ where $p_j^L$ and $p_j^D$ denote the raw (unadjusted) $P$-values for $H_0^L$ and $H_0^D$, respectively, at SNP $j$. By virtue of the Holm adjustment (Holm 1979), see Section 1.3.2 for an introduction, the smallest of the $P_j^M$ will be multiplied by a factor of $m$. Let $k$ denote the index of the SNP at which the smallest of the $P_j^M$ occurs. Then, for any rejections to occur under the IUT method with the Holm adjustment it must hold that $mP_k^M \leq \alpha$. Assume then that at least one rejection occurs so that $P_k^M$ is the smallest of the maximum $P$-values from the IUT approach and that $mP_k^M \leq \alpha$.

Under the GBA approach, the adjusted $P$-values for $H_0^D$ will always be at least as large as the value corresponding to $H_0^L$ for the same SNP. This is due to the logical structuring of the hypotheses and was demonstrated in the previous section. Denote by $\tilde{P}_j^L$ and by $\tilde{P}_j^D$ the GBA adjusted $P$-values for $H_0^L$ and $H_0^D$, respectively, for SNP $j = 1, \ldots, m$. So long as the $H_0^D$ hypothesis has not been rejected for any SNP, then $\tilde{P}_j^L = mp_j^L$ for all $j$. Thus, for
the GBA approach to reject both hypotheses for SNP $k$, it must hold that

$$m p_k^L = \hat{P}_k^L \leq \hat{P}_k^D = m \max \{p_k^L, p_k^D\} = m P^M \leq \alpha. \quad (3.6)$$

Inequality (3.6) identifies that the two procedures will agree for all rejections. This is because of the step down nature of both methods. In the case that SNP $k$ is rejected (as in the above argument), then SNP $k$ is essentially removed from analysis and the factor $m$ becomes $m_1 = m - 1$ with the exact argument from above reapplied to the data. The process is continued until no more rejections can occur. Say after $s$ rejections, $m_s = m - s$, no more rejections can occur in the IUT approach with the Holm adjustment. Then it follows from Inequality (3.6) that $\hat{P}_k^D = m_s \max \{p_k^L, p_k^D\} = m_s P^M > \alpha$ so that no more rejections of $H_0^D$ can occur in the GBA approach. Note however that it is possible for more rejections of some $H_0^L$ to still occur under the GBA approach as $\hat{P}_k^L \leq \hat{P}_k^D$. However, this is only of interest in the case that $H_0^L$ is of interest even when $H_0^D$ is not significant.

In conclusion, the proposed GBA will always be more powerful than the traditional Bonferroni procedure and is identical to the IUT approach under Bonferroni-Holm correction in the case that only the rejection of both $H_0^L$ and $H_0^D$ is of interest. In addition, both the GBA approach and the IUT (coupled with the Holm correction) address the unidentifiable issue of $D$ under the null hypothesis $H_0': \mu_1 = \mu_2 = \mu_3$ by only allowing for the rejection of $H_0^D$ when $H_0^L$ is also found significant. Finally, by setting $H_0^L$ to be the primary test over $H_0^D$ in the GBA approach, we allow for both the traditional consideration of only $H_0^L$ independent of $H_0^D$, as well as the simultaneous consideration of both $H_0^L$ and $H_0^D$ to establish the existence of linked QTL. In the remainder of this article, we attempt to quantify, both through simulation studies and real data, the magnitude of the power improvement of the GBA (and the identical IUT approach) over the standard Bonferroni correction, and that the improvements lead to greater scientific discovery while maintaining strong control of the FWER.
3.4 Results

3.4.1 Power Simulation

We investigated a simulation study to quantify the power advantage of the proposed GBA (and IUT) over the standard Bonferroni adjustment within the LD-based QTL mapping model of Fu et al. (2013). The QTL, phenotype, and SNPs were generated under the assumptions of the alternative hypotheses in (3.2) and (3.3). The QTL was generated using an assigned probability of $q = 0.7$ for the major allele. For each individual $i$, $Q_i = l$ with $l \in \{1, 2, 3\}$ was used to code the QTL genotypes of $aa$, $Aa$, and $AA$, respectively. The normally distributed phenotype dependent on the value of the QTL is generated as $Y_i(Q_i = l) \sim N(\mu_l, \sigma)$. The means for the phenotype $Y$ corresponding to the values of the QTL were set at $\mu_1 = 8$, $\mu_2 = 10$ and $\mu_3 = 12$. SNPs were then generated using the conditional probability of the SNP genotype given the value of the QTL genotype for each individual. In general, for an LD-based QTL mapping model, researchers genotype the SNP first and then use the SNP to generate a QTL based on the conditional probability of QTL genotype given SNP genotype as given in Table 3.1. However, for our purposes, we are interested in extending from single SNP mapping to multiple SNPs mapping. Therefore, we derive the conditional probability of SNP genotype given QTL genotype (see Table 3.2) from the Bayes Rule in Equation (3.7) and Table 3.1.

$$P(M|QTL) = \frac{P(QTL|M)P(M)}{P(QTL)}.$$  \hfill (3.7)

Table 3.2. The theoretical conditional probabilities of SNP genotype (columns) given QTL genotype (rows).

<table>
<thead>
<tr>
<th></th>
<th>$MM$</th>
<th>$Mm$</th>
<th>$mm$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AA$</td>
<td>$\frac{p_{11}^2}{q^2}$</td>
<td>$\frac{2p_{11}p_{10}}{q^2}$</td>
<td>$\frac{p_{10}^2}{q^2}$</td>
</tr>
<tr>
<td>$Aa$</td>
<td>$\frac{2p_{11}p_{10}}{q(1-q)}$</td>
<td>$\frac{2(p_{11}p_{00}+p_{10}p_{11})}{q(1-q)}$</td>
<td>$\frac{2p_{10}p_{00}}{q(1-q)}$</td>
</tr>
<tr>
<td>$aa$</td>
<td>$\frac{p_{10}^2}{(1-q)^2}$</td>
<td>$\frac{2p_{10}p_{00}}{(1-q)^2}$</td>
<td>$\frac{p_{00}^2}{(1-q)^2}$</td>
</tr>
</tbody>
</table>
Fig. 3.5. Power comparison between the graphical Bonferroni adjustment (GBA) and standard Bonferroni adjustment under different sample size, number of SNPs, and heritability (A: $H^2 = 0.1$, B: $H^2 = 0.4$).

Sample sizes of $n = 100$, 300, and 500 were used to represent small, medium, and large sample sizes, respectively. The number of SNPs per simulation was set at $m = 1$, 10, 50, 100, 500, and 1,000 to show the initial power under the single SNP scenario and the corresponding decreasing power trend as the number of SNPs increases. Finally, the heritability was set at two values, $H^2 = 0.1$ and 0.4, corresponding to high and low error variance (Wang and Wu 2004). The model error variance $\sigma^2$ was computed using the heritability and genetic variance of the QTL. Power estimates were averaged over 1,000 simulations.

The simulation results, shown in Table 3.3 and depicted in Figure 3.5, demonstrate the power comparison of the proposed graphical Bonferroni adjustment (GBA) with the traditional Bonferroni adjustment. (Note that the IUT with Holm correction is identical to the GBA results presented in Figure 3.5 as only the power for the rejection of both hypotheses is plotted. Table 3.3 demonstrates the power gains in $H_0^L$ that the GBA approach achieves. GBA (L), as compared to the results of both hypotheses being rejected, GBA (D).) These results provide an experimental reference for researchers about how power varies among different sample size $n$, the number of SNPs $m$, and the degree of heritability ($H^2$). As expected, the power under high heritability (B: $H^2 = 0.4$) is much higher than that
Table 3.3. The results of the power simulation as depicted in Figure 3.5. Here, Bon is the Bonferroni adjustment and GBA (D) is the graphical Bonferroni adjustment when rejecting both nulls of interest (equivalent to IUT with the Holm adjustment), and GBA (L) is the graphical Bonferroni adjustment where rejecting $H_0^L$ only is meaningful.

<table>
<thead>
<tr>
<th>m</th>
<th>n = 100</th>
<th></th>
<th>n = 300</th>
<th></th>
<th>n = 500</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bon.</td>
<td>GBA (D)</td>
<td>GBA (L)</td>
<td>Bon.</td>
<td>GBA (D)</td>
<td>GBA (L)</td>
</tr>
<tr>
<td>H^2 = 0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.094</td>
<td>0.094</td>
<td>0.105</td>
<td>0.299</td>
<td>0.299</td>
<td>0.299</td>
</tr>
<tr>
<td>10</td>
<td>0.073</td>
<td>0.081</td>
<td>0.153</td>
<td>0.178</td>
<td>0.201</td>
<td>0.219</td>
</tr>
<tr>
<td>50</td>
<td>0.038</td>
<td>0.041</td>
<td>0.053</td>
<td>0.091</td>
<td>0.099</td>
<td>0.102</td>
</tr>
<tr>
<td>100</td>
<td>0.018</td>
<td>0.019</td>
<td>0.028</td>
<td>0.061</td>
<td>0.064</td>
<td>0.074</td>
</tr>
<tr>
<td>500</td>
<td>0.008</td>
<td>0.008</td>
<td>0.016</td>
<td>0.028</td>
<td>0.030</td>
<td>0.040</td>
</tr>
<tr>
<td>1000</td>
<td>0.005</td>
<td>0.005</td>
<td>0.009</td>
<td>0.021</td>
<td>0.023</td>
<td>0.031</td>
</tr>
<tr>
<td>H^2 = 0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.839</td>
<td>0.839</td>
<td>0.840</td>
<td>0.937</td>
<td>0.937</td>
<td>0.937</td>
</tr>
<tr>
<td>10</td>
<td>0.352</td>
<td>0.417</td>
<td>0.442</td>
<td>0.835</td>
<td>0.918</td>
<td>0.918</td>
</tr>
<tr>
<td>50</td>
<td>0.208</td>
<td>0.233</td>
<td>0.266</td>
<td>0.715</td>
<td>0.834</td>
<td>0.835</td>
</tr>
<tr>
<td>100</td>
<td>0.144</td>
<td>0.161</td>
<td>0.181</td>
<td>0.639</td>
<td>0.746</td>
<td>0.748</td>
</tr>
<tr>
<td>500</td>
<td>0.076</td>
<td>0.082</td>
<td>0.100</td>
<td>0.482</td>
<td>0.557</td>
<td>0.560</td>
</tr>
<tr>
<td>1000</td>
<td>0.060</td>
<td>0.065</td>
<td>0.080</td>
<td>0.427</td>
<td>0.489</td>
<td>0.493</td>
</tr>
</tbody>
</table>

of the low heritability ($A: H^2 = 0.1$) and the power under large sample size ($n = 500$, blue curves) is much higher than that of the small sample size ($n = 100$, green curves). Under high heritability ($H^2 = 0.4$) and a larger sample size ($n = 500$), the power of the multiplicity adjustment remains high even as the number of SNPs becomes large ($m = 1,000$). However, in practice it is often expensive to collect so many sample measurements, so these results are useful in deciding the opportunity costs in power for smaller sample sizes. It is worth noting that for moderate numbers of SNPs, the power increase of the GBA over the Bonferroni adjustment allows for the possibility of maintaining the power level of the Bonferroni adjustment while decreasing the sample size of the study or increasing the number of SNPs, a great advantage for researchers.

