Tel-Aviv University
The Raymond and Beverly Sackler Faculty of Exact Sciences
Department of Statistics and Operations Research

Controlling the False Discovery Rate
for Multiple Testing Involved
in the Statistical Analysis of Gene Expression Data

Thesis submitted for the degree “Doctor of Philosophy”

By

Anat Reiner Benaim

Submitted to the Senate of Tel-Aviv University

August 2006
This work was carried out

at the Department of Statistics and Operations Research

Tel-Aviv University

under the supervision of

Prof. Yoav Benjamini
Acknowledgements

I would first like to thank my advisor, Prof. Yoav Benjamini, under whose guidance I have been working for several years, during both Masters and PhD studies. Apart from his excellent and sincere guidance, I deeply value his insight in adjoining me with the exciting topics appealing to my interest.

I would also like to thank Dr. Daniel Yekutieli for the collaborative work, as well as his help in introducing me to important concepts and results related to his work.

I would like to thank the team of research on genomics and behavior:

Maryland Psychiatric Research Center (MPRC), University of Maryland:
Prof. Gregory I. Elmer
Dr. Neri Kafkafi

The Institute for Genomic Research (TIGR), Maryland:
Dr. Norman Lee
Noah Letwin

Dept. of Zoology, Tel-Aviv University:
Prof. Ilan Golani,

Thank you for the foundations supporting my research:

National Institute of Health (NIH)
FIRST foundation of the Israeli Academy of Science & Humanities
Abstract

Statistical challenges in multiple testing of gene expression data

Statisticians dealing with large multiple comparisons problems initially confronted mainly post-hoc pairwise comparisons. However, recent developments in high throughput biotechnology dramatically changed the scale of multiplicity, as in the case of gene expression measured by microarray technology. Here, thousands of genes are measured simultaneously, in search for changes in expression levels across different conditions, as it may reflect gene activity and involvement in a biological process of interest.

The multiplicity problem here is challenging not only by its sheer size, but also by several characteristics typical of microarray experiments. First, expression levels of different genes may be dependent due to co-regulations of the genes, as well as the measurement errors of these expression levels that are subjected to several technical factors. Multiple testing of such data faces the challenge of correlated test statistics. It is therefore necessary to develop methodological tools that will address multiplicity while accounting for this dependency.

Second, statistical analysis of gene expression concerning a wide range of research questions is called for, from a simple two-group comparison, to a more complex analysis, such as pairwise comparisons, interaction contrasts, or correlation analysis, all on a gene level. Thus the study may contain multiple research questions leading to a dramatic increase in multiplicity. Here, new approaches may be adopted for dealing with multiplicity by viewing our different research questions as separate families of hypotheses.

The False Discovery Rate (FDR) criterion

When many hypotheses are tested simultaneously, the probability that a type I error is committed increases sharply with the number of hypotheses. One criterion representing the type I error is the family-wise type I error rate (FWER), which is the probability of committing even one error in the family of hypotheses. However, this
criterion may be too conservative for microarray experiments, which are typically used to screen genes for further, more focused, research. Instead, it may be more appropriate to emphasize the proportion of errors among the identified differentially expressed genes. The expectation of this proportion is the False Discovery Rate (FDR) of Benjamini and Hochberg (1995) that introduced the linear step-up procedure (BH), which was shown to control the FDR at a desired level $q$, under some conditions. Controlling this FDR criterion in the simultaneous testing of gene expression is the focus of this thesis, since it admits more powerful procedures.

**Objectives**

Recent advances of FDR methodology are modified and adapted to the setting of gene expression analysis. Four procedures are studied: The BH procedure, resampling–based FDR controlling procedures that estimate either the marginal or joint distribution of the test statistics, and FWER controlling procedures. Their FDR control is examined, as well as power.

The applicability of the BH procedure to control the FDR in two-sided tests is further assessed by examining the case of correlated test statistics, as arises when the data measurement errors tend to be positively correlated due to common latent factors involved. The problem is studied first in the simplest possible setting: two normally distributed correlated test statistics. The behavior of the FDR is examined as a function of the correlation and the distance of one of the hypotheses from the null. The findings are extended into the case of more than two tests, where the proportion of true null hypotheses varies, as well as the dependency structure.

The identification of differentially expressed genes in two biologically different situations is the simplest setting in which the statistical analysis of gene expression is called for, and where the multiplicity problem is evident. When conducting a more complex analysis, such as pairwise comparisons, interaction contrasts, or correlation analysis on a gene level, a further complication occurs due to the multiplicity of genes. Thus, it is a very real possibility to face a multiplicity problem of testing millions of hypotheses. Terming a group of hypotheses to be tested simultaneously a family of hypotheses, it becomes clear that the entire set of hypotheses that interests the researcher does not necessarily form a single family of hypotheses. Westfall and Young
(1993) suggest that correction for multiplicity should be applied within each family. Nevertheless, multiplicity of families of hypotheses can be a serious problem, and a more appropriate solution for this problem should be offered. In this thesis two approaches are presented and illustrated with real data examples. (a) Selecting subsets from a large pool of hypotheses in a few of steps, using test statistics that are independent from one step to the other, (b) Organizing the families of hypotheses in a hierarchical tree structure, and using the hierarchical testing approach developed by Benjamini and Yekutieli (2002).

The next level of refinement in research questions arises when there is another factor that may affect expression levels. In particular, when the information on the phenotypic level on individuals can be measured, an opportunity to study the connection between the expression level and functionality arises. A new dimension of complexity is added, as phenotypic information can be measured per individual, while expression data can typically be measured only over a pool of individuals, in order to produce a reliable quantity of cRNA (the genetic substance). Correlating the two variables as they are yields artificially increased correlations, due to the elimination of the within-animal expression variance. A reliable estimate of this correlation and its significance is therefore offered. As before, in such analyses multiple testing remains a major concern. Even within the FDR framework, the more complex studies pose new challenges. The strategy introduced for complex studies is taken a step further, by showing that one can lower the demands raised by existing theoretical bounds.

**Summary of results**

Comparative analysis of FDR and FWER controlling procedures showed that, in addition to the higher power retained by all FDR controlling procedures, the use of resampling achieves additional considerable gain in power. Thus, computational efforts should be invested in accordance to the importance of maximizing the number of genes selected for the next stage.

Examining the case of two-sided testing with dependent normally distributed test statistics, we identified the least favorable data properties bringing the BH procedure to the maximal FDR. Considering common correlation $\rho$ between the test statistics, it was shown that the least favorable distance between means of true and false hypotheses is the sum of the two normal critical values corresponding to each subset of null hypotheses.
when there is perfect correlation between the test statistics. However, unlike the case of independence and one-sided positive dependent test statistics, the BH procedure does not control the FDR at \( \frac{m_0}{m} q \). We found that the deviation of FDR from \( \frac{m_0}{m} q \) reaches a global maximum when the proportion of true null hypotheses is \( \frac{1}{2} \). Using this characterization we got that for any \( \rho \), the BH procedure would control the FDR at level \( q \). Using the BH at level \( q' \leq \frac{q}{1+\frac{1}{2}\left(1-\frac{m_0}{m}\right)} \), where \( m \) is the number of null hypotheses and \( m_0 \) is the number of true null hypotheses, would control the FDR at level \( \frac{m_0}{m} q \). A simulation study indicated that these results apply also in the case of spatial dependence, where the correlation is not constant.

The control of the FDR was examined for a large complex research, consisting of several questions to be answered based on analysis of hierarchical layout. It was shown that while in some cases the direct application of BH is the only approach that can be used due to clear dependency between research questions, in other cases, where none or moderate dependency can be assumed, the hierarchical tree-testing approach yields more rejections while controlling the FDR at the desired level.

A pioneering functional genomics study, associating genotypic and phenotypic data, was analyzed using the two approaches introduced for analyzing complex research. Thus they were examined under some degree of dependency between the test statistics attributed to different stages of the analysis. It was found that if all screened families are approximately the same in terms of proportion of true null hypotheses and p-values distribution, the subset selection method is superior as the testing is done at a higher \( q \) and the behavior of BH will be little affected by the amalgamation of the families. If a few families have high proportion of correlations while many others have none (or close to none) the second method has the advantage in spite of lower \( q \) at both stages. Even when theoretical bounds do not guarantee control of FDR within one single stage, it seems that a correction whose value can be estimated will provide control at the desired level.
The correlation analysis raised the problem of genotypic information measured for pooled samples, while phenotypic information was measured on individual subjects. A solution was offered based on jittering the expression data by an amount estimated from the biological replicates. This solution avoids the artificial increase in the correlation coefficient and its significance.
Table of Contents

List of Figures .................................................................................................................. 5

List of Tables ..................................................................................................................... 7

1. Introduction .................................................................................................................. 9

1.1. Statistical challenges in multiple testing of gene expression data ......................... 9
1.2. Multiple testing procedures for identifying differentially expressed genes .......... 10
1.3. Inter-correlations within gene expression data ....................................................... 11
1.4. Complex research .................................................................................................... 13
1.5. Functional genomics analysis .................................................................................. 14
1.6. Outline and objectives .............................................................................................. 16

2. Microarray Technology ............................................................................................... 21

2.1. Overview ................................................................................................................ 21
2.1.1. What are microarrays? .................................................................................. 21
2.1.2. Therapeutic motivation ................................................................................. 22
2.2. Genetic terminology ............................................................................................... 23
2.3. Expression profiling techniques ............................................................................ 24
2.3.1. cDNA microarrays ....................................................................................... 24
2.3.2. High density synthetic oligonucleotide microarrays ................................ 27
2.4. Data analysis questions .......................................................................................... 29
2.4.1. Detecting differential expression .................................................................. 29
2.4.2. Pattern discovery ............................................................................................ 30
2.4.3. Class prediction .............................................................................................. 32
2.4.4. Inferring regulatory pathways and networks .............................................. 33
2.5. Preprocessing ........................................................................................................ 33
2.5.1. Expression ratios ............................................................................................ 34
2.5.2. Normalization ................................................................................................ 35
2.5.3. Spatial effects removal ................................................................................... 38

3. False Discovery Rate Methodology ........................................................................... 39

3.1. The False Discovery Rate criterion ....................................................................... 39
3.2. Economic interpretation of the FDR criterion .................................................... 40
3.3. The Linear Step-Up procedure (BH) .................................................................... 41
3.4. Adaptive approaches ............................................................................................. 42
3.5. Multiplicity adjusted p-values ................................................................. 43
3.6. Resampling FDR adjustments ................................................................. 44

4. **Recent work on False Discovery Rate related to gene expression analysis** ................................................................. 47

4.1. Overview ................................................................................................... 47
4.2. False positive Criteria ................................................................................. 48
   4.2.1. Frequentist motivation ......................................................................... 48
   4.2.2. Bayesian motivation ............................................................................ 48
4.3. Multiple testing procedures ...................................................................... 51
   4.3.1. Frequentist approach .......................................................................... 51
   4.3.2. Bayesian Approach ............................................................................. 52

5. **Comparative analysis of FDR controlling multiple testing procedures** ............................................................................. 55

5.1. Overview ................................................................................................... 55
5.2. Comparative analysis on real microarray data ............................................. 55
   5.2.1. The data ................................................................................................ 55
   5.2.2. Deriving the test statistics .................................................................... 56
   5.2.3. Multiple testing ..................................................................................... 56
   5.2.4. Results ................................................................................................ 57
5.3. Comparative analysis on simulated data ...................................................... 58
   5.3.1. Simulated data configuration ............................................................... 58
   5.3.2. Results ................................................................................................ 59
5.4. Conclusions ............................................................................................... 61

6. **FDR control for dependent two-sided test statistics** .................................. 63

6.1. Overview ................................................................................................... 63
6.2. Simulation study, \( m=2 \) ........................................................................... 63
   6.2.1. Methods .............................................................................................. 63
   6.2.2. Results ................................................................................................. 65
6.3. Theoretical study, \( m=2 \) .......................................................................... 68
6.4. Simulation study, \( m>2 \) ............................................................................ 75
6.5. Theoretical study, \( m>2 \) .......................................................................... 84
6.6. Generalized dependence structure ............................................................. 89
   6.6.1. Spatial dependence ............................................................................. 90
6.7. Conclusions ............................................................................................... 91
7. Multiplicity issues related to complex research questions ............ 93

7.1. Overview ............................................................................................................. 93
7.2. Complex research questions ............................................................................. 94
   7.2.1. The experiment ............................................................................................... 94
   7.2.2. Notation and ANOVA model .......................................................................... 95
   7.2.3. Question 1: Brain ANOVA per gene ............................................................... 95
   7.2.4. Question 2: Strainwise comparisons in Brain analysis per gene ...................... 96
   7.2.5. Question 3: Strain*Region interactions in Brain analysis per gene ................. 97
7.3. FDR controlled testing of complex research questions ....................................... 97
   7.3.1. The direct approach using the procedure in BH ............................................. 97
   7.3.2. The direct approach for selected subset ......................................................... 98
   7.3.3. Hierarchical testing and associated false discovery rates ............................... 99
   7.3.4. FDR controlling hierarchical testing ............................................................. 100
7.4. Analysis and Results .......................................................................................... 101
   7.4.1. Brain ANOVA per gene: The need for repeated median removal of chip layout effects .................................................. 101
   7.4.2. Brain ANOVA per genes ............................................................................... 104
   7.4.3. Identifying specific Strain-wise differences in Brain analysis per gene .......... 105
   7.4.4. Brain region interactions with Strain per gene .............................................. 106
7.5. Conclusions ........................................................................................................ 108

8. Functional genomics: Correlating complex behavioral traits with expression data ................................................................................................................. 111

8.1. Overview .............................................................................................................. 111
8.2. The experiment ..................................................................................................... 112
8.3. The research question .......................................................................................... 113
8.4. Dependence between the tests of the two stages ................................................ 114
8.5. Methods ............................................................................................................... 115
   8.5.1. Proposed Methods: Subset selection and hierarchical testing ........................ 115
   8.5.2. Single-Step Methods ..................................................................................... 115
   8.5.3. Jittering to address pooling effect .................................................................. 118
   8.5.4. Additional details of the statistical analysis .................................................. 120
   8.5.5. Simulation study .............................................................................................. 121
8.6. Analysis and results ............................................................................................. 123
   8.6.1. Simulation study results .................................................................................. 123
   8.6.2. Data analysis results ...................................................................................... 128
8.7. Conclusions ........................................................................................................... 129
9. Discussion and further research .................................................. 133

9.1. Summary of results.......................................................................... 133
9.2. Further research................................................................................ 137

References .......................................................................................... 141

Appendix: Computer programs .............................................................. 153

A1. Identifying differentially expressed genes between two independent
group using FDR controlling procedures (Chapter 5)............................. 153
A2. N-way median polish (Chapter 7)....................................................... 159
A3. Complex research analysis (Chapters 7-8)......................................... 165
   A3.1 Analysis of variance ................................................................. 165
   A3.2 Pairwise comparison ............................................................... 167
   A3.3 Identifying interaction .............................................................. 169
   A3.4 Simple effect analysis of variance ............................................ 171
   A3.5 Identifying correlation ............................................................. 173
A4. Accompanying SAS macro programs (sorted alphabetically)............. 175
List of Figures

Figure 1: The Central Dogma of molecular biology .......................................... 25
Figure 2: cDNA microarray experiment schema .............................................. 26
Figure 3: Quantification from two-color hybridization .................................... 26
Figure 4: Light directed oligonucleotide synthesis .......................................... 27
Figure 5: Expression probe and array design ............................................... 28
Figure 6: Hierarchical clustering diagram .................................................... 31
Figure 7: M-A plot ...................................................................................... 36
Figure 8: Variance stabilizing transformation ............................................... 37
Figure 9: Raw and FDR adjusted p-values ..................................................... 44
Figure 10: FDR and FWER adjusted p-values – original data ......................... 57
Figure 11: FDR and FWER adjusted p-values – simulated data .................... 60
Figure 12: Power – by configuration ............................................................. 60
Figure 13: |X_1| vs. |X_2| ........................................................................ 65
Figure 14: The FDR deviation from q/2 vs. ρ, m=2 ....................................... 66
Figure 15: The simulated FDR deviation from q/2, vs. μ_2, m=2 .................. 67
Figure 16: Joint distribution of |X_2|, |X_1| when |ρ|=1 - cases of Q area cross 69
Figure 17: Joint distribution of |X_2|,|X_1| - FDR calculation .......................... 70
Figure 18: The FDR deviation from q/2, ρ=1 – computed function ................ 72
Figure 19: Least favorable μ_2 vs. q ............................................................. 73
Figure 20: FDR vs. μ_2, m=3 .................................................................... 76
Figure 21: FDR vs. ρ, m=3 ........................................................................ 78
Figure 22: FDR vs. ρ, m=4 ....................................................................... 80
Figure 23: FDR vs. ρ, m=6 ....................................................................... 82
Figure 24: Maximal FDR and FDR deviation vs. m_0/m ............................... 87
Figure 25: Transformation from spatial dependence to perfect correlation ....... 91
Figure 26: P-values before and after spatial effects removal – strain effect ...... 102
Figure 27: A schematic structure of a microarray ........................................... 104
Figure 28: Raw and FDR adjusted p-values – strain effect ............................. 105
Figure 29: Frequency of genes with significant pairwise difference ................ 106
Figure 30: Genes mapped into a network linked with seizure activity .......... 108
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>Jittered data vs. original data – two runs</td>
<td>119</td>
</tr>
<tr>
<td>32</td>
<td>FDR for full-tree and outer-node schemes</td>
<td>124</td>
</tr>
<tr>
<td>33</td>
<td>FDR multiplier for the level restricted scheme</td>
<td>125</td>
</tr>
<tr>
<td>34</td>
<td>FDR for the level restricted scheme</td>
<td>126</td>
</tr>
<tr>
<td>35</td>
<td>P-values for expression-behavior correlation analysis</td>
<td>130</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Intersection points with $Q$ area boundaries</td>
<td>71</td>
</tr>
<tr>
<td>Table 2</td>
<td>Simulation configurations</td>
<td>122</td>
</tr>
<tr>
<td>Table 3</td>
<td>Power comparison of proposed methods and single-step methods</td>
<td>127</td>
</tr>
</tbody>
</table>
1. **Introduction**

1.1. **Statistical challenges in multiple testing of gene expression data**

The multiple comparisons problem raised the attention of the statistical community since the 1950’s, with the understanding that when many hypotheses are tested, the probability that a type I error is committed increases sharply with the number of hypotheses. Multiple testing generated by post-hoc pairwise comparisons was the type of multiplicity initially confronted. However, the multiplicity problem gains a much larger dimension in the case of gene expression analysis, recently encountered within the newly emerging discipline of bioinformatics. Moreover, it poses a new scale of complication due to the nature of the data and analysis, and thus demands new approaches and solutions.

Gene expression is monitored for thousands of genes simultaneously by means of the microarray technology. Changes in gene expression levels across different conditions, denoted differential gene expressions, reflect gene activity and involvement in a biological process of interest. Gene expression analysis aims to identify the genes with differential expression, and thus the required inference is on the gene level. This directly leads to multiple testing of a very large magnitude, which may grow by several scales if every per-gene testing is further composed of several tests.

The multiplicity problem here is challenging not only by its size, but also by several characteristics typical of microarray experiments. First, expression levels of different genes may be dependent due to co-regulations of the genes, as well as the measurement errors of these expression levels, that are subjected to several technical factors. Multiple testing of such data faces the challenge of correlated test statistics. It is therefore necessary to develop methodological tools that will address multiplicity while accounting for this dependency. Second, statistical analysis of gene expression concerning a wide range of research questions is called for, from a simple two-group comparison, to a more complex analysis, such as pairwise comparisons, interaction contrasts or correlation analysis, all on a gene level. Thus the study may contain
multiple research questions leading to a dramatic increase in multiplicity. Here, new approaches may be adopted for dealing with multiplicity by viewing the different research questions as separate families of hypotheses.

### 1.2. Multiple testing procedures for identifying differentially expressed genes

The number of comparisons of gene expression level studied in a single experiment has been growing literally at an exponential rate since the beginning of the 1990’s. Based on a sample survey of published studies, the number of genes investigated per study prior to the introduction of microarray technology was typically less then 10. It leaped to a few dozens with the entry of the early versions of microarrays into use. Subsequent to the rapid development of this technology, it reached 4,000 genes by 1995, and 10,000 genes by 2001. During the years that followed, the recognition of the scientific potential of these experiments, as well as the accumulating technical expertise, gradually increased. As a result, the degree of complexity of the experiments has increased as well, as more factors were inserted into the experimental design. This drove up the number of the hypotheses tested to new orders of magnitude, reaching up to millions (see Chapter 7 and Chapter 8 in this thesis).

Plainly, the simultaneous testing of so many genes created a very large multiplicity problem. Yet, while numerous methods were available for controlling the family-wise type I error rate (FWER), which is the probability of committing even one error in the family of hypotheses, attention to the multiplicity problem in gene expression analysis has been virtually null until the work on this thesis began, in 2001, as reflected by published work in scientific literature. The scientific publication database ISI Web of Knowledge ([http://portal.isiknowledge.com](http://portal.isiknowledge.com)) records Callow et al (2000) as the earliest work published on microarray data analysis with adjustment to multiple testing. Dudoit et al (2002b) appears to be the first study to call attention to the importance of the multiplicity problem as one of the key statistical issues arising in microarray data analysis. The Westfall and Young step-down algorithm, herein WY (Westfall and Young, 1989), a permutation-based procedure, is used in that work to adjust for multiplicity by controlling the FWER, without assuming $t$ distribution of the test statistics of each gene’s differential expression.
While in some cases FWER control is needed, the multiplicity problem in microarray data does not require a protection against even a single type I error, so that the severe loss of power involved in such protection is not justified. Instead, it may be more appropriate to emphasize the proportion of errors among the identified differentially expressed genes. The expectation of this proportion is the False Discovery Rate (FDR) of Benjamini and Hochberg (1995). Controlling this FDR criterion in the simultaneous testing of gene expression is the focus of this thesis, since it admits more powerful procedures. Benjamini and Hochberg (1995) introduce the linear step-up procedure (BH) that is shown to control the FDR at a desired level $q$.

Advances in FDR methodology offer improved ways of incorporating FDR control in gene expression analysis. The results of Benjamini and Yekutieli (2001) extend the scope of applicability of the BH procedure to dependency situations, and Yekutieli and Benjamini (1999) introduce resampling-based procedures that control the FDR under dependency. Procedures that are adaptive to the proportion of true null hypotheses have been proposed to increase power (Hochberg and Benjamini, 1990, Storey, 2002, Storey and Tibshirani, 2003, Pounds and Cheng, 2003, Jiang, 2004). The first part of this thesis, namely Chapter 5, introduced these ideas to microarray analysis (IMS-NSF conference, 2001). Later, the hierarchical testing approach developed by Benjamini and Yekutieli (2002) for multi-stage analytical schemes facilitates the control of the FDR over more than one stage of hypotheses testing in the research. The above FDR methods will all be discussed in detail in the next chapters and will be studied within the context of microarray experiments.

1.3. **Inter-correlations within gene expression data**

Characterization of the dependency between the expression levels of different genes attributed to common co-regulation is one of the main research targets of microarray experiments. Its study is especially challenging due to the high complexity of biological functional pathways. Technical factors involved in the experiment draw our attention to dependency also among the measurement errors of the expression levels. Geographical location on the chip may generate clusters of spatially correlated measurements. RNA source and sample pooling, as well as the use of different printing tips for different areas on the chip, can result in clusters of
common correlation. Thus, when the focus is shifted to per-gene inference, the test statistics calculated for each gene may be correlated due to these dependencies. In such a case, the control of the FDR is not straightforward. The challenge is to understand how dependency in general and its structure in particular, affect this criterion.

The FDR controlling BH procedure has been investigated with regard to test-statistics dependency. Benjamini and Yekutieli (2001) show that the FDR is controlled for positively dependent test statistics as well. The technical condition under which the control holds is that of positive regression dependency on each test statistic corresponding to the true null hypotheses (as defined there). In particular, the condition is satisfied by positively correlated normally distributed one-sided test statistics, and their studentized t-statistics. Such a case may rise when the data measurement errors tend to be positively correlated, due to common latent factors involved. Note that when the null hypotheses are all true, these are the main sources of variability. Furthermore, when the alternative positive or negative directions are about equally likely to occur, the property of FDR control can be extended to two-sided tests. Yekutieli (2002) shows that the absolute values of multivariate normal and t-statistics satisfy the positive regression dependency condition when all null hypotheses are true. He emphasizes the need to further investigate the scope of problems for which the two-sided tests control the FDR at the desired level.

In fact, initial results regarding the performance of related multiple testing procedures under dependency started to accumulate in much simpler settings in earlier studies. Simes (1986) proposed an improved Bonferroni procedure, based on the same linear set of critical values as for the FDR controlling procedure BH and used to test the intersection null hypothesis that all hypotheses tested are true, versus the alternative that at least one is not. This test was shown to control the FWER in the weak sense, i.e. when all null hypotheses are true under independence. Hochberg and Rom (1995) and Samuel-Cahn (1996) examined the conservativeness of the Simes procedure with regard to correlated bivariate normal test-statistics, i.e. when the number of tests $m$ is merely 2. For one-sided tests, the procedure was found to control the FWER in the case of positive dependency but not in the case of negative dependency. For two sided-tests based on bivariate normal test-statistics, the procedure was found to always control the FWER. Only later, Sarker (1998) extended some of the results to the general $m$. 
Note that for general cases, in which the positive dependency conditions do not apply, Benjamini and Yekutieli (2001) prove that the effect of general dependency may increase the FDR of the linear step-up procedure from level $q$ up to $q^* = q \sum_{i=1}^{m} \frac{1}{i}$. Modifying the procedure by using $q^*$, which is smaller than $q$ by the same factor, will therefore guarantee FDR control under all types of dependency. However, this amount of increase takes place for a very unusual joint distribution, which is not even unimodal. Such modification may be too conservative for many practical problems.

1.4. **Complex research**

The identification of differentially expressed genes in two biologically different situations is the simplest setting in which the statistical analysis of gene expression is called for, and where the multiplicity problem is evident. Given two experimental conditions, possibly a treatment group versus a control, comparing the expression level of each gene in the two groups can be easily done. Estimating the difference, as well as testing for its statistical significance, can also be done easily at the individual gene level. Each of the problems faced by the researcher conducting a more complex analysis, such as pairwise comparisons, interaction contrasts, or correlation analysis on a gene level is further complicated by the multiplicity aspect stemming from the very large number of genes involved in the analysis. Thus, it is a very real possibility to face a multiplicity problem of testing millions of hypotheses because of the simultaneous treatment of many genes.

The usual concern among researchers developing multiple testing procedures is about the level of FWER or FDR in the set of rejected hypotheses when testing a family of null hypotheses. But what is exactly the family of hypotheses tested? Westfall and Young (1993) suggest three guidelines for determining whether a set of tests should be considered a “family” of tests, thus tested in a multiple testing procedure. (a) The questions asked form a natural and coherent unit. (b) All tests are considered simultaneously. (c) It is considered a priori probable that many or all of members of the family of null hypotheses are in fact true. It is important to note that the guidelines are subjective in the sense that a family of hypotheses is determined not only by the nature of the hypotheses tested but also by the goal of the study and by the prior knowledge of the researcher.
It becomes clear that the entire set of hypotheses that interest the researcher do not necessarily form a single family of hypotheses. In this thesis, such a study, comprising of several families of hypotheses, will be termed a complex study. Complex studies are not limited to microarray analysis. The novelty is in the scale of the problem. The specification of research directions and the choice of families of null hypotheses influence the discoveries made. Therefore, the interpretation of the results of a complex study is impossible without a detailed account of the choice of families of tested hypotheses, the considerations guiding the researcher in constructing the families of hypotheses, and the multiplicity corrections applied.

Westfall and Young (1993) suggest that correction for multiplicity should be applied within each family. Nevertheless, multiplicity of families of hypotheses can be a serious problem. Take as an example a study consisting of 1000 families of hypotheses, each one comprising 100 hypotheses, all tested by independently distributed test statistics according to their true null hypotheses. Testing each family of hypotheses using the linear step-up procedure in BH at level 0.05 solves the multiplicity problem at the family level: in 950 out of 1000 families no false discoveries are expected to occur. The problem lies at the complex study level, where it is expected to make in each family \( \frac{0.05}{(1-0.05)^2} = \frac{0.05}{0.9025} \) false discoveries and thus total of \( \frac{50}{0.9025} \) false discoveries in all 1000 families (Finner and Roters, 2002). A more appropriate solution for this problem should be offered.