Although the power increase of the GBA improves moderately over the standard Bonferroni adjustment for the case of low heritability ($H^2 = 0.1$) when the sample size is small ($n = 100$), these findings are comparable to seminal results found by previous multiplicity improvements over their competitors (Benjamini and Hochberg 1995; Holm 1979). All in all, our proposed GBA method generally shows a valuable increase in power over the Bonferr-
roni adjustment under all 12 circumstances with the different combinations of sample size, number of SNPs, and heritability, with this difference tending toward zero as the sample size approaches infinity. In addition, the GBA method allows interest in identifying QTL existence (rejecting $H_0^L$ only) without doubly confirming linkage through $H_0^D$.

### 3.4.2 Poplar Leaf Shape QTL Mapping Project

To show how the power advantage of the GBA approach leads to increased scientific discovery over the Bonferroni adjustment for moderate numbers of SNPs, we apply it to a real poplar leaf shape QTL mapping study, where the Bonferroni adjustment was previously used for multiple correction (Fu et al. 2013). The study design used a representative leaf from each of 106 poplar trees (i.e., *Populus szechuanica* var. *tibetica* belonging to the Tacamahaca section) that was randomly selected and photographed for shape QTL analysis. The trees were also genotyped for a panel of 29 microsatellite markers (only 16 of which were usable), which are another type of genetic marker as opposed to the now more common SNP. An RCC (radius centroid contour) approach was used to represent the leaf shape (phenotype) with a high dimensional curve. The first six principal components (PCs) were selected to capture the majority variation of leaf shape from six orthogonal directions. Significant QTLs affecting the shape variability were mapped through the statistical LD-based QTL mapping model. The standard Bonferroni adjustment was applied within each PC separately to control the FWER for the multiple tests resulting from considering multiple microsatellites simultaneously (Fu et al. 2013).

The GBA successfully detects all significant microsatellites that the Bonferroni adjustment located. However, in addition to the previously established results, the GBA detects 3 more microsatellites within PC 4 (responsible for 5.1% of the total variation) that were not detected previously by the Bonferroni adjustment. These previously undetected significant results correspond to markers 1, 12, and 15. Figure 3.6 shows the leaf shape profile curves for the QTL genotypes as identified by marker 1. The profiles for markers 12 and 15 are equally subtly different in the aa genotype from the AA and Aa genotype, suggesting that multiple QTL may contribute to this small effect on leaf shape. Such a theory is consis-
Fig. 3.6. The control of leaf shape according to the different genotypes (AA, Aa, aa) of the QTL identified by marker 1 on PC 4. This shows the increased sensitivity of the GBA approach as the effect of the aa QTL genotype is subtly different from that of genotypes AA and Aa, but nevertheless, corresponds to information which was undetected under the Bonferroni correction.

tent with QTL mapping assumptions generally, that small effects are potentially effected by many QTL (Doerge 2002). Fu et al. (2013, Figure 7) demonstrates the leaf shape effect curves for PC’s 1 and 3, where the genotypic effects of the QTL on the phenotype are more pronounced, and hence detected by both the GBA and Bonferroni approaches.

3.4.3 Mouse HDL Cholesterol QTL Mapping Project

We also applied the introduced LD-based QTL model (Fu et al. 2013) and the proposed GBA multiplicity correction approach to an outbred mouse HDL cholesterol genome data set to compare our findings with some highly validated discoveries in the current literature. After summarizing the study details, we demonstrate that the GBA approach performs consistently with previously established results.

Epidemiological studies have consistently shown that the level of plasma high density lipoprotein (HDL) cholesterol is negatively correlated with the risks of coronary artery disease and gallstones (Lyons et al. 2003; Mehrabian et al. 2000; Su et al. 2009a,b; Wang et al. 2003, 2004). Because of the inverse relationship between HDL and cardiovascular disease, there has been considerable interest in understanding genetic mechanisms contributing to
variations in HDL levels. HDL levels vary considerably in different people, which are affected by interactions of multiple genes and environmental factors, and up to 70% of this variation in humans is genetically determined (Rader and Maugais 2000; Wang et al. 2003). Because of the concordance between human QTLs regulating HDL and corresponding mouse loci and many easily controlled experimental advantages, mouse has become an animal model in HDL research. Numerous findings in HDL QTL associations are obtained from crosses between different inbred mouse strains. By crossing inbred strains that significantly differ in HDL levels and subsequently testing for association between HDL levels and genetic markers in the progeny, numerous significant QTLs involved in HDL have been identified in mouse (Korstanje et al. 2004; Lyons et al. 2003; Machleider et al. 1997; Mehrabian et al. 2000; Su et al. 2009a,b,c, 2010; Wang et al. 2003, 2004; Wergedal et al. 2007).

Compared to the inbred mice strains with coarse mapping resolution, the QTL research on wild-caught and commercial stocks of outbred mice, as resources for genetic fine mapping, is far underdeveloped. Zhang et al. (2012) published an open resource outbred mouse database (available at http://cgd.jax.org/datasets/datasets.shtml) with 288 Naval Medical Research Institute (NMRI) mice and 44,428 unique SNP genotypes. Three hundred 4-to-6-week-old male NMRI mice were purchased and individually housed with the same diet and environmental conditions. The blood samples of each mouse were measured by submandibular puncture after a 4-hr fast. Then plasma samples were frozen for measurement of HDL cholesterol. There were 10 mice removed because the standard deviation of individual blood pressure is greater than two. Another two mice were also discarded for their 99% identity of SNP genotypes. This caused the final sample size to be 288. A total of 581,672 high density SNP were initially genotyped by the Novartis Genomics Factory using the Mouse Diversity Genotyping Array (Yang et al. 2009). In order to guarantee promising data for association mapping studies (Yalcin et al. 2010), only polymorphic SNPs with minor allele frequency greater than 2%, Hardy-Weinberg equilibrium $\chi^2 < 20$, and missing values less than 40% were retained. Moreover, identical SNPs within a 2Mb interval were collapsed. This left 44,428 unique SNP genotypes for their resulting analysis using
Fig. 3.7. The negative log of the GBA-adjusted $p$-values for $H_0^D$ for each SNP in the mouse HDL cholesterol QTL mapping project. The red reference line corresponds to a 0.05 family-wise error rate.

From Zhang’s work, adjustments for multiplicity at the genome-wide association level were made using a simulation approach (Knijnenburg et al. 2009) as well as the permutation approach (Churchill and Doerge 1994). They identified three loci as significant, with two loci on Chromosome1 (Chr1) and a single locus on Chromosome5 (Chr5) (Zhang et al. 2012, Figure 3).

Recalling the detailed adjustment structure of the GBA, it can be seen that the adjusted $p$-value obtained from GBA for the test of $H_0^D$ will never be smaller than that of $H_0^L$ (as demonstrated in Section 3.3.5). Hence, reporting the significant adjusted $p$-values for $H_0^D$ is sufficient for demonstrating those SNPs that show strongest evidence of linkage to a true QTL. Figure 3.7 depicts the negative log of the adjusted $p$-values for $H_0^D$ for each SNP as a function of the location (in Mb) of each SNP for the 19 autosomal chromosomes and the X chromosome of mouse. The threshold for the adjusted $p$-values of $-\log(0.05) \approx 2.9957$ supports two dramatically significant findings, one on Chr1 at Mb173 and Mb182, and the other on Chr5 at Mb125. These significant discoveries are the same as the findings in current outbred mouse literature, compare to Figure 3 of Zhang et al. (2012). The two other spikes depicted in Figure 3.7, which are not significant at the 0.05 level, are located on Chr1 at Mb181, Chr2 at Mb169, and Chr4 at Mb150.

In Table 3.4 we summarize our findings as compared to results obtained from inbred mouse crosses using very different approaches (Su et al. 2009b, Table 2). Three QTLs have been reported coincident with candidate genes, of which our study finds the two most
Table 3.4. The significant results of the outbred mice HDL cholesterol QTL mapping project depicted in Figure 3.7. SNPs are ordered by significance level. Corresponding concurrence candidate gene and QTL from previous inbred crosses studies are shown.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position (Mb)</th>
<th>Adjusted P</th>
<th>Raw P</th>
<th>Candidate Gene</th>
<th>QTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1***</td>
<td>173,155,512</td>
<td>5.7 x 10^{-15}</td>
<td>1.3 x 10^{-10}</td>
<td>Apoa2</td>
<td>c, g, h, i, k</td>
</tr>
<tr>
<td>5***</td>
<td>125,530,593</td>
<td>5.2 x 10^{-10}</td>
<td>1.2 x 10^{-14}</td>
<td>Scarb1</td>
<td>b, d, e, f, i</td>
</tr>
<tr>
<td>1</td>
<td>181,681,689</td>
<td>2.2 x 10^{-01}</td>
<td>4.9 x 10^{-06}</td>
<td>Hdlq69</td>
<td>f</td>
</tr>
<tr>
<td>2</td>
<td>169,427,598</td>
<td>1.6 x 10^{-01}</td>
<td>3.7 x 10^{-06}</td>
<td>Hdlq64</td>
<td>f</td>
</tr>
<tr>
<td>4</td>
<td>150,473,153</td>
<td>1.1 x 10^{-01}</td>
<td>2.5 x 10^{-06}</td>
<td>Hdlq64</td>
<td>f</td>
</tr>
<tr>
<td>16</td>
<td>30,270,738</td>
<td>1.0 x 10^{-00}</td>
<td>2.9 x 10^{-05}</td>
<td>Apod</td>
<td>a, f</td>
</tr>
<tr>
<td>19</td>
<td>54,506,048</td>
<td>1.0 x 10^{-00}</td>
<td>4.9 x 10^{-05}</td>
<td>Hdlq48</td>
<td>d, f</td>
</tr>
</tbody>
</table>

*** Significant at the FWER 5 x 10^{-10} level

Major, as did another study using these same data (Zhang et al. 2012). Of those two we located, the Chr1 locus at Mb173, the highest peak in Figure 3.7, is the major determinant of HDL, which has been detected as QTL Hdlq15 in inbred mouse strains multiple times (as referenced in Table 3.4). Combining mouse crosses with haplotype analysis for the HDL QTL located on Chr 1 locus at Mb173 reduced the list of candidates to a small amount. Numerous mouse crosses have linked HDL to this region, and Apoa2 has been identified as the gene underlying the QTL (Machleider et al. 1997; Su et al. 2009a,b,c; Wang et al. 2004); this gene has been highlighted in Nature Reviews Genetics (Flint and Eskin 2012). Chr5 locus at Mb125, the second highest peak in Figure 3.7, is located in the same locus as QTL Hdlq1 found by Su et al. (2009c) and Korstanje et al. (2004) (as referenced in Table 3.4). In addition, they conclude that Scarb1 (a well known gene involved in HDL metabolism) is the causal gene underlying Hdlq1 by haplotype analysis, gene sequencing, expression studies, and a spontaneous mutation (Su et al. 2010; Wergedal et al. 2007).