1.5. **Functional genomics analysis**

The development of microarray technology opened a new era for functional genomics, which is concerned with understanding the connection between biological function and the information in DNA sequences. The initial and simplest statistical question related to such analysis is the identification of genes differentially expressed between two classes that differ in functionality. The two classes may involve healthy versus sick cells, one region in the brain versus another, hearing versus deaf animals, or high activity strain versus low one. The statistical setting becomes somewhat more complicated when more than two classes of function exist, that is the functionality is classified into a few levels. Here, the effort may initially involve the identification of genes that are not similarly expressed across all
functionality class-levels, a problem that leads into one-way analysis of variance (ANOVA). Next, the researcher may be interested in comparisons among (possibly all) pairs of class-levels in order to specifically identify where the difference lies. Alternatively the test of interest we may be linear contrasts that compare a weighted average of gene expression over one subset of levels with a weighted average of another subset.

The next level of refinement, and its associated statistical complexity, arises when there is another factor that may affect expression levels. The same factor may either be another subject of study by itself, or it may affect functionality and confound the findings of interest, thereby masking the differences of interest. In both cases, an appropriate model for identifying differential expression is two-way ANOVA. ANOVA modeling has been used by several researchers for the initial screening of genes through testing for class differences (Pavlidis and Noble, 2001, Kerr and Churchill, 2001, Kerr et al, 2000, 2002, Pavlidis, 2003, Smyth et al, 2005). The interaction between the two factors at particular combinations of their levels may be tested as well (Yekutieli et al, 2006). However, a test for class difference within each level of the additional factor may be preferred. This may pose a problem of sample size, as not many replicates are available at this resolution. Simple effect ANOVA (Winer, 1971), where the standard deviation is estimated from the entire data per gene, may be used in this case, as will be discussed later.

A better opportunity to study the connection between the expression level and functionality is when the information on the phenotypic level on individuals can be measured. Thus, instead of classifying activity into high and low, measurements on quantitative traits that reflect activity might be used in order to associate expression levels with measured activity. Such association may be affected by the other factors as well, so the study of association does not come instead of the strategy described above, but rather on top of it. Furthermore, this opportunity adds a new dimension of complexity: while phenotypic information can be measured per individual, expression data can typically be measured only over a pool of individuals, in order to produce a reliable quantity of RNA. Correlating the two variables as they are yields artificially increased correlations, due to the elimination of the within-animal expression variance. A reliable estimate of this correlation and its significance is therefore offered, based on jittering the expression data, which amounts to adding noise to it, by a quantity that is estimated from the pooled biological replicates.
Another major concern in such analyses is multiple testing. The one-way ANOVA considered for a single gene has been the foundation of the theory and practice of multiple comparisons procedures that control the family-wise error-rate (FWER) - the probability of making even one type I error, because of the many tests involved. The number increases dramatically, when every gene is tested using ANOVA (for instance, in the experiment discussed in this chapter, there are over two millions potential tests). As a result, the control of the overall type I error is complicated, and it has become quite common to confront the multiplicity problem in studies by using the False Discovery Rate (FDR) methodology. In functional genomics, for instance, Getchell et al (2004) study differential expression between different strains of aging mice, and classify them into different functional categories. Similar classification is done by Xiao et al (2002), which identifies differentially expressed genes related to human peripheral nerve injury, under different biological states. Luan and Li (2004) apply FDR controlling procedures to identify genes whose time course expressions are synchronized to certain periodic biological process.

However, even within the FDR framework, the more complex studies pose new challenges. If the researcher is interested in first indentifying a pool of genes according to some initial screen, the analysis will involve more than one stage. Here, a new approach to control the FDR based on hierarchical scheme of hypothesis testing may be more appropriate. One important aspect when several analytical stages are considered is the dependence of the test statistics of one stage on the test statistics of another stage. The available theory for the hierarchical testing scheme relies on the independence between the test statistics of the different stages. For the correlation analysis conducted in this thesis, an initial screening test that all means are the same is done first. In this case independence between this first stage and the correlation analysis of the next stage does not hold exactly, since the test statistic for correlation analysis has a weak dependence on the test statistics of the first test. It is necessary to investigate FDR control when the hierarchical approach is used under these conditions.

1.6. Outline and objectives

The general objective of this thesis is to study the FDR control in multiple testing involved in statistical analyses of microarray data. Two main questions are being examined: First, how characteristics of the data affect FDR behavior. Second,
how complexity of the analysis suggests the adoption of new approaches for FDR control.

Chapters 2-4 present the necessary background supporting and motivating our work. In chapter 2 the microarray technology for monitoring gene expression levels is summarized. The motivation of developing and using this facility will be clarified, along with the necessary biological background. Principles of its operations will be explained, including experimental design perspective. The nature of its output will be discussed with its preprocessing considerations. Potential questions that can be answered by statistical analysis will be outlined, as well as popular methodologies adopted.

Chapter 3 and chapter 4 present theoretical background on FDR. Chapter 3 discusses the FDR as a concept, within the initial line of investigation on which Benjamini and Hochberg (1995) is based. Chapter 4 continues the discussion on FDR methodology developed later on, which was highly motivated by multiplicity in the analysis of microarray data, and extends the discussion on adaptive approaches of FDR control. Different lines of FDR interpretation offered in literature are discussed in this chapter. The discussions in Chapter 3 and Chapter 4 motivate the use of particular methodologies in this work, as detailed in the chapters that follow.

The next four chapters 5-8 present the main original contributions of this thesis. In chapter 5 advances of FDR methodology are modified and adapted to the setting of gene expression analysis, considering four procedures and studying their properties using simulation. The first procedure considered is the BH as applied to the p-values corresponding to the \( t \) tests. The second uses the same BH, as applied to the marginal p-values estimated by resampling and then pooling the resampling distributions over genes. The two other procedures are based on estimating the joint distribution of the p-values and the FDR at a given potential threshold using resampling. They differ by the way this distribution is summarized. It is first shown that all four procedures control the FDR at the desired level and that all four are also more powerful than their FWER counterparts. Their gain in power is then compared, since not all four are equally easy to implement. The results of this chapter have been published jointly with D. Yekutieli and Y. Benjamini in Bioinformatics (Reiner et al, 2003).

In chapter 6 the effect of dependency structure between the test statistics in the case of two-sided testing is assessed, taking similar first steps as in the studies of
Simes procedure. That is, the problem is studied first in the simplest possible setting: two normally distributed correlated test statistics. The relevant expectations are calculated first by using simulations, and these results serve us as data to study and understand the behavior of the FDR as a function of the correlation and the distance of one of the hypotheses from the null. The analysis of the simulation results serves us as a mean to generate theoretical hypotheses as to the behavior of the BH procedure, and the results regarding the behavior are then given theoretical support. Finally, the findings are extended into the case of more then two tests, where different cases of proportion of false null hypotheses can be examined, as well as different dependency structures. The results of this chapter have been presented by the author of this thesis at the Multiple Comparisons Procedure Conference, Shanghai, China, 2005, and are under revision for the proceedings journal (Reiner, 2005).

In chapter 7, the problem of complex research is raised and studied. Approaches of dealing with the multiplicity issue by means of FDR control are suggested. These approaches are demonstrated by analyzing the results of a complex experiment involving the study of gene expression levels in different brain regions across multiple mouse strains.

Although a direct approach towards FDR control, where the entire set of null hypotheses is tested simultaneously using the procedure in BH, is feasible, the direct approach is found to be limited. The approach investigated is to divide the statistical analysis into several main research directions, or research questions, and control the FDR for each research direction separately. Furthermore, even the large families associated with each research direction need not be studied merely by the above-mentioned direct approach. In Chapters 7 and 8, in particular in Chapter 7, two other approaches of controlling the FDR of a direction of research will be presented and illustrated with real data examples. (a) Selecting subsets from a large pool of hypotheses in a few of steps, using test statistics that are independent from one step to the other, (b) Organizing the families of hypotheses in a hierarchical tree structure, and using the hierarchical testing approach developed by Benjamini and Yekutieli (2002) for such tree-structured analytical schemes. The results of this chapter have been written jointly with D. Yekutieli, G. I. Elmer, N. Kafkafi, N. E. Letwin, N. H. Lee and Y. Benjamini and are in press for Statistica Nederlandica (Yekutieli et al, 2006).
In chapter 8, recent advances in FDR methodology (Benjamini and Yekutieli, 2002, Yekutieli et al, 2005) are used to make well-founded inference on gene-level functionality. A new testing scheme is proposed, allowing FDR control either over all stages of the analysis or over a single stage, depending on the research emphasis. Rather than testing all possible end-hypotheses, one can: (a) Screen and select hypotheses from a larger pool at the first stage based on ANOVA testing at each gene, and continue testing the selected ones using test statistics that are independent of those used for screening. (b) Organize the required inferences in tree-structured families of hypotheses, and test separately each family of hypotheses by implementing a hierarchical testing approach. This strategy is taken a step further by showing that one can lower the demands raised by existing theoretical bounds. At the same time the indiscriminate use of test statistics that are dependent across stages of the analysis is warned against. An appropriate strategy for dealing with differences in measurement units between genotypic and phenotypic data is suggested. This methodology is applied for the purpose of searching for correlations between gene expression in the brain and behavioral traits (endpoints) measured in the open field test, while considering multiple mouse strains and the different brain regions where gene expression is measured.

The scope of the problems and their proposed solutions, as detailed in the Methods section of Chapter 8, will be more comprehensible with the help of a concrete example. Such experiments are not common yet, so the experiment analyzed will be described in some detail. In addition, FDR control through the two approaches is studied with the aid of simulations, accounting for the possible potential dependence between the test statistics used in two stages of analysis. These results teach us of the nature of the FDR criterion using each of the two approaches, and thus help in determining the appropriate threshold to set in our real data. In order to further investigate the strength of the two proposed approaches, they are compared to single-step methods existing in literature, some of which were considerably motivated by microarray data analysis. The results of this chapter have been written jointly with D. Yekutieli and Y. Benjamini and submitted for publication in Bioinformatics (Reiner et al, 2005).

Chapter 9 is the closing chapter of this thesis. It is aimed to stress the major contributions of the research, both in Statistics and Genomics. It first summarizes the results of the work and their contributions and implications in both
fields. It concludes with suggested directions for further research, discussing questions remaining open for investigation as well as new ones emerging from this thesis.
2. Microarray Technology

2.1. Overview

2.1.1. What are microarrays?

Microarrays are a recent and innovative biotechnology, which allows the monitoring of genetic activity in cells for thousands of genes simultaneously. Microarray are being used increasingly within experiments involved in biological and medical research to address a wide range of problems, such as classification of tumors (e.g. Golub et al, 1999, Alizadeh et al, 2000), tracking of patterns of co-regulated genes and monitoring response to different treatments or environmental stress conditions (e.g. Alon et al, 1999, Perou et al, 1999, Gasch et al, 2000, Ross et al, 2000) and identification of biological pathways (e.g. Butte et al, 2000, Friedman et al, 2000, ideker et al, 2000, Pe’er et al ,2001, Hartemink et al, 2002). The development of the microarray technology was motivated by the increasing rate at which genomes were being sequenced. With the DNA sequences of many genomes already completed and the recent release of the sequenced human genome (Science Magazine, 2001), microarray technology serves as an essential tool in defining the role of each gene and understanding how the genome functions as a whole.

A microarray is used to measure gene activity in a genetic sample extracted from the subjects of interest. It is a coated microscopic glass slide on which DNA of different genes is printed. The genetic material in the sample hybridizes to the DNA on the glass and its abundance is measured. The technological principles of this process are further detailed in section 2.3, preceded by an introduction to essential genetic terminology in section 2.2. The output of a microarray experiment is a dataset containing a gene activity measurement for each gene. Based on this data, research questions concerning the biolobical conditions examined in the experiment can be studied. Data processing and analysis methodologies are reviewed in sections 2.4 and 2.5.

There are several types of microarray systems, including the cDNA microarrays developed in the Brown and Botstein labs at Stanford (DeRisi et al, 1997,
Brown and Botstein, 1999 and Hughes et al, 2001) and the high-density oligonucleotide chips from the Affymetrix Company (Lockhart et al, 1996). Although this thesis uses the former for its application, its results are relevant for both types. They are both based on hybridization - pairing between two complementary strands of nucleic acids (one of which is immobilized on a matrix). Complementarity provides a core capability of molecular biology, as it possess high sensitivity and specificity of detection as a consequence of mutual selectivity between the two strands.

2.1.2. **Therapeutic motivation**

The Human Genome Project has recently identified the genes that make up the coding system for the manufacturing of proteins in the human body. The proteins make up a catalogue of possible therapeutic targets for medicine (Science Magazine, 2001). This has drawn us to the next big frontier in genome-based medicine, the identification of the functions of these genes and their interaction pathways in health and disease. This current emphasis on ‘functional genomics’ has been advanced greatly by the emergence of microarray technology, and consequently, gene expression analysis using microarrays has become a key component in the drug development pipeline (e.g. Golub et al, 1999, Pemoroy et al, 2002, Gordon et al, 2002).

As reviewed by Gerhold et al (2002), the promising potential of microarrays is based on their inherent ability to monitor *gene* *expression* of thousands of genes simultaneously. These expressions characterize the responses of cells or tissues to the biological state they are in. Thus microarrays facilitate medical research by documenting detailed responses of cells and tissues to both disease and effects of drug treatments. Consequently, pharmaceutical and biotechnology companies integrate gene expression analysis into their existing processes for drug discovery and development. At the stage of early target discovery and validation, investigators use microarrays to prioritize a few genes leading to potential therapeutic targets on the basis of various criteria. Subsequently, in the compound discovery and optimization stage, microarray data analysis assists in drug discovery and toxicology by eliminating poor compounds and optimizing the selection of promising leads.
2.2. Genetic terminology

A deoxyribonucleic acid or DNA molecule (Watson and Crick, 1953) is a nucleic acid that contains the genetic instructions specifying the biological development of all cellular forms of life. DNA is often referred to as the molecule of heredity as it is responsible for the genetic propagation of most inherited traits. In humans, these traits can range from hair color to disease susceptibility. During cell division, DNA is replicated and can be transmitted to offspring during reproduction. In complex eukaryotic cells such as those from plants, animals and fungi, most of the DNA is located in the cell nucleus. By contrast, in simpler cells called prokaryotes, such as bacteria, DNA is not separated from the cytoplasm – a jelly-like material that fills the cells - by a nuclear envelope.

The DNA molecule is a double-stranded chain composed of four basic molecular units called nucleotides. Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases. The four different bases found in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). The two chains of the DNA molecule are held together by hydrogen bonds between nitrogen bases, forming a double helix, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. A gene consists of a segment of DNA, which codes for a particular protein, the ultimate expression of the genetic information. Proteins are molecular compounds that are essential to the structure and function of all living cells. They are sequences of twenty different types of amino acids.

The expression of the genetic information stored in the DNA molecule occurs in three stages known as the central dogma of molecular biology (Crick, 1958 and Crick, 1970). The central dogma describes the flow of genetic information from DNA for the purpose of making proteins. This process is broken down into: transcription, translation and replication (Figure 1). During transcription, DNA is transcribed into messenger ribonucleic acid or mRNA, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine. The mRNA is then transported out of the nucleus into the cytoplasm. During translation, mRNA is translated to produce a protein. The correspondence between DNA’s four-letter alphabet and a protein’s twenty-letter alphabet is specified by the genetic code, which relates nucleotide triplets to amino acids. A series of 3 nucleotides, or a codon, is responsible for coding one amino acid. The amino acids are carried to the site of translation by transfer RNA (tRNA). Finally,
replication is the use of existing DNA as a template for the synthesis of new DNA strands. Replication is carried out by using the enzyme *DNA polymerase* and its associated proteins, to copy, or replicate, the master template itself so that the cycle can repeat DNA → RNA → protein. DNA can self-replicate by using one strand of the double helix as a template. In eukaryotes, this happens in the nucleus. In prokaryotes, replication occurs in the cytoplasm.

Different aspects of gene expression can be studied using microarrays, such as expression at the transcription or translation level, and subcellular localization of gene products. To date, attention has focused primarily on expression at the transcription stage, i.e., on mRNA or transcript levels. The level, or abundance, of mRNA can be measured with the help of the following process: A single strand of DNA is synthesized in the lab to complement the bases in a given strand of mRNA. This strand is called *complementary DNA* or *cDNA*. cDNA is most often synthesized from mRNA using the enzyme *reverse transcriptase*. A strand of cDNA will bind to its exact complement. As described in section 2.3, microarrays derive their power and universality from this key property of DNA molecules, *complementary base-pairing*, where the DNA nitrogen bases pair to their complements. The term *hybridization* is used to refer to the annealing of DNA strands from different sources according to the base-pairing rules.

### 2.3. Expression profiling techniques

#### 2.3.1. cDNA microarrays

As reviewed by Duggan et al (1999), in a cDNA experiment, DNA copies, called clones, of genes of interest are obtained (Figure 2). They are amplified by a technology called PCR and purified, and then printed on a coated glass microscope slides using a computer controlled, high-speed robot, thereby forming the ‘probe’ array. RNA from test and reference samples, called ‘target’, is fluorescent labeled with either green (Cye3-dUTP) or red (Cye5-dUTP) fluorescent dyes. It is transformed back to its complementary DNA (cDNA) representation through a reverse transcription. The dyed samples are pooled and allowed to hybridize to the clones on the array. Laser excitation of the incorporated targets yields an emission, which is measured using a scanning laser microscope. Monochrome images from the scanner are imported into software in which the images are pseudo-colored and merged, such that the color intensity quantifies expression.
Figure 1: **The Central Dogma of molecular biology**

The three steps of the central dogma are (i) Transcription, during which DNA is transcribed into mRNA, (ii) Translation, during which mRNA is translated to produce a protein. (iii) DNA strands are repeatedly synthesized to repeat the process. Figure taken from http://users.ugent.be/~avierstr/principles/centraldogma.html, the public website of Gent University, Belgium.

Data from a single hybridization experiment is viewed as a normalized ratio indicative of the relative abundance of the two samples. An internal control is thus provided for each measurement (spot on the array). Each spot on the microarray represents a mix of treatment and control expression of one gene. Figure 3 displays...
an example of color intensities representing different level of differential expression in treated cells relative to non-treated cells.

**Figure 2: cDNA microarray experiment schema**

Clones of genes of interest are obtained, amplified using PCR and purified, and then printed on a coated glass microscope slides using a robot. RNA from both test and reference samples is fluorescent labeled with either green or red dyes. It is transformed back to its cDNA representation through a reverse transcription. The dyed samples are pooled and hybridized to the clones on the array. Laser excitation measures expression levels through color intensities represented on an image. Figure taken from Duggan et al (1999).

**Figure 3: Quantification from two-color hybridization**

Each spot on the microarray represents a mix of treatment and control expression of one gene. The two spots represent genes that show different directions of high differential expression in treated cells relative to non-treated cells.

Figure taken from Duggan et al (1999).
2.3.2. **High density synthetic oligonucleotide microarrays**

The fabrication of hundreds of thousands of polynucleotides (DNA chains composed of multiple nucleotides) at high spatial resolution in precise locations on a surface is the key technology underlying this method. As explained in Lipshutz et al (1999), light-directed synthesis helps construct high-density DNA arrays using two techniques: photolithography and solid-phase DNA synthesis (Figure 4). Synthetic linkers terminated with photochemically removable protecting groups are attached to a glass substrate. Light is directed through a photolithographic mask to specific areas on the surface to produce localized photodeprotection and to activate selected sites. The first of a series of nucleotides is incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. Next, light is directed to different regions of the substrate by a new mask, and the chemical cycle is repeated. Highly efficient strategies can be used to synthesize arbitrary polynucleotides at specified locations on the array in a minimum number of chemical steps.

![Figure 4: Light directed oligonucleotide synthesis](image)

Oligonucleotide arrays for expression monitoring are designed and synthesized based on sequence information alone, without the need for physical intermediates such as clones, PCR products and cDNAs. The key to their use is the
targeted design of sets of probes to specifically monitor the expression levels of as many genes as possible. Using as little as 200 to 300 bases of gene, independent chains of 25 bases, called oligonucleotides, are selected to serve as sensitive, unique, sequence-specific detectors. Probe design is based upon complementarity to the selected gene reference sequence, and its uniqueness relative to other genes. The arrays are designed \textit{in silico}, and as a result, it is not necessary to prepare, verify, quantitate and catalogue a large number of cDNAs, PCR products and clones, and there is no risk of a misidentified tube, clone, cDNA or spot.

Key to this approach is the use of probe redundancy. First, multiple oligonucleotides of different sequence are designed to hybridize to different regions of the same RNA. The use of multiple independent detectors for the same molecule greatly improves signal-to-noise ratios (due to averaging over the intensities of multiple array features). An additional level of redundancy comes from the use of mismatch (MM) control probes that are identical to their perfect match (PM) partners except for a single base difference in a central position (Figure 5). The MM probes act as specificity controls that allow the direct subtraction of background signals, and allow discrimination between ‘real’ signals and those due to non-specific or semi-specific hybridization (see some suggested improvements regarding the use of MM as background, e.g. Binder et al, 2004 and He et al, 2005).

\textbf{Figure 5: Expression probe and array design}

Mismatch (MM) control probes are identical to their perfect match (PM) partners except for a single base difference in a central position. Figure taken from Lipshutz et al (1999).
2.4. **Data analysis questions**

As reviewed by Slonim et al (2002), many different biological questions have been routinely studied using transcriptional profiling on microarrays. It has become increasingly clear that simply generating the data is not enough; one must be able to extract from them meaningful information about the system being studied. As transcriptional profiling has grown in popularity, statistical methods for interpreting the data have proliferated. A wide range of approaches is available for gleaning insights from the data obtained from microarray experiments. Some of the common themes in microarray data analysis include differential expression detection, pattern discovery, class prediction and network and pathway identification. Chapter 8 of this thesis focuses on the more recent experiments enclosing higher level of complexity. These are being conducted as part of the effort to adopt a system perspective and achieve more targeted conclusions that can link specific genes with biological settings and functional attributes.

2.4.1. *Detecting differential expression*

As pointed by Slonim et al (2002), the most basic question one can ask in a microarray experiment is which genes’ expression levels changed significantly from one condition to the other. Answering this question involves many considerations. There may be two experimental conditions or many, the conditions may be independent or related to each other in some way (as in a time series), or there may be many different combinations of the levels of the experimental factors. Replicates, if present at all, might be samples from different animals or repeated hybridizations of the same samples. Reflecting this variety, many different methods are commonly used for identifying significant changes. While at earlier stages the costly replication of arrays was rare, as researchers recognized their importance, experiments with replicates became more common.

Researchers typically rely on variants of common statistical tests. Standard statistical tests are used for detecting significant change between repeated measurements of a variable in two groups, such as the \( t \)-test (e.g. Dudoit et al, 2002b), its generalization to multiple groups via the ANOVA \( F \) statistic (e.g. Pavlidis, 2003), non-parametric rank-based statistics (e.g. Zhan et al, 2002) and variations on these tests (e.g. Tusher et al, 2001).
2.4.2. Pattern discovery

As reviewed by Slonim et al (2002), pattern discovery provides a high-level overview of a data set and may be the first analysis step in a study that ultimately involves other analytical methods. Such techniques include a variety of dimension-reduction methods, as well as various clustering techniques designed for finding groups of genes or of biological conditions within the data, that show similar patterns of expression. What these methods have in common is that they simplify the data set, ideally in ways that impart additional information about its structure, and that they are considered ‘unsupervised’, meaning that the reduction is derived solely from the data rather than reflecting any previous knowledge or classification scheme.

Principal components analysis (Landgrebe et al, 2002), singular value decomposition (Alter et al, 2000) and multidimensional scaling (Bittner et al, 2000), are related dimension reduction techniques that can be used for visualizing large data sets such as gene expression data. The term ‘clustering’ applies to a wide variety of unsupervised methods for organizing multivariate data into groups with roughly similar patterns. Clustering has many applications in expression-data analysis. Clues to unknown gene function may be inferred from clusters of genes similarly expressed across many samples (Eisen et al, 1998). Clustering samples over the expression levels of multiple genes has been proposed as a way of defining new disease subclasses (Golub et al, 1999). Cluster analysis may be used primarily for data reduction and visualization, or it may be used to generalize or predict the categorization of new samples (Yona, 1999). To solve any of these problems, researchers can choose from a vast library of techniques for grouping multivariate data.

Perhaps most familiar to biologists are the hierarchical clustering methods (two of the first studies that used it for microarray data are Eisen et al, 1998 and Spelman et al, 1998). In this family of techniques, all data instances start in their own clusters, and the two clusters most closely related by some similarity metric are merged. The process of merging the two closest clusters is repeated until a single cluster remains. This arranges the data into a tree structure that can be broken into the desired number of clusters by cutting across the tree at a particular height. Tree structures are easily viewed and understood, and the hierarchical structure provides potentially useful information about the relationships between clusters. Figure 6
demonstrates hierarchical clustering result on data of 11 genes characterizing sensitivity to compound cytochalasin D in 60 cancer categories (Staunton et al, 2001).

Another common family of clustering methods is that of partition or centroid algorithms (Jain and Dubest, 1988). These methods generally require specification of the number, \( k \), of clusters, and start with \( k \) data points that may be chosen either randomly or deliberately. These \( k \) points are used as the ‘centroids’ — the multidimensional center points — of an initial set of clusters. The algorithm then partitions the samples into the \( k \) clusters, optimizing some objective function (such as within-cluster similarity) by iteratively assigning samples to the nearest centroid’s cluster and adjusting the centroids to represent the new clusters’ center points. A variation that allows samples to influence the location of neighboring clusters is known as the self-organizing map. Such maps are particularly valuable for describing the relationships between clusters.

Other techniques seek to optimize a measure of within-cluster similarity or separation between clusters, but avoid specifying the number of clusters ahead of time. Instead, they specify information-theoretic bounds on cluster membership (e.g. Sharan and Shamir, 2000, De Smet et al, 2002). Methods based on statistical modeling are another mean of making inference on gene expression data (Yeung et al, 2001, Fraley and Raftery, 2002). Model-based methods assume the data can be generated by a specified statistical model, such as a mixture of Gaussian distributions. Then, parameters that best fit the data under the assumed model are searched for.

Figure 6: Hierarchical clustering diagram
11 genes clustered by cancer categories. The red and green colors signify gene down-regulation and up-regulation, respectively. Figure taken from Slonim et al (2002).
2.4.3. Class prediction

In contrast to pattern discovery, class prediction methods are techniques specifically designed to classify objects into known groups. A wealth of machine-learning literature describes computational techniques for classifying multidimensional data (Ben-Hur et al, 2002). Most such methods include a training phase run on samples whose classes are already known, and a testing phase, in which the algorithm generalizes from the training data to predict classifications of previously unseen samples. Because of this directed training phase, prediction methods are referred to as ‘supervised’ classification methods. For microarray data, prediction generally refers to the classification of patients’ samples by characteristics such as disease subtype or response to treatment. The goal may be diagnostic, offering a new way to distinguish similar-looking diseases (Golub et al, 1999, Khan et al, 2001, Armstrong et al, 2002), or it may be a true effort to predict the clinical outcome (Alizadeh et al, 2000, Pomeroy et al, 2002).

As with clustering, choosing a prediction method requires selecting from a vast range of techniques. Linear and quadratic discriminant methods are a straightforward way to classify and identify classifiers among the genes (Dudoit et al, 2002a). Another powerful approach is the $k$-nearest neighbor prediction, in which the prediction for a test sample $x$ is the most common class label among the $k$ training samples most similar to $x$ (Pomeroy et al, 2002, Dudoit et al, 2002a). Simple neural networks may be effective at learning the complex functions often inherent in multi-class diagnostic problems (Mitchell, 1997).

Two other well-studied classes of algorithms are of growing interest for microarray prediction problems: support vector machines (SVMs) and decision tree classifiers. SVMs (Brown et al, 2000, Furey et al, 2000) are a family of statistical machine-learning methods that try to draw a hyperplane in $n$-dimensional gene-expression space between the training examples from two classes. If no separating hyperplane exists, the samples are mapped into a higher-dimensional space where such a separator does exist. The algorithms minimize potential overfitting problems by choosing the separator farthest from the training samples, thus leaving room for generalization. Decision tree algorithms (Mitchell, 1997) classify samples by filtering them through a tree-like structure, testing at node some simple attribute of that sample. Single decision trees are particularly prone to overfitting. However, as tree models are easily built, easily understood, and able to model quite complex functions,
there are many modified tree-based techniques for avoiding overfitting and improving performance, such as ‘pruning’ the tree.

2.4.4. **Inferring regulatory pathways and networks**

Analysis of differential expression may provide new information about the biological pathways involved in a process. Looking specifically for information on gene interactions indicated by expression data may ultimately suggest new pathways and associations. Even simple pairwise comparisons can indicate novel interactions (Butte et al, 2000). More complex gene relationships may be discovered as researchers learn to combine the data in more complex ways. Although it is optimistic to assume that expression data alone will be sufficient for the inference of complete regulatory pathways, several recent studies successfully tackle parts of the problem. Bayesian network models and variations can help discover novel interactions, information dependencies and regulatory relationships from expression data (Liang et al, 1998). Whereas the posterior probabilities of all models are likely to be very low, repeated random resampling of the data, i.e. bootstrapping, can help in identifying ‘high-probability’ genes.

More recent modifications of Bayesian network methods focus further on finding probabilistically supported gene interactions, on combining these into subnetworks (e.g. Pe'er et al, 2001, Hartemink et al, 2002), on modeling ‘latent’ or hidden variables representing biological information unavailable to the model (Segal et al, 2001, Yoo et al, 2002) and on incorporating prior biological knowledge or annotation (Hartemink et al, 2002). It has proved more difficult to infer the direction of causal relationships successfully directly from transcriptional data. In general, models that incorporate existing constraints from other data sources seem to produce hypotheses that agree better with existing biological knowledge than do models learned from the expression data alone (Hartemink et al, 2002).

2.5. **Preprocessing**

As reviewed by Quackenbush (2002), common strategies for labeling the RNA samples include the use of a single label and independent arrays for each sample, or a single array with distinguishable fluorescent dye labels for the individual RNAs. Regardless of the approach chosen, the arrays are scanned after hybridization and independent grayscale images are generated for each pair of samples to be
compared. These images must then be analyzed to identify the arrayed spots and to measure the relative fluorescence intensities for each element. There are many commercial and freely available software packages for image quantification. Although there are minor differences between them, most give high-quality, reproducible measures of hybridization intensities. They provide methods to deal with the type of measurement reported (mean, median or integrated intensity, or the average difference), the background correction performed and the spot-quality assessment and trimming used.

Assuming a high-quality intensity measurement is available for each biological sample we assay, the hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene-by-gene basis. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low-quality measurements and to facilitate comparisons, by adjusting the measured intensities for biases stemming from technological factors.

2.5.1. **Expression ratios**

As summarized by Quackenbush (2002), suppose an array has \( N \) distinct elements. A query and a reference sample are compared, which for convenience will be called \( R \) and \( G \), respectively (for the red and green colors commonly used to represent array data). Then the ratio for gene \( i \), \( 1,...,N \) can be written as \( \frac{R_i}{G_i} \). This definition does not limit us to any particular array technology: the measures \( R_i \) and \( G_i \) can be made on either a single array or on two replicate arrays.

Although ratios provide an intuitive measure of expression changes, they have the disadvantage of treating up- and down-regulated genes differently. Genes up-regulated by a factor of 2 have an expression ratio of 2, whereas those down-regulated by the same factor have an expression ratio of \((-0.5)\). Conversely the logarithm transformation of the ratio has the advantage of producing a continuous spectrum of values and treating up- and down-regulated genes in a similar fashion. This is due to the fact that logarithms treat numbers and their reciprocals symmetrically, and a gene expressed at a constant level (with a ratio of 1) has a
log(ratio) equal to zero. Therefore the log-transformed ratio has quite a symmetric distribution, which is often close to the normal.

2.5.2. Normalization

Normalization reduces non-biological variations from the expression data. This implies adjustment of the individual hybridization intensities in order to balance them, so that meaningful biological comparisons can be made. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels.

There are many approaches to normalizing expression levels. The simplest one is the total intensity normalization (Quackenbush, 2002), which assumes equal quantities of RNA for the two compared samples and random sampling representation of the genes in the organism. Therefore, the total hybridization intensities summed over all elements in the arrays should be the same for each sample. Using this approach, a normalization factor is calculated by summing the measured intensities in both channels

$$\frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} G_i}$$

In addition to total intensity normalization, there are a number of alternative approaches to normalizing expression ratios, including linear regression analysis (Chaterjee, 1991), log centering, rank invariant methods (Tseng et al, 2001) and Chen’s ratio statistics (Chen et al, 1997), among others. However, none of these approaches takes into account systematic biases that may appear in the data. Several reports have indicated that the log(ratio) values can have a systematic dependence on intensity, which most commonly appears as a deviation from zero for low-intensity spots. Locally weighted linear regression (lowess) (Cleveland, 1979) analysis has been proposed (Yang et al, 2002) as a normalization method that can remove such intensity-dependent effects in the log(ratio) values.
Figure 7: **M-A plot**

The log of ratios of the two dyes intensities (M) is plotted versus the log of their product (A), for each element on the array.

- a. Intensity-specific bias is visible through non-constant deviations from the M=0 line.
- b. Local weighted linear regression (LOWES) eliminates these biases for each element.

Figure taken from Quackenbush (2002).

An easy way to visualize intensity-dependent effects, and the starting point for the lowess analysis described here, is to plot the measured $M = \log(R_i/G_i)$ for each element on the array as a function of $A = \log(R_i \cdot G_i)$ (Figure 7). This M-A plot can reveal intensity-specific artifacts in the log(ratio) measurements (Figure 7a). Lowess detects systematic deviations in the plot and corrects them by carrying out a local weighted linear regression as a function of the log(intensity) and subtracting the calculated best-fit average log(ratio) from the experimentally observed ratio for each data point (Figure 7b). Lowess uses a weight function that de-emphasizes the contributions of data from array elements that are far from each point.

Huber et al (2002) generalize the model offered by Chen et al (1997), by assuming quadratic variance-versus-mean dependence proposed by Rocke and Durbin (2001). Based on this, they derive a parametric family of transformations of the measured intensities, such that the variance of the transformed intensities becomes approximately independent of the mean. In addition, they take into account the fact that many image quantification methods produce a certain fraction of non-positive intensities, for which ratios make no sense and the real-valued logarithm is not defined. Thus they offer transformed intensities that may be viewed as a generalization of the log-ratio, as the two coincide for the highly expressed genes.

For gene $k$ with expression $Y_i$ where $E(Y_i) = u_i$ and $Var(Y_i) = v_i$, assume $u_i$ depends on $k$ through a quadratic function of the mean $u_i$ of the following form:
\[ v_k = v(u_k) = \left(c_j u_k + c_2\right)^2 + c_3, \]

with \( c_j > 0 \). Their transformation is given by (Tibshirani, 1988),

\[
h(y) = \int \frac{1}{\gamma \sqrt{v(u)}} du
\]

which is a result of a linear approximation of \( h(Y_i) \) around \( h(u_i) \). Inserting the variance expression into the transformation expression yields

\[
h(y) = \gamma \cdot \arcsin(h(a + by)),
\]

where \( \gamma = \frac{1}{c_i}, \ a = \frac{c_2}{\sqrt{c_j}}, \) and \( b = \frac{c_1}{\sqrt{c_j}} \).

A plot of this function is represented by the solid line in Figure 8, super-positioned on the intensity histogram. It is compared to the log ratio transformation, represented by the dotted line. As can be seen, the two transformations become more similar as intensity increases, but the proposed transformation has the advantage of no singularity at zero, and it continues to be smooth and real-valued in the range of small or negative intensities. The parameters of the model are estimated from the data with a robust variant of the of maximum likelihood estimation.

Figure 8: Variance stabilizing transformation

The log ratio and arcsine transformations are super-positioned on the intensity histogram. They become similar as intensity increases, but the arcsine transformation has the advantage of no singularity at zero, and it continues to be smooth and real-valued for small or negative intensities. Figure taken from Huber et al (2002).
2.5.3. Spatial effects removal

The physical layout of the chips on individual expression levels should also be considered, since they impact the measurement error. While non-systematic spatial biases as local smears are not simple to detect and restrict, factor-attributed effects such as block, row and column effect on the chip can be located due to the massive amount of observations each of them contain. The median polish method suggested by Tukey (1977) (see also Emerson and Wong, 1985), may be applied as described in Chapter 7, to identify and remove such spatial effects, and thus increase power. This algorithm iteratively subtracts the median signal for each category of each factor, until the median of the residuals is 0 in all directions. The “polished” signals are then used for data analysis. The use of medians rather than means makes this procedure resistant to real differential expressions that otherwise will be smoothed out.
3. False Discovery Rate Methodology

3.1. The False Discovery Rate criterion

The common approach in simultaneous testing is to construct a procedure that controls the FWER (Tukey, 1953, Hochberg and Tamahane, 1987). Benjamini and Hochberg (1995) offer another measure for the erroneous rejection of a number of true null hypotheses, the false discovery rate (FDR). The FDR is the expected proportion of erroneously rejected null hypotheses among the rejected ones. When some of the tested hypotheses are in fact false, FDR control is less strict than FWER control, and thus FDR controlling procedures are potentially more powerful. While some situations require FWER control, such as when the result of rejecting hypotheses yields an action (e.g. a drug is approved), in other cases FDR control is sufficient. The analysis of gene expression data is such a case, as its purpose is to extract genes that are potential candidates for further investigation. Several erroneous rejections will not distort the conclusions at this stage of the investigation, as long as their proportion is small. Such errors do incur economical cost in that pursuing them at later stages will result in loss of time and money. Controlling the probability of at least one such rejection appears to be over-conservative and will result in reduced experimental efficiency due to unnecessary loss of power. Controlling the FDR instead allows control of the proportion of effort invested in vain, on the average, at the next stage of the investigation.

FDR is defined as follows. Consider a family of $m$ simultaneously tested null hypotheses of which $m_0$ are true. For each hypothesis $H_i$ a test statistic is calculated along with the corresponding p-value $P_i$. Let $R$ denote the number of hypotheses rejected by a procedure, $V$ the number of true null hypotheses erroneously rejected, and $S$ the number of false hypotheses rejected. Now let $Q$ denote $V/R$ when $R>0$ and 0 otherwise. Then the FDR is defined as

$$FDR=E(Q).$$

As shown in Benjamini and Hochberg (1995), the FDR of a multiple comparison procedure is always smaller than or equal to the FWER, where equality holds if all null hypotheses are true. Thus control of the FDR implies control of the FWER when all hypotheses are true. In the context of gene expression analysis, this
result means that if in reality no genes are differentially expressed and the FDR is controlled at some level $q$, then the probability of erroneously detecting any differentially expressed genes is less than or equal to $q$.

3.2. Economic interpretation of the FDR criterion

Microarray experiments, like other kinds of high throughput experiments, typically serve as means of screening for the purpose of supplying the researcher with an initial pool of candidates. Therefore, statistical considerations that limit the power to generate candidate hypotheses should not be taken at this stage. This argument is acceptable in the sense that protection is not needed against even a single type I error, so that the severe loss of power involved in such protection is not justified. However, the proportion of errors in the pool of candidates is of great economical significance since follow-up studies are costly, and thus avoiding multiplicity control is costly. Indeed, the FDR criterion is economically interpretable; when considering a potential threshold, the adjusted FDR gives the proportion of the investment that is about to be wasted on false leads. Here FDR screening essentially serves as an initial dimension reduction technique. Thus the choice of the FDR level $q$ is an economical one. It is for these reasons that multiplicity should be controlled when testing differently expressed genes in microarray analysis, and that it is best done using the FDR criterion.

Controlling the FDR at the screening stage of the research carries a benefit for the next research stages, as shown by Benjamini and Yekutieli (2005). Consider a study with $R_1$ significant results, while controlling the FDR at level $q_1$. A follow-up study is conducted on the pool identified in the first stage, and a level $\alpha$ FWER controlling procedure is applied. It has been shown that the FWER of the combined two-stage study is $\alpha q_1$. Alternatively, if a level $q_2$ FDR controlling procedure is used in the second stage, the combined two-stage FDR is shown to satisfy $E(V_2 / R_2) \leq q_1 q_2$. Either way, the initial chosen level, $q_1$, can be allowed to be quite high. For example, assume that in a microarray experiment involving 10,000 genes, 100 genes were identified using an FDR controlling procedure with $q=0.2$. Next, using Bonferroni with $\alpha=0.25$ at the second stage, that is, assessing individual significance by comparing to $0.25/100$, controls the FWER at the level of 0.05.
3.3. **The Linear Step-Up procedure (BH)**

This procedure makes use of the ordered p-values \( P(1) \leq \ldots \leq P(m) \). Denote the corresponding null hypotheses \( H(1), \ldots, H(m) \). For a desired FDR level \( q \), the ordered p-value \( P(i) \) is compared to the critical value \( q \cdot \frac{i}{m} \). Let \( k = \max \{ i : P(i) \leq q \cdot \frac{i}{m} \} \).

Then reject \( H(1), \ldots, H(k) \), if such a \( k \) exists. Benjamini and Hochberg (1995) show that when the test statistics are independent, this procedure controls the FDR at the level \( q \).

Actually, the FDR is controlled at level \( q \cdot \frac{m_0}{m} \leq q \).

Benjamini and Yekutieli (2001) further show that \( FDR \leq q \cdot \frac{m_0}{m} \) for positively dependent test statistics as well. The technical condition under which the control holds is that of positive regression dependency on each test statistic corresponding to the true null hypotheses (as defined there). In particular, the condition is satisfied by positively correlated normally distributed one-sided test statistics, and their studentized \( t \) tests. The studentized form applies to the cDNA microarray data structure as a result of the tendency of the measurement errors of gene expressions to be positively correlated, due to common latent factors involved. When no real differential expression exists, these are the main sources of variability. Furthermore, since up-regulation and down-regulation are about equally likely to occur, the property of FDR control can be extended to two-sided tests (Yekutieli, 2002).

For more general cases, in which the positive dependency conditions do not apply, Benjamini and Yekutieli (2001) prove that replacing \( q \) with \( q \cdot \frac{m_0}{m} \) in the linear step-up procedure will provide control of the FDR. However, this modification may be too conservative for the microarray problem. In fact, the simulation study detailed in Chapter 5 presents the claim that working with \( q \) already controls the FDR.
3.4. **Adaptive approaches**

Since the BH procedure controls the FDR at a level too low by a factor of $\frac{m}{m_0}$, it is natural to try to estimate $m_0$ and use $q^* = q \cdot \frac{m}{\hat{m}_0}$ instead of $q$ to gain more power. Estimating $m_0$ from a set of p-values goes back to Schweder and Spjøtvoll (1982). Hochberg and Benjamini (1990) formalized their approach and synthesized an adaptive procedure that controls the FWER (see Turkheimer, 2001, for further progress). Benjamini and Hochberg (2000) suggest the adaptive procedure that combines the estimation of $m_0$ with the BH procedure. The procedure first uses the linear step-up procedure at level $q$, and stops if no hypotheses are rejected. Otherwise, $m_0$ is estimated by the following algorithm:

1. Compute $m_0[k] = \frac{m + I - k}{I - P(k)}$.
2. Starting with $k = 2$, stop when for the first time $m_0[k] > m_0[k-1]$.
3. Estimate $\hat{m}_0 = \text{Ceiling}(\min(m_0[k], m))$.
4. Apply the linear step-up procedure with $q^* = q \cdot \frac{m}{\hat{m}_0}$.

Benjamini et al (2001) suggest a two-stage procedure with proven FDR controlling properties under independence. They also show in simulations that their procedure is the only adaptive procedure offering FDR control under positive dependency. The procedure is as follows: apply the BH procedure at level $q' = \frac{q}{I + q}$, let $r_i$ be the number of rejected hypotheses; if $r_i = 0$ reject no hypotheses and stop; if $r_i = m$ reject all $m$ hypotheses and stop; otherwise: Let $\hat{m}_0 = m - r_i$ and apply the BH procedure with $q^* = q \cdot \frac{m}{\hat{m}_0}$.

Adaptive methods offer better performance only by utilizing the difference between $\frac{m}{m_0}$ and 1. If the difference is small, i.e. when the potential proportion of differentially expressed genes is small, they offer little advantage in power while their properties are not well established under dependency. As more specific genes are pre-selected to the microarray experiments, such that the proportion of differentially
expressed genes is not small, \( \frac{m}{m_0} \) gets smaller, and the adaptive procedures will offer
a more detectable advantage. Much of the more recent work on FDR methodology in
microarray analysis involves adaptive methods. The literature survey on adaptive
approaches is delayed to the next chapter.

3.5. Multiplicity adjusted p-values

The results of a multiple testing procedure can be reported as multiplicity
adjusted p-values. As with the regular p-value, each adjusted p-value is compared to
the desired significance level, and if smaller, the hypothesis is rejected. Therefore,
the way adjusted p-values are used and interpreted remains conveniently familiar,
regardless of the adjustment procedure complexity.

For an FWER controlling procedure, the adjusted p-value of an individual
hypothesis is the lowest level for which \( FWER \leq \alpha \). For instance, for the Bonferroni
procedure, the adjusted p-value is simply \( P_j \cdot m \). For Holm’s procedure (Holm, 1979),
where \( k \) is set to be the smallest \( i \) that satisfies \( \frac{\alpha}{m+1-i} \), and reject all
hypothesis \( H_{(i)} \), \( i=1,\ldots,k-1 \), the adjusted p-value can be calculated as
\( P_{(j)}^{Holm} = \max_{1 \leq j \leq i} \{ P_{(j)} \cdot (m+1-i) \} \). For an FDR controlling procedure, the adjusted p-
value of an individual hypothesis is the lowest level of FDR for which the hypothesis
is first included in the set of rejected hypotheses. Thus the adjusted p-value of \( P_{(j)} \)
using the BH procedure, is \( P_{(j)}^{BH} = \min_{j \leq i} \{ P_{(i)} \cdot \frac{m}{i} \} \). Figure 9a displays typical
distribution of raw p-values versus their rank, while Figure 9b zooms into FDR
adjusted p-values relative to the raw values. The adjusted p-values are compared to
the desired significance level, here 0.05.
Figure 9: Raw and FDR adjusted p-values

a. Raw p-values corresponding the true null hypotheses are distributed U[0,1], and thus follow the straight line while the smaller p-values, corresponding to the false null hypotheses do not follow this distribution, and fall below the line.

b. Zooming in: FDR adjusted p-values are obviously higher than the raw values, and should be compared directly with the desired FDR level. For 0.05 significance level, FDR control yields around 970 rejections.

3.6. Resampling FDR adjustments

For data containing high inter-correlations, a generally designed multiple comparison procedure may be over-conservative under specific dependency structures. Resampling-based multiple testing procedures, introduced by Westfall and Young (1989), utilize the empirical dependency structure of the data to construct more powerful FWER controlling procedures. In p-value resampling, the data is
repeatedly resampled under the complete null hypotheses, and a vector of resampling-based p-values is computed. The underlying assumption is that the joint distribution of p-values corresponding to the true null hypotheses, which is generated through the p-value resampling scheme, represents the real joint distribution under the global null hypothesis. Thus, for each value of $p$, the number of resampling-based p-values less than $p$, denoted by $V^*(p)$, is an estimated upper bound to the expected number of p-values corresponding to true null hypotheses less than $p$. The WY procedure estimates the FWER by

$$FWE^{est}(p) = \frac{\#\{V^*(p) > 0\}}{N},$$

where $N$ is the number of resampling iterations. Then $H_0$ is rejected if $FWE^{est}(p_i) \leq \alpha$.

Yekutieli and Benjamini (1999) follow a similar path to achieve FDR p-value adjustments, but, unlike the FWER, the FDR is also a function of the number of false null hypotheses rejected. Therefore, for each value of $p$, they first conservatively estimate the number of false null hypotheses, denoted by $\hat{s}(p)$, using the BH procedure, and then estimate the FDR adjustment by

$$FDR^{est}(p) = E_{V^*(p)} \frac{V^*(p)}{V^*(p) + \hat{s}(p)}.$$

Two estimation methods are suggested differing by their strictness level. The FDR local estimator is conservative on the mean, and the FDR upper limit bounds the FDR with probability $(1 - \alpha)%$.

The two above methods use resampling to estimate the joint distribution of the p-values. A third alternative uses the BH procedure to control the FDR, but rather than using the raw p-values, employ the resampling scheme to compute the marginal p-values (‘point estimates’)

$$p_i^{est} = \frac{E_{V^*(p)} V^*(p)}{m}.$$

It estimates the p-values by resampling from the marginal distribution and collapsing over all hypotheses in the following way, assuming exchangeability of the marginal distributions: For gene $g$, with an observed test statistics $t_g$, the estimated p-value is

$$p_g^{est} = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{1}{J} \sum_{j=1}^{J} I_{t_i^* \leq |t_g|} \right].$$
where $N$ is the number of genes, $J$ is the number of resamplings, $t_{ij}^*$ is the statistics computed for gene $i$ in resample $j$, and $I$ is the indicator function. Next, the estimated p-values in the BH procedure are used to easily obtain the BH point estimate for gene $g$:

$$P_{BH}^{(g)} = \min_{j} \frac{P_{(j)}^{est} \cdot m}{i}$$

All above FDR adjustments can now be used to test the null hypotheses at some arbitrary value $q$. But rather than adhering to $q$, all p-value adjustments can be plotted simultaneously as a function of any monotone transformation of $p$ (for example, the test statistic). Such a plot, suggested by Yekutieli and Benjamini (1999) and by Storey (2003), allows the researcher to decide on a meaningful rejection region while being warned of the overall type I error in terms of the FDR. Its use is demonstrated in Chapter 5.
4. Recent work on False Discovery Rate related to gene expression analysis

4.1. Overview

FDR methodology received more attention following the introduction of microarray technology. It is apparent that a great deal of the research effort into this new statistical concept was motivated by the increasing demand for powerful procedures aimed to confront the large multiple testing problem on a relatively small number of replications in the analysis of gene expression data. Extensive collaboration among statisticians and biologists provided opportunities for exploration, and resulted in a considerable number of published studies that offer novel techniques. This chapter reviews the body of work as it has been emerging along two main conceptual paths: (i) the frequentist approach, which is the settings of the original FDR and is the focus of this thesis, and (ii) a Bayesian approach.

Before delving into FDR methodology, it might be appropriate to mention that the need to increase power in microarray analysis brought forth another effort to relax FWER. This attempt was made by Van der Laan et al (2004). They offer procedures to control the generalized family-wise error rate (gFWER), defined as the probability of committing at least \( k+1 \) type I errors:

\[
gFWER(k) = Pr(V \geq k + 1)
\]

for a user supplied \( k \). Thus when \( k=0 \), the gFWER is the usual FWER. They establish finite sample control and exact asymptotic control results for such augmentation procedures, under general distributions, with arbitrary dependence structures among test statistics. Adjusted p-values for the gFWER controlling augmentation procedures are shown to be simple functions of the adjusted p-values for the original FWER-controlling procedure.

The actual number of errors is also considered by Korn et al (2004). They offer a permutation-based step-down procedure that is proved to control the number of false discoveries. Nevertheless, unlike research into FDR methodology, and in spite of allowing user flexibility in number of errors, these criteria have not gained popularity. In fact the reason may be arbitrariness in \( k \), which does not provide other
researchers with clear enough information as to the worthiness of the list of rejected hypotheses.

4.2. False positive Criteria

4.2.1. Frequentist motivation

*Probability of False Discovery Proportion*

In contrast to FDR-controlling procedures that focus on the expected value of the proportion of false positives among the rejected hypotheses, Genovese and Wasserman (2002) consider tail probabilities for this proportion. Under the assumption that the test statistics are independent, these authors provide procedures that control the probability, FDP(q), that the proportion of false positives exceeds a user-supplied value q:

\[
FDP(q) = Pr\left(\frac{V}{R} > q\right)
\]

Van der Laan et al (2004) argue that in studies in which investigators wish to have high confidence that the set of rejected null hypotheses contains at most a specified proportion q of false positives, FDP(q) control is the appropriate form of Type I error control. They propose simple augmentations of FWER-controlling procedures for control of FDP(q). Additional investigators proposed to control FDP related criteria: Recently, Lehmann and Romano (2005) gave some other nonparametric procedures to control FDP(q) that can be used more generally. Korn et al (2004) offer a permutation-based procedure that is heuristically argued to control the actual false discovery proportion.

4.2.2. Bayesian motivation

*Empirical Bayes local FDR*

Efron et al (2001) offer a simple nonparametric empirical Bayes model, based on a mixture density of two gene populations – those that are differentially expressed and those that are not. Under this model, the inference produces a posteriori probabilities of effect for the individual genes, with a minimum of a priori assumptions. Let \(\pi_0, \pi_1\) be the probabilities that a gene is affected or unaffected, respectively, and \(f_0(z), f_1(z)\) be their corresponding densities with regard to the
assigned relative expression score $Z$. Then the mixture density of the two populations is defined as

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

The aposteriori probabilities $\pi_0(Z), \pi_1(Z)$ are obtained by application of Bayes’ Rule to the mixture model:

$$\pi_0(Z) = \pi_0 f_0(Z) / f(Z),$$

where $f_0(Z) / f(Z)$ is estimated through logistic regression and $\pi_0$ is then obtained through the upper bound estimate

$$\pi_0 \leq \min_Z \{ f(Z) / f_0(Z) \}.$$  

This posterior probability is interpreted as a local FDR, denoted $\text{fdr}(Z)$, which has a slight conservative bias over the FDR.

Datta and Datta (2005) adopt the empirical Bayes approach as well. They offer an estimate of the posterior probability based on an estimate of the derivative of the logarithmic marginal density. This density will have a fatter left tail than the standard normal density. A step-down permutation-based algorithm is offered to obtain it.

**Positive FDR**

Storey (2002, 2003) proposes to make use of the positive FDR (pFDR), where the expectation in the definition of FDR is conditioned on at least one rejection made: $pFDR = E(V/R \mid R > 0)$. Rather than controlling this criterion at a specified level, he suggests to estimate it for a given rejection threshold. Using the mixture model formulation,

$$pFDR = \frac{\pi_0 \gamma}{Pr(P \leq \gamma)}$$

where $\gamma$ denotes a rejection region of the form $[0, \gamma]$, for some $\gamma \geq 0$, and $P$ is the random p-value resulting from the test. This can be estimated by

$$\hat{\pi}_0 \gamma \frac{\hat{\pi}_0 \gamma}{\hat{F}(\gamma)} = \frac{\hat{\pi}_0 \gamma}{\#(p_i \leq \gamma)}$$

where $F(\gamma) = Pr(P \leq \gamma)$. The pFDR analogue of the p-value, defined for a statistic $T$ with an observed value $t$ as $Pr(T \geq t \mid H = 0)$, is termed “q-value”.
Conditional FDR

Tsai et al (2003) discuss the use of the conditional FDR (cFDR), a measure of the number of false positives among the \( r \) most significant findings. It is defined as

\[
cFDR = E \left( \frac{V}{R \mid R = r} \right) = \frac{E(V \mid R = r)}{r}
\]

Nevertheless they show that under Storey’s (2002) mixture model,

\[
cFDR(\alpha) = pFDR(\alpha)
\]

Posterior expected FDR

Muller et al (2004) formulate the posterior expected FDR within the context of microarray data. Let \( d_i \) denote an indicator of rejection for gene \( i \) and \( D = \sum d_i \) be the number of discoveries. Let \( z_i \) denote an indicator of differential expression for gene \( i \), and \( v_i = P(z_i = 1 \mid x) \) be the marginal posterior probability for that gene, with observed expression level \( x \). Define

\[
Q(d, z) = \sum d_i \frac{(1 - z_i)}{D + \epsilon}.
\]

Conditioning on \( x \) and marginalizing with respect to \( z \), the posterior expected rate, i.e. the FDR, is obtained:

\[
FDR(d, x) = \int Q(d, z) dp(z / x)
\]

\[
= \sum d_i \frac{(1 - v_i)}{D + \epsilon},
\]

and the optimal decision will take the form \( d_i = I(v_i \geq t^*) \), where \( t^* \) is the optimal cutoff.

This criterion is used by Muller et al (2004), who consider the multiple comparisons problem in the above framework in order to evaluate required sample sizes. Here, the two competing goals of maximizing power and controlling the type I error are combined into a single problem. They first define the above posterior expected FDR, which is a function of \( t^* \). Next, they formulate a loss function in terms of the false negative rate. The bound on the FDR is implicitly attended to by substituting \( t^* \) into \( Q(d, z) \), which is now a function of the sample size. They follow an expectation-minimization policy to choose the appropriate sample size that will both control the FDR and achieve the desired power.
4.3. Multiple testing procedures

4.3.1. Frequentist approach

Parametric procedures

Pounds and Morris (2003) rely on the fact that the distribution of the set of p-values can be expressed as a mixture consisting of a uniform [0,1] component, attributed to the p-values arising from the null hypothesis, and another component arising from the alternatives, modeled by beta distribution. Thus they approximate the mixture distribution through a beta-uniform mixture (BUM) defined as

$$f(x | a, \lambda) = \lambda + (1 - \lambda)ax^{a-1}$$

for $0 < x \leq 1$, $0 < \lambda < 1$ and $0 < a < 1$. They estimate the parameters using maximum likelihood. An estimate of the proportion of p-values arising from the null hypothesis is given by an upper bound to the mixture density

$$\hat{\pi}_{ab} = \hat{\lambda} + (1 - \hat{\lambda})\hat{a}.$$ 

Once a threshold $\tau$ is selected, the estimated density $\hat{f}(x | a, \lambda)$ can be partitioned into four regions, each region corresponds to a unique outcome, and the FDR can be formulated through the corresponding areas under the curve. Thus the FDR that results from using $\tau$ is estimated by

$$FDR_{ab} = \frac{\hat{\pi}_{ab}\tau}{\hat{\lambda}\tau + (1 - \hat{\lambda})\hat{a}}.$$

Allison et al (2002) too offer a Beta-uniform mixture model, and allow more than one Beta component in the model. They rely on the fact that any distribution on the interval [0,1] can be modeled as a mixture of $V$ separate component distributions where the $j$th component, $j=1,\ldots,V$, is a Beta distribution with parameters $r_j$ and $s_j$. The log of the likelihood for the collection of $k$ p-values from a model with $v+1$ components is expressed as

$$L_{v+1} = \sum \ln \beta(1,1)(x_i) + \sum \lambda_j \beta(r_j, s_j)(x_i),$$

where $x_i$ is the p-value for the $i$th test, $\lambda_{ij}$ is the probability that a randomly chosen test from the collection of tests is for a non-differentially expressed gene, and $\lambda_j$ is the probability that a randomly chosen test from the collection of tests is for a
differentially expressed gene, the distribution of which is \( \beta(r_j, s_j) \). The sum from 1 to \( k \) is due to an assumption of independence.