One thing deserving mention is that the records of QTL in inbred mouse studies use the coarser scale cM, while outbred mouse studies use the finer scale Mb. We are able to approximate the cM-to-Mb rate based on the fact that QTL Hdlq15 located in Chr1 locus cM85 (Table 2 of Su et al. (2009b)) in inbred mouse is the same one located on Chr1 at Mb173 (Table 3 of Su et al. (2009c)). We thus mapped the inbred mouse results to those of outbred mice by performing a detailed comparison of map positions with that of inbred
mouse to obtain the comparison results depicted in Table 3.4.

3.5 Discussion

Detecting significant genes that cause disease (for example the inverse relation between human cholesterol and cardiovascular disease) or regulate biological traits through LD-based QTL mapping has been popular in many disciplines (Ardlie et al. 2002; Dawson et al. 2002; Fu et al. 2013; Goldstein and Weale 2001; Martin et al. 2000; Morris and Kaplan 2002; Reich et al. 2001; Sachidanandam et al. 2001; Service et al. 1999; Skipper et al. 2004; Terwilliger and Weiss 1998; Wang et al. 2011; Weiss and Clark 2002). These new techniques can simultaneously consider tens of thousands of SNPs, bringing substantial challenges for multiple testing. In addition, high dimensional biological traits, often reduced to multiple PC components, have been widely used and add yet another demand for a powerful and computationally efficient approach to adjust for multiple tests (Drake and Klingenberg 2010; Fu et al. 2013; Langlade et al. 2005).

These multiple tests require an adjustment on the resulting $P$-values in order to preserve control of the family-wise error rate (FWER) at a pre-specified level $\alpha$. In some cases, follow up work on the significant findings may justify using the false discovery rate (FDR) as the error rate of interest. Typically however, the significant results are directly reported and therefore the FWER is the more desirable form of error rate to control (Goeman and Solari 2014). The current standard approach in LD-based QTL mapping is to apply a Bonferroni adjustment to correct for multiplicity and preserve the FWER. As is well known, the Bonferroni correction is overly conservative for large numbers of tests, but the advantages of simplicity without independence assumptions on the corresponding family of tests continue to make it popular.

In this article, we tailored a multiple correction approach, based on graphical weighted-Bonferroni methods (Bretz et al. 2009), which allows for the logical order among the two hypotheses in (3.2) and (3.3) to be structured into the multiplicity correction. As in the LD-based QTL mapping model of Fu et al. (2013), we need to test two hypotheses for each SNP, one with $H^L_0$ (3.2) about whether or not an association exists between QTL and
phenotype, and the other with $H_0^D$ (3.3) about whether or not LD exists between SNP and QTL. Among these two tests, the existence test has higher priority because the LD test will not be applicable if a QTL does not exist, and the existence of QTL is the ultimate goal in real applications. Although the logical structure of the two tests is known, none of the current LD-based QTL literature considers this priority structure when performing these two tests (Das and Wu 2008; Fu et al. 2013; Lou et al. 2003; Wang and Wu 2004). Accounting for this structure using GBA provides the novel ability to identify QTL existence (rejecting $H_0^I$ only) even when linkage (rejecting $H_0^D$) is not detected.

The significance of the power advantage of the proposed method over the Bonferroni method, established theoretically, through simulations, and finally on real data, is such that we advocate its use whenever multiple tests are needed for the LD-based QTL mapping design, where both $H_0^I$ and $H_0^D$ tests are considered.

The R code for our GBA adjustment approach and help file can be downloaded for free from www.stat.usu.edu/gsaunders, and is also included in Appendix A of this dissertation.
CHAPTER 4

QTL MAPPING: HYPOTHESES AND APPROACHES

4.1 Introduction

With the advent of genetic maps and the development of statistical models and analysis techniques, the secrets of the genetic involvement in characters of a quantitative nature are steadily being uncovered. This effort to identify regions along the genome which are associated with quantitative traits of interest is known as quantitative trait loci (QTL) mapping. As reviewed in Doerge (2002), early successes in QTL mapping ranged from the location of the cystic fibrosis gene in humans (Kerem et al. 1989), to the identification of a gene affecting horn development in cattle (Georges et al. 1993), and further studies have continued to reveal findings as diverse as QTL impacting fruit texture in apples (Longhi et al. 2013). While many approaches to QTL mapping exist (see Doerge (2002) for a review) this work focuses on single-marker mapping (Knott and Haley 1992; Luo and Kearsey 1989; Luo and Suhai 1999) which is both simple and useful for identifying candidate lists of significant markers. Two disadvantages of the method are the larger sample size that is required as compared to other QTL mapping approaches, and the multiple testing issues inherent to performing many separate statistical tests simultaneously (Doerge 2002). The main contribution of Chapter 3 was to investigate a simple and powerful multiplicity adjustment approach to in part remedy the impact of multiple testing in single-marker QTL mapping.

The introduced multiplicity adjustment of Chapter 3 is independent from the issue of calculating the raw $P$-values (pre-adjustment values) for each hypothesis test. However, the validity of the underlying (raw) $P$-values is an important aspect of the developed multiplicity adjustment approach. In Chapter 3, the asymptotic chi-square distribution provided by Wilks (1938) was utilized to obtain the raw $P$-values from the likelihood ratio test statis-
tic (Section 3.3.2). Churchill and Doerge (1994) critique such a choice stating that, “In most cases, the regularity conditions that ensure an asymptotic chi-square distribution for the likelihood ratio test statistic are not satisfied.”

The aim of this chapter is three-fold. First, to discuss the appropriateness of asymptotic chi-square distributions in single-marker QTL mapping. Second, to present the various advantages and disadvantages of each of several methods suggested in the literature for the computation of the raw $P$-values for the two QTL mapping hypotheses of Chapter 3 (Section 3.3.2). These include permutations (Churchill and Doerge 1994), a Monte Carlo approach to simulating the null distribution (McLachlan 1987; Wu et al. 2007), an empirical approach for large-scale testing (Efron 2004), and the previously applied theoretical asymptotic chi-square distributions (Knott and Haley 1992; Wilks 1938). Third, a novel null kernel estimation approach of the joint (bivariate) null distribution for single-marker QTL mapping based on simulations is also presented and compared to each of the univariate approaches just mentioned.

4.2 The QTL Hypotheses

In the linkage disequilibrium (LD) QTL mapping framework of Chapter 3 (Fu et al. 2013), two hypothesis tests concerning (1) the existence of a QTL and (2) its linkage to a given genetic marker, were conducted using the mixture model likelihood given by

$$L(\omega, \mu, \sigma|Y, M) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_{g|M_i} f(Y_i|\mu_g, \sigma).$$

(4.1)

This likelihood (4.1) assumes each individual’s phenotype $Y_i$, $i = 1, \ldots, n$, is a random variate resulting from their latent QTL genotype $g$. The density $f(Y_i|\mu_g, \sigma)$ denotes the corresponding normal distribution functions for the distinct QTL genotypes $g \in \{1, \ldots, G\}$ with $\mu = (\mu_1, \ldots, \mu_g)$. The mixing proportions $\omega_{g|M_i}$ denote the conditional probabilities of individual $i$ having QTL genotype $g$ given their SNP genotype $M_i$, and is a function of the genetic linkage between the marker and QTL. Likelihood (4.1) has been used extensively in QTL mapping (Churchill and Doerge 1994; Fu et al. 2010, 2013; Knott and Haley 1992;
Lander and Botstein 1989; Wang and Wu 2004; Wu et al. 2007). However, the use of Likelihood (4.1) as presented in Chapter 3 differs importantly from the traditional use as found in the literature.

Knott and Haley (1992) detail the traditional use of Likelihood (4.1), corresponding to the alternative hypothesis of a linked QTL, as well as the null hypotheses of interest in single marker QTL mapping, one of either a no QTL or an unlinked QTL model. These traditional hypotheses and corresponding likelihoods are as follows.

- $H_A$: a linked QTL. The hypothesis here is that a QTL exists and is linked to the marker under consideration. The likelihood for this hypothesis is the same as the likelihood in (4.1), which we have been considering previously in Chapter 3. For completeness, the likelihood as given in Knott and Haley (1992) is written as

$$L = \prod_{i=1}^{n} \sum_{g=1}^{G} \text{trans}(g|M_i) f_g(Y_i|\mu + a_g - d_g, \sigma).$$

Here, $\text{trans}(g|M_i)$ is the “transmission probability of offspring $i$ being [QTL] genotype $g$ given that it has marker genotype $[M_i]$ at the marker being considered” (Knott and Haley 1992). Further, the transmission probability is a function of the recombination fraction $r$ between the marker and QTL. The values $a_g$ and $d_g$ are the additive and dominance effects of the QTL, respectively.

- $H_0^2$: an unlinked QTL. This null model assumes that there is in fact a QTL underlying the phenotype, but that the QTL is not linked with the marker under consideration. The corresponding likelihood is the standard mixture model likelihood (McLachlan and Peel 2000) and is given by Knott and Haley (1992) as

$$L = \prod_{i=1}^{n} \sum_{g=1}^{G} \text{trans}(g) f_g(Y_i|\mu + a_g - d_g, \sigma).$$

Here, “$\text{trans}(g)$ is the transmission probability of the offspring being genotype $g$” (Knott and Haley 1992). (Recall that concluding $H_0^2$ is a novel ability of the GBA
method proposed in Chapter 3; see Section 3.5.)

- **$H_0$:** no QTL. Under this null model, all phenotypic variation is assumed to be the sole result of random effects. The likelihood thus reduces to (Knott and Haley 1992)

$$L = \prod_{i=1}^{n} f(Y_i|\mu, \sigma).$$

Note that the mean of the phenotype under both $H_A$ and $H_0^2$, $\mu + a_g - d_g$, which is dependent on the QTL genotype, can be written as simply $\mu_g$, where the additive and dominance effects ($a_g$ and $d_g$, respectively) can be absorbed into $\mu_g$. This results in the more typical notation of mixture models as found in say McLachlan and Peel (2000) and will be used in the remainder of this work.