*Adaptation to sparse signal recovery*

Abramovich et al (2006) show that the linear step-up procedure BH can also be used to adapt to unknown degree of signal sparsity, as often is the case. Consider \( x_i \sim N(\mu_i, \sigma) \) with \( i=1,\ldots,n \) and independent from each other. Without loss of generality, take \( \sigma \) as known. Let \( t_k \) be the right tail quantile of the standard normal distribution for gene \( k \),

\[
t_k = \sigma \left( q k / 2n \right),
\]

where \( q \) is the desired FDR. Let \( k_{FDR}^* \) be the largest index for which \( |y_{i(k)}| > t_k \). Then the thresholding scheme for the location parameter \( \theta \)

\[
\hat{\theta}_{FDR}^k = \begin{cases} y_k, & |y_k| > t_{k, \text{ross}} \\ 0, & \text{otherwise} \end{cases}
\]

is inherently adaptive to unknown degree of sparsity. These results suggest the optimality of FDR thresholding for estimating a sparse vector of means by its adaptation to the degree of sparsity. Based on these results, Sabatti et al (2002) argues in favor of FDR control in microarray data analysis.

### 4.3.2. Bayesian Approach

*Non-parametric procedures*

(i) **Significance Analysis of Microarray (SAM)**

The first study to contemplate the use of the FDR controlling approach in microarray analysis was done by Tusher et al (2001). It refers to the problem of multiplicity and offers estimation of criteria interpretable as FDR. In order to analyze the data, a procedure named Significance Analysis of Microarray (SAM) is implemented. SAM identifies genes with statistically significant changes in expression through a set of gene-specific t tests. A small positive constant \( s_0 \) is added to the denominator to overcome the smaller variance in the case of low intensities. A two-stage p-value adjustment is then applied.

Genes with scores greater than a fixed threshold are deemed potentially significant. The percentage of such genes identified in error merely is the estimated FDR. The threshold can be adjusted to identify smaller or larger sets of genes, and
FDRs are calculated for each set. In order to estimate the FDR, the proportion of erroneously identified genes is averaged over permutations of the measurements, while assuming that all null hypotheses are true. The use of permutations allows the possibility of dependent tests. As pointed out by the authors, it seems plausible that this estimated FDR approximates the strongly controlled FDR when any subset of null hypotheses is true.

However, the authors noted that due to the limited number of possible distinct permutations, the number of distinct values that the p-value can take is limited. Consequently, the FDR estimate turns out to be too “granular”, so that either zero or 300 significant genes are identified, depending on how the p-value was defined. A similar result was obtained using the adaptation to dependent tests suggested by Benjamini and Yekutieli (2001).

(ii) Adaptive estimates

A method of estimating \( m_0 \) in the context of the Bayesian FDR appeared in Storey (2001, 2002) and is implemented in SAM (Storey and Tibshirani, 2003). In their notation \( \frac{m_0}{m} = \pi_0 \), and it is interpreted as the a priori probability of each hypothesis to be a true null hypothesis. Then the pFDR=FDR/P(R>0) is the posterior probability that the null hypothesis is true given that its test statistic is rejected, and the pFDR is estimated. Given \( \lambda \), a tuning parameter specifying a fixed initial rejection region, \( \pi_0 \) is estimated by \( \hat{\pi}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda) \cdot m} \), where \( R(\lambda) \) is the number of hypotheses with test statistics in the initial rejection region. Storey et al (2004) prove that for a slightly modified BH procedure with \( m \) replaced by \( \hat{\pi}_0(\lambda) \cdot m \), the adaptive procedure controls the FDR at \( q \). A bootstrap method is proposed to choose \( \lambda \). Black (2004) and Jiang (2004) note that while \( \hat{\pi}_0(\lambda) \) is unbiased when all null hypotheses are true and the p-values are distributed U[0,1], it has an upward bias when the p-values come from a mixture distribution corresponding to the true and alternative hypotheses. This bias increases as the distance between the two distributions decreases.

Storey and Tibshirani (2003) present FDR control by means of maximizing rejections while observing the estimated q-value, using a smoothing spline technique to estimate \( \pi_0 \). This suggestion completed the circle, in the sense that the latter is the same as working with the FDR adjusted p-values using an
adaptive BH procedure with their estimator $\hat{\pi}_0(\lambda)$. Thus the two approaches are in fact similar. Yet, Jiang (2004) shows by simulation that while the BH adaptive procedure overestimates $\pi_0$ and thus guarantees control of the FDR, Storey’s (2002) and Storey and Tibshirani’s (2003) methods underestimate $\pi_0$. In the former this downward bias becomes larger when $\pi_0$ is larger, and in the latter this bias tends to be very small. For this reason and for their large variation, these two procedures produce FDR estimate smaller than the true one set in the simulation, and controlling the estimated FDR to be below $q$ does not assure FDR control at $q$. Jiang (2004) suggests instead another solution to the underestimate of $\pi_0$ that works in conjunction with the BH. It uses the average of $\hat{\pi}_0(\lambda)$ over a range of $\lambda$ to estimate $\pi_0$ with a relatively small bias and a small variance.

Pounds and Cheng (2004) add that even when Storey’s (2002) estimator is monotonized by the proposed cumulative minimization operation, it underestimates the pFDR. They claim that smooth estimates of the cFDR will have less tendency to be biased downward by such an operation than rough estimates. A smooth estimate of the cFDR could be obtained by using a smooth $\hat{\ell}(\alpha)$. Thus they propose the spacing LOESS histogram (SPLOSH) to estimate the cFDR. This method treats $\hat{\ell}(\alpha)$ as an estimate of the $p$-value CDF, $\hat{F}(\alpha)$. A smooth estimate could be obtained by integrating an estimate of its derivative, the $p$-value PDF $f(\alpha)$.

**Parametric procedures**

Newton et al (2004) fits a semi-parametric hierarchical mixture model to generate the posterior probabilities of differential expression for each gene. Given the data and the fitted model, a score is assigned to each gene $i$,

$$\beta_i = 1 - P(H_i \mid x_i)$$

where $H_i$ is the indicator that the null hypothesis holds for gene $i$ and $x_i$ is its observed gene expression levels. Then, the expected number of false discoveries for a given list with values $\beta_j$ less than some bound $\tau$ is

$$C(\tau) = \sum_{j} \beta_j I[\beta_j \leq \tau]$$

Thus for a list of genes with size $J$, $C(\tau)/J$ is the expected rate of false detections, i.e. the FDR. Do et al (2005) suggest a similar formulation, but estimate the posterior probabilities through a Dirichlet Process mixture model.
5. Comparative analysis of FDR controlling multiple testing procedures

5.1. Overview

Based on the argument as detailed in the introduction, suggesting the advantage of FDR over FWER control in the case of microarray data, the two previous chapters reviewed recent advances of FDR methodology and described several approaches to estimate FDR. This chapter focuses on the study of the properties of the BH procedure relative to FDR controlling procedures that estimate the FDR based on resampling the data. The gain in power is compared to FWER methodology and evaluated relative to the computing resources invested and ease of implementation. The procedures are studied on simulated data, under several configurations of differential expression occurrences.

Note: The results of this chapter have been written jointly with D. Yekutieli and Y. Benjamini and published in Bioinformatics (Reiner et al, 2003).

5.2. Comparative analysis on real microarray data

5.2.1. The data

We analyzed the part of the cDNA microarray data used in Dudoit et al (2002b) that was made publicly available on the web. The data consists of gene expression measurements from a study of lipid metabolism in mice (Callow et al, 2000). The goal of the experiment was to identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels (treatment groups) compared to inbred control mice. Of the two models considered, the one whose data was available for our examination was the apolipoprotein AI knockout. The apolipoprotein AI gene is known to play a pivotal role in HDL metabolism.

The treatment group consisted of 8 mice with the apolipoprotein AI knockout and the control group consisted of 8 “normal” C57B1/6 mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and...
labeled using a red fluorescent dye (Cy5). The reference sample for all hybridizations was prepared by pooling cDNA from the 8 control mice and labeling it with green fluorescent dye (Cy3). In each experiment, target cDNA was hybridized to microarray containing 6359 cDNA probes, including 200 related to lipid metabolism.

5.2.2. \textit{Deriving the test statistics}

We applied the normalization described in Dudoit et al (2002b), through lowess smoother of the log intensity ratio $\log_2(\text{Red}/\text{Green})$ versus the mean log intensity $\log_2\sqrt{(\text{Red} \times \text{Green})}$. We obtained roughly the same high $t$ statistic values as those obtained in the original analysis. Before proceeding to the resampling procedure, we examined the p-values obtained directly from the “raw” $t$ statistics with 14 degrees of freedom. Ignoring multiplicity, the actual number of raw p-values larger than 0.05 was 568. On the other extreme, the Bonferroni adjustment pointed to 8 rejections.

5.2.3. \textit{Multiple testing}

We consider four FDR controlling procedures and study their properties. The first one is the BH, described in section 3.3, as applied to the p-values corresponding to the $t$ tests. The other three are based on resampling and were described in Section 3.6: the BH point estimate, that uses the same BH, as applied to the marginal p-values estimated by resampling and then pooling the resampling distributions over genes; the two other procedures are based on estimating the joint distribution of the p-values and the FDR at a given potential threshold using resampling. They differ by the way this distribution is summarized, that is by the local FDR estimator used. The resampling point estimate is conservative on the mean, and the resampling upper limit bounds the FDR with probability 95%.

Applying the FDR controlling linear step-up procedure in Benjamini and Hochberg (1995) on the raw p-values, we came up with the same 8 rejected null hypotheses obtained in the original analysis. This may not be a surprising result, since the subgroup of genes was identified by the FWER controlling procedure in the original analysis through extremely low p-values. The FWER adjusted p-values of that subgroup were all smaller then 0.01 while the rest were all greater then 0.6. However, we may be missing further discoveries if the actual distribution of the test statistics is not quite the same as the $t$ distribution underlying the derivation of the p-
values. We therefore estimated the distribution of the $t$ statistics using 1000 resampling iterations. Adjusted p-values were also calculated using the resampling-based procedures (see section 3.6 for details): Westfall and Young algorithm, the two local FDR estimators of Yekutieli and Benjamini (1999), namely the resampling-based point estimator and the resampling-based 1-0.05 upper limit, and the BH point estimator.

5.2.4. Results

Figure 10 is a plot of the adjusted p-values versus the test statistics. The ten highest absolute $t$ values (except the largest one, 20.6, which is too far to the right) are marked on the plot. As seen, the FDR local estimators and the BH point estimator consistently produce much lower adjusted p-values than those produced by the WY algorithm. The WY adjusted p-values decrease more slowly than the FDR adjusted p-values. As implied by the plot, at the 0.05 significance level, we may still reject the same 8 hypotheses by all procedures. Increasing the FDR level to 0.1 allows rejection of only one more hypothesis. Using the WY algorithm leaves us with the initial 8 genes.

![FDR and FWER adjusted p-values – original data](image)

**Figure 10:** FDR and FWER adjusted p-values – original data

The FDR local estimators and the BH point estimator consistently produce much lower adjusted p-values than those produced by the WY algorithm. At the 0.05 significance level, we may still reject the same 8 hypotheses by all procedures. Increasing the FDR level to 0.1 allows rejection of only one more hypothesis. Using the WY algorithm leaves us with the initial 8 genes.
The FDR controlled and FWER controlled results for this experiment are very close, both being very different from the unadjusted results. This should come as no surprise since the most significant 8 genes are separated from the others, as discussed earlier. In fact, it is reassuring that the reduced conservativeness of FDR controlling procedures does not trigger discovery of artifacts. In other cases typical of microarray data, where there is no clear distinction between differentially expressed genes and similarly expressed ones, we would expect to find that controlling the FDR allows the identification of more genes than controlling the FWER. We thus proceed to comparatively examine the performance of the multiple testing procedures under a controlled (simulated) occurrence of differential expression.

5.3. Comparative analysis on simulated data

5.3.1. Simulated data configuration

We fixed the number of differentially expressed genes to 70, roughly 1% of the total number of genes in the experiment. Differential expression was generated using the weak $l_p$-ball model described in Abramovich et al (2006), by which a sparse signal pattern was generated:

$$r \cdot n^{1/p} \cdot i^{-1/p}, \quad i=1,...,n$$

where $p$ is the decay rate parameter, $r$ is the decay function maximum and $n$ is the number of values generated, here 70. We used $p = 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2$.

We modified the original data, which contained 6359 genes. For each one of the 70 genes with the top differential expression measurement, the mean difference was subtracted from the group with the greater mean, thereby removing potential differences not attributed to noise. This modified data set served as the raw data for our simulation, where on each simulation repetition, the experiment and control groups were shuffled. No shuffling of the genes was performed, so that the original dependency structure was preserved. Next, we added the simulated sparse differential expression values to 70 randomly selected genes, thereby getting a single repetition in the simulation. We then applied the multiple testing procedures described earlier on each repetition, this time with 100 resampling iterations. We repeated the simulation 400 times, calculating the average FDR and power over the repetitions.
In spite of the relatively small number of repetitions, a reasonable estimate of FDR was possible since a variance reduction technique was used. The technique is based on the fact that there is a correlation between the BH procedure and the other procedures, since they are used on the same samples and the same sorted p-values, and make use of the p-values in a similar way. Since the FDR of the BH procedure is exactly \( \frac{m_0}{m} q \), the FDR of any other procedure, \( \hat{FDR}_{proc} \), can be estimated by \( \frac{m_0}{m} q + \hat{FDR}_{proc} - \hat{FDR}_{BH} \). Thus we used the difference \( \hat{FDR}_{proc} - \hat{FDR}_{BH} \) to compare the procedures. Due to the high correlation between the procedures, this difference has a variance smaller than the variance of \( \hat{FDR}_{proc} \). For power calculations we do not have an exact absolute value for any procedure. Still, the comparison of the different methods enjoys a similar variance reduction.

5.3.2. Results

Figure 11 presents the mean curves of the adjusted p-values versus the test statistics, with FDR level below 0.25. The maximal standard error of the estimated FDR was below 0.003. The plot also includes the “true FDR”, which is the proportion of the absolute values of the \( t \) statistics that exceed a reference point due to chance, out of the total number of absolute values of the \( t \) statistics that exceed the same reference point. As seen in the plot, for all FDR controlling procedures, the adjusted p-values are larger then the corresponding true FDR, indicating guaranteed FDR control. This result holds for FDR level smaller then 0.5. As expected, all FDR controlling procedures produce FDR adjusted p-values much closer to the true FDR than the FWER adjusted p-values obtained by the WY algorithm.
Figure 11: FDR and FWER adjusted $p$-values – simulated data

For all FDR controlling procedures, the adjusted $p$-values are larger than the true FDR, indicating FDR control. Their corresponding adjusted $p$-values are much closer to the true FDR than the FWER adjusted $p$-values obtained by the WY algorithm.

Figure 12: Power – by configuration

All FDR controlling procedures obtain substantially more power than the FWER controlling procedures. The resampling point estimator is the most powerful procedure, with the other two resampling estimators following very closely behind, with no consistent advantage of one over the other. The upper-limit resampling estimator does not allow increase of power over the BH resampling estimator due to its relative conservativeness. The BH procedure performs relatively well in spite of it assuming either independent or positive dependent $t$'s and not using resampling. Holm’s procedure, which controls the FWER, performs relatively poorly.

Figure 12 displays the power of the various multiple testing procedures, for each configuration of effects. Control level is 0.05. Here we also include Holm’s non-resampling multiple testing procedure, which controls the FWER (introduced in
section 3.5). As seen, all FDR controlling procedures obtain substantially more power than the FWER controlling procedures. The resampling point estimator is the most powerful procedure, with the other two resampling estimators following very closely behind, with no consistent advantage of one over the other. Although the upper-limit resampling estimator estimates the joint distribution of the test statistics, its relative conservativeness does not allow increase of power over the BH resampling estimator, which estimates only the marginal distribution. The resampling upper-limit estimator does supply more protection, in that it further controls the empirical FDR with probability 0.95. The BH procedure performs relatively well, in spite of it assuming $t$ distributed test statistics that are either independent or positively dependent, and not using resampling. Holm’s procedure, performing the most poorly, reconfirms the advantage of resampling for FWER control under dependency.

5.4. **Conclusions**

All FDR controlling procedures retain higher power than FWER controlling procedures, and are therefore highly useful for the discovery of differential genetic expression. The choice among the four is a matter of buying more power and better properties at the expense of more complicated computations.

It should be emphasized that a substantial increase in power is already gained when the p-values are estimated by resampling, and then used in the BH linear step-up procedure. Shuffling the control and experiment groups created permutations while adhering the original dependency structure among the genes. Collapsing the distributions of the test statistics for the genes to a single distribution, and using it to estimate the p-value at each gene, overcomes the discrete nature of the permutation distribution of a test statistics based on few observations, a problem described by Tusher et al (2001). This procedure can be implemented in any statistical software that enables resampling. Still, the researcher may be better off using the more powerful resampling point estimate of Yekutieli and Benjamini (1999). The resampling upper limit estimator offers both FDR control (which holds on the average) and a control on the empirical FDR level (up to probability $1-\alpha$). Thus one gives up (very little) in terms of power relative to the resampling point estimator, gaining further assurance on the actual proportion of false discoveries.
6. **FDR control for dependent two-sided test statistics**

6.1. **Overview**

The previous chapter clarified and demonstrated the advantage of FDR methodology for microarray data in terms of gain in power. One of its results was that the BH procedure, the simplest procedure to use, is not too far behind more sophisticated procedures. This chapter investigates how this procedure performs in the case of dependent two-sided test statistics. In particular, the effect of the proportion of true null hypotheses, the level of the correlation between the test statistics and the structure of the correlation, are assessed in this chapter. The FDR is first examined for the case of two multiple hypotheses by a simulation study that offers first insight into the behavior of the FDR. A theoretical analysis follows in order to obtain explicit upper bounds. These results are extended for more than two multiple tests, thereby offering a better perspective on the effect of the proportion of false null hypotheses, as well as the structure of the test statistics correlation matrix.

*Note:* The results of this chapter have been presented by the author of this thesis at the Multiple Comparisons Procedure Conference, Shanghai, China, 2005, and are under revision for the proceedings journal (Reiner, 2005).

6.2. **Simulation study, \( m=2 \)**

6.2.1. **Methods**

Recall that it is known from Yekutieli (2002) that the positive regression dependency is satisfied by positively correlated normally distributed one-sided test statistics, and their studentized t-statistics. In addition, it is shown there that the absolute values of multivariate normal and t statistics satisfy the positive regression dependency condition when all null hypotheses are true. Here, the scope of problems for which the two-sided tests control the FDR is further investigated. In the first stage, the effect of dependency on FDR is examined for two test
statistics $X_i \sim N(\mu_i, \sigma^2)$, $i=1,2$. The hypotheses tested are $H_{0i}: \mu_i = 0$ versus $H_{0i}: \mu_i \neq 0$. The mean of one population, denoted $\mu_i$, was set fixed to 0. The standard deviation of both test statistics was set to 1 and treated as known. Note that when both null hypotheses are true, it is known from the previous works that the FDR, which in this case equals the FWER, is controlled at level $q$ (Yekutieli and Benjamini, 2001). On the other hand, when both hypotheses are false, the FDR is 0 by definition. It is thus necessary to examine closely the effect of the correlation on the behavior of the FDR for two-sided tests, one of which is true and the other not.

Three parameters varied in this study. The first one was the correlation coefficient $\rho = cor(X_1, X_2)$, varying between –1 and +1 in steps of 0.1. The second one was the difference between the mean of the population for which the null hypothesis is not true and the hypothesized mean – from which the observations were drawn. This difference, denoted $\mu_2$, varied over the range of 0 to 6 in steps of 0.5. The third parameter was the level $q$ at which the linear step-up procedure was conducted, varying over the values 0.05, 0.1, 0.2 and 0.4.

Each configuration was studied using the same 10,000 repetitions. For each simulated value, we computed the p-value, denoted “raw p-value” and defined as the probability of obtaining a larger absolute value for an observation drawn from a standard normal distribution. Based on the raw-p values, multiplicity adjusted p-values were computed using the Benjamini and Hochberg linear step-up procedure.

In each repetition, $q$, $\mu_2$, and $|\rho|$ were modified per same vectors of simulated normal deviates. This allowed the use of variance reduction technique, by taking the difference between the observed proportion in each configuration and the observed proportion under $|\rho| = 0$. The average of the latter is known to have expectation of $\frac{q}{2}$, for any $\mu_2$, so the average of the proportions difference (“FDR deviation”) is a more accurate estimate of the FDR at each configuration when added back to $\frac{q}{2}$. The FDR values and deviations were then visually and theoretically studied as a function of the parameters.

Note that even though the joint distribution of $(X_1, X_2)$ is unimodal, the joint distributions of their absolute values, which are the test statistics of the two-sided
hypotheses, are not necessarily so. Vastly different forms of joint distributions of the absolute values are evident, as the cases in Figure 13 demonstrate.

Figure 13: $|X_1|$ vs. $|X_2|$

Different forms of joint distributions of normally distributed test statistics, $X_1$ and $X_2$, and the absolute values of the test statistics, $|X_1|$ and $|X_2|$, depending on the correlation coefficient and the distance between the means of the true and false null hypotheses.

6.2.2. Results

Positive and negative dependence show symmetric patterns of relationship between the FDR deviation and the correlation coefficient. This is obvious as the joint distribution of $|X_1|$ and $|X_2|$ is the same as that of $-X_1$ and $X_2$, and the two bivariate distributions have opposite $\rho$. $\mu_2$ appears to have an interesting non-monotonic effect on the relations between FDR and $\rho$. As seen in Figure 14, for values of $\mu_2$ up to a certain level, the FDR deviation decreases as $|\rho|$ increases. Around this value (about 4.5 for $q=0.05$ and 3.5 for $q=0.1$), the relations become opposite: the FDR deviation increases as $|\rho|$ increases. Beyond this level, the effect
of $\mu_2$ gradually weakens and eventually decays. The value around which the relation changes is also the one where the FDR deviation reaches its maximum.

![Graphs showing FDR deviation vs. $\rho$ for different $q$ and $\mu$ values](image)

**Figure 14:** The FDR deviation from $q/2$ vs. $\rho$, $m=2$

For values of $\mu$, up to a certain level, the FDR deviation decreases as $|\rho|$ increases (thin broken lines). Around this level (continuous line), in which the FDR deviation reaches its maximum, the relations become opposite. Thus maximal FDR deviation is obtained at perfect correlation.
Figure 14 indicated that the maximum FDR level is reached for a certain level of $\mu_2$ when $|\rho|$ is 1. Therefore, in Figure 15 we plot the deviation of FDR from $q/2$ as a function of $\mu_2$, for different cases of $q$, emphasizing the case of $|\rho|=1$ by a continuous line. The maximal FDR is reached for some least favorable $\mu_2$ that gets smaller as $q$ increases, and seems to be close to $2\Phi^{-1}\left(\frac{m_0}{m}, \frac{q}{2}\right)$. The minimum FDR deviation is reached at a level of $\mu_2$ that is close to 0. Note that the maximal deviation of the FDR in the liberal direction, from the value of $\frac{q}{2}$ expected under independence, is not greater than $q^2$ - which is a very small increase.

**Figure 15:** The simulated FDR deviation from $q/2$, vs. $\mu_2$, $m=2$

The maximal FDR deviation is reached for some least favorable $\mu_2$, which gets smaller as $q$ increases. The minimum FDR deviation is reached at a level of $\mu_2$ that is close to 0.
6.3. **Theoretical study, \( m=2 \)**

A theoretical analysis will be necessary in order to establish FDR control under dependence with regard to the dependence parameters. It is of interest to identify the values of \( \rho \) and \( \mu_2 \) that maximize the FDR before moving to higher dimensions. In what follows we take from the simulation study the observation that the maximal FDR is achieved when \( |\rho| = 1 \) and proceed to analyze this case.

Figure 16 is a plot of the joint distribution of the absolute values of two normal test statistics, denoted \( X_i, i=1,2 \), in the case of \( |\rho| = 1 \). Their critical values, denoted \( Z_i, i=1,2 \), correspond to the two constants of the linear step-up procedure for the case of two-sided tests, \( \frac{q}{m} \), where \( i \) is the order index of corresponding \( P_{(i)} \).

Thus for \( m=2 \),

\[
\left(z_1, z_2\right) = \left\{ \Phi^{-1}\left(1 - \frac{q}{2}\right), \Phi^{-1}\left(1 - \frac{q}{4}\right) \right\}.
\]

Recalling that we test the null hypothesis that \( \mu_i = 0 \), while only \( \mu_2 \neq 0 \), the proportion of false discoveries, denoted \( Q \), takes one of three values \( Q = \frac{i}{2} \), \( i=0,1,2 \). These values have the probabilities visualized here as areas corresponding to the location of the test statistics relative to their respective critical values. Note that since \( Z \) index correspond to the p-value order index, here \( X_{(i)} \) is compared to \( Z_2 \) and \( X_{(2)} \) is compared to \( Z_1 \). Thus, for both \( X_{(i)} \) greater then their respective critical values, the two null hypotheses are rejected and \( Q \) is 0.5. If both \( X_{(i)} \) are smaller then their respective critical values, no rejection is made and \( Q \) is 0. If only \( X_{(i)} \) is greater then its critical value, only the true null hypothesis is rejected, and thus \( Q \) is 1. If only \( X_{(2)} \) is greater then its critical value, only the false null hypothesis is rejected and \( Q \) is 0.

The distribution satisfies \( X_2 = X_1 + \mu_2 \) and thus its intersection with the axes is at \( \mu_2 \). The graphic representation of the relative location of \( \mu_2 \) displayed in Figure 16 and the critical values with the resulted false discovery proportions allow writing down the relations between \( \mu_2 \) and FDR.
Joint distribution of $|X_2|$, $|X_1|$ when $|\rho|=1$ - cases of $Q$ area cross

The proportion of false discoveries $Q$ takes one of the values $0, \frac{1}{2}, 1$. These values have the probabilities visualized here as areas corresponding to the location of the test statistics relative to their respective critical values. Joint distributions with different cases of $\mu_2$ have different area-crossing locations. One case is emphasized in bold line as an example.

We first easily evaluate the coordinates of all possible points of intersection with the boundaries of the areas of $Q$. The calculations of the separating points relied on the fact that the joint distribution creates $45^\circ$ angles every time it collides with the axes, due to its equal intersection coordinate with the two axes, $\mu_2$. Figure 17 represents the area crossings, according to the different joint distributions varying in $\mu_2$. Table 1 lists the point coordinates.
Figure 17: Joint distribution of $|X_2|,|X_1|$ - FDR calculation

Four different ranges of $\mu_2$ generate four expressions that compose the FDR function.
Table 1: **Intersection points with Q area boundaries**

The coordinates of six points represent all possible locations of intersection with the boundaries of the areas of Q.