An important difference between the likelihood (4.1) and the likelihood given for $H_A$ by Knott and Haley (1992) is the use of the parameter $\omega_g|M_i$, which is the probability of QTL genotype $g$ given the marker genotype of individual $i$, as compared to the more traditional $\text{trans}(g|M_i)$ used by Knott and Haley (1992). While the transmission probability, $\text{trans}(g|M_i)$, is a function of the recombination fraction $r$ between the marker and QTL, the parameter $\omega_g|M_i$ is a function of the linkage disequilibrium $D$ between the marker and QTL, as well as the specific allele probabilities for the marker ($p$ and $1-p$) and QTL ($q$ and $1-q$). Thus, the difference between $\text{trans}(g|M_i)$ and $\omega_g|M_i$ is not only in the number of parameters involved, but also in the fact that $r$ relates to the genetic distance along the chromosome. This is in contrast to the linkage disequilibrium $D$, which is a direct measure of association between the marker and QTL, no matter their genetic loci. This separates the QTL detection problem from the typical requirements of an *a priori* genetic map that traditional QTL mapping is dependent upon. Such a scenario is advantageous to mapping QTL in natural segregating populations (Fu *et al.* 2013), but also performs well in traditional experimental crosses (Chapter 3). This work considers only the parameter $\omega_g|M_i$. 
The conditional probabilities $\omega_{g|M}$ are given in Table 4.1 where, as noted in Chapter 3, the values $p_{11}$, $p_{10}$, $p_{01}$, and $p_{00}$ are defined as

$$p_{11} = pq + D, \quad p_{10} = p(1 - q) - D, \quad p_{01} = (1 - p)q, \quad p_{00} = (1 - p)(1 - q) + D.$$ 

The case of $D = 0$ corresponds to no linkage between the marker and QTL, in which case the conditional probability $\omega_{g|M}$ reduces to $\omega_g$ as the QTL genotype is now independent of the marker genotype. Specifically, in the case of $D = 0$, the different QTL genotypes have probabilities $\omega_1 = (1 - q)^2$, $\omega_2 = q(1 - q)$, and $\omega_3 = q^2$, assuming a two-allele co-dominant QTL model. In light of this, the likelihoods for each of the hypotheses of interest in traditional QTL mapping (Knott and Haley 1992) are most appropriately written for the purposes of this work as follows.

- **$H_A$:** a linked QTL.

$$L(p, q, D, \mu_1, \ldots, \mu_G, \sigma|Y, M) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_{g|M}(p, q, D) f(Y_i|\mu_g, \sigma). \quad (4.2)$$

- **$H_0^2$:** an unlinked QTL.

$$L(q, \mu_1, \ldots, \mu_G, \sigma|Y) = \prod_{i=1}^{n} \sum_{g=1}^{G} \omega_g(q) f(Y_i|\mu_g, \sigma) \quad (4.3)$$

- **$H_0^1$:** no QTL.

$$L(\mu, \sigma|Y) = \prod_{i=1}^{n} f(Y_i|\mu, \sigma). \quad (4.4)$$

Knott and Haley (1992) suggest that “the evidence for a QTL is primarily obtained from differences in the mean effects of different marker genotypes, differences which will only be observed for QTL linked to markers being considered. This suggests that the use of ‘no QTL’ as the null hypothesis will not bias the results when an unlinked QTL is present.” Similarly, Lander and Botstein (1989) also use a ‘no QTL’ model for their null hypothesis. However, in Appendix 4 of their paper, Lander and Botstein (1989) state that the use
Table 4.1. The theoretical conditional probabilities of QTL genotype (columns) given marker genotype (rows).

<table>
<thead>
<tr>
<th></th>
<th>AA (g = 3)</th>
<th>Aa (g = 2)</th>
<th>aa (g = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>( \frac{p_1^2}{p^2} )</td>
<td>( \frac{2p_1p_0}{p^2} )</td>
<td>( \frac{p_0^2}{p^2} )</td>
</tr>
<tr>
<td>Mm</td>
<td>( \frac{2p_1p_0}{p(1-p)} )</td>
<td>( \frac{2(p_1p_0+p_1p_0)}{p(1-p)} )</td>
<td>( \frac{2p_0p_0}{p(1-p)} )</td>
</tr>
<tr>
<td>mm</td>
<td>( \frac{p_0^2}{(1-p)^2} )</td>
<td>( \frac{2p_0p_0}{(1-p)^2} )</td>
<td>( \frac{p_0^2}{(1-p)^2} )</td>
</tr>
</tbody>
</table>

of the null hypothesis \( H_0^2 \), an unlinked QTL, would be more appropriate than using the \( H_0^1 \) no QTL hypothesis, in the scenario of a segregating QTL with large effects. “When the phenotypic distribution is bimodal due to the segregation of a QTL with large effects somewhere in the genome, it is no longer possible to use a simple normal distribution as the null hypothesis. (The fit would be so bad that one would always reject the null hypothesis in favor of the presence of a QTL, even at positions unlinked to any QTL.)” This is why the work in Chapter 3 used the two hypothesis framework of testing first for the existence of a QTL, and then following up with a test for the linked QTL. While the approach in Chapter 3 is more comprehensive than the current literature, there are some important subtleties.

At first glance, it would appear that in Chapter 3, \( H_0^1 \) was tested first against \( H_A \), and if it was rejected, \( H_0^2 \) was tested and required to show significance before concluding \( H_A \). However, this is not exactly the case, and the subtle difference is important. Like the current literature, the first test in Chapter 3 (concerning the existence of a QTL) used the ‘no QTL’ hypothesis \( H_0^1 \) against the alternative of a ‘linked QTL’ (\( H_A \)). However, the second hypothesis test of linkage between marker and QTL, while not exactly \( H_0^2 \) against \( H_A \), performs a similar inspection concerning the linkage of the marker and QTL. The aim is to verify that the reason \( H_0^1 \) was rejected was because there is indeed evidence of the QTL being linked to the marker, and not just a QTL segregating in the genome (\( H_0^2 \)). This is done by testing the hypothesis that \( D = 0 \) in the model corresponding to \( H_A \) by means of the linkage test of Brown (1975). Even though the QTL has not technically been genotyped, predicted genotypes are obtained by virtue of the maximum likelihood estimates of \( q \) and \( D \) (\( \hat{q} \) and \( \hat{D} \), respectively). The genotyped marker also allows for a maximum likelihood
estimate $\hat{p}$ of $p$, the major marker allele frequency. Specifically, the hypotheses and test statistic of interest are given as (Brown 1975)

$$H_0^D: D = 0 \text{ vs } H_1^D: D \neq 0.$$  \hspace{1cm} (4.5)

$$X_D^2 = \frac{n\hat{D}^2}{\hat{p}(1 - \hat{p})\hat{q}(1 - \hat{q})} \sim \chi_1^2$$  \hspace{1cm} (4.6)

The linkage test of Brown (1975) was originally designed to test for linkage between two genotyped markers. When the test is performed as designed, on two genotyped markers, the corresponding test statistic follows asymptotically a $\chi_1^2$ distribution. However, as suggested in Luo et al. (2000), who studied a very similar likelihood based statistic to that of $X_D$, there is evidence that the distribution of their test statistic in the context of a single genotyped marker and unknown QTL is not that of a $\chi_1^2$. Luo et al. (2000) state that “it is not clear what the distribution of the likelihood ratio test statistic is under the null hypothesis” and they suggest using permutations (Churchill and Doerge 1994) to obtain the critical thresholds for the test statistic. In our own simulation study we found evidence that $X_D^2$, when applied to a genotyped marker and unknown QTL, may be distributed as $\chi_4^2$ under the null hypothesis $H_0^2$. (Figure 4.1 depicts this result and the details of the simulation are provided at the end of this section.) However, there is currently no theoretical support for such a conclusion, and this result differs from Luo et al. (2000) whose results are more suggestive of a $\chi_2^2$ distribution, but with too great a variance to reach definitive conclusions. Interestingly, the likelihood ratio for the model of Luo et al. (2000), under the asymptotic theory of Wilks (1938), would be distributed as a $\chi_2^2$.

The null and alternative hypotheses corresponding to the QTL existence test of Section 3.3.2 were given as

$$H_0^L: \mu_1 = \mu_2 = \mu_3 \equiv \mu \text{ vs }$$

$$H_1^L: \text{one of the equalities above does not hold.}$$
In can be seen that under $H_0^L$ the likelihood is given by that of $H_0^1$, i.e., Likelihood (4.4).

This follows from the fact that under $H_0^L$, $f(Y_i|\mu_g, \sigma) = f(Y_i|\mu, \sigma)$ for $g = 1, \ldots, G$ which then eliminates $p, q,$ and $D$ from the model as $\sum_{g=1}^{G} \omega_{g|M}(p, q, D) = 1$. Hence, so far as the likelihood is concerned, $H_0^L$ is identical to the traditional hypothesis $H_0^1$. As for the alternative $H_1^L$, which suggests only a separate means model, the likelihood could be either of Likelihood (4.3) or Likelihood (4.2), corresponding to $H_0^2$ (an unlinked QTL) or $H_A$ (a linked QTL), respectively. While the likelihood corresponding to $H_0^2$ may appear to be the best match for $H_1^L$, that of $H_A$ is more appropriate for the two stage testing approach of Chapter 3, and was the likelihood of choice in that chapter as shown in Likelihood (4.1).

This allows for what might be termed as a ‘two steps forward and one step back’ testing approach to mapping QTL. In the first test, as is traditionally done, $H_0^1$ is tested against $H_A$, with significance concluding that there was sufficient evidence to discredit the ‘no QTL’ hypothesis. The follow up test, $H_0^D$ of Brown (1975), then allows to step back to $H_0^2$, and although not directly through a likelihood ratio test, tests for more conclusive evidence against linkage equilibrium ($D = 0$) before officially concluding $H_A$. The advantage of this approach is that both tests can be performed through a single application of the EM algorithm. This is computationally more efficient than performing a separate run of the EM algorithm to estimate the model parameters of $H_0^2$ as well.

Under traditional likelihood theory (Wilks 1938), the likelihood ratio test statistic (LRTS) used to test $H_0^1$ against $H_A$ would be distributed asymptotically as a $\chi_5^2$ random variable. This is because the difference in free parameters between the null ($\mu, \sigma$) and alternative models ($\mu_1, \mu_2, \mu_3, \sigma, p, q, D$) is five. Churchill and Doerge (1994) critique using the asymptotic chi-square distribution for this LRTS stating, “In most cases, the regularity conditions that ensure an asymptotic chi-square distribution for the likelihood ratio test statistic are not satisfied.” They propose using permutations in place of the asymptotic chi-square distribution to obtain critical thresholds. This approach has become very popular in the QTL mapping literature, as witnessed by the fact that currently their paper has been cited over 3,800 times. Further clarifying the difficulty with asymptotic chi-square
distributions, Doerge (2002) states, “Because the likelihood is usually a function of mixtures of (normal) distributions and, when maximized under both the null and alternative hypotheses, leads to test statistics that fail to follow standard statistical distributions, it is difficult to declare a QTL with confidence.”

Generally speaking, the failure of the likelihood ratio statistic to follow standard asymptotic distributions when testing a mixture model against a model of homogeneity has been well established in the literature (Chen et al. 2001; Cheng and Traylor 1995; Ghosh and Sen 1985; McLachlan and Peel 2000; Self and Liang 1987; Titterington et al. 1985). However, these results apply directly only to testing the hypothesis of no QTL, $H_0^1$, i.e., homogeneity, against the hypothesis of an unlinked QTL, $H_0^2$, i.e., a mixture alternative, which we never perform. While it is true that $H_A$ is also a mixture model, none of the mixture model literature has dealt explicitly with the likelihoods for the testing of $H_0^1$ against the alternative of the form $H_A$. This scenario differs from the standard mixture models (like that corresponding to $H_0^2$) as the mixture proportions $\omega_{g|M_i}$ are conditional upon each individual’s phenotype. The asymptotic behavior of the likelihood ratio statistic corresponding to the testing of $H_0^1$ to $H_A$ has thus, to our knowledge, not been explicitly explored theoretically for single-marker mapping. (This differs from interval mapping approaches where the topic has been explored more thoroughly, see Rebai et al. (1994) for details.) On open problem, deserving of future research, would be in extending the Davies Approximation (Davies 1987), that is calculated in Rebai et al. (1994) for Interval Mapping, to the case of single-marker mapping that is the emphasis of the model of Fu et al. (2013). In this work, after demonstrating through our own simulation study the failure of the test statistics to follow the asymptotic distribution of Wilks (1938), we take the approach of applying a bivariate simulation based null distribution to determine appropriate joint thresholds for the test statistics of interest.

A simulation study of Knott and Haley (1992) (the only study we have been able to find bearing any evidence on the asymptotic distribution of the LRTS for the single-marker case) leaves doubt as to the validity of applying the asymptotic thoery (Wilks 1938) to
testing $H_0^1$ against $H_A$. Commenting on this study, Churchill and Doerge (1994) state, “Their results suggest that the chi-square approximation to the distribution of likelihood ratio test statistics is not reliable in many cases and is at least questionable in every case.”

Knott and Haley (1992) simulated an F2 intercross to determine the goodness of fit of the empirical distribution of the test statistics to the corresponding theoretical asymptotic distributions. (It is important to note as we did before, that they used likelihoods containing the recombination fraction $r$ rather than the linkage disequilibrium parameters $p, q,$ and $D$ that we use.) They determined that for testing $H_0^1$, no QTL, against the alternative $H_A$, a linked QTL, that there was insufficient evidence to reject the hypothesis that the corresponding likelihood ratio test statistic was distributed as $\chi^2_\nu$, with $\nu$ as designated by the theoretical results from Wilks (1938). Quoting Knott and Haley (1992), “With a single marker, comparing linked versus no QTL the mean and variance of the statistics are higher than expected and likewise the number significant, but nonetheless the test statistic distribution is very similar to a $\chi^2$.” Lander and Botstein (1989) also accept the asymptotic theory of Wilks (1938) for a Backcross design when they refer to “a general result about maximum likelihood estimation in large samples” concerning the distribution of the corresponding likelihood ratio statistic for testing $H_0^1$ against $H_A$. Further, even Churchill and Doerge (1994), in discussing the results of their simulation study used to demonstrate the performance of permutations for testing $H_0^1$ against $H_A$, state, “Note that the comparisonwise values are fairly constant throughout the entire genome, and agree fairly well with the threshold values based upon a chi-square distribution.”

We performed a simulation study based on that of Knott and Haley (1992) in order to assess the distribution of the likelihood ratio statistic under the null hypothesis of no QTL ($H_0^1$) against the alternative of a linked QTL ($H_A$). Data was also generated under $H_0^2$ to explore the distribution of $\chi^2_D$ for the test of $H_0^D$ under both $H_0^1$ and $H_0^2$. A phenotype $Y \sim N(0,1)$ and single marker $M$ with major allele probability of $p = 0.5$ were generated independently, 1,000 times under a sample size of $n = 100$ (Knott and Haley (1992) used $n = 1,000$ in their simulation). Note here that $Y = (Y_1, \ldots, Y_n)$ and $M = (M_1, \ldots, M_n)$. 
Fig. 4.1. A) Demonstration of the empirical cumulative density function (blue dots) for the Likelihood Ratio Test of $H_0^1$ against $H_A$ for synthetic data simulated under the null hypothesis of no QTL, $H_0^4$. B) The empirical cumulative density functions corresponding to the test of $D = 0$ for (i) synthetic marker and QTL data where both genotypes are known (large gray dots) which is consistent with a $\chi^2_1$ distribution, (ii) synthetic data simulated under $H_0^2$ (blue dots), and (iii) synthetic data simulated under $H_0^1$ (small gray dots). In both panels A and B, the cumulative density functions (underlaid in gray) for several $\chi^2_\nu$ distributions are also plotted for $\nu = 1, \ldots, 10$ for reference.

Likelihood ratio test statistics (LRTS) were calculated for each simulated pair of $Y$ and $M$ using the alternative hypothesis $H_A$ of a linked QTL with $G = 3$ separate genotypes, representing an F2 intercross. The test of $H_0^D$ was also performed in each case, providing values for $\chi^2_D$. Also, a separate simulation was conducted where the marker and QTL genotypes were known exactly and only a test of $H_0^D$ was performed to explore the distribution of $\chi^2_D$ under the original setting of the test (Brown 1975).

The empirical cumulative density function of the LRTS obtained from the simulation are shown in Figure 4.1A overlaid on several $\chi^2_\nu$ cumulative density functions for $\nu =$
Our study provides similar results to those of Knott and Haley (1992), but with more pronounced evidence that the distribution of the LRTS under the null hypothesis $H_0^1$ differs from the theoretical asymptotic result ($\chi^2_5$) of Wilks (1938). For example, the mean, variance, and 95th quantile of the 1,000 LRTS obtained from the simulation were respectively 6.9, 28.6, and 17.8. The 95th quantile of the $\chi^2_5$ distribution is roughly 11.1 (with a mean and variance of 5 and 10, respectively) so that far more than 5% of the LRTS corresponding to the data simulated under the null hypothesis $H_0^1$ would be in the critical region (roughly 16.5%) when using the $\chi^2_5$ distribution to obtain the threshold.

As for the test of $H_0^D$ (Figure 4.1B), our simulations suggest that the $\chi^2_4$ assumption appears to be correct in the case that both the marker and QTL genotypes are known (Brown 1975). However, when the QTL genotype is unknown (as is always the case in QTL mapping) the $\chi^2_4$ assumption appears to no longer be appropriate, which is consistent with the simulation study of Knott and Haley (1992) who found overwhelming evidence that the likelihood based version of this test ($H_0^A$ against $H_A$) using the recombination fraction $r$ instead of linkage disequilibrium $D$ was not distributed as a $\chi^2_4$. These results were confirmed later on for the likelihood version of the test using the linkage disequilibrium $D$ (Luo et al. 2000). Interestingly, the distribution of $\chi^2_4$ under our simulation study appears to be somewhat consistent with a $\chi^2_4$ when the data is generated under $H_0^2$, see Figure 4.1B. However there is currently no theoretical reason that this should be the case. On the other hand, it is clear that the test is extremely poorly behaved when the data is generated under $H_0^3$, emphasizing the importance of the structured GBA approach of Chapter 3 so that $H_0^D$ is tested only when it is safe to assume that either of $H_0^2$ or $H_A$ holds. In any case, it is evident that the test statistics fail to follow the suggested asymptotic $\chi^2$ distributions suggested by Wilks (1938) under either $H_0^3$ or $H_0^2$.

### 4.3 Comparison of Univariate Approaches

As stated previously, the theoretical asymptotic sampling distributions suggested by Wilks (1938) and Brown (1975) were used in Chapter 3 to obtain $P$-values for the test of each of the two hypotheses of interest: $H_0^1$, no QTL, and $H_0^2$, an unlinked QTL, against
the alternative of a linked QTL, $H_A$. While the asymptotic theory is the simplest method for computing raw $P$-values for these hypothesis tests, the violation of regularity conditions for the mixture model likelihoods leads to doubt as to their validity. Fortunately, other more robust approaches have been suggested in the literature. Among these, permutations (Churchill and Doerge 1994) and simulations (Lander and Botstein 1989) were selected for consideration as they are by far the most popular approaches. We also consider a large-scale testing approach used to empirically approximate the null distribution (Efron 2004). Originally introduced for genomics and image processing applications, the empirical approach can be adapted to any large-scale testing scenario, including QTL mapping, as we demonstrate below.

The classic simulation from Churchill and Doerge (1994) was recreated here to study the performance of each of these methods for the computation of the raw $P$-values in testing $H_0^L$ (or, equivalently $H_0^L$) and $H_0^P$. The simulation was performed using the \texttt{sim.map} and \texttt{sim.cross} functions of the R/qtl package (Broman et al. 2003). As in Churchill and Doerge (1994), four chromosomes were simulated under a sample size of $n = 100$ individuals, with the first and third chromosomes having 50 markers each and the second and fourth having 10 markers each. All chromosomes were assigned a length of 100 cM. Two QTL were simulated, one on the first chromosome at 44.4 cM (from the left end) and the other on the second chromosome at 61.6 cM (from the left end). The first QTL was given an additive effect of 0.75 ($\sigma = 1$) and the second an additive effect of 1 ($\sigma = 1$). The EM algorithm was applied to the simulated data as described in Fu et al. (2013) and raw (unadjusted) $P$-values obtained by each of the following methods. The comparison of the results follows the description of each of the methods.

4.3.1 Permutations

Perhaps the most widely used approach to identifying markers linked to a QTL ($H_0^L$ against $H_A$) is the permutation method of Churchill and Doerge (1994). The method is robust and, as stated by Cheverud (2001), has several advantages. "Primary among these is that it draws the threshold directly from the data being analysed. Peculiarities of the
observed data, such as deviations of the phenotype from a normal distribution, biased allele frequencies, and patterns of missing data are maintained in the permuted data sets and are included in estimation of the thresholds obtained." Further, the permutation approach avoids most assumptions about the null distribution from which the data was generated, requiring only the assumption of exchangeability under the null hypothesis (Churchill and Doerge 1994). While computationally expensive, the method is implemented relatively simply. After performing the initial analysis, the values of the phenotype $Y$ are permuted while leaving the marker values fixed for each individual. The maximum likelihood analysis is rerun on the permuted values (again using the EM algorithm) and then the process is repeated, typically either 1,000 or 10,000 times depending on the desired level of significance.

$P$-values for the original test statistics are obtained by calculating the percent of permuted values more extreme than those observed. If the observed data is consistent with the null hypothesis, then the test statistics from the analyses on the permuted phenotypes will be similar in value to those computed on the original data. As an alternative to $P$-values, which are computationally intensive to obtain even for just a few significant digits, significance thresholds for the LRTS are typically obtained. This is done on either the chromosome or whole genome level by taking the 95th quantile of the distribution of the maximum LRTS for each permutation, where the maximum is computed either for all markers on the chromosome or all markers on the genome.