<table>
<thead>
<tr>
<th>POINT</th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$z_1$</td>
<td>$\mu_2 - z_1$</td>
</tr>
<tr>
<td>B</td>
<td>$z_1$</td>
<td>$z_1 - \mu_2$</td>
</tr>
<tr>
<td>C</td>
<td>$\mu_2 + z_1$</td>
<td>$z_1$</td>
</tr>
<tr>
<td>D</td>
<td>$z_2$</td>
<td>$z_2 - \mu_2$</td>
</tr>
<tr>
<td>E</td>
<td>$z_2$</td>
<td>$\mu_2 - z_2$</td>
</tr>
<tr>
<td>F</td>
<td>$\mu_2 - z_1$</td>
<td>$z_1$</td>
</tr>
</tbody>
</table>

We now calculate the FDR for the four ranges of $\mu_2$:

**Case I:** $0 < \mu_2 < z_2 - z_1$

\[
P(Q = 1/2) = 1 - \Phi(z_1) + 1 - \Phi(\mu_2 + z_1)
\]

\[
= 2 - \Phi(z_1) - \Phi(\mu_2 + z_1)
\]

, and

\[
P(Q = 1) = 0,
\]

resulting in

\[
FDR = 1 - \frac{1}{2} [\Phi(z_1) + \Phi(\mu_2 + z_1)]
\]

**Case II:** $z_2 - z_1 \leq \mu_2 < 2z_1$

\[
P(Q = 1/2) = \text{same as in case I}, \quad \text{and}
\]

\[
P(Q = 1) = \Phi(\mu_2 + z_1) - \Phi(z_2),
\]

resulting in

\[
FDR = 1 - \Phi(z_2) + \frac{1}{2} [\Phi(\mu_2 + z_1) - \Phi(z_1)]
\]

**Case III:** $2z_1 \leq \mu_2 < z_1 + z_2$

\[
P(Q = 1/2)
\]

\[
= 1 - \Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(z_1) + 1 - \Phi(\mu_2 + z_1)
\]
\[ = 2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(\mu_2 + z_1), \text{ and} \]
\[ P(Q = 1) = \text{same as in case II}, \]
resulting in
\[ FDR = 1 - \Phi(z_1) - \Phi(z_2) + \frac{1}{2}[\Phi(\mu_2 + z_1) + \Phi(\mu_2 - z_1)] \]

**Case IV:** 
\[ z_1 + z_2 \leq \mu_2 \]
\[ P(Q = 1/2) = \text{same as in case III}, \]
and
\[ P(Q = 1) = \Phi(\mu_2 + z_1) - \Phi(\mu_2 - z_1), \]
resulting in
\[ FDR = 1 - \Phi(z_1) + \frac{1}{2}[\Phi(\mu_2 + z_1) - \Phi(\mu_2 - z_1)] \]

**Figure 18:** The FDR deviation from \( q/2, \rho = 1 \) – computed function

The FDR as a function of \( \mu_2 \) is displayed for various values of \( q \). The behavior of the least favorable \( \mu_2 \) as a function of \( q \) is evident, and summarized in Figure 19.
We now turn to prove that this function takes the same shape for all cases of $\mu_2$ in the relevant interval, in terms of monotonic trends and global minima and maxima. The proof relates to $\mu_2 > 0$, and we shall conclude the same for $\mu_2 < 0$ due to the symmetry of $\Phi$.

**Case I:** $FDR = 1 - \frac{1}{2} \left[ \Phi(z_1) + \Phi(\mu_2 + z_1) \right]$ is monotone decreasing if $\Phi(z_1) + \Phi(\mu_2 + z_1)$ is monotone increasing. $\Phi(z_1)$ is a constant, and $\Phi(x)$ is monotone increasing in $x$.

**Case II:** Similarly, $FDR = 1 - \Phi(z_2) + \frac{1}{2} \left[ \Phi(\mu_2 + z_1) - \Phi(z_1) \right]$ is monotone increasing if $\Phi(\mu_2 + z_1)$ is monotone increasing, since all other terms are constants. $\Phi(x)$ is monotone increasing in $x$.

**Case III:** $FDR = 1 - \Phi(z_1) - \Phi(z_2) + \frac{1}{2} \left[ \Phi(\mu_2 + z_1) + \Phi(\mu_2 - z_1) \right]$ is monotone increasing if $\Phi(\mu_2 + z_1) + \Phi(\mu_2 - z_1)$ is monotone increasing. As in the previous cases, each one of the $\Phi$ terms is monotone increasing since $\Phi(x)$ is monotone increasing in $x$, up to a limit. The sum of monotone increasing components is monotone increasing too.
Case IV: \( FDR = 1 - \Phi(z_i) + \frac{1}{2} [\Phi(\mu_z + z_i) - \Phi(\mu_z - z_i)] \) is monotone decreasing if its first derivative with respect to \( \mu_z \) is negative. Differentiating yields
\[
\frac{\partial FDR}{\partial \mu_z} = \frac{1}{2} [f(\mu_z + z_i) - f(\mu_z - z_i)].
\]

In addition, \( \mu_z \geq z_1 + z_2 \), thus \( \mu_z + z_i \) and \( \mu_z - z_i \) are both positive. Since for positive \( x \) the normal density \( f(x) \) is monotone decreasing in \( x \), \( f(\mu_z + z_i) < f(\mu_z - z_i) \), and thus the derivative is negative.

It is left to pinpoint the location of the maximum. Following the proof of relative monotone behavior of all function parts, it is left to show that the maximum of the function in case III is greater than or equal to the maximum of the function in case I, and thus is the global maximum of the whole function. Let us denote the FDR at case i \( F_i(\mu_z) \). \( F_1(\mu_z) \) is at its maximum at \( \mu_z = 0 \), while \( F_3(\mu_z) \) is at its maximum at \( \mu_z = z_1 + z_2 \). Evaluating the two functions at their maximal points, we obtain
\[
F_i(0) = 1 - \frac{1}{2} [\Phi(z_i) + \Phi(z_i)] = 1 - \Phi(z_i)
\]
\[
F_3(z_1 + z) = 1 - \Phi(z_1) - \Phi(z_2) + \frac{1}{2} [\Phi(z_1 + z_1 + z_2) + \Phi(z_1 + z_2 + z_1)]
\]
\[
= 1 - \Phi(z_1) + \frac{1}{2} [\Phi(2z_1 + z_2) - \Phi(z_2)]
\]

Thus it is left show that
\[
1 - \Phi(z_1) + \frac{1}{2} [\Phi(2z_1 + z_2) - \Phi(z_2)] \geq 1 - \Phi(z_i)
\]
or
\[
\frac{1}{2} [\Phi(2z_1 + z_2) - \Phi(z_2)] \geq 0
\]
which is true, since \( 2z_1 \geq 0 \) and \( \Phi(x) \) is monotone increasing in \( x \).

Result: Let us summarize the above discussion as following: For two-sided test statistics which are normally distributed with correlation 1, the maximal increase in FDR, when the BH procedure is used, occurs when the non-zero expectation is the sum of the two critical values used in the procedure, \( \mu_z = z_1 + z_2 \).
Let us compute the maximal FDR deviation. We substitute the least favorable value $z_1 + z_2$ for $\mu_2$ in the FDR expression corresponding to the function interval containing the maximal FDR (case III or case IV). We obtain

$$FDR_{\mu_2=\mu_2^*} = 1 - \Phi(z_i) + \frac{1}{2} [\Phi(\mu_2 + z_i) - \Phi(\mu_2 - z_i)]$$

$$= 1 - \Phi(z_i) + \frac{1}{2} [\Phi(z_1 + z_2 + z_i) - \Phi(z_1 + z_2 - z_i)]$$

$$= 1 - \Phi(z_i) + \frac{1}{2} [\Phi(2z_1 + z_2) - \Phi(z_2)]$$

Subtracting $\frac{q}{2}$ from this expression will give the maximal deviation:

$$1 - \Phi(z_i) + \frac{1}{2} [\Phi(2z_1 + z_2) - \Phi(z_2)] - \frac{q}{2}$$

$$= 1 - \left(1 - \frac{q}{2}\right) + \frac{1}{2} \left[\Phi(2z_1 + z_2) - \left(1 - \frac{q}{4}\right)\right] - \frac{q}{2}$$

$$= \frac{1}{2} \left[\Phi(2z_1 + z_2) - \left(1 - \frac{q}{4}\right)\right]$$

$$\leq \frac{1}{2} \left[1 - \left(1 - \frac{q}{4}\right)\right] = \frac{q}{8}$$

Thus the FDR deviation is bounded by $\frac{q}{8}$.

### 6.4. Simulation study, $m>2$

If the configuration of $\mu_2$ obtained for $m=2$ turns out to be the least favorable in higher dimensions, there is a good prospect for getting a solid result that bounds the increase of the FDR in two-sided tests. We examined the cases of $m=3,4,6$ for the purpose of assessing the behavior of (i) $\mu_2$, which is assumed here equal for all false null hypotheses, and (ii) the FDR deviation as a function of $m$ and $\frac{m_0}{m}$ and as a function of $\rho$, assuming $\text{corr}(X_i, X_j) = \rho$ for all $i \neq j$. Figure 20 displays further evidence that the FDR deviation as a function of $\mu_2$ remains similar across $m$ larger than 2. It includes here only the case of $m=3$, but represents the behavior observed also for higher $m$, through simulations.
Figure 20:  FDR vs. $\mu_2, m=3$

The FDR deviation as a function of $\mu_2$, displayed here for different cases of $q$, remains similar across $m=3$.

Figures 21-23 present the FDR deviation as a function of $\rho$. We observe that for a specific $q$ and a given proportion $\frac{m_0}{m}$, the least favorable $\mu_2$ remains the same across $m$. For instance, for $q=0.05$ and $\frac{m_0}{m} = \frac{1}{2}$, the maximal FDR deviation is obtained for $\mu_2 = Z_1 + Z_2 = 4.2$ ($\mu_2$ was analyzed in 0.1 resolution), where $Z_1 = Z_{\frac{1-0.05}{2}} \approx 1.96$ and $Z_2 = Z_{\frac{1-0.05}{4}} \approx 2.24$ (see Figures 14a, 22b and 23b). Similarly, for $q=0.1$ and $\frac{m_0}{m} = \frac{1}{2}$, the least favorable $\mu_2$ is 3.6, where $Z_1 = Z_{\frac{1-0.1}{2}} \approx 1.64$ and $Z_2 = Z_{\frac{1-0.1}{4}} \approx 1.96$ (see Figures 22e and 23e).
In addition, we notice that, for a given $m$, the least favorable $\mu_2$ across
\[ \frac{m_0}{m} \text{ is always obtained at } \frac{m_0}{m} = \frac{1}{2} \] (see Figures 22 and 23).

The implication of these two observations, stated simply, is that the least
favorable $\mu_2$, which is the sum of the two critical values $z_1, z_2$ associated with the
case of $m=2$, is the least favorable $\mu_2$ for any $m$, and is the local least favorable $\mu_2$
for the case of $\frac{m_0}{m} = \frac{1}{2}$.

However, we shall take from this body of simulations only the conclusion
that the maximal FDR is obtained for $\rho = 1$, and proceed to derive the other results
theoretically.

**Figures 21-23:**
The FDR deviation from $q m_0 / m$ reaches its maximum at the perfect correlation case. The point of
changing relations between the FDR deviation and $\rho$, which is also the least favorable $\mu_2$ obtain the
same values for all values of $m$ checked by the simulations, given the same $m_0 / m$. 
Figure 21: FDR vs. $\rho, m=3$
FDR vs. $\rho, m=3$ (contd.)
Figure 22: FDR vs. $\rho$, $m=4$
FDR vs. $\rho$, $m=4$ (contd.)

- false hypothesis proportion = 0.25
- $q=0.2$
- $\rho$ deviation

- false hypothesis proportion = 0.5
- $q=0.4$
- $\rho$ deviation

- false hypothesis proportion = 0.75
- $q=0.4$
- $\rho$ deviation
Figure 23: FDR vs. $\rho$, $m=6$
FDR vs. $\rho, m=6$ (contd.)
6.5. Theoretical study, \( m > 2 \)

Consider a set of \( m \) p-values \( P_1, \ldots, P_m \), corresponding to the null hypotheses \( H_1, \ldots, H_m \), \( m_0 \) of them correspond to the subset of true null hypotheses, and \( m_1 = m - m_0 \) correspond to the subset of false null hypotheses. Given that these two subsets are correlated with \( \rho = 1 \), all p-values in each subset must have the same value. We can thus represent these \( m \) p-values by two p-values, denoted \( P^0 \) and \( P^1 \), with respective weights \( w_0 = m_0 \), \( w_1 = m_1 \). Implementing the BH procedure to test each of the hypotheses, we use the sorted \( m \) p-values through the sorted \( P^0 \) and \( P^1 \).

Recall that in the BH procedure, p-value \( P_{(i)} \) in the case of two-sided test is compared to \( \frac{q}{m} \cdot i \), starting from the highest p-value. Two cases are possible:

(i) \( P^0 > P^1 \).

Start with \( P^0 \): if \( P^0 \leq \frac{q}{2} \), reject all hypotheses, and thus \( Q = \frac{m_0}{m} \).

Else, if \( P^1 \leq \frac{q}{2} \cdot \frac{m_1}{m} \), reject \( H_{(1)}, \ldots, H_{(m_0)} \), and thus \( Q = 0 \).

Else, stop. Thus \( Q = 0 \).

(ii) \( P^1 > P^0 \).

Start with \( P^1 \): if \( P^1 \leq \frac{q}{2} \), reject all hypotheses, and thus \( Q = \frac{m_0}{m} \).

Else, if \( P^0 \leq \frac{q}{2} \cdot \frac{m_0}{m} \), reject \( H_{(1)}, \ldots, H_{(m_0)} \), and thus \( Q = 1 \).

Else, stop. Thus \( Q = 0 \).

Accordingly, the joint distribution of \( P^0 \) and \( P^1 \) has the same structure as in the case of \( m = 2 \) (recall Figure 15), with the general term \( \frac{m_0}{m} \) replacing \( \frac{1}{2} \), and thus

\[
(z_1, z_2) = \left( \Phi^{-1}\left(1 - \frac{q}{2}\right), \Phi^{-1}\left(1 - \frac{q}{2} \cdot \frac{m_0}{m}\right) \right).
\]

Recall the assumption that \( \mu \) of all false null hypotheses are equal, and denote them \( \mu_2 \). Therefore, in order to obtain FDR as a function of \( \mu_2 \), we can use all points computed and probability formulas used in the case of \( m = 2 \). We obtain:
Case I: \(0 < \mu_2 < z_2 - z_i\)

\[
P\left(Q = \frac{m_0}{m}\right) = 1 - \Phi(z_1) + 1 - \Phi(\mu_2 + z_1)
\]

\[
= 2 - \Phi(z_1) - \Phi(\mu_2 + z_1)
\]

\[
P(Q = 1) = 0
\]

\[
\text{FDR} = \frac{m_0}{m} \left[2 - \Phi(z_1) - \Phi(\mu_2 + z_1)\right]
\]

Case II: \(z_2 - z_1 \leq \mu_2 < 2z_1\)

\[
P\left(Q = \frac{m_0}{m}\right) = \text{same as in case I.}
\]

\[
P(Q = 1) = \Phi(\mu_2 + z_1) - \Phi(z_2)
\]

\[
\text{FDR}
\]

\[
= \frac{m_0}{m} \left[2 - \Phi(z_1) - \Phi(\mu_2 + z_1)\right] + \Phi(\mu_2 + z_1) - \Phi(z_2)
\]

\[
= \frac{m_0}{m} \left[2 - \Phi(z_1)\right] + \left(1 - \frac{m_0}{m}\right) \Phi(\mu_2 + z_1) - \Phi(z_2)
\]

Case III: \(2z_1 \leq \mu_2 < z_1 + z_2\)

\[
P\left(Q = \frac{m_0}{m}\right) = 1 - \Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(z_1) + 1 - \Phi(\mu_2 + z_1)
\]

\[
= 2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(\mu_2 + z_1)
\]

\[
P(Q = 1) = \text{same as in case II.}
\]

\[
\text{FDR}
\]

\[
= \frac{m_0}{m} \left[2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(\mu_2 + z_1)\right] - \Phi(z_2) + \Phi(\mu_2 + z_1)
\]

\[
= \frac{m_0}{m} \left[2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1)\right] + \left(1 - \frac{m_0}{m}\right) \Phi(\mu_2 + z_1) - \Phi(z_2)
\]
**Case IV:** \( z_1 + z_2 \leq \mu_2 \)

\[
P \left( Q = \frac{m_0}{m} \right) = \text{same as in case III.}
\]

\[
P(Q = 1) = \Phi(\mu_2 + z_1) - \Phi(\mu_2 - z_1)
\]

**FDR**

\[
= \frac{m_0}{m} [2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1) - \Phi(\mu_2 + z_1)] + \Phi(\mu_2 + z_1) - \Phi(\mu_2 - z_1)
\]

\[
= \frac{m_0}{m} [2 - 2\Phi(z_1)] + \left( 1 - \frac{m_0}{m} \right) \Phi(\mu_2 + z_1) - \Phi(\mu_2 - z_1)
\]

Thus the general expression for least favorable case of \( \mu_2 \) for two-sided tests becomes

\[
\mu_{2F} = Z_1 + Z_2 = \Phi^{-1}(1 - \frac{q}{2}) + \Phi^{-1}(1 - \frac{m_0}{m} \cdot \frac{q}{2})
\]

We can now obtain the expression for the maximal FDR by substituting the expression for the least favorable case of \( \mu_2 \), in the expression corresponding to case III or IV. We obtain:

\[
FDR_{\mu_2=\mu_{2F}} = \frac{m_0}{m} [2 - 2\Phi(z_1) + \Phi(\mu_2 - z_1)] + \left( 1 - \frac{m_0}{m} \right) \Phi(\mu_2 + z_1) - \Phi(z_2)
\]

\[
= \frac{m_0}{m} [2 - 2\Phi(z_1) + \Phi(z_1 + z_2 - z_1)] + \left( 1 - \frac{m_0}{m} \right) \Phi(z_1 + z_2 + z_1) - \Phi(z_2)
\]

\[
= \frac{m_0}{m} [2 - 2\Phi(z_1) + \Phi(z_2)] + \left( 1 - \frac{m_0}{m} \right) \Phi(2z_1 + z_2) - \Phi(z_2)
\]

In order to obtain the maximal FDR deviation from the expected FDR under independence, we subtract from the above expression \( \frac{m_0}{m} q \). Figure 24 shows the FDR and the FDR deviation as a function of \( \frac{m_0}{m} \), for \( q=0.05 \).
Figure 24: Maximal FDR and FDR deviation vs. \( m_0/m \)

a. The FDR is always controlled at \( q \). The maximum deviation from the line representing the case of independence is obtained for \( m_0/m = \frac{1}{2} \).

b. The FDR deviation as a function of \( m_0/m \) clearly shows that its maximum is obtained near \( m_0/m = \frac{1}{2} \).

Since \( \Phi(2z_1 + z_2) \) is very near 1 for realistic values of \( q \), say \( q \leq 0.2 \), and 1 is the maximum value obtainable for this expression, we may come up with an upper bound for the maximal FDR deviation, setting \( \Phi(2z_1 + z_2) = 1 \):

\[
FDR_{dev}\big|_{\mu_1=\mu_1'} \leq \frac{m_0}{m} \left[ 2 - 2 \left( 1 - \frac{q}{2} \right) + 1 - \frac{m_0}{m} \cdot \frac{q}{2} \right] + I - \frac{m_0}{m} - I + \frac{m_0}{m} \cdot \frac{q}{2} - \frac{m_0}{m} \cdot \frac{q}{2} \\
= \frac{m_0}{m} - \left( \frac{m_0}{m} \right)^2 \cdot \frac{q}{2} - \frac{m_0}{m} \cdot \frac{q}{2} \\
= \frac{q}{2} \cdot \frac{m_0}{m} - \frac{q}{2} \left( \frac{m_0}{m} \right)^2
\]
\[
= \frac{q}{2} \cdot \frac{m_0}{m} \left(1 - \frac{m_0}{m}\right)
\]

This function has its maximum at \( \frac{m_0}{m} = \frac{1}{2} \). Indeed, substituting \( \frac{m_0}{m} = \frac{1}{2} \) in the maximal deviation expression yields \( \frac{q}{8} \) as obtained before for the case of \( m=2 \). The deviation, as we see, is independent of \( m \) and depends only on \( \frac{m_0}{m} \) and \( q \). There is no deviation when all null hypotheses are true, as we already know from Yekutieli (2002), and when all null hypotheses are false, which is essentially true since no false discoveries are possible in such a case.

Note that an upper bound to the maximal FDR can be expressed by

\[
FDR_{\mu^r=\mu^*} = q \cdot \frac{m_0}{m} + \frac{q}{2} \cdot \frac{m_0}{m} \left(1 - \frac{m_0}{m}\right),
\]

\[
= q \cdot \frac{m_0}{m} \left[1 + \frac{1}{2} \left(1 - \frac{m_0}{m}\right)\right].
\]

If we would like to offer \( q^* \) such that the FDR will be controlled at \( \frac{m_0}{m} q \), which is important if one wishes to use adaptive procedures that make use of the factor in the BH procedure, then we use

\[
q^* \cdot \frac{m_0}{m} \left[1 + \frac{1}{2} \left(1 - \frac{m_0}{m}\right)\right] \leq \frac{m_0}{m} q,
\]

and thus using

\[
q^* \leq \frac{q}{1 + \frac{1}{2} \left(1 - \frac{m_0}{m}\right)}
\]

will guarantee FDR control at level \( \frac{m_0}{m} q \). When \( \frac{m_0}{m} = \frac{1}{2} \), we obtain \( q^* = \frac{4}{3} q \). Note that when \( \frac{m_0}{m} = 1 \), we obtain \( q^* = q \), which proves that when all null hypotheses are true, the FDR is controlled at level \( q \), regardless of dependence. When \( \frac{m_0}{m} = 0 \), \( q^* = \frac{2}{3} q \) will guarantee control at level \( q \) of adaptive procedures. Thus, adaptive
methods that do not underestimate $m_0$ (see Jiang, 2004, and Section 4.3.2 here) are good candidates for controlling the FDR under the high correlation conditions described here. Note that Jiang (2004) shows that Storey’s method (2002) and Story and Tibshirani’s method (2003) underestimate $m_0$ and are thus not expected to control the FDR under high correlation.

6.6. **Generalized dependence structure**

The generalization in the previous section to $m$-dimensional multiplicity problem was based on the simulation finding that test statistics dependency relations characterized by perfect correlation are the least favorable case of dependence. However, at the starting point for the simulation study all correlations were set equal. The study was based on the search over a correlation matrix with some common $\rho$ to a correlation matrix with $\rho = 1$:

\[
\begin{bmatrix}
1 & \rho \\
\rho & \ddots & \rho \\
\rho & \cdots & 1
\end{bmatrix}
\]

Given test statistics with the general correlation matrix,

\[
\begin{bmatrix}
1 & \rho_i^0 \\
\rho_i^0 & \ddots & \rho_i \\
\rho_i & \cdots & 1
\end{bmatrix}
\]

it is of interest to find out whether the same conclusion can be made regarding the FDR in this case: Is it smaller than the FDR corresponding to the perfect correlation case?
This question will be studied again using simulation. For the purpose of simulation variance reduction, the comparison will be made by averaging the test statistics within subsets of null hypotheses, thereby generating test statistics with correlation matrix of all $\rho$’s equal 1. This averaging transformation can serve as a mean to compare the original dependence structure to the perfect correlation structure with reduced noise due to dependence between the two cases.

6.6.1. Spatial dependence

While the case of identical correlation coefficients represents dependency of the measurement errors due to pooled common control in microarray experiment, we would now like to examine spatial dependency, which is another possible case of dependency structure on a microarray. Geographical clusters formed by local factors, such as print-tips or dye smears will imply correlation between neighboring wells that diminishes as the distance between the wells increases. We thus conduct a simulation in order to evaluate the effect of spatial dependency on FDR control.

We simulate a set of $m$ normally distributed test statistics, $m_0$ of them having mean $\mu = 0$ and $m_1$ having mean $\mu = \mu_i$, $\mu_i \geq 0$, with standard deviation 1 and spatial correlation matrix with $\rho_{ij} = \rho^{i-j}$, where $\rho$ ranges from -1 to 1. We average the test statistics separately for those corresponding to mean 0 and those that are not. We construct a test statistics representing the transformation from the original test statistics to their average. For the $k$th test statistics, let

$$Z_k = \frac{(1-t)x_k + tx_k}{\sqrt{\text{var}(1-t)x_k + tx_k}}$$

where $I$ indicates correspondence to subset of false null hypotheses, and $t$ ranges from 0 to 1, such that as $t$ increases, the weight of the averaging transformation increases. We repeat this simulation 1,000 times and compute the FDR. If the case of $\rho = 1$ is indeed the least favorable, we expect the FDR to increase with $t$ for $\rho < 1$, while by definition the FDR remains fixed for $\rho = 1$.

Figure 25 shows the case of $\frac{m_0}{m} = \frac{1}{2}$, with $m$ chosen to be 6, as it is the minimum required to create a spatial dependence structure in both subsets of hypotheses. We used several values for $\mu_2$, and only the one found to be least favorable for the case of equal $\rho$, showed a coherent behavior of increase with $t$. 
This was evident for relatively large values of $|\rho|$, starting from around 0.7. Recall that the relationship between the FDR and $|\rho|$ were not monotonic across all $|\rho|$, although the general increase was clearly evident (see for example Figure 14). Therefore only $|\rho| \geq 0.7$ are shown here. This behavior was more easily observed for larger $q$, due to the smaller noise, and therefore we chose to show here the case of $q=0.2$.

\[ q=0.2 \]

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig25.png}
\caption{Transformation from spatial dependence to perfect correlation}
\end{figure}

An increase in $t$ is observed when $\mu_2$ is set to the least favorable value found for the case of common $\rho$.

### 6.7. Conclusions

In this chapter we discussed the previously non-resolved case of two-sided testing with the BH procedure of normally distributed test statistic, which are dependent. We identified the least favorable configurations yielding the maximal FDR. We considered common correlation between the normally distributed test statistics, where a subset of $m_0$ hypotheses are true out of a total of $m$ null hypotheses, and $\mu_2$ is the distance between the means corresponding to the true and
false null hypotheses. For some range of \( \mu_2 \), the FDR tends to concave symmetrically in \( \rho \), and thus reaches its maximum at \(|\rho| = 1\). For a certain \( \mu_2 \), the concavity reaches maximum over other values of \( \mu_2 \), and thus it is the least favorable \( \mu_2 \). We take notice of the fact that each subset of null hypotheses has the same critical value for all its hypotheses when there is perfect correlation between them. We showed that the least favorable \( \mu_2 \) is the sum of the two normal critical values corresponding to each subset of null hypotheses.

The FDR is always bounded by \( q \). Looking at the deviation of FDR under dependence from \( \frac{m_0}{m} q \), which is the value under independence, we showed that this deviation reaches a global maximum of \( \frac{q}{8} \) when \( \frac{m_0}{m} = \frac{1}{2} \). We found that for any common \( \rho \), using \( q^* \leq \frac{q}{1 + \frac{1}{2} \left( 1 \frac{m_0}{m} \right)} \) will guarantee control of the FDR at level \( \frac{m_0}{m} q \). The increase of the FDR in \(|\rho|\) was indicated by simulation study to be valid for the case of spatial dependence, which is another dependence structure characteristic of microarray data. Our conclusion is that while the case of perfect correlation remains the most difficult for FDR control, the proof of FDR control for this case is not likely to be found by merely parametrizing the problem using a single parameter and maximizing over it, as is the case for the perfect correlation.
7. Multiplicity issues related to complex research questions

7.1. Overview

The two previous chapters focused on the relatively simple setting of one test, or research question, per gene, in particular comparison of two samples. This chapter deals with more complicated cases, in which several questions are posed in the research, such that multiple tests are to be conducted for each gene. Here, the multiplicity problem can rise to a very large size. Controlling directly the FDR for all hypotheses tested may involve millions of tests, and while sometimes feasible, it is usually not very appealing. Correction for multiplicity for each gene separately does not deal with multiplicity of genes. Thus a more appropriate solution for this problem should be offered.

This chapter investigates two approaches of controlling the FDR of such a complex study, based on dividing the analysis into research questions. The two approaches are discussed and illustrated by a study of gene expression levels in different brain regions across multiple mouse strains. The results of the two previous chapters indicated that the BH procedure controls the FDR under dependence: In Chapter 5 it was shown by simulation, and in Chapter 6 it was proved for the case of two-sided correlated tests where not all null hypotheses are true. This chapter relies on these results in that it examines the use of theory developed for independent test statistics for dependent situations faced in microarray analysis.

Note: The results of this chapter have been written jointly with D. Yekutieli, G. I. Elmer, N. Kafkafi, N. E. Letwin, N. H. Lee and Y. Benjamini and are in press for Statistica Nederlandica (Yekutieli et al, 2006).
7.2. Complex research questions

The identification of differentially expressed genes in two biologically different situations, which was the subject of Chapter 5, is a relatively simple direction of research in which there is a single null hypothesis for each gene and entire set of null hypotheses make up a single family of hypotheses. More complicated directions of research are typically problems that have a traditional statistical solution, but include a separate family of hypotheses for each gene. The very large number of genes involved in the analysis compounds this difficulty. In order to describe these problems, we turn to a specific large-scale experiment on brain, behavior, and genetics.

7.2.1. The experiment

In the current experiments, 10 adult males taken from each of 10 different strains of inbred mice were tracked for 30 minutes in a novel large arena (“open field”) and analyzed. The strains include the major inbred strains traditionally used for research, including inbred strains derived from wild mice. Note that inbred strains are homogeneous in terms of genetic background, since all animals from the same strains are homozygote as it relates to each gene in the genome. The strains exhibit a wide range of behavior, as measured by different endpoints – quantitative summaries of the behavior. Chapter 8 will discuss the behavioral part of the experiment.