4.3.2 Simulation

McLachlan (1987), Lander and Botstein (1989), and van Ooijen (1999) support generating a null distribution using simulated data consistent with the model assumptions of the null hypothesis $H_0$, no QTL. Doerge (2002) comments that such an approach is "indeed useful if the model used to simulate the data is the true model. However, the model rarely describes the complicated relationships that occur in the genome." However, van Ooijen (1999) demonstrates that the thresholds from permutations and simulations are very similar. The method is performed by first obtaining the maximum likelihood estimates for the parameters under the null hypothesis. Then, similar to permutations, the analysis is
performed first for the observed data, and then for the phenotype $Y$ simulated under the null hypothesis using the maximum likelihood estimates of the corresponding parameters obtained previously. This process is repeated, say 1,000 times, and $P$-values for the observed test statistics are obtained as the percent of simulated test statistics more extreme than those observed. The computation time for this approach is thus comparable to the permutation method, but utilizes the full distributional assumptions of the null hypothesis (Doerge 2002). Significance thresholds instead of $P$-values can be computed for each chromosome or the entire genome in the same manner as for permutations.

### 4.3.3 Large-Sample Empirical

In writing generally about large-scale testing scenarios, Efron (2004) states that, "Permutation methods are popular... as a way of avoiding assumptions and approximations... but they do not automatically resolve the question of an appropriate null [distribution]." He suggests further that even when individual test statistics can be assumed to behave according to their classical one-sample theoretical distributions, the same does not typically hold for the multiple testing case. Efron (2004) suggests resolving the issue by using the data to calculate an empirical null distribution and demonstrates the method with two applications, one to genomics and another to image processing. Hence, while not developed specifically for QTL mapping, the large-scale setting of QTL mapping makes itself amenable to the method.

The empirical approach to estimating the null distribution (Efron 2004) utilizes the theoretical one-sample distribution to obtain $P$-values for the observed test statistics. Then, the $P$-values are transformed to $z$-scores, and a kernel density is fit to a counts histogram of those values using Poisson regression. Applying some calculus to the main peak of the data, he fits a normal distribution to the main portion of the data which is then implemented as the empirical null hypothesis (details can be found in Remark D of Efron (2004)). Finally, $P$-values are newly obtained according to this distribution. An important assumption of the method is that only a small percentage of the data are in truth interesting, i.e. alternatively expressed. Efron (2004) suggests no more than 10% of the data being non-null for the
method to work. It is also important to have more than say 100 observed values, with the method being more useful the greater the number of (multiple) tests performed.

4.3.4 Simulation Results and Discussion

Figure 4.2 demonstrates the results of the Churchill and Doerge (1994) simulation for each of the permutation, simulation, and theoretical approaches. (The results of the empirical approach (Efron 2004) were not included in this figure as the method failed to identify any markers as significantly linked with a QTL.) In keeping with the tradition of QTL mapping (Lander and Botstein 1989) Figure 4.2 shows the LOD (log₁₀ of the likelihood ratio) across each marker and each chromosome for the test of \( H_0 \) (solid black line). LOD scores greater than the significance threshold identified by each of the permutation, simulation, and theoretical approaches show the regions on each chromosome (if any) where the respective test identified markers as significantly linked to a QTL. All but the empirical approach correctly identified the simulated QTL on Chromosomes 1 and 2. It is not surprising that the simulation approach showed the narrowest regions (highest threshold) surrounding the true locations of the synthetic QTL since the data were generated under the precise conditions assumed by the simulation method. Interestingly, the permutation, simulation, and theoretical approaches agree rather well on their selected thresholds, despite the mentioned difficulties of the theoretical asymptotic distributions. This is confirmed in Churchill and Doerge (1994) when they state, in speaking of their own simulation results, “the [obtained permutation threshold] is slightly greater than the chi-square critical value.”

To explore more deeply the performance of each method, it is worth considering the \( P \)-values provided under each method. Panel (a) of Figure 4.3 shows the behavior of the \(- \log_{10} \) of the \( P \)-values corresponding to the testing of \( H_0 \). The results are quite comparable between the permutation, simulation, empirical, and theoretical approaches. (A technical artifact is present in the graphs of panel (a), where for both the simulation and permutation approaches any \( P \)-value with a \(- \log_{10} \) value greater than 3 is identically zero. This is due to the 1,000 (i.e., \( 10^3 \)) replications of both the theoretical and permutation approaches. To allow for the visualization of the \(- \log_{10} \) transformation, a value of \( 10^{-10} \) was first added...
Fig. 4.2. The resulting genome-wise significance thresholds for the LOD (log_{10} of the likelihood ratio, solid black line) according to each of the simulation, permutation, and theoretical approaches. The empirical approach did not identify any markers as supporting a linked QTL and was thus not included in the figure. Triangles demarcate the simulated QTL at 44.4 cM on the first chromosome and at 61.6 cM on the second chromosome.

...to all P-values for all methods.) All approaches capture the significant QTL located on Chromosome 1 and 2 with the highest peaks in the graph near those areas.

Especially enlightening are the plots in panel (b) of Figure 4.3, which visualize the P-values (−log_{10} scale) corresponding to the test of \( H_5 \). These values were not used in the computation of the threshold for both the permutation and simulation approaches in keeping with their traditional applications to just \( H_{01} \) against \( H_A \) (Churchill and Doerge 1994; Lander and Botstein 1989). The theoretically obtained P-values shown in panel (b) show extreme significance for many markers, including markers that are not near any QTL. On the other hand, the other methods fail to detect any markers showing a significantly linked QTL under the test of \( H_{01}^2 \).

The tentative conclusion here is that the test of \( H_{02}^2 \) against \( H_A \) is currently providing very little extra insight as to which markers are significantly linked to a QTL. In fact, if the standard Holm adjustment (Holm 1979) is applied to the P-values resulting from the theoretical approach to testing \( H_{01}^1 \) against \( H_A \), the obtained significance threshold is nearly identical to that obtained from the graphical Bonferroni adjustment (GBA) of Chapter 3. This can be seen from the fact that in each case considered in this simulation study, \( H_{01}^1 \) is rejected in favor of \( H_A \) whenever \( H_{01}^1 \) was first rejected. All this suggests that the test of
Fig. 4.3. (a) Graphical demonstration of the $-\log_{10}$ of the $P$-values corresponding to the testing of $H_0^1$ for each of the four univariate methods under consideration: permutations, simulations, empirical, and theoretical. A technical artifact is present in both the permutation and simulation graphs, where to allow the log transformation of zero (occurring for both the permutation and simulation methods), the value of $10^{-10}$ was first added to all raw $P$-values. (b) The $-\log_{10}$ of the $P$-values corresponding to the test of $H_0^2$.

$H_0^2$ is poorly behaved, so that under any of these univariate approaches there is relatively little added benefit to considering this secondary test.

The following section introduces a bivariate approach which will allow consideration of both the test statistics obtained for testing each of $H_0^b$ and $H_0^D$. This approach allows for more information to be gleaned from the secondary test of $H_0^D$, thus focusing in on the true QTL. The bivariate approach results in a single joint $P$-value for each SNP, and thus supports the use of a single Bonferroni-Holm correction as in the IUT approach, but does not require a union hypothesis, an advantage similar to that of the GBA approach. Further, the computational burden, while more than the theoretical asymptotic distributions, is
significantly less (1,000 or 10,000 times less) than either of the simulation or permutation methods.

4.4 Null Kernel P-value Method

Let $T$ and $U$ denote two test statistics of interest, not necessarily independent, that bear on a joint hypothesis $H_0$. In other words, $H_0$ is rejected only if the joint value of $T$ and $U$ shows significance. (Such a scenario occurs in the two hypothesis test approach to QTL mapping as detailed in Section 4.2.) The unknown joint sampling distribution of $T$ and $U$ can be approximated in a non-parametric manner by the following technique, which we call the Null Kernel method.

1. Simulate $s$ data sets, each of size $n$, based on the model assumptions of the union hypothesis $H_0$.

2. Calculate $T_i$ and $U_i$ for $i = 1, \ldots, s$.

3. Estimate the joint density $\hat{f}$ of $T$ and $U$ using a kernel density estimation technique on the $T_i$ and $U_i$.

4. Compute the cdf $\hat{F}$ of $\hat{f}$ by $\hat{F}(c) = \int_{A(c)} \hat{f}$, where $A(c) = \{(t, u)|f(t, u) \geq c\}$.

5. The joint $p$-value for the calculated statistics $\hat{t}$ and $\hat{u}$ can then be obtained by the formula $p = 1 - \hat{F}(\hat{f}(\hat{t}, \hat{u}))$.

The $p$-value obtained in this manner thus represents the probability under $H_0$, as estimated by $\hat{f}$, that an observed joint test statistic $(t, u)$ would be more extreme than, or less likely to occur than, $(\hat{t}, \hat{u})$.

4.4.1 Location Testing for a Bivariate Normal

We first explored the performance of the Null Kernel approach on the test of location for the bivariate normal distribution as compared to the well established Hotelling's $T^2$ statistic.
(Johnson and Wichern 2002). Hotelling’s $T^2$ statistic (for two dimensions) is calculated by

$$T^2 = n(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0)$$

(4.7)

where $\mathbf{X} = (X_1, X_2)$, $\mu_0 = (\mu_{01}, \mu_{02})$ represents the hypothesized mean of the bivariate normal distribution, and $S$ denotes the sample variance-covariance matrix. The main contribution of Hotelling is in demonstrating that the distribution of $T^2$ under $H_0$ is $2(n - 1)/(n - 2)F_{2,n-2}$, where $F_{2,n-2}$ is the $F$-distribution with 2 numerator and $n - 2$ denominator degrees of freedom. This provides the corresponding $P$-value as $p = P(T^2 > 2(n - 1)/(n - 2)F_{2,n-2})$ (Johnson and Wichern 2002).

To compare the Hotelling and Null Kernel methods, 100 simulations of a test of location for a sample size of $n = 20$ were performed. In each simulation, the data were generated from a bivariate normal distribution with mean $\mu = (u, v)$ with $U, V \sim \text{Unif}(0,1)$ and variance-covariance matrix $\Sigma = ((1, 0.3)', (0.3, 1)')$. The null hypothesis $H_0 : \mu = (0, 0)$ was tested in each case against the alternative $H_A : \mu \neq (0, 0)$.

$P$-values under the Null Kernel approach were obtained for this comparison study by the five steps introduced in the previous section, as detailed below.

1. $s = 1,000$ data sets were simulated, each of size $n = 20$, based on the model assumptions of the hypothesis $H_0 : \mu = (0, 0)$, with the sample variance-covariance matrix $S$ computed from the data.

2. The test statistics $T_i = \sqrt{n}(\bar{x}_1 - \mu_1)/S_{11}$ and $U_i = \sqrt{n}(\bar{x}_2 - \mu_2)/S_{22}$ were calculated for $i = 1, \ldots, s$.

3. The joint density $\hat{f}$ of $T$ and $U$ was estimated using a kernel density estimation technique (bivariate.density of the sparr package (Davies et al. 2011) in R (R Core Team 2013)) on the $T_i$ and $U_i$ where the tuning parameter was selected so that the size of the test was maximized while still being less than or equal to $\alpha = 0.05$.

4. The CDF $\hat{F}$ of $\hat{f}$ was computed by $\hat{F}(c) = \int_{A(c)} \hat{f}$, where $A(c) = \{(t, u)|\hat{f}(t, u) \geq c\}$. 