The gene expression part of the experiment involved harvesting tissue from the mice used in the behavioral assessment protocol 7-12 days following the experiment. Five tissue areas were dissected (prefrontal cortex, ventral striatum, temporal lobe, periaquaductal gray and cerebellum). The tissue from all mice from the same strain was divided to two groups providing two biological replications per strain. cDNA two-color spotted microarrays were used to measure gene expression. The level of gene expression was measured versus a pooled control, swapping the dyes between the two replicates. In total, ~27,000 genes were analyzed. The printing robot included 48 printing tips, and thus the printed array contained 4*12 square blocks, each printed with a different pin and including 24*24=576 spots. This experimental design and the effects of block location and within-block location are discussed in section 7.4.1, which deals with data preprocessing related to experimental design.
7.2.2. Notation and ANOVA model

Let us introduce some notations for the parameters of the model describing the above experiment. $M_{grsmk}$ is the level of the expression of gene $g$, for dye $k$, mouse $m$, strain $s$ and region $r$, and its expectation given the strain is $\mu_{grs}$. Denote the average expression level over the brain by $\mu_{g+}$, the average expression level over the strains in one brain region by $\mu_{gr+}$, and the average expression level over the strains and over the brain regions by $\mu_{g++}$. In addition, several technical factors related to the spot location on the slide affect expression level: $I_{ranks(g)}$ is the effect of within-block row $i$ (where gene $g$ is located), $i = 1, \ldots, 24$ and $J_{ranks(g)}$ is the effect of within-block column $j$ (where gene $g$ is located), $j = 1, \ldots, 24$. Similarly, $MI_{ranks(g)}$ is the effect of block row $u$, $u = 1, \ldots, 12$ and $MJ_{ranks(g)}$ is the effect of block column $v$, $v = 1, \ldots, 4$.

The ANOVA model adding up the biological sources of variations along with the experimental design and the technical sources of variation is

$$M_{grsmk} = \mu_{g+} + \mu_{g+s} + \mu_{gr+} + \mu_{grs} + I_{ranks(g)} + J_{ranks(g)} + MI_{ranks(g)} + MJ_{ranks(g)} + e_{grsmk},$$

where $\sum_s \mu_{g+s} = 0$, $\sum_r \mu_{gr+} = 0$, $\sum_{s,r} \mu_{grs} = 0$ and the errors $e_{grsmk}$ are distributed $N(0, \sigma^2)$. Since $e_{grsmk}$ for $k = 1,2$ may be correlated we work with measurements averaged over dye, i.e. $M_{grsm} = \frac{1}{2}(M_{grsm1} + M_{grsm2})$. However, we do not simultaneously fit all parameters. Instead, $I, J, MI$ and $MJ$ are estimated per slide using repeated median removal of effects, as described in Section 7.4.1, and the estimates are subtracted from $M_{grsmk}$.

7.2.3. Question 1: Brain ANOVA per gene

The first question in this study is whether genes are differently expressed in the brain between the strains. This is the immediate extension of the question about differently expressed genes from two groups to more groups, in this case to the 10 strains. The desired inferences are about the genes only, so such a question may be easily posed as a testing problem of the intersection hypothesis for each gene:

$$H_0^*(g): \mu_{g1+} = \mu_{g2+} = \ldots = \mu_{g10+}.$$
Each hypothesis of this form can be tested by one-way ANOVA, using F-test, or Tukey’s test (Tukey, 1977). A discovery in this analysis is that there is some difference in brain expression levels of gene $g$ between the strains. The multiplicity issue that stems from testing many genes is typically addressed by controlling the FDR over the entire family. Using SAM (Tusher et al, 2001), F-statistics are calculated except that the denominator has an added “fudge factor” $s_0$ and their distribution is resampled. But both their visualization and their adaptive procedure for estimating $m_0$ run into complications because F-tests are involved rather than estimated differences.

While in usual applications of ANOVA rarely does an analysis end with rejecting the intersection null hypothesis only, it may be an end by itself in microarray significance analysis. As an important example take the case where the analysis is done as a dimension reduction preliminary step, to be followed by clustering analysis or discriminant analysis. More commonly in ANOVA, the rejection of the intersection hypothesis is followed by an analysis of the pairwise comparisons. This question is discussed next.

In the current study the emphasis is on differences in expression levels related to the studied strains. In another phase of this study, the parallel question regarding the testing across brain regions for differences in expression levels of genes, as reflected by $H'_0(g)$ (in obvious notation), may be of similar interest. This certainly constitutes a different research question as well.

### 7.2.4. Question 2: Strainwise comparisons in Brain analysis per gene

A question of immediate interest to researchers is the identification of which specific pairs of strains differ in their expression level in a gene where difference exists. This is the classical pairwise comparison problem, where the strainwise comparison of strains $s'$ and $s''$, in terms of their brain expression level in gene $g$ is expressed in the hypothesis

$$H_{0_{s's''}}(g): \mu_{gs'+} = \mu_{gs''+}$$

Pairs of strains identified to differ in the expression level of a gene of importance, may serve as the breeding source for a backcross experiment that tries to identify the genomic locations of other genes that caused the difference in their expression. Clearly discovering such pairs is of scientific importance. Note that there are $10^9/2 = 45$ such pairwise comparisons per gene, and with the ~27,000 genes
analyzed we reach the multiplicity problem of about 1.2 million pairwise comparisons.

7.2.5. **Question 3: Strain*Region interactions in Brain analysis per gene**

While identifying strains that differ in their expression level of a certain gene over the brain is of obvious interest, the availability of data on five brain regions offers the possibility to search the data for interesting findings in another direction as well: identifying specific strain*region interactions in the expression level of a gene. The relevant hypothesis is

\[ H_{0sr}(g): \mu_{gsr} - \mu_{gs+} + \mu_{g+r} + \mu_{g++} = 0, \]

and finding a significant interaction means that the expression level of gene \( g \), in strain \( s \), is different than what would be expected under a typical additive pattern for that gene. This is clearly a data mining operation, looking for interesting clues to pursue.

Two approaches can be taken. The first argues that such an interaction is interesting only after brain analysis showed strain differences in that gene. The second argues that any identification of interaction is of interest, but only after the interaction was shown to exist in the gene at large. Taking the first approach, any finding of strainwise difference in a gene is of interest, and then finding the interaction is another discovery. In the second approach genes are first screened for significant interaction, and then the specific interactions become the only interesting discoveries. This is actually parallel to the relationship between the testing of the strain factor in the ANOVA that is followed by the identification of specific pairwise comparisons. Both approaches are legitimate, and we follow the first one.

7.3. **FDR controlled testing of complex research questions**

7.3.1. **The direct approach using the procedure in BH**

Consider the entire set of tested hypotheses, even if it consists of several separate families, as a single family, and use the BH procedure on the p-values. FDR is controlled if the two conditions in Benjamini and Yekutieli (2001) are satisfied: (i) the test statistics are Positive Regression Dependent on the subset of true null hypotheses; (ii) the distribution of each p-value corresponding to a true null
hypothesis is either $U[0,1]$ or stochastically larger than $U[0,1]$. As a result of (ii), if the entire family of tests is not at our disposal and the missing is at random, we may put the unknown p-values at 1, and get valid (if somewhat conservative) inference.

As we noted in the introduction, the FDR criterion is quite accommodating when families of hypotheses are combined. It is important to note that the BH procedure has some asymptotic properties that make it very effective (Abramovich et al, 2006).

7.3.2. **The direct approach for selected subset**

The problem of testing a potentially very large family of hypotheses can be alleviated by screening out a large proportion of the hypotheses, and testing the remaining ones using a multiple testing procedure. However, for any procedure used to control for type I error, the distribution of the test statistics for the remaining hypotheses under the null hypotheses has to be $U[0,1]$ (or stochastically larger). One obvious way to achieve both these goals is to choose families of the null hypotheses according to prior knowledge. When the choice of families is data dependent extra care must be taken.

One important example of data dependent choice where the condition is satisfied is the case where the *same hypotheses are tested again* using independent data in both stages. Under such circumstances, if FDR is first controlled at level $q_1$ to select a subset of hypotheses, and then controlled at $q_2$ on the selected subset, the joint procedure enjoys FDR level of $q_1 * q_2$ (Reiner et al 2003, Benjamini and Yekutieli 2005). They show that this approach can be much more powerful than testing indiscriminately the original family of hypotheses.

This is in sharp contrast to the case we face. Notice for example that in research Question 3, a strain*region interaction in gene expression levels is not possible if there is no strain effect in the expression level. We therefore test each gene for a significant strain effect, and then apply the direct approach to the hypotheses belonging to families of hypotheses with significant strain effects. The test statistics we use to test the strain effect is independent of the interaction term test statistics. This means that any selection criterion applied to strain effect test statistics does not violate the conditions needed for FDR control. In section 5 the selection criterion used is the BH procedure at level 0.05. Notice, that this approach cannot be used for research question 2. The dependency between the strain effect test statistics
and the individual strain pairwise test statistics rules out the possibility of using the strain effects to choose the set of families used in the direct approach.

7.3.3. **Hierarchical testing and associated false discovery rates**

In the general case, a set of $m$ hypotheses $H_1,\ldots,H_m$ is arranged in a tree with $L$ levels, (in addition to the root hypothesis). Each hypothesis $H_i$ is on some level $L(i)$, and, except the root hypothesis, has a parent of level $Par(i)$ at level $L(i)-1$. The $m$ hypotheses can be divided into $T+1$ families: $\mathcal{I}$ is the family of hypotheses in level 1; for $t=1,\ldots,T$, let $\mathcal{I}_t = \{H_i : Par(i) = t \}$. $m_t$ and $m_t^0$ denote the total number and number of true null hypotheses in $\mathcal{I}_t$, respectively. It is assumed that while the progeny of a false null hypothesis can be either true or false, the progeny of a true null hypothesis is always true. The number of hypotheses per parent at each layer or across layers need not be the same, nor does the depth of the tree.

In hierarchical testing a hypothesis is tested only if its parent hypothesis is rejected (except the root hypothesis, which is always tested). All tested hypotheses with the same parent are tested simultaneously. Testing the tree of hypotheses begins by applying the BH procedure at level $q$ to the hypotheses in $\mathcal{I}_0$; at the second stage the BH procedure is separately applied at level $q$ to each family of hypotheses corresponding to hypotheses rejected in $\mathcal{I}_0$; at the next stage the BH procedure is separately applied to each family of hypotheses corresponding to a second stage discovery; this continues as long as additional rejections are made and the rejected hypotheses are parent hypotheses.

Once the testing at all layers is completed, the investigator ends up with a set of discoveries (rejected hypotheses) of interest. How can the FDR be controlled over the set of important discoveries? There is no single answer to this question since under the tree structure different sets of discoveries may be of interest, corresponding types of false discovery rates can be defined, and thus the procedure that control each may be different as well.

We shall now briefly summarize the definitions and upper bounds of the FDR criterion under hierarchical testing introduced by Benjamini and Yekutieli (2002). Given a subfamily $S$ of hypotheses of interest, the definition of the FDR is as usual,
\[
FDR_s = \begin{cases} 
E \left[ \frac{V_s}{R_s} \right], & R_s > 0 \\
0, & R_s = 0 
\end{cases}
\]

**Full-tree FDR**

The discoveries of interest are any of the entire set of hypotheses tested at all layers \(H_1, \ldots, H_m\). Any rejection in the tree is of equal interest.

**Level-l restricted FDR**

The discoveries of interest are those hypotheses that reside at some specific level \(l\), chosen in advance. Rejections along the path leading to this layer are of no importance by themselves, and only their implications at the chosen level are of interest.

**Outer (end) node FDR**

The subfamily of interest is the hypotheses on the outer nodes of the tree reached by the hierarchical testing procedure. It is between the above two schemes: As in the \(l\)-level testing scheme, if a hypothesis is in the set of interest, its parent, parent's parent, and so on, up to the root, are not. But like the full-tree scheme, the hypotheses of interest are not restricted to belong to one level only. Yet this scheme is also very different from the above two, for the set of hypotheses of interest is not chosen in advance, but is rather the result of the hierarchical testing procedure.

### 7.3.4. **FDR controlling hierarchical testing**

Benjamini and Yekutieli (2002) provide upper bounds for the FDR of the three testing schemes. To compute the upper bounds they assume that the p-values are independently distributed; \(P \sim U[0,1]\) for all true null hypotheses; the conditional marginal distribution on the interval \(U[0,\alpha]\) of the p-values, corresponding to false null hypotheses, is stochastically smaller than \(U[0,\alpha]\) for all \(\alpha \in [0,q]\).

They use the BH procedure at level \(q\) whenever a subfamily is tested. Under the assumptions above they show that when a large proportion of the hypotheses rejected are parent hypotheses, the FDR of the full-tree testing scheme is less than \(2q\delta^*\). The \(\delta^*\) is a family-specific multiplicative factor that emerges out of using the formulation in Benjamini and Yekutieli (2001) for the FDR of the BH procedure in a family, as it stands by itself, to a testing procedure in which the BH is applied to several families and the FDR is calculated jointly. It turns out that \(\delta^*\) is
the ratio between the expected number of discoveries made by the BH procedure when one specific true null is set at 0, and the expected number of discoveries otherwise made plus one. The theoretical analysis is conducted under independence of the test. This leads to a calculated upper bound shown to be smaller than 1.44, but more typically \( \delta^* = 1 \). A further refinement to the bound on the tree FDR is possible: if most of the rejected hypotheses are terminal nodes, the bound on FDR is nearly \( q \cdot \delta^* \cdot \frac{\mu_0}{m} \), where \( \frac{\mu_0}{m} \) is the weighted mean of the proportion of the true null hypotheses in each tested family.

The FDR of the outer-node testing scheme for a tree with \( L \) layers is shown to be bounded by \( 2Lq \delta^* \). However, this bound may be much too high. Since a rejection in this scheme does not necessarily imply the rejection is taken into account in the FDR calculation, the total number of rejections is smaller than in the case of the full-tree testing scheme, while the expected number of false rejections remains the same. Thus the FDR tends to be greater.

It is shown for the level-restricted case, that

\[
\text{FDR}_{\text{level}=l} \leq q \cdot \delta^* \cdot B(R_l, J),
\]

where \(|J|\) is the number of families tested at any level, \( R_l \) is the number of rejections at level \( l \) and \( B(R_l, J) = E\left[ \frac{R_l + J}{R_l + J + 1} \right] \). Note that if the number of discoveries at level \( l-1 \) (which is the number of families tested at level \( l \)), is large in relation to the total number of rejections in level \( l \), there is no universal bound to \( \text{FDR}_{\text{level}=l} \). On the other hand, if many rejections are made in level \( l \), \( \text{FDR}_{\text{level}=l} \approx q \cdot \delta^* \cdot \frac{\mu_0}{m} \). The behavior of \( B(R_l, J) \) and the use of \( \left[ \frac{R_l + J}{R_l + J + 1} \right] \) as its estimator will be studied in Chapter 8.

7.4. Analysis and Results

7.4.1. Brain ANOVA per gene: The need for repeated median removal of chip layout effects

Our approach is to use an F-test at each gene to test the hypothesis

\( H_0^g (\cdot): \mu_{g1} = \mu_{g2} = \ldots = \mu_{g10} \)
No pooling across genes was used (with no fudge factor as in SAM), so the distribution of the test statistics under the null hypothesis should be F distribution with 9 and 40 degrees of freedom. Figure 26a presents the results of close to 25,600 p-values using their quantile plot, where the ordered p-values are plotted versus their rank. This is equivalent to a p-p plot under the assumed F-distribution. We expect to see in the upper right part of the plot a straight line, corresponding to genes where the hypothesis of no strain differences is true. Instead the line curves concavely, indicating that the p-values are bigger than they should be.

![Quantile-Quantile Plot](image)

**Figure 26:** P-values before and after spatial effects removal – strain effect

a. Before spatial effects removal – raw p-values distribution is clearly not U[0,1], as indicated by the deviation of the ranked raw p-values from the straight line. 730 genes are identified as differentially expressed.

b. After spatial effects removal – raw p-values distribution is closer to U[0,1], deviating from the straight line only as p-values are lower. Now nearly 1000 genes are identified. Thus effect removal improves the sensitivity of the testing procedure.

We therefore revisited the preprocessing stage and studied the possible effects of the physical layout of the chips on individual expression levels. Effects of
block row and column and within-block row and column were calculated for log base 2 of the two microarray channels signals. A four-way median polish model, which is an extension of the two-way median polish (Tukey, 1977 and Emerson and Wong, 1985), was applied for the purpose of identifying block and slide spatial effects. The structure of an array is schematically described in Figure 27, implying that the printing structure of the array comprises the effects of block row, block column, within block row, and within block column. The median polish algorithm iteratively subtracts the median signal for each category of each effect (block row, block column, within-block row, within-block column), until the median of the residuals is 0 in all directions. The “polished” signals are then used for data analysis. It is essential that this resistant procedure be used, and not the usual ANOVA that is based on means, so that real differences in expression levels will not be smoothed out.

Figure 26b present the results of the same ANOVA tests as those displayed in Figure 26a, except after the four-way median polishing. The improvement is noticeable, with the p-values on the right side approximately on the line of slope 1. Before spatial effects removal, the raw p-values distribution is clearly not U(0,1), as indicated by the deviation of the ranked raw p-values from the straight line. Controlling the FDR at 0.05, 730 genes are identified as differentially expressed. After spatial effects removal, the raw p-values distribution is closer to U(0,1), deviating from the straight line only as p-values are lower. Now nearly 1000 genes are identified. Thus effect removal improves the sensitivity of the testing procedure.
Figure 27: **A schematic structure of a microarray**

The printing structure of the array comprises the effects of block row, block column, within block row and within block column.

We therefore proceed with the parametric analysis, relying on the p-values calculated by the ANOVA to make as many gene discoveries while controlling the FDR at 0.05.

7.4.2. *Brain ANOVA per genes*

A two-way ANOVA was applied to identify genes that could distinguish between strains, while adjusting for the effect of brain region. Each gene was analyzed separately. Testing the main factor of strain, 2676 p-values lower than 0.05 (10.5%) were obtained. Adjusting the p-values for multiplicity using the BH linear step-up procedure, as explained in Reiner et al (2003), 957 genes were found statistically significant. Figure 28 presents the FDR adjusted p-values using the BH procedure for the most significant genes. The procedure also amounts to drawing a line through the points \((1, q/m)\) and \((m, q)\), and looking for the largest value under this line in Figure 28. Zooming in for FDR less or equal to 0.1, the multiplicity effect on the significance threshold is visible.
Figure 28: Raw and FDR adjusted p-values – strain effect

Zooming in for FDR less or equal to 0.1, the multiplicity effect on the significance threshold is visible. After FDR adjustment, nearly 1000 genes are found to have a differential expression at the 0.05 level.

7.4.3. Identifying specific Strain-wise differences in Brain analysis per gene

Comparing strains by pairs through the same model described above involved approximately 25,600*45=1.15 million comparisons. The hypotheses have natural and intuitive tree structure where the parent of all pairwise hypotheses in a gene is the intersection hypothesis for that gene residing at level 1. It makes sense to follow hierarchical testing at this stage, but the test statistics will clearly not be independent. We do not know so far how the FDR is inflated under such circumstances, so we turn to direct testing of all 1.15 million hypotheses, relying on the many studies indicating that the BH procedure controls the FDR under the pairwise correlation structure (Yekutieli, 2002a and 2002b, Kesselman et al, 1999, Benjamini et al, 1999). Controlling the FDR at the 0.05 level yields 7771 significant pairs. Figure 29 shows the frequency of significant pairwise comparisons for each pair of strains. Although each of the strains was found to have at least 10 genes at which it differs from any other strain, the pattern of significant pairs indicated that two of the strains, CAST/Ei and SPRET/Ei, are different in a substantial way from all other strains as well as from each other (as reflected by hundreds of gene expression levels differences.)
Figure 29: Frequency of genes with significant pairwise difference.

The number of genes found to have a different expression between strains is visibly higher when comparing to strain 10 (SPRET/Ei), implying a large-scale change in genetic activity for this strain.

Interestingly, the SPRET/Ei belongs to the Mus Spretus species and the CAST genotype belongs to the subspecies Mus m. castaneus, while the eight other inbred strains belong to the subspecies Mus m. musculus (Bonhomme and Guenet, 1996). It is thus interesting to find that the divergent speciation of SPRET and CAST is reflected in the large gene-expression distinction consistently recognized in the pair-wise comparisons in the brain. This raises the possibility that differences in behavior that are known to exist between the strains will be associated with differences in levels of gene expression in the brain.

7.4.4. Brain region interactions with Strain per gene

Standardized residuals from the above analysis of variance model were used to calculate a test-statistic for each combination of strain and brain region, for each gene. Testing those interactions only in those 957 genes that were identified at the first stage of the analysis, still involves the testing of about 50,000 interactions. We can restrict our testing to these interactions only by the subset selection method, because the tests of the interactions are independent of the tests of strain effects in
ANOVA. Only 13 significant interactions of strain and brain region were found this way at FDR level 0.05.

Turning to the hierarchical testing, it is natural to arrange all hypotheses of strain effects for the genes at level I, with the family of interactions per gene being its progeny at level II. If any such discovery amounts to a meaningful discovery, it calls for a full tree-testing scheme. Thus, if each one of the first level analysis is done at the FDR level of 0.017, and at each rejected gene separately the testing of the 50 relevant hypotheses is done at FDR level of 0.017, the overall FDR level is bounded by $2 \times 1.44 \times 0.017 \sim 0.05$.

Some 758 genes and 76 interactions are found significant using the hierarchical testing approach. The interaction component of the hierarchical analysis was sufficiently powerful to detect the brain region by strain interactions. Of these 76 interactions, two-thirds (47) were associated with the CAST genotype a non-additive pattern against the model suggested by the parent species *Mus m. musculus*. A possible explanation for this disproportionate representation, which reflects the deviation from the additive pattern formed by the eight *Mus m. musculus* genotypes, lies in either a cis- or trans-acting regulatory polymorphism whose consequence affects a gene-network that is tissue-specific (Cowles et al, 2002). An evidence for such a functional explanation of the brain-region by strain interactions can be found in the mapping of these genes into networks (Figure 30). For example, 5 of the 47 genes (Casp8AP2, Tnfsf6, Trip15, Fkbp3, Nrf1) exhibiting significant interactions associated with CAST are part of a gene network that has been linked with seizure activity. Interestingly, CAST is particularly resistant to beta-carboline-3-carboxylate induced seizures (Le Roy et al, 1998). A possible explanation to the fact that the CAST is involved in such interactions more than other strains lie in the increased incidence in which such wild-derived genotypes naturally carry Robertsonian chromosomes, but we shall not dwell into this topic here.
**Figure 30:** Genes mapped into a network linked with seizure activity

Genes identified in the analysis (filled in red) exhibit significant gene interactions associated with CAST. Map prepared by Ingenuity Pathway Analysis, a web based program for finding gene, protein and small-molecule interactions.

7.5. Conclusions

The purpose of this chapter was to demonstrate how our understanding the way by which the FDR progresses when testing hypotheses hierarchically, can enable us to test complex and extremely large families of hypotheses. Generally, we have used the theory developed for independent test statistics in the dependent situations faced in microarray analysis. This is based on our experience, described in Chapter 5, in using the BH procedure in microarray analysis, where under the type of dependency encountered in practice the FDR is controlled.

The multiple testing FDR controlling procedures discussed in this chapter were implemented using the MULTTEST procedure of SAS software. The input to this procedure is the set of p-values derived from an ANOVA model, or from tests of correlations. The flexibility of this approach lies in the fact other procedures, such as
those based on mixed model, repeated measurements and the like, can all be easily implemented. Resampling-based procedures have also been implemented using R programming language (and S Plus), so the above approach should be easy to implement for other settings.

The biological implications of the results obtained here are not within the scope of this thesis and will be detailed elsewhere, but some first conclusions of importance were discussed here. In particular it was found that even though we started with two different research questions, they turn out to be related. Moreover, the results reported here are the first stage of an even more complex study, where interest further lies in the relationship between measured behavioral aspects of the mouse strains and the gene expression levels in the brain of those strains, in those genes where strain differences are evident. In particular we have interest in correlating gene expression levels in the brain with the 17 measures of behavior, which quantify exploratory behavior giving rise to a the multiplicity problem to about 2.2 million tests of hypotheses. While the approach taken here is applicable there as well, the solutions developed will be discussed in the next chapter.
8. **Functional genomics: Correlating complex behavioral traits with expression data**

8.1. **Overview**

The previous chapter introduced the notion of complex study, where several research questions are posed. It investigated the implementation of new approaches for FDR control, focusing in the case where all research questions concern gene expression data. This chapter moves up in the scale of complexity to cases where another type of data is combined with the genotypic data, in order to infer on gene-level functionality through correlation analysis between the phenotypic and the genotypic information.

The correlation test statistics of the final stage of the analysis may have some degree of dependency in the early screening test statistics. Thus the new FDR controlling approaches, which are known in theory to bound the FDR under independence, are now evaluated under an assumption of slight dependence. They are also compared to single-step methods existing in literature. Data with the same level and structure of dependence is simulated, and the FDR is calculated for several configurations of true null hypotheses proportion in each stage. The FDR multiplier $B(R_i,J)$ is estimated and $q^*$ is determined accordingly. Based on these findings, the real data introduced in the previous chapter is analyzed, this time in search for correlations between gene expression and phenotypic behavior end-points.

Another problem present in the data and typical of microarray experiments is the difference in measurement units between genotypic and phenotypic data. Here, genotypic information is measured over a pool of mice, while phenotypic information is measured per mice. A solution is offered based on jittering the expression data by a quantity estimated from the biological replicates.

In section 7.5 it was observed that several genes identified in the analysis were part of the same network. Indeed any identified list of genes of interest can contain clusters or networks of functionally associated genes, which are dependent of each other. Independent and dependent genes are biologically two different conditions. While independent genes are likely to be related to different mechanism,
dependent genes have more of a potential to belong to the same mechanism or at least to several related ones (Letwin et al, 2006). Therefore, in order to properly interpret the potential biological implication of the identified genes, it is important to understand that they are not necessarily independently regulated, even though the statistical problem of multiple comparisons treats each gene with a separate hypothesis.

Note: The results of this chapter have been written jointly with D. Yekutieli, G. I. Elmer, N. E. Letwin, N. H. Lee and Y. Benjamini and is being submitted for publication in Bioinformatics (Reiner et al, 2005).

8.2. The experiment

The aim of this experiment is to find a connection between open field exploratory behavior of mice and the level of gene expression in different brain regions. On the behavioral side, the purpose was to use the highly informative and quantitative characterization of open-field exploratory behavior, as encompassed by the recently developed SEE (Software backed Strategy for Exploring Exploration, Drai and Golani, 2001, Benjamini et al, 2001, Kafkafi et al, 2005). This strategy attempts to capture highly structured behavioral patterns using ethologically relevant measures (tests or behavioral endpoints). For that purpose, the path of the animal in a large open field is automatically tracked and digitized for half an hour. It is then smoothed using a special adaptation of LOWESS, and statistically segmented into discrete behavioral units with proven ethological relevance for rodents: stops (lingering episodes) and progression segments (Drai et al, 2000, Kafkafi et al, 2005). The quantitative properties of the segments, such as their length, duration, maximal speed and spatial spread constitute a large number of the endpoints studied. More traditional endpoints include global measures such as total activity and time spent at center. All behavioral endpoints studied exhibited high broad sense heritability with significant strain differences (Kafkafi et al, 2005).

In the current experiments, 10 adult males taken from each of 8 different strains of inbred mice were tracked for both behavioral and genotypic information. We note that two of the 10 strains whose gene expression was measured were not tracked for behavioral data since at the time of the behavioral experiment, which was
conducted first, they were not available. This large number of strains used is important since it increases the specificity of the results.

The gene expression part of the experiment involved harvesting tissue from the mice used in the behavioral assessment protocol 7-12 days following the experiment. The procedure was described in Chapter 7. Let us just recall that five tissue areas in the brain were dissected and that the tissue from mice of the same strain was pooled into two groups in order to have enough quantity for measuring expression levels, thereby providing two biological replications per strain. The gene expression part of the experiment, conducted as a piggyback experiment on part of the multi-laboratory behavioral study has been analyzed in Yekutieli et al, 2005.

8.3. The research question

Denote the value of the behavioral endpoint \(b\) of mouse \(m\) from strain \(s\) by \(B_{bsm}\) and its expectation given the strain by \(\beta_{bs}\). \(M_{grsm}\) is the level of the expression of gene \(g\) (averaged over dyes) in mouse \(m\), strain \(s\) and region \(r\), and its expectation given the strain is \(\mu_{grs}\). As in Chapter 7, denote the average expression level over the brain by \(\mu_{g+}\), the average expression level over the strains in one brain region by \(\mu_{gr+}\), and the average expression level over the strains and over the brain by \(\mu_{g++}\).

The major goal of the current analysis is to study the relationship between measured behavioral aspects of the strains of mice, and the gene expression levels of those strains, for genes for which strain differences are evident, in specified brain regions. In particular we have an interest in correlating gene expression levels in the brain with the 17 measures of behavior, which quantify exploratory behavior. It should be mentioned at this stage that all 17 behavioral measures show significant difference between the strains (Kafkafi et al, 2005), so we use the 8 strains as 8 genetic opportunities to measure both gene expression and behavior. We are interested in testing

\[ H_0^b(g,r): \text{cov}(B_{bsm},M_{grsm}) = 0 \]

By itself this testing problem involves 17*5*27,000 hypotheses of the above form. We would like to approach this mega-family of hypotheses by first screening for potential families of hypotheses for which the correlation is non-zero. If expression level of a particular gene in a particular brain region is the same for all strains there is no reason to believe it can be correlated with behavior.
Therefore, the screening question is whether genes are differently expressed between the strains in a region of the brain. Such a question may be easily posed as a testing problem of the intersection hypothesis for each gene $g$ and each brain region $r$:

$$H_0^\prime(g, r): \mu_{gr1}=\mu_{gr2}=\ldots=\mu_{grk}$$

As a reminder, in the previous chapter we tested the less specific hypothesis $H_0(g)$, which did not include inference about the specific brain region where the difference exists. The ANOVA model adding up the sources of variations is similar to the model introduced in Chapter 7 (Section 7.2.2).

As to the behavior, for each trait $b$ in strain $s$ and mouse $m$,

$$B_{bsm} = \beta_{bs} + \beta_{bm} + \varepsilon_{bsm},$$

where $\beta_{bs}$ is the characteristic value for strain $s$ and $\beta_{bm}$ is the characteristic bias for mouse $m$, $\beta_{bm} \sim N(0, \sigma^2)$. The errors $\varepsilon_{bsm}$ are independent and distributed $N(0, \sigma^2)$ across $s$ and $m$, but may be correlated within one behavioral trait, i.e $\text{corr}(\varepsilon_{bsm}, \varepsilon_{b'm}) = 0$ but $\text{corr}(\varepsilon_{bsm}, \varepsilon_{bsn}) \neq 0$.

### 8.4. Dependence between the tests of the two stages

The strategy of testing a family of hypotheses only if another (parent) hypothesis is rejected at the screening stage can be found in the previous chapter, involving gene expression in brain regions across strains of mice. While in one study the test statistics in the second stage were almost independent of those in the screening stage (testing interactions after testing main effects) they were highly dependent in the other study (testing pairwise comparisons following the testing of main effects.)

The available theory relies on the independence between the statistics in the two stages. However, the testing of pairwise differences between means has obviously a strong dependence on the results of the test that all means are the same. In the situation we consider here, independence does not hold exactly, but the test statistic for correlation analysis has much weaker dependence on the screening test that all means are the same. We shall use a simulation to study the effect of this dependency on the FDR, for the hierarchical testing scheme we use here.
8.5. Methods

8.5.1. Proposed Methods: Subset selection and hierarchical testing

The theoretical background and terminology for these two methods were given in Chapter 7.2. In this subsection, we shall explain the motivation of implementing the two methods in our analysis. Our analysis consists of two stages: (1) Different hypotheses are tested at the screening stage and at the second stage; in our case $H^s_{0}(g, r)$ regarding strain differences and $H^b_{0}(g, r)$ regarding correlation. (2) The same data is used in both stages. Because of the latter, the test statistics used in the second stage may be dependent of the test statistics of the first stage, and thus violate the U[0,1] condition for the p-value needed for type I error control. Because of the former, even if the test statistics at both stages are independent, the overall FDR is merely $q_2$. 

Our problem can be imbedded in the general scheme for hierarchical testing of trees of hypotheses using the BH procedure described in Benjamini and Yekutieli, 2002 and Yekutieli et al, 2005. In our case, the tree has two layers, (besides the root): At the first layer are the hypotheses regarding the question of strain differences within brain regions for each gene $\{ H^s_{0}(g, r): g = 1,2,...,27000; r = 1,2,...,5 \}$. At the next layer, each selected subfamily of hypotheses $\{ H^b_{0}(g, r): b = 1,2,...,17 \}$ has as its parent the single hypothesis $H^s_{0}(g, r)$ from the layer above it – in this case layer 1.

In our case, the analysis is mainly concerned with identifying combinations of correlated genes and behavioral traits within brain regions, leading to concern about level-2 restricted FDR. Taking a somewhat wider perspective, we might want to emphasize the importance of the first level discoveries as well, since they indicate strain differences. If this perspective is taken, the full-tree FDR should be of concern to the investigator. The outer-node FDR does not seem to have a meaningful interpretation in our case.

8.5.2. Single-Step Methods

In order to further evaluate the strength of the above two proposed approaches, they will be compared to several single-step methods existing in literature, which control the FDR while adapting to the proportion of true null
hypothesis, termed $\pi_0$ by them. These methods retain considerable power relative to other methods since they use lower threshold achieved by estimating $\pi_0$, as was discussed in section 3.4. In particular, as was shown in the literature survey (section 4.3.2), the average estimate proposed by Jiang (2004) and the SPLOSH algorithm proposed by Pounds and Cheng (2003) are found to better control the FDR than other proposed methods such as Storey's (2002) method and Storey and Tibshirani's (2003) method, that both underestimate $\pi_0$ and thus do not always control the FDR. In addition, Hochberg and Benjamini's (1990) adaptive procedure was shown by Jiang (2004) to overestimate $\pi_0$ and thus always control the FDR. Thus the methods selected for comparisons with the proposed methods are the adaptive BH, SPLOSH and the average estimate. While Hochberg and Benjamini's (1990) adaptive procedure was introduced in Chapter 3, the other two methods were discussed in Chapter 4 and their procedures are now given below. These three methods are applied within the simulation study and their FDR and power are compared. The BH procedure results are given as well.

**Average estimate method**

1. Define $0 = t_1 < \ldots < t_B < t_{B+1}$ as equally spaced points in the interval $[0,1]$.
2. Define $NB_i = \# \{ p_i : p_i > t_i \}$ and $NS_i = \# \{ p_k : t_i \leq p_k > t_{i+1} \}$ for some arbitrary $k$.
3. For each $t_i$ ($i = 1, \ldots, B$), $\hat{\pi}_0(t_i)$ is an estimate of $\pi_0$ by using Storey's (2002) estimate, taking $\lambda = t_i$,

$$\frac{\# \{ p_i : p_i > \lambda \}}{(1 - \lambda)m} = \frac{NB_i}{(1 - t_i)m}.$$  

4. Since p-values corresponding to the false null hypotheses are smaller than those corresponding to the true null hypothesis, for small index $i$ there are more p-values in the interval $[t_i, t_{i+1})$. Thus, let $i = \min \{ i : NS_i \leq \frac{NB_i}{B - i + 1} \}$, and then

$$\hat{\pi}_0 = \frac{1}{B - i + 1} \sum_{j=i}^{B} \frac{NB_j}{(1 - t_j)m}.$$  

In the simulation here $B$ is set to 10, which is one of the values recommended by Jiang's (2004) study.
**SPLOSH method**

Suppose there are \( g \) unique p-values. Let \( \tilde{p}_{(1)} < \tilde{p}_{(2)} < \ldots < \tilde{p}_{(g)} \) be the \( g \) ordered unique p-values. Let \( a(i) = (i - 1/2)/g \) be their adjusted ranks. Let \( \tilde{\alpha}(j) \) be the average of \( a(i) \) for all \( i \) such that \( p(i) = \tilde{p}(j) \).

1. For \( j = l, \ldots, u - 1 \), define and compute:
   \[
   m_j = \frac{\tilde{p}_{(j+1)} + \tilde{p}_{(j)}}{2},
   \]
   \[
   \tilde{x}_j = \arcsin \left[ 2 \left( m_j - \frac{1}{2} \right) \right],
   \]
   \[
   \delta_j = \frac{\tilde{\alpha}(j+1) - \tilde{\alpha}(j)}{\tilde{p}(j+1) - \tilde{p}(j)}
   \]
   and
   \[
   \tilde{y}_j = \log(\delta_j).
   \]

2. Apply Lowess (Cleveland, 1979) to \( (\tilde{x}_j, \tilde{y}_j) \) for \( j = l, \ldots, u - 1 \) to obtain an estimated curve \( \hat{y}(\cdot) \).

3. For \( j = l, \ldots, u \), let \( \hat{f}^*(\tilde{p}_{(j)}) = \exp[\hat{y}(\tilde{x}_{(j)})] \) be an estimate of \( f(\tilde{p}_{(j)}) \) up to a constant \( c \).

4. Let \( \hat{f}(\tilde{p}_{(i)}) = 1/c \hat{f}^*(p_{(i)}) \) estimate the PDF at \( p_i \) for \( i = l, \ldots, G \), where
   \[
   c = \frac{1}{2} \sum_{j=l}^{u-1} [\hat{f}^*(\tilde{p}_{(j)}) + \hat{f}^*(\tilde{p}_{(j+1)})] \Delta_j
   \]
   is determined by trapezoid rule integration (Anton, 1992).

5. Let \( \hat{F}(\tilde{p}_{(i)}) = 0 \) and for \( k = l + 1, \ldots, u \) let
   \[
   \hat{F}(\tilde{p}_{(k)}) = \frac{1}{2} \sum_{j=l}^{k-1} [\hat{f}(\tilde{p}_{(j)}) + \hat{f}(\tilde{p}_{(j+1)})] \Delta_j
   \]
   be an estimate of \( F(\tilde{p}_{(j)}) \) obtained by trapezoid rule integration.

6. Following Efron et al (2001), let
   \[
   \hat{\pi} = \min_{1 \leq i \leq g} \hat{f}(p_i).
   \]

7. For \( i = l, \ldots, g \), obtain \( r_{(i)} \equiv \hat{r}(p_{(i)}) \) by substituting, \( \hat{\pi} \), \( p_{(i)} \) and \( \hat{F}(p_{(i)}) \) in
   \[
   \hat{r}(\alpha) = \frac{\hat{\pi} \alpha}{\hat{F}(\alpha)}.
   \]
   Additionally, use L'Hospital's Rule (Anton, 1992).
\[
\lim_{\alpha \to 0} \hat{\alpha} = \lim_{\alpha \to 0} \frac{\hat{\pi} \alpha}{\hat{F}(\alpha)} = \lim_{\alpha \to 0} \frac{\hat{\pi}}{f(\alpha)}
\]

to motivate \( \frac{\hat{\pi}}{f(0)} \) as an estimate of the cFDR for p-values that equal 0.

8. Following Storey (2002), define
\[
h_{(i)} = \min_{k \geq i} (r_{(k)})
\]
as a monotone quantity based on the cFDR estimates \( r_{(i)} \) for \( i = 1, ..., g \).

8.5.3. \textit{Jittering to address pooling effect}

Recall that we have two means for the expression levels and individual data on behavioral traits. We lack information on the individual expression levels. One option was thus to reduce all information to strain averages, having only 8 points, which amounts to using strains as the unit of correlation analysis. This type of correlation has been termed \textit{ecological} correlation (e.g. Robinson, 1950, Greenland and Robins, 1994), having limitations when it comes to interpretation, as it is the individual animal, behavior and expression level that is studied. Furthermore, this approach yields high correlations but also high p-values (not very significant) because of the small number of strains available. This option was discarded early on.

It may still be argued that since strains are genetically identical it is reasonable to assume that the individual gene expressions for each strain are all at the same (mean) level while still making use of the individual behavioral information. Such an approach leads to correlation that is always too high and too significant due to the artificially reduced variation of the expression level. For instance, consider the case of the gene noted as r045c007, within the cerebellum brain region, and the proportion of time the mouse spends in the center of the wide arena (after logistic transformation). Figure 31a displays by the bold dots the proportion of time at center for individual animals versus expression levels all set at their strain average. The estimated correlation with the behavioral trait is -0.62, with p-value <0.0001. In the calculation of Spearman’s correlation, the ranks of expression levels are being used. Thus all expression levels of animals from the same strain will receive the same rank, artificially reducing the real variability of expression levels.

We may simulate individual values of gene expression for mice, under the null hypothesis of no correlation between expression level and behavior, by jittering
the mean expression values. Jittering refers to the addition of a noise factor to a variable, due to this variable remaining constant across different levels of another variable. Here, gene expression measurements remain constant across different animals in the same strain. How much jittering is appropriate? From the two means we have for each strain, we can estimate the between-mice individual variability in expression levels: for gene g and brain region r, 

$$\hat{\sigma}^2_{gr} = \frac{\hat{\sigma}^2_{b,gr}}{k_{gr}} \cdot n_{gr},$$

where \( \hat{\sigma}^2_{b,gr} \) is the estimated variance of the biological replicates, \( n_{gr} \) is the number of animals for which gene expression in gene g and brain region r was measured (here 10), and \( k_{gr} \) is the number of biological replicates taken to measure this expression level (here 2). This is the amount by which we randomly jitter the mean values and assign the random values to the individual animals. Now the correlation and the significance are more realistic: 2 different runs for the above combination yield correlations of -0.31 and -0.44, and the p-values are 0.029 and 0.001, respectively, as depicted in Figures 31a-b. Both are weaker than before and less significant, but closer to reality.

**Figure 31:  Jittered data vs. original data – two runs**

After jittering the mean values of the expression levels and assigning the random values to the individual animals, the correlations are weaker and the p-values are less significant.
However, it is also immediately evident now that both correlations and their significance vary much between one such simulated data to the other. We thus simulated such data 1000 times and averaged the correlations and the logarithm of the p-values over the simulation. For the above example, the simulation-averaged correlation is -0.36 and the back transformed p-value is 0.01.

8.5.4. Additional details of the statistical analysis

Analysis of variance will be used to test the difference between strains on the gene-brain region level. In assessing the strain differences as a first stage of research, we face the problem of too small a sample size, as the data consists only 2 biological replicates for each gene in each brain region. This problem is quite common in microarray experiments that include more than one factor, for the mere reason of high economical cost. Pavlidis et al (2003) show that statistical analysis of experiments containing less than five biological replicates may result in poor power and reproducibility. Thus, rather than using the conventional 1-way ANOVA F-test, we will test the simple effect of strain within brain region, as proposed by Winer (1971). For each brain region, the F-ratio will be based on the between-strain variation estimated from the 1-way model only for that specific brain region (nominator), and the within-strain variation estimated from the full 2-way model with interaction (denominator). Thus for a gene $g$ and brain region $r^*$,

$$F(g, r^*) = \frac{MS_{\text{between-regions}}}{MS_{\text{within}}} = \frac{n \sum_{s=1}^{S} (\bar{X}_{gsr^*} - \bar{X}_{g^{\cdot}r^{\cdot}})^2}{(S-1) \sum_{s=1}^{S} \sum_{r=1}^{R} \sum_{k=1}^{n} (X_{gsrk} - \bar{X}_{gs^{\cdot}})^2 / SR(n-1)}$$

where $n$ is the number of replicates and $S$ and $R$ are the numbers of strains and brain regions, respectively. A discovery in this analysis is that there is some difference in expression levels of gene $g$ in brain region $r^*$.

Correlation coefficients are calculated in the next stage of the analysis. Our choice of the Spearman correlation rather than Pearson correlation in order to assess the association is motivated by the fact that the rank-based Spearman correlation assesses the strength of a monotone relationship while Pearson’s assesses the strength of a linear relationship. It is true that through the use of suitably chosen transformations for each combination Pearson’s correlation may be almost as useful. However, with little-monitored screening of many correlations this becomes a formidable task. The use of Pearson correlation and the effort involved cannot be
avoided when the number of observations available for the correlation analysis is very small, where Spearman’s correlation cannot achieve low p-values. This is not our case.

8.5.5. Simulation study

A simulation study of the FDR level achieved by hierarchical testing was conducted on data simulated to have the same characteristics as our experiment data (except for number of genes). Expression data at the gene and brain region level was set to include non-zero differences between strains for a proportion  of the cases. Of them, a proportion  was set to correlate with some of the behavioral traits. A correlation of any gene was either for all of its brain regions (pa) or for only one of them (1- pa) (but note that pc and not pa governs π0). pa was set to 0.5 for all configurations except 7, for which it was set to 0.95. p of the traits were correlated with expression. It was set to 0.2 for all configurations. Effect and noise sizes were chosen so that the power under the alternative at the first stage will be high - from 0.5 to 1.

Behavioral traits data simulation

Since traits were found in the real data to be correlated with each other in a clustered form, they were constructed in 3 clusters in the simulation. Thus, a correlation could be only with none or all the traits in one cluster. Trait data was generated in such clusters through independent sets of orthogonal vectors containing entries -1,0, and 1. Partial sums of the vectors were used to create dependent traits, and N(0,4) orthogonal normal deviation was added independently to each trait. Formally, for each cluster made out of J assigned orthogonal vectors, the signal for trait b, strain s and mouse m is

\[ B_{bsm} = \sum_{j=1}^{J} I_j [V_j + \varepsilon_{s_{mj}}] + \varepsilon_{bsm}, \]

where Vj, j=1,...,J, is one vector in the set of orthogonal vectors, and Ij is an indicator fixed on 1 for a specific j within a cluster, in order to create within-cluster dependence.

Gene expression data simulation

The expression of correlated combinations of genes and brain regions was generated by linear combinations of the orthogonal vectors belonging to one
randomly selected cluster of traits. Again, independent \( N(0,4) \) distributed noise was added. Formally, the signal for gene \( g \), brain region \( r \), strain \( s \) and mouse \( m \) is

\[
M_{grsm} = \sum_{j=1}^{J} a_j V_j + \epsilon_{grsm},
\]

where \( a_j \) is a linear multiplier. The expression levels for cases with difference between strains but no correlation with behavioral data were generated through an additional vector of effects orthogonal to all the vectors that generated the trait data.

**Simulation configurations**

Data was simulated 500 times on 1000 genes, 5 brain regions, 20 behavioral traits and 12 strains, each with sample size of 2 mice. Table 2 lists the selected configurations of proportion of false null hypotheses in each stage along with \( \pi_0 \). Each configuration represents a different possible magnitude and ratio of the two. The last configuration, run on 10,000 genes, represents a problem more similar to our experiment data in terms of multiplicity size, with no correlations and nearly no strain differences.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Number of Genes</th>
<th>( p_d )</th>
<th>( p_c )</th>
<th>( \pi_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.998</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>0.5</td>
<td>0.1</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>1000</td>
<td>0.5</td>
<td>0.5</td>
<td>0.95</td>
</tr>
<tr>
<td>7</td>
<td>10000</td>
<td>0.95</td>
<td>0.95</td>
<td>0.82</td>
</tr>
<tr>
<td>8</td>
<td>10000</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 2: Simulation configurations**

Different combinations of proportions of strain difference in the first stage and correlations in the second stage comprise a range of possible scenarios. The last configuration represents a case more similar in sample size to the experiment data, with practically no effect of interest.

The comparison of the proposed methods to the single-step methods will take place on three configurations. Two configurations, 2 and 3, have very high \( \pi_0 \) - 0.998 and 0.99, respectively, as in a typical case of correlations between genotype and
phenotype, which are rare. The third configuration is 7, where $\pi_{0}$ is lower - 0.82, but still relatively high, and may therefore also be realistic in microarray data.

The simple effect F-test was applied at the first stage to select combinations of gene and brain region where strain differences in expressions were found. Next, the expression levels of the selected combinations were tested for Spearman’s correlation with behavioral traits. The FDR controlling procedure in BH was applied in two ways: using the subset selection method, and using the hierarchical testing scheme. The FDR level was calculated by averaging the false discovery proportion for each configuration.

8.6. **Analysis and results**

We first describe the simulation study results, since they affect our choice of thresholding decisions in the real data analysis.

8.6.1. **Simulation study results**

*Proposed methods*

For the subset selection method, the FDR is controlled at the desired level of 0.05. For the hierarchical testing method, we first examine the value of $\delta^{\ast}$. We found that it receives values very close to 1, and in fact keeps stable, around 1.04-1.06. Running additional simulations on configuration 3 with $q$ set to 0.01 and 0.1, which yields $\delta^{\ast}$ of 1.03 and 1.06, respectively, we may reasonably conclude that $\delta^{\ast}$ keeps stable for any $q$. This result of $\delta^{\ast} = 1$ (or maybe 1+$q$?) complies with the results obtained through previous simulations.

The full-tree FDR should be controlled at the level of $2q\delta^{\ast}$. Consequently, the conservative bound on the FDR level is still $2q$, and as seen in Figure 32, it is achieved for all configurations. We observe that the full-tree testing scheme FDR increases as the overall proportions of false null hypotheses decreases, which stems from fewer hypotheses tested. When the proportion of false null hypotheses is small, most rejected hypotheses are parent hypotheses, indeed the case in which the full-tree FDR reaches its upper limit.

Next we examine the FDR of the end-node testing scheme, which under independence should be controlled at level $2Lq\delta^{\ast} = 4q\delta^{\ast}$. Again, $\delta^{\ast} = 1$, and the conservative bound of $4q$ still applies. This is evidently much too conservative
bound, as the actual value that can be inferred from Figure 32 lies well below $2q$. This is in spite of the observation shown by Figure 32 that the outer-node FDR level tends to exceed the full-tree FDR level, as expected. The degree of this excess depends on the number of correlations. Where it is relatively small, as in configurations 2 and 5, this gap is relatively small. Where there are no correlations at all, as in configurations 1 and 4, the two criteria are nearly equal. Here, the expected total number of discoveries remains the same as that of the full-tree scheme.

In view of the above, for both full-tree and end-node FDR control we use

$$q^* = \frac{q}{2}$$

in order to control the FDR at level $q$.

---

**Figure 32:** FDR for full-tree and outer-node schemes

The bound of $2q$ on the FDR level is achieved for all configurations. The full-tree testing scheme FDR increases as the overall proportion of false null hypotheses decreases. The outer-node FDR level tends to exceed the full-tree FDR level, where the degree of the excess depends on the number of correlations: where it is relatively small, as in configurations 2 and 5, this gap is relatively small. Where there are no correlations at all, as in configurations 1 and 4, the two criteria are nearly equal. Thus for both full-tree and end-node using $q^* = q/2$ will control the FDR at level $q$. 

---
Restricting our interest to the correlation-testing layer, we focus on how much this FDR is increased due to the restriction. This FDR multiplier, previously denoted by $B(R, J)$, increases in the number of families visited relative to the number of rejections within them. Indeed, as seen in Figure 33, our simulation results show that the $FDR^{level}$ multiplier, averaged over a specific configuration, can take very large values. Here, it reaches around 56 and 77, when there are no correlations (configurations 1 and 4), and thus very few rejections relative to the number of gene and brain region combinations selected in the first stage. However, when there are correlations in the data, this multiplier is much smaller. For the cases with the smallest number of correlations (configuration 2 and 5), the multiplier is around 5.3 and 3.1. For cases with a larger number of correlations (configurations 3 and 6), this number is already less than 2, here around 1.6 and 1.8. For the case with the largest number of correlation (configuration 7, not depicted), the factor moves further towards 1, here 1.3.

Figure 33: **FDR multiplier for the level restricted scheme**

The $FDR^{level}$ multiplier can take very large values when there are no correlations (configurations 1 and 4) in the data, but is much smaller when there are.
Large values that the FDR multiplier can take are responsible for the large values that the $FDR_{level=1}$ can take, as depicted in Figure 34, where $FDR_{level=1}$ of the uncorrected method may reach 1. Nevertheless, dividing $FDR_{level=1}$ by the estimated factor $B(R,J)$, will reduces $FDR_{level=1}$ to the desired level. Thus, for the level-restricted testing scheme, using a $q^* = \frac{q}{B(R,J)}$ will achieve control at level $q$. $B(R,J)$ can be calculated a priori based on the data, setting $q$ to the desired final level. Then, the data should be tested again, this time using $q^*$.

**Figure 34:** FDR for the level restricted scheme

When $FDR_{level=1}$ is large, so will be the FDR multiplier. Dividing $FDR_{level=1}$ by the estimated factor $B(R,J)$ will reduces $FDR_{level=1}$ to the desired level. Thus, for the level-restricted testing scheme, using a $q^* = \frac{q}{B(R,J)}$ will achieve control at level $q$. 
While $\delta^* \cdot \frac{m_0}{m}$ can in principle get as high as 1.44, we saw from Figure 32 that over the configurations studied, it is just around 1. Hence, searching for the minimal $q^*$ for which we can control $q^* \cdot B(R, J) \leq q$, yields $\text{FDR}^{\text{level}} \leq q$.

**Comparison to single-step methods**

All single-step methods were found to control the FDR at the desired level of 0.05. Table 3 displays the power of the proposed methods to identify correlations compared to the single-step methods, for the three data configurations. As can be seen, among the single-step methods, the adaptive BH method and the average estimate method do equally well in terms of power. The SPLOSH method usually yields considerably lower power, except in the case of the highest $\pi_0$. The hierarchical testing procedure has the highest power, even though it requires the use of a lower $q$. The subset selection method retains power similar to the two more powerful single-step methods.

<table>
<thead>
<tr>
<th>Method Type</th>
<th>Method</th>
<th>Power $\pi_0 = 0.82$ (conf. 7)</th>
<th>Power $\pi_0 = 0.99$ (conf. 3)</th>
<th>Power $\pi_0 = 0.998$ (conf. 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-step methods</td>
<td>BH</td>
<td>0.76</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>adaptive BH</td>
<td>0.78</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>SPLOSH</td>
<td>0.45</td>
<td>0.34</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>average estimate</td>
<td>0.78</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td>proposed methods</td>
<td>subset selection</td>
<td>0.59</td>
<td>0.44</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>hierarchical tree</td>
<td>0.59</td>
<td>0.49</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**Table 3: Power comparison of proposed methods and single-step methods**

Estimated power from simulation study (standard errors of simulation estimates are smaller than 0.006 for all configurations). When $\pi_0$ is near 1, the hierarchical tree testing scheme retains the highest power while controlling the FDR, in spite of requiring a lower $q$ relative to other methods. When $\pi_0$ is much smaller than 1, the single-step methods adaptive BH and average estimate retain the highest power.
8.6.2. Data analysis results

Brain ANOVA per gene

The simple effect ANOVA was applied to identify combinations of gene and brain region that distinguish between strains. The analysis yielded 167 combinations of gene and brain region with differences between strains, using the threshold of 0.025. Within this list, 92 different genes were included, and all five brain regions represented.

Correlation analysis

The gene and brain region combinations identified in the first stage were now each tested for correlation with each of the 17 behavioral traits. Thus there are 17 hypotheses times 167 combinations of gene and brain region, which brings the multiplicity problem in this stage to over 2800 tests of hypotheses. We implement the two testing approaches to control the overall FDR at level 0.05, adopting the strategy of which appropriateness was further supported by the simulations. For the subset selection method, using this threshold in the second stage implies overall control due to only slight possible dependence between the two stages. For the hierarchical tree method, using threshold of \( q/2 \) guarantees overall control at level \( q \). We thus ran the tree-testing scheme on coarse grid of 0.01, 0.025 and 0.05, taking two times the minimal p-value of all runs as an adjusted p-value, for each combination.

Correlation analysis at the 0.05 level of the jittered data through the subset selection method identifies 225 triplicates for which the gene and brain region combination is correlated with a specific trait. Analysis through the hierarchical tree method identifies 230 triplicates. At this level \( B=(167+230)/231=1.7 \), justifying the somewhat conservative value of 2 we chose. While the hierarchical tree method only slightly exceeds the subset selection method in the overall number of triplicates, we observe that there are several interesting differences in the content of the two lists. First, the subset selection method identified 3 additional genes over the other list and 6 additional gene and brain region combinations. However, it yielded only 8 additional triplicates, while the hierarchical tree method identified 24 additional triplicates (and no additional genes or gene and brain region combinations). This reflects the fact that the subset selection method is more powerful in identifying the combination of gene and brain region in the first layer. On the other hand, once a nuclear family is identified, the hierarchical tree method will be able to make more discoveries inside the family. This is evident in Figure 35, where the p-values are
displayed for each of the methods. The color displays the direction of the correlation (positive by green or negative by red), and is brighter for higher significance. Clearly, within a nuclear family (a column in the matrix) selected by both methods, the p-values generated by the hierarchical tree method are smaller, as can be seen by the much brighter colors of Figure 35b. Consequently, for the hierarchical tree method there are relatively many additions of gene and brain region expression by behavioral endpoints.

Note that we chose not to display at all p-values above 0.05. While for the significant cases we can make a conclusion regarding the direction of the correlation, we cannot do so for the non-significant cases, even though their sign is available. This is the motivation for blackening them in the matrix. For the significant correlations, the display may further benefit from including a color code that displays the size of the correlation on top of the information about their significance.