Fig. 4.4. Visualization of the Null Kernel method as applied to a sample of 1,000 $T$ and $U$ statistics simulated under the bivariate normal null distribution with zero mean, unit variances, and covariances of 0.3. The contours of the estimated null density $\hat{f}$ are superimposed for $1 - \hat{F}$ values of 0.05, 0.01, 0.001, and 0.0001.

5. The joint p-value for the calculated statistics $\hat{t}$ and $\hat{u}$ were then obtained by the formula $p = 1 - \hat{F}(\hat{f}(\hat{t}, \hat{u}))$.

A total of 100 $P$-values were obtained (in pairs) for both Hotelling's $T^2$ statistic and the Null Kernel method. The Null Kernel $P$-values were obtained by comparing the observed $\hat{t}$ and $\hat{u}$ to an estimated $\hat{F}$, which differed slightly for each simulation. An example of the Null Kernel null density $\hat{f}$ for 1,000 $T$ and $U$ statistics with contours corresponding to various values of $1 - \hat{F}$ are plotted in Figure 4.4. Figure 4.5 shows how the $P$-values from the Hotelling and Null Kernel methods compare. The horizontal and vertical lines are drawn at the critical threshold $-\log_{10}(0.05)$ so that the resulting Quadrants I and III depict regions of agreement between the methods while Quadrants II and IV show regions of discord. Quadrant I, showing tests declared significant by both methods (i.e., the power of the methods) contains 79 of the 100 points. Quadrant III, showing tests where the null hypothesis was retained (Type II Errors) by both methods, contains 19 of the 100 points. Thus, for 98 of the 100 points the methods agree on their testing decisions. The remaining discordant points are in Quadrant II, containing 1 point, and Quadrant IV, containing 1 point.

An advantageous property of the Null Kernel method is the rapid change in the magnitude of the $P$-values from marginally significant to extremely significant over just a short
Fig. 4.5. Comparison of $P$-values ($-\log_{10}(p)$) obtained from either the Null Kernel method or Hotelling’s $T^2$ test. The four quadrants, I, II, III, and IV demonstrate the regions of agreement (I and III) and discord (II and IV) between the two methods.

distance into the critical region. This property is apparent in Figure 4.5 where the points leave the line of equality ($y = x$) due to the more extreme values of the Null Kernel method. These more extreme values lead to advantages in the multiple testing framework where, in exchange for greater protection against Type I Errors, marginally significant $P$-values are often made non-significant after adjustment for all simultaneous tests. Another advantage of the Null Kernel method is that it extends easily to cases where the null sampling distribution is much more difficult (or impossible) to obtain analytically, as in the case of the QTL mapping hypothesis $H_5^2$, presented in Section 4.2.

It is important to note that the smoothing parameter, i.e., the bandwidth, used to calculate $\hat{f}$ has a large impact on the magnitudes of the resulting $P$-values for the Null Kernel approach. To remedy the arbitrary selection of this tuning parameter, we suggest using the parameter to define the size of the critical region of $\hat{f}$ to be $\alpha$. At times an exact $\alpha$-level test is not possible under this approach. In this case, a unique bandwidth can still be obtained by maximizing the size of the test such that the level is still less than $\alpha$.

**Type I Error Control**

A second simulation analyzing the Type I Error rate of the Null Kernel method was performed similar to the first, except that the data were generated under the null model, i.e., $\mu = (0,0)$. All other parameters were as in the previous section. Figure 4.6 depicts the
Fig. 4.6. The $-\log_{10}$ of the $P$-values from the Null Kernel and Hotelling's $T^2$ methods for data simulated consistent with the null hypothesis. Both methods properly control the Type I Error with less than 5% of the data in the critical region. Specifically, the Null Kernel method shows a 4% Type I Error rate while Hotelling's $T^2$ method shows a 3% rate.

results which demonstrate that the Null Kernel method properly controls the Type I Error rate for this simulation study.

4.4.2 QTL Mapping Simulation Revisited

The main reason for developing the Null Kernel method was for its application to the two hypothesis QTL mapping approach presented in Section 3.3.2 in an effort to make better use of the information contained in the second hypothesis test of linkage between the marker and QTL. Hence, we explore the performance of the Null Kernel method on the simulated QTL mapping data of Section 4.3. The results are compared to those established in Section 4.3 for the univariate permutation, simulation, and theoretical approaches.

To generate the null kernel density $\hat{f}$ under the hypothesis $H_0^1$, that there are no QTL segregating in the genome, 1,000 simulations of a randomly generated phenotype ($Y_i \sim N(0,1)$) and independently generated markers ($p = 0.5$) were generated with a sample size of $n = 100$ corresponding to the actual study design. The EM algorithm was then applied to each simulated marker-phenotype pair to determine the maximum likelihood estimates of all parameters under both the null, $H_0^1$, and alternative, $H_A$. These maximum likelihood estimates provided for the calculation of the two test statistics of interest, the LRTS testing $H_0^1$ against $H_A$, and the $\chi^2_D$ statistic testing the hypothesis of no linkage.
between marker and QTL, $H_0^D : D = 0$. The resulting 1,000 pairs of test statistics were used as the basis for the simulated bivariate distribution. The density $\hat{f}$ was fit to these bivariate data, and $P$-values calculated according to the Null Kernel approach. The level curves of the Null Kernel estimated null, $\hat{f}$, with the 1,000 simulated values of $\chi^2_D$ and the LRTS ($\chi^2_L$) underlaid are shown in Figure 4.7. The resulting $P$-values were adjusted using Holm's procedure (Holm 1979), which is briefly explained in Section 1.3.2, to control (strongly) the probability of any Type I Errors at $\alpha = 0.05$, i.e., the FWER.

The $-\log_{10}$ of the adjusted $P$-values obtained from the Null Kernel method are plotted in Figure 4.8. Also included in Figure 4.8 are the adjusted $P$-values computed previously in Section 4.3 for the same synthetic QTL mapping data. The Null Kernel approach provides for tighter regions surrounding the true locations of the QTL on both Chromosome 1 (44.4 cM) and Chromosome 2 (61.6 cM) than any of the univariate methods. As mentioned previously, the Null Kernel approach does well at contrasting between significant and non-significant markers. This can be seen by the frequency with which the $-\log_{10}$ of the adjusted $P$-values are zero (corresponding to an adjusted $P$-value of 1). The only places where the $-\log_{10}$ of the Null Kernel adjusted $P$-values are greater than zero are in the regions immediately surrounding the true locations of the QTL. While the Theoretical approach shares this property to some degree, it is not as pronounced as in the Null Kernel method. Neither of the permutation or simulation approaches manifest this property. The
Fig. 4.8. The panels above compare the resulting adjusted *P*-values from each of the permutation, simulation, and theoretical approaches (Section 4.3) against the results of the Null Kernel method for the simulated QTL mapping data of Section 4.3. For ease of reference, the results for each method are highlighted individually in their own set of panels for each Chromosome (1-4), as labeled in the panel corresponding to Chromosome 4. The rug plots along the bottom of each panel show the locations of the synthetic markers.

Computation time of the Null Kernel approach was greater than that of the theoretical and empirical approaches, but substantially faster than either of the simulation and permutation approaches due to the ability to simulate all the data under the null hypothesis, i.e., both marker and phenotype, instead of using the actual marker data as in both the simulation and permutation methods.
4.4.3 Mouse HDL Cholesterol QTL Mapping Revisited

Returning to the mice HDL QTL mapping data of Section 3.4.3, we explore the performance of the Null Kernel approach on real data. Recall that these data (publicly available at http://cbd.jax.org/datasets/datasets.shtml) contain 44,428 distinct SNPs spanning all 19 autosomal chromosomes and the X chromosome of the mouse genome for 288 individual outbred mice. The mice were obtained from the Naval Medical Research Institute (NMRI). Measurements of High Density Lipoprotein (HDL) cholesterol were obtained for each mouse with the intention of mapping QTL responsible for HDL. A summary of the study design and measurement details can be found in Section 3.4.3.

To perform the Null Kernel method, the null density $\hat{f}$ was generated similarly to the previous Section (4.4.2), but under a sample size of $n = 288$. Hence, despite the fact that this study contained 44,428 SNPs, the computation time of the Null Kernel approach was similar to the previous section where there were just 120 markers (SNPs). The $P$-values resulting from the Null Kernel approach on these data were adjusted with the Holm adjustment to control the FWER at the $\alpha = 0.05$ level. The negative log of the adjusted $P$-values is plotted in Figure 4.9 for each SNP. As in Section 3.4.3, there is a strong signal on Chromosome 1 at the 173 Mb position (172.9 Mb to 173.7 Mb) and on Chromosome 5 at the 125 Mb position (124.5 Mb to 125.8 Mb). Other significant results were located at Chromosome 5 at 79 Mb and 122 Mb, Chromosome 6 at 20.1 Mb, and Chromosome 15 at 78.4 Mb, however these are potentially loci exhibiting linkage with the true QTL rather than representing independent QTL.

Figure 4.10 shows the bivariate view of the two test statistics, the LRTS and $\chi_D^2$, for each of the 44,428 SNPs analyzed for mice HDL QTL data. The black nodes in this plot demonstrate those SNPs which were found significant after the Holm adjustment. For comparison, the quartile to the upper right of the two dashed lines shows those SNPs (points) which were identified as significant under the GBA approach of Chapter 3. It is interesting to note that the QTL detected by the Null Kernel method on Chromosomes 6 and 15 and Chromosome 5 at the 79 and 122-Mb positions form the collection of points in
Fig. 4.9. The negative log of the Holm adjusted $P$-values for the Null Kernel approach applied to the 44,428 SNPs from the mice HDL QTL mapping study.

Figure 4.10 with $\chi^2_D$ values greater than 250 but with relatively small LRTS values ($\chi^2_L$). This may suggest that these SNPs are linked to true QTLs, in other words, locations along the genome which appear to be QTL but are in fact loci exhibiting strong linkage with the real QTL. Such a conclusion would be consistent with those of the original study on these data (Zhang et al. 2012) where it was determined, after several extensive analyses, that only the two major QTL on Chromosomes 1 and 5 were unarguably real QTL, and that other locations were linked to these QTL. Despite such evidence, their strong degree of linkage would certainly require further analysis before any definitive conclusions could be reached as it could well be possible that these are true QTL with only moderate effects on the phenotype Luo et al. (2000). It should be noted that those points which appear distinctly separated from the main body of data, but were not identified by the Null Kernel approach as significant, had Null Kernel adjusted $P$-values smaller than one, but greater than the cut-off of 0.05.

In comparing the significance results of the GBA and Null Kernel approaches, it is evident that the GBA favors consideration of $H^I_0$ more so than does the Null Kernel. This is shown in Figure 4.10 by the rejection of smaller $\chi^2_L$ values by the GBA than by the Null Kernel. This is consistent with the theoretical basis of the GBA approach, which considers $H^I_0$ as the primary test, and does not consider $H^D_0$ unless the primary test is first significant. This hierarchical approach was established to preserve the identifiability of parameters under $H^D_0$. Hence, the GBA would not allow for the rejection of the cluster of values with extremely large $\chi^2_D$ values, but small $\chi^2_L$ values that were rejected by the
Fig. 4.10. The joint plot of the observed test statistics for the mouse HDL QTL mapping data. Black points denote SNPs that were found significant after the Null Kernel $P$-values were adjusted for multiplicity using the Holm adjustment. For comparison, all points in the upper right quartile demarcated by the dashed lines were declared significant by the GBA method of Chapter 3 (where both hypotheses $H_0^D$ and $H_0^L$ were found significant).

Null Kernel approach. However, the identifiability issues inherent to the GBA approach are overcome by the Null Kernel approach through the simulation of a null distribution rather than relying on asymptotic approximations, which require the identifiability protections. From a computational standpoint, the Null Kernel approach is more computationally demanding than the GBA approach. However, in light of the distributional difficulties associated with the GBA approach (see Section 4.2) and the very similar results of the Null Kernel approach, the Null Kernel approach should be preferred.

4.5 Discussion

The performance of the Null Kernel approach on both simulated and real data shows consistently tighter intervals surrounding the detected QTL than other methods. However, the performance of the Null Kernel approach on the mice HDL QTL mapping data as compared to the Graphical Bonferroni Adjustment (GBA) of Chapter 3 shows that the Null Kernel approach shows greater ability to consider information from the test of $H_0^D$, as explained in the previous paragraph. Interestingly, apart from this artifact of the Null Kernel approach, the two methods perform very similarly, which is a slightly unexpected
result due to the evidence that the test statistics of the GBA approach do not follow their standard asymptotic distributions.

Perhaps the most attractive property of the Null Kernel approach is in the strict dichotomy between significance and non-significance in the resulting adjusted $P$-values. As demonstrated in Figure 4.8, the only loci for which any significance is found in the adjusted $P$-values relate very well to the true (simulated) QTL. Every other loci was identically 1 in the adjusted $P$-value. However, another attractive property of the Null Kernel approach is what may be termed a post-hoc inspection of the test statistics as was done for example in the mice HDL QTL data. This allowed us to determine which results were likely due to significance in just one coordinate of the test statistics (such as would be the case in a union intersection approach to testing) as opposed to significance in both coordinates (as would be the case in an intersection union test).

As mentioned previously, the Null Kernel approach is often conservative in that an exact $\alpha$-level test is often not possible. This difficulty can likely be remedied by simulating more than 1,000 values of the test statistics $T$ and $U$ under the null hypothesis, as was done explicitly in this work. Generating say 10,000 variates would allow a greater chance of observing more extreme chance observations, better approximating the tails of the bivariate distribution, and providing a greater chance for achieving an exact $\alpha$-level test. The computational burden will certainly be increased under such an approach, in both the simulation and computation of the statistics $T$ and $U$, but more importantly, in the fitting of the kernel density estimate $\hat{f}$ to more data. In any case, a test with conservative Type I Error control, which is also powerful enough to detect true QTL effects is certainly an attractive option.
CHAPTER 5
DISCUSSION

Over the past century multiple comparison procedures (MCPs) have grown from essentially non-existent to having entire books and conferences dedicated to their study and advancement. In an article summarizing John W. Tukey’s contributions to MCPs (Benjamini and Braun 2001), Tukey is credited as having emphasized that, “Professional statisticians... have much to learn from the methods of good scientists and also bear an obligation to offer alternatives (or entirely new approaches) that meet real needs and are practical as well.”

The contributions of this work resolve three real needs of researchers. First, the computational burdens of the Focus Level method for gene set testing on GO graphs (which limited its use in real world applications (Liang and Nettleton 2010)) were overcome through an extension of graphical weighted Bonferroni procedures (Bretz et al. 2009) to the case of restricted hypotheses (Chapter 2). The improvement allows the root node of the GO graph to be used as the focus level, freeing the resulting adjusted $P$-values to be interpreted apart from the GO graph rather than only in context of the significant GO graph. Second, the need for a more powerful multiplicity adjustment approach in LD-based QTL mapping was accomplished (Chapter 3) by newly applying a graphical Bonferroni adjustment (Bretz et al. 2009). This was shown to control for a model identifiability issue inherent to the two hypothesis LD-based QTL mapping model of Fu et al. (2013) and that in certain scenarios is equivalent to a conceptually simpler intersection union test, when it is adjusted for multiplicity through the Bonferroni-Holm adjustment (Holm 1979). Third, distributional difficulties with the hypotheses of QTL mapping were detailed and a bivariate approach surmounting these difficulties, the Null Kernel method, were presented in Chapter 4.

While current needs have been met with the contributions of this work, future work remains to be done. Within the QTL mapping framework of Fu et al. (2013) there is
still an open question as to the precise theoretical distributions of the test statistics, either asymptotic or exact. While much work has been performed in this area for interval mapping and composite interval mapping (Rebai et al. 1994), little work has been completed in the single-marker design. Further, extending the Null Kernel method (Chapter 4) to an empirical approach based on the ideas in Efron (2004), rather than a simulation approach, appears a promising avenue. In any case, future work includes implementing the Null Kernel approach into a generalized R package (R Core Team 2013).

As for gene set testing (with special focus on Gene Ontology graphs) work has been done to compare the power of testing methods such as Fisher's exact test and Goeman's Global Test (Fridley et al. 2010). However, other $P$-value combination methods such as Stouffer's method or the min-$P$ approach (Liang and Nettleton 2010) have yet to be similarly studied. A power analysis similar to that of Owen (2009) could provide not only power considerations as in Fridley et al. (2010) but also the various alternatives for which each of the methods is most powerful. This would aid researchers in deciding which method is most powerful and most appropriate for their specific analysis. Also needed is a study detailing the alternatives for which the Short Focus Level (Chapter 2) and Focus Level (Goeman and Mansmann 2008) procedures are each most powerful. Further, it would be highly valuable if there were a way to select the focus level by selecting the level maximizing the number of rejections over all possible focus levels in either the Focus Level or Short Focus Level procedures. Strict control of the FWER (or some other error rate) would be the difficulty in such an approach, but some starting ideas leading towards a potential solution to this problem can be found in Goeman and Solari (2011).
REFERENCES


Davies, R.B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1), 33–43.


APPENDICES
APPENDIX A
GBA Source Code and Help File

A.1 GBA Help File

p.adjust.GBA {source file}

Adjust p-values for Multiple Comparisons
Using the Graphical Bonferroni Approach

DESCRIPTION

Given a two-column data frame or matrix of p-values the
method returns the adjusted p-values according to the
Graphical Bonferroni Approach (GBA) of Saunders, G., Fu,
G., and Stevens, J. R.

USAGE

p.adjust.GBA(p, fiName = NULL)

ARGUMENTS

p - Data frame, numeric matrix (or vector) containing the
    unadjusted p-values. If data frame or matrix, the
first column contains the higher level hypotheses as described by Saunders et al. with the lower level hypotheses in the second column. If vector, say of length 2n, the first level hypotheses should correspond to the first n components with the second level hypotheses corresponding to the last n.

**fileName** - If NULL, results are returned directly. If a character name "FOO" is provided, the results are written in the working directory under "FOO.csv".

**DETAILS**

For details on the graphical Bonferroni approach (GBA) see the paper by Saunders et al. "A Power Improving Correction for Multiple SNPs Selection Used in Linkage Disequilibrium QTL Mapping."

**VALUE**

If fileName = NULL then a two-column matrix is returned with colnames = c("padj.D","padj.LR"). If fileName = "FOO" then the two columns are written as a data.frame to "FOO.csv".
SEE ALSO

gMCP. For small data sets (p-value matrices with less than 50 rows) the package gMCP, using the format of BauerEtAl2001(), will provide more details and plotting options. For larger data sets, p.adjust.GBA is recommended.

EXAMPLES

# Load Function
source("http://math.usu.edu/gsaunders/p.adjust.GBA.R")

# Simulated data
set.seed(1234)
p <- matrix(rbeta(1000,.1,1),500,2)
colnames(p) <- c("chisD","LR")
padj <- p.adjust.GBA(p)
head(padj)

# or similarly, using a vector
p <- as.vector(p)
names(p) <- c(paste("chisD",1:500),paste("LR",1:500))
padj <- p.adjust.GBA(p)
head(padj)

# Significant SNPs
sig <- which(apply(padj<0.05,1,prod)==1)
length(sig)
# write results to file
p.adjust.GBA(p, fiName="FOO")

A.2 GBA Source Code

p.adjust.GBA <- function(p.data,fiName=NULL,trace=FALSE){

  p <- c(as.numeric(p.data[,1]),
         as.numeric(p.data[,2]))

  m <- length(p)
  w <- m/2
  g <- w-1

  #-- The following steps come from the Algorithm 2

  #-- Step 0.
  I <- 1:m; pmax <- 0
  R <- S <- logical(m)
  R[1:(m/2)] <- TRUE

  while( !all(!R) ){  
    #-- Step 1.
    j <- I[R][order(p[R]*w)[1]]
##-- Step 2.

\[ p[j] \leftarrow \min(\max(p[j]w,p_{max}),1); \quad p_{max} \leftarrow p[j] \]

##-- Step 3.

\[ R[j] \leftarrow \text{FALSE} \]

\[ S[j] \leftarrow \text{TRUE} \]

\[ \text{if}(j \leq m/2)\{ \]

\[ R[j+m/2] \leftarrow \text{TRUE} \]

\} else { \]

\[ w \leftarrow w-1 \]

\[ g \leftarrow g-1 \]

\} 

\[ \text{if} (\text{trace})\{ \]

\[ \text{cat}("j = ",j," \mid \text{pmax} = ",p_{max},"/n") \]

\[ \text{flush.console}() \]

\} 

##-- Early terminate if \( p_{max} \) obtains 1 as all remaining
## \( p \)-values must necessarily be adjusted to 1.

\[ \text{if} (\text{pmax} == 1)\{ \]

\[ p[I\neg S] \leftarrow 1 \]

\[ R[I] \leftarrow \text{FALSE} \]
message("Function completed successfully before \"I\" was empty.")
}

}###-- End while loop.

padj <- matrix(p,m/2,2)
colnames(padj) <- c("padj.LR","padj.D")

if (is.null(fiName)){
    return(padj)
}else{
    write.table(padj,
        file=paste(fiName,"_padj.csv",sep=""),
        sep="",
        row.name=FALSE, col.names=TRUE)
}

}###-- End function.
CURRICULUM VITAE

Garrett Saunders

PhD in Mathematical Sciences, Statistics Specialization, Utah State University, Logan, Utah. MS in Mathematics, Boise State University, Boise, Idaho. BS in Applied Mathematics, Brigham Young University-Idaho, Rexburg, Idaho. Research interests include multiple comparison procedures, statistical applications within genomics and genetics, and statistics education.