8.7. Conclusions

In this chapter we suggested two approaches to control the FDR in a two-staged analysis. In both methods only families corresponding to genes that passed the FDR screening are tested for correlation. In the first method all these families are combined to a single family over which the FDR is controlled. In the second method each screened family is tested separately while controlling the FDR within it. Under the second scheme, the overall FDR can still be controlled by lowering the level at which the FDR is controlled. The amount by which the level should be lowered depends on the goal of the researchers, whether in full-tree, fixed-layer or end-nodes.

If all screened families are approximately the same in terms of \( \frac{m_0}{m} \) and p-values distribution, the first method is superior as the testing is done at a higher \( q \) and the behavior of BH will be little affected by the amalgamation of the families. If a few families have high correlations while many others have none (or close to none) the second method has the advantage in spite of lower \( q \) at both stages. In our case, indeed, the subset selection method was as expected more powerful in the first stage, but the second method is more powerful in identifying correlations for the families selected in the first stage. Even when theoretical bounds do not guarantee control of FDR within one stage \( l \), it seems that at least in the case discussed, a correction using \( B(R_l, J) \) will provide control at the desired level.
Figure 35: P-values for expression-behavior correlation analysis

The raw p-values are displayed for (a) the subset selection method and (b) the hierarchical testing method. The color displays the direction of the correlation (positive by green or negative by red), and is brighter for higher significance. The subset selection method is more powerful in identifying the combination of gene and brain region in the first layer. Once a nuclear family is identified, the hierarchical tree method will be able to make more discoveries inside the family, as seen by the brighter colors of (b).
The single-step adaptive methods are more powerful than the hierarchical method when \( \frac{m_0}{m} \) is substantially smaller than 1, and lose their advantage as \( \frac{m_0}{m} \) nears 1. However, \( \frac{m_0}{m} \) substantially smaller than 1 is not the typical case in microarray data or in most other studies in which the number of hypotheses is large, since researchers would naturally look for discoveries which are not abundant. In contrast, the hierarchical method has the advantage of reducing the number of tested hypotheses, and thus reduces the payment in power due to multiplicity. Thus, if the researcher is interested in an initial screening of the genes that would successfully eliminate cases with no chance of being selected in the second stage, the hierarchical testing scheme will yield the largest number of discoveries. However, if a sensible initial filter does not exist and \( \frac{m_0}{m} \) is away from 1, as, for example, in focused arrays, an adaptive FDR controlling single-step procedure, such as the adaptive BH (Hochberg and Benjamini, 1990) or the average estimate (Jiang, 2004), can be used.

We emphasized the importance of avoiding ecological correlations that take into account only pooled or averaged data. We offered a solution based on jittering the expression data by an amount estimated from the biological replicates. We showed how this solution avoids the artificial increase in the correlation coefficient and its significance. This solution may be extended to the case where the measured functional variable is given by average, as is typical when the search for correlation is conducted over publicly available databases. In such searches standard deviation information on the phenotype, which is usually available as well, can be used to jitter both phenotype and expression level. Nevertheless, whenever feasible we recommend measuring phenotypic information at the same resolution as the genotypic information is measured.
9. Discussion and further research

9.1. Summary of results

This thesis aimed to establish the relevance and applicability of FDR control in the analysis of gene expression data. Several important perspectives were investigated, emphasizing the nature of microarray technology and research objectives. FDR estimation approaches and controlling procedures were examined and compared, the scope of FDR applicability under dependence was generalized, and FDR control subjected to the ever-growing complexity of research goals and experimental design was addressed.

This thesis offers genomics researchers FDR technologies for addressing the multiplicity problems inherent in the statistical analysis of gene expression data. This data challenges the analyzer since it is subjected to unique characteristics related to experimental design, dependencies between gene expression measurements, complex analytical process, small sample size and pooling of subjects. From the statistical aspect, the thesis establishes the appropriateness and benefits of using FDR controlling procedures for this analysis: It provides the theoretical justification to extend the use of these simple procedures under extreme dependency cases. Some of the tools offered were novel when the work on this thesis started and by now have become standard. Other powerful novel methodologies are offered to control the FDR in complex analyses, at a scale not confronted before (for all we know).

Our first step taken in order to inquire the appropriateness and advantages of FDR methodology for microarray data analysis was a comparative study of four FDR controlling procedures. One procedure considered was the BH linear step-up procedure as applied to the p-values corresponding to the per-gene statistical test. The second procedure used the same BH, as applied to the marginal p-values estimated by resampling and pooling over genes. The two other procedures were based on estimating the distribution of the p-values and the FDR at a given threshold using resampling: One threshold was conservative on the mean, and the other on the upper limit of the confidence interval on the mean. Their properties were examined in terms of conservativeness, gain in power, relevance of their assumptions to gene expression data and simplicity of implementation. In addition to the higher power retained by
these procedures relative to FWER controlling procedures, our results implied that the use of resampling achieves considerable gain in power. Thus, computational efforts should be invested according to the researcher’s emphasis on maximizing the number of genes selected for the next stage.

The applicability of the FDR was then extended to the case of two-sided testing with dependent test statistics, which is highly relevant to microarray data analysis. The least favorable data properties bringing to the maximal FDR were identified, considering common correlation between the normally distributed test statistics, where a subset of \( m_0 \) hypotheses are true out of a total of \( m \) null hypotheses, and \( \mu_2 \) is the distance between the means corresponding to the true and false null hypotheses. It was shown that the least favorable \( \mu_2 \) is the sum of the two unique normal critical values corresponding to each subset of null hypotheses when there is perfect correlation between the test statistics. It was found that the FDR reaches a global maximum when \( \frac{m_0}{m} = \frac{1}{2} \). Assuming the combination of all above least favorable conditions, it was concluded that for any common \( \rho \), using \( q \) would control the FDR at level \( q \), and using \( q^* \leq \frac{q}{1 + \frac{1}{2} \left( \frac{m_0}{m} \right)} \) would control the FDR at level \( \frac{m_0}{m} q \). This yields \( q^* \leq \frac{4}{5} q \) when \( \frac{m_0}{m} = \frac{1}{2} \) and \( q^* \leq \frac{2}{3} q \) when \( \frac{m_0}{m} = 0 \). It was shown by simulation that FDR increases with correlation also in cases of dependency structure other than common correlation, when the correlation is high.

The next direction was to examine the FDR control from a point of view that takes into consideration the complexity of a typical microarray research. A research that consisted of several research questions was considered. When this type of analysis is conducted on microarray data, where each gene is examined separately for several hypotheses, the multiple hypotheses can escalate to very large numbers. Instead of controlling the FDR for all the hypotheses at once, two approaches for dealing with the multiplicity problem were examined: The subset method, that controls the FDR at each stage, and the hierarchical tree method, that organizes the hypotheses in a tree structure, comprising of families of hypotheses, and controls the FDR for each family.
An algorithm based on the nature of FDR progression in a hierarchical tree-testing scheme was used in order to efficiently control it in a study consisting of extremely large families of hypotheses. It was shown that while in some cases the subset method is the only one that can be used, due to clear dependency between research questions, in other cases, where none or moderate dependency can be assumed, the hierarchical tree-testing approach yields more rejections while controlling the FDR at the desired level.

Finally, solutions were introduced and proposed to two important problems encountered in studies of correlation between genes’ expression levels and phenotypic data, where genes are first screened for showing differential expression across a factor of interest. Such experiments are a natural extension of current experimental methodologies, and because of their increased specificity are expected to become an important component of functional genomics research. The control of the increased false discoveries stemming from the many hypotheses tested and the hierarchical way by which the testing is done, was now addressed while being subjected to two complications: (i) some degree of dependence between the test statistics of the two different stages of the analysis, and (ii) the difference in measurement units between genotypic and phenotypic data stemming from limitations of current microarray technology.

Correlating behavior and gene expression in specific brain regions was of interest, but it was unknown ahead of time what regions will give rise to interesting findings. Thus the five brain regions were included in the same ANOVA, as well as the 8 strains, making use of the increased replicates on one hand, and dealing with the resulting increase in the number of hypotheses in a new way: Genes were first screened for strain differences in expression levels, since these strains are known to have different behavioral phenotypes and thus may indicate correlation with behavior. Then, correlation between expression and behavioral endpoints were tested for.

The use of both the subset method and the hierarchical testing method was offered for controlling the FDR, in such multistage analysis. In both methods only families corresponding to genes that passed the FDR screening are tested for correlation. In the first method all these families are combined into a single family over which the FDR is controlled. In the second method each screened family is tested separately while controlling the FDR within it. Under the second scheme, the overall FDR can still be controlled by lowering the level at which the FDR is
controlled at each level. The amount by which the level should be lowered depends on the goal of the researchers, whether in full-tree, fixed-layer or end-nodes.

If all screened families are approximately the same in terms of \( \frac{m_0}{m} \) and \( p \)-values distribution, the first method is superior as the testing is done at a higher \( q \) and the behavior of BH will be little affected by the amalgamation of the families. If a few families have high correlations while many others have none (or close to none) the second method has the advantage in spite of lower \( q \) at both stages. In our case, indeed, the subset selection method was as expected more powerful in the first stage, but the second method is more powerful in identifying correlations for the families selected in the first stage.

The available theory (Benjamini and Yekutieli, 2002) calculates the needed tightening of \( q \) under the assumption of independence between the test statistics used for screening and those used for testing in the second stage. The theoretical bounds were used under the weak dependency existing between the F-statistic and the Spearman correlation, checking their adequacy by simulations. This need not be the case for stronger types of dependence, as may exist in post-hoc pairwise comparisons. Even when theoretical bounds do not guarantee control of FDR within one stage \( l \), it seems that at least in the case discussed, a correction using \( B(R_j, J) \) will provide control at the desired level.

The importance of avoiding ecological correlations that take into account only pooled or averaged data was emphasized. A solution based on jittering the expression data by an amount estimated from the biological replicates was offered. It was shown how this solution avoids the artificial increase in the correlation coefficient and its significance. This solution may be extended to the case where the measured functional variable is given by average, as is typical when the search for correlation is conducted over publicly available databases (e.g. Jackson Labs mouse strain resources - [http://www.jax.org/resources/mouse_resources.html](http://www.jax.org/resources/mouse_resources.html) and MIAME microarray experiment resources - [http://www.mged.org/miame](http://www.mged.org/miame)). In such searches standard deviation information on the phenotype, which is usually available as well, can be used to jitter both phenotype and expression level. Nevertheless, whenever feasible it is recommended to measure phenotypic information at the same resolution as the genotypic information is measured.
9.2. Further research

Over the last two years, increasing number of active research groups, in numerous theoretical and applied papers, suggest new interpretations and estimators of the FDR, along with controlling procedures, helpful software and supportive application studies. The first part of this thesis, described in Chapter 5 and published in Reiner et al (2003), has had its own timely effect on the research in the field. The two main philosophies regarding statistical inference are also presented in regard to microarray data analysis: the Bayesian and the frequentist. While they do not contradict each other, they suggest different perspectives on the FDR and on how it can be estimated from the data. This is a cause to conduct a methodological comparative study of these two approaches. Moreover, both lines of research contemplate the use of permutation-based algorithms to estimate the FDR, on the already established grounds of effective estimation capability, and thus many alternative procedures have been proposed within the two lines of research. Therefore, one of the means of comparing the two approaches is through the working characteristics of the procedures. Such comparisons are still missing.

This suggestion for further investigation also applies for sample size estimation in microarray experiments while controlling the FDR. This answer seems to be still in its infancy. So far, only one work offers sample size calculation based on sound arguments involving both power and FDR consideration (Muller et al, 2004). While this work adopts the Bayesian perspective and formulates the question as a decision problem, another natural direction would be through the frequentist approach, starting by producing simulation-based sample sizes as a function of FDR and power, subjected to different dependency scenarios.

As it was found that simple FDR controlling procedures perform well on correlated test statistics, it was further shown that the use of common correlation coefficient to bound from above the type I error in the case of general correlation coefficient is a valid and convenient approach supported by our results and contemplated in literature. However, it is not clear why the FDR increases with the level of common correlation. Based on the computational reasoning used in Chapter 6, this implies that the joint distribution of the test statistics absolute values covers the maximal expected region of false rejections when correlation is maximized. This could be a direction for a proof. In addition, the results suggest that the further it is from independence, the higher the potential impact on the FDR. This could result in a
decrease of the FDR relative to the case of independence, as was seen in the shape of
the FDR function versus $\mu_2$ and versus $\rho$, and show where other procedures may
gain much. It is plausible, therefore, to guess that the minimum FDR is obtained at
maximum correlation, as was the case for the maximal FDR. Our arguments were
carried for the case of normally distributed test statistics. Other types of test statistics,
such as t, F and rank statistics may show different FDR behavior, and thus may have
different least favorable properties that are to be explored.

It was shown in this thesis that the case of common correlation 1 yields the
highest FDR. This might suggest that understanding the case of common correlation
facilitates the establishment of type I error conservativeness. In fact, FWER was
examines several methods for approximating multivariate normal and t probabilities,
where part of them relies on replacing the true correlation matrix with a more
convenient matrix. He finds that these methods approximate the true values with
satisfactory accuracy for practical purposes. One method is based on Tukey and
Kramer’s conjecture (Tukey, 1953, Kramer, 1957), that the use of balanced critical
values instead of the true ones is always conservative (i.e. the covariance matrix $V$
is replaced by the identity matrix). Iyengar (1988) and Iyengar and Tong (1989)
investigated replacing $\rho_{ij}$ by their common average $\bar{\rho}$ (sometimes called generalized
Tukey’s conjecture). While a theoretical proof for FWER control for any correlation
matrix was not provided, no counter example has been found for the types of
correlation structure arise with multiple comparison problems.

Following these promising evidences, the question is whether a similar
conjecture holds for FDR control. Investigators have been already addressing FDR
control as well, for the case of common correlation. Troendle (2000) offers step-
down and step-up procedures, which asymptotically control the FDR when the test
statistics are the t statistics from consistent multivariate normal estimators of the
tested parameters. Determination of the necessary critical constants is achieved using
numerical integration when the correlations are equal. Somerville (2004) extends the
use of the common correlation assumption by offering suitable step-down and step-up
FDR controlling procedures along with tables of critical values corresponding to
possible given $\rho$’s. He also observes that underestimating $\rho$ leads to a conservative
rejection procedure, an observation which is in accordance with the finding here that
the FDR is monotone in $|\rho|$. Dunnett et al (2001) develop step-down and step-up methods for calculating the power when testing for treatments vs. control. They assume common correlation of \( \frac{1}{(1+n_0/n)} \) between the test statistics, where \( n_0 \) is the sample size of the control group, and \( n \) is the sample size of each one of the treatment groups. So now it remains to be shown that using these FDR controlling procedures with the average correlation remains valid.

It follows that our approach on FDR control under dependence may be extended to give support to the conjecture. FDR control can be examined when replacing general correlation matrix by a common correlation matrix whose value is the average correlation. An increase in FDR will imply that the common correlation case offers an upper bound to the FDR. Interestingly, it would also be illuminating to examine the scope of the conjecture for FWER control, using similar means of exploration as we did for the FDR, thereby strengthening our results on conservativeness under dependence.

While the hierarchical testing scheme offers an attractive solution to the one of the more challenging recent problems in microarray analysis, its limitation is the assumption of independence between stages in the hierarchical layout. In many cases, dependency exists by definition between the results of two research questions. The two stages composing the studies discussed in Chapters 7 and 8 are possibly dependent to some degree, but this dependence is shown by simulation to have limited effect on FDR control. However, this situation can change for other structures of dependency. Further verification of the validity of the procedure is being planned.

Several additional issues in hierarchical testing are of interest for further examination: First, experience at QTL analysis using hierarchical testing (Benjamini and Yekutieli, 2002) revealed that the upper bounds offered by the current theory are somewhat high. Additional theoretical and simulation work may allow us to lower them. Second, as our case contained two stages of analysis, these were the only trees studied. FDR control over a higher number of stages should be examined as well. Third, it is of interest to turn the hierarchical methods into methods adaptive to \( \frac{m_0}{m} \), thereby increasing their power even further. Finally, an additional issue to be further
investigated is the role of the $\delta^*$ component in the control level, subjected to the data configuration, in particular its interaction with $\frac{m_0}{m}$.
References


procedures. Recent Developments in Multiple Comparisons Procedures, Institute of Mathematical Statistics, Lecture Notes – Monograph Series, 47, 24-32.


Appendix:  Computer programs

A1.  Identifying differentially expressed genes between two independent group using FDR controlling procedures (Chapter 5)

<table>
<thead>
<tr>
<th>Language:</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective:</td>
<td></td>
</tr>
</tbody>
</table>
This R program adjusts p-values generated in multiple hypothesis testing of gene expression data obtained by cDNA microarray experiment. It assumes treatment and control groups with replications.

| Reference: |

| Input Data: |
Example format for input data is in a lipid metabolism study described in:

(see data file rg.alko.morph in: http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html)

The input file should contain the following columns, with headers:
- gene id
- for each subject: reference signal followed by treatment signal

| Output: |
1. Plot of p-values vs. reference |t| (as described in Reiner et al, 2003). The |t| values should represent the observed absolute t's. A suggested algorithm for the constructing the vector is included in the program.
2. Vectors of adjusted p-values for each |t|, for each method.

| Data Pre-processing: |
data normalization using lowess smoother.

| Hypothesis Testing: |
The null hypothesis of no difference in differential expression between treatment and control group was tested, for each gene.

| Function arguments: |
- file.name:  name of R object containing the data
- control.size:  control group size (number of subjects)
- treat.size:  treatment group size (number of subjects)
- perm:  number of permutations for resampling
- ref.length:  length of reference |t| vector
fdr.ma <- 
function(file.name, control.size, treat.size, perm = 1000, ref.length = 200) 
{

######################
#  Begin Program  #
######################

# Sample Size Definitions #
genes.num <- dim(file.name)[1]
n.control <- control.size
n.k <- treat.size
n.total <- n.control + n.k

# Definitions for Normalization #
log.ratio.array <- matrix(NA, genes.num, n.total)
a.array <- matrix(NA, genes.num, n.total)
m.array <- matrix(NA, genes.num, n.total)
norm.m.array <- matrix(NA, genes.num, n.total)
lowess.m <- vector("numeric", genes.num)

# Definitions for Computing No-Resampling P-Values #
t <- vector("numeric", genes.num)
mean.control <- vector("numeric", genes.num)
mean.k <- vector("numeric", genes.num)
var.control <- vector("numeric", genes.num)
var.k <- vector("numeric", genes.num)

# Definitions for Computing Resampling P-Values #
perms.num <- perm
perm.array <- matrix(NA, genes.num, n.total)
t.array <- matrix(NA, genes.num, perms.num)

# Normalization #
k <- 3
for (j in 1:n.total)
{
  a.array[, j] <- ifelse((file.name[, k - 1] == 0 | file.name[, k] == 0), 0 + rnorm(1, 0, 1), log(file.name[, k], base = 2) + log(file.name[, k - 1], base = 2))
  m.array[, j] <- ifelse((file.name[, k - 1] == 0 | file.name[, k] == 0), 0 + rnorm(1, 0, 1), log(file.name[, k], base = 2) - log(file.name[, k - 1], base = 2))
  lowess.m <- lowess(a.array[, j], m.array[, j], f = 1/5)
  uns <- match(rank(a.array[, j]), unique(sort(rank(a.array[, j]))))
  norm.m.array[, j] <- m.array[, j] - lowess.m[uns]
  k <- k + 2
}

# Compute t-statistic #
mean.control <- apply(norm.m.array[, 1:n.control], 1, mean)
mean.k <- apply(norm.m.array[, (n.control + 1):n.total], 1, mean)
var.control <- apply((norm.m.array[,1:n.control] - (apply(norm.m.array[,1:n.control],1,mean)))^2)/(n.control-1),1,sum)
var.k <- apply((norm.m.array[,n.control+1:(n.total)] - (apply(norm.m.array[,n.control+1:(n.total)],1,mean)))^2)/(n.control-1),1,sum)
num.t <- mean.k - mean.control
den.t <- sqrt((var.k/n.k)+(var.control/n.control))
t <- num.t/den.t

# Construct reference |t| vector - suggested algorithm #
sub.size <- ref.length
interval.length.vector <- c(3,3,4)  # length of intervals - sum of elements eq. max
dense.ratio.vector <- c(1,6,3)     # tick marks density ratio between intervals
max.ref.vector <- sum(interval.length.vector)
interval.num.element.vector <- dense.ratio.vector*(sub.size/(sum(dense.ratio.vector)))
from.vector <- vector(“numeric”,length(dense.ratio.vector))
to.vector <- vector(“numeric”,length(dense.ratio.vector))
for (i in 1:length(dense.ratio.vector)) {
  if (i == 1) {
    to.vector[i] <- max.ref.vector - interval.length.vector[i]
    from.vector[i] <- max.ref.vector
    ref.vector <- seq(from=from.vector[i],to=to.vector[i],length=interval.num.element.vector[i])
  } else {
    to.vector[i] <- to.vector[i-1] - interval.length.vector[i]
    from.vector[i] <- to.vector[i-1] + (to.vector[i-1] - to.vector[i]) / interval.num.element.vector[i]
    ref.vector <- c(ref.vector,seq(from=from.vector[i],to=to.vector[i],length=interval.num.element.vector[i]))
  }
}

# Adjust Non-Resampling P-Values by BH Linear Step-Up Procedure #
pt.vector <- 2*(1-pt(abs(t),(n.total-2)))
sort.pt <- sort(pt.vector)
hyp.num <- length(sort.pt)
pt.adjust <- vector(“numeric”,sub.size)
r.vector <- vector(“numeric”,sub.size)
for (m in 1:sub.size) {
  r.vector[m] <- sum(abs(t)>ref.vector[m],na.rm=T)
  pt.adjust <- sort(2*(1-pt(ref.vector,(n.total-2))))*hyp.num/r.vector
}

# Create Sets of Permutations of the Original Data Set #
for (j in 1:perms.num) {
perm.col.vector <- sample(n.total, n.total, replace = F)
perm.array <- norm.m.array[, perm.col.vector]
t.array[, ] <- (((perm.array[, (n.control+1):n.total])%*%rep(1,n.k))/n.k)

((perm.array[, 1:n.control])%*%rep(1,n.k))/n.control) /
sqrt((as.vector(((perm.array[, (n.control+1):n.total])-
    as.vector((perm.array[, (n.control+1):n.total])%*%rep(1,n.k))/n.k)^2)/(n.k-1)%*%
    rep(1,n.k))/n.k)

# Adjust Resampling P-Values #

v.array <- matrix(NA, sub.size, perms.num)
begin.time <- proc.time()
for (i in 1:sub.size)
{
  count.array <- abs(t.array) > ref.vector[i]
count.array[is.na(count.array)] <- 0
v.array[i, ] <- rep(1, genes.num)%*%count.array
}
est.mean.exc <- apply(v.array, 1, mean)
est.95qu.exc <- apply(v.array, 1, quantile, prob = .95)
s.hat <- r.vector - est.mean.exc
s.hat[s.hat < est.95qu.exc] <- 0
su.hat <- r.vector - est.95qu.exc
su.hat[su.hat < 0] <- 0
q.array <- v.array/(v.array-0.0001+s.hat)
qu.array <- v.array/(v.array-0.0001+su.hat)
bh.value <- est.mean.exc/r.vector
bh.value[bh.value > 1] <- 1
bh.value[r.vector == 0] <- 0

# Monotone Adjustment #
for (u in (sub.size-1):1)
{
  pt.adjust[u] <- ifelse(pt.adjust[u]>pt.adjust[u+1], pt.adjust[u+1], pt.adjust[u])
bh.value[u] <- ifelse(bh.value[u]>bh.value[u+1], bh.value[u+1], bh.value[u])
qu.value[u+1] <- max(qu.value[(u+1):1])
q.value[u] <- min(q.value[u:sub.size])
}
# Plot P-Values

plot(ref.vector,qu.value,col=1,type="l",ylab="Adjusted P-Values",xlab="Reference |t|",main="FDR Adjusted P-
Values")
legend(5.5,1.5,legend=c("Res. upper limit","BH point-estimate","Res. point-estimate","BH LSU - No
Resampling"),lty=1,col=1:4,cex=.8)
lines(ref.vector,bh.value,col=2)
lines(ref.vector,q.value,col=3)
lines(ref.vector,pt.adjust,col=4)
abline(0.25,0,lwd=0.01,lty=2)
abline(0.2,0,lwd=0.01,lty=2)
abline(0.15,0,lwd=0.01,lty=2)
abline(0.1,0,lwd=0.01,lty=2)
abline(0.05,0,lwd=0.01,lty=2)
return(qu.value,bh.value,q.value,pt.adjust)

#########################################################################
# End of Program         #
#########################################################################
A2. N-way median polish (Chapter 7)

**Language:**
SAS Macro

**Objective:**
This program corrects expression signals for systematic biases caused by location on the chip by median polish.

**Reference:**

**Input Data:**
The input file should contain the following columns:
- gene id
- reference signal
- treatment signal
- factors of which effects are to be removed by median polish

**Output:**
1. input file with the addition of polished signals
2. Effect plots for each polished variable

**Macro arguments:**
- dataset: input file name
- dim: list of factors of which effects should be removed
- signal: list of variables to polish
- max_iter: maximum num. of iterations to estimate effect
- init_conv: initial convergence value
- eps: minimum convergence value
- plot: indicator for graphical output
%macro med_pol(dataset,dim,signal,max_iter,init_conv,eps,plot);
   %local i j;
data for_plot;
run;
/* find num. of variables to polish and num. of effects to remove */
%words("&signal.");
%let signal_num=&word_num.;
%words("&dim.");
%let dim_num=&word_num.;
/* start variable loop */
%do i=1 %to &signal_num.;
   %let iter=1;
   %let conv=&init_conv.;
   /* define current variable */
data _null_;
call symput('current_signal',scan("$signal.",&i.));
run;
%local conv_list_&current_signal.;
/* initialize residuals and common effect for current variable */
data &dataset.;
set &dataset.;
resid_&current_signal.=&current_signal.;
common_&current_signal.=0;
run;
/* initialize effects of all factors for current variable */
%do j=1 %to &dim_num.;
data _null_;
call symput('current_dim',scan("$dim.",&j.));
run;
data &dataset.;
set &dataset.;
&current_dim._effect_&current_signal.=0;
run;
%end;
/* start iteration loop */
%do %while (&iter. le &max_iter. and &conv. gt &eps.);
   %let conv=&eps.;
   /* start effect loop */
   %do j=1 %to &dim_num.;
      /* define current factor */
data _null_;
call symput('current_dim',scan("$dim.",&j.));
run;
data &dataset.;
set &dataset.;
&current_dim._effect_&current_signal.=0;
run;
%end;
/* start iteration loop */
%do %while (&iter. le &max_iter. and &conv. gt &eps.);
   %let conv=&eps.;
   /* start effect loop */
   %do j=1 %to &dim_num.;
      /* define current factor */
data _null_;
proc means data=&dataset. max
noprint;
var &current_dim.;
output out=dim_file max=max;
run;
data _null_;    
set dim_file;
call
symput("&current_dim._num",max);
run;
/* start level loop */
%do k=1 %to &current_dim._num.;
    /* compute median of
    variable for current level 
    */
    proc univariate
    data=&dataset. noprint;
    where &current_dim.=&k.;
    var residual_&current_signal.;
    output out=med_file
    median=median;
    run;
data _null_;    
set med_file end=last;
if last then call
symput('median',median);
run;
/* update convergence
value */
data _null_;    
call
symput('conv',max(&conv.,abs(&median.)));
run;
/* update residuals
(subtract median) and
effect (add median) */
data &dataset.;
set &dataset.;
if &current_dim.=&k. then
do;
    resid_&current_signal
    al.=resid_&current_signal.-median.;

    &current_dim._effect_
    t_&current_signal.=
    &current_dim._effect_
    t_&current_signal.+       
    &median.;
end;
/* k: dim_num */
/* compute overall median of effect of current factor */
proc sort data=&dataset.;
    by &current_dim.;
run;
data eff_file;
    set &dataset.;
    by &current_dim.;
    if last.&current_dim.;
run;
proc univariate data=eff_file
    noprint;
    var
        &current_dim._effect._&current_sign;
    output out=med_file median=median;
run;
data _null_;"n
    set med_file end=last;
    if last then call
        symput('median',median);
run;
/* update common effect (add median) */
/* and current factor effect (subtract median) */
data &dataset.;
    set &dataset.;
    common_&current_signal.=common_&current_signal.+&median.;
    &current_dim._effect._&current_signal.=&current_dim._effect._&current_signal.-&median.;
run;
%end; /* k: dim_num */
/* j: dim */
data _null_;"n
    call
        symput("conv_list_&current_signal.","&&conv_list_&current_signal."||"||left("&conv.");
    call symput('iter',&iter.+1);"n
run;
%end; /* j: dim */
/* while iteration */
/* initialize polished value of current variable to common effect */
data &dataset.;
    set &dataset.;
    if &current_signal. ne . then
        &current_signal._pol=common_&current_signal.;
run;
/* record iteration num. and convergence value for current variable */
%let iter_&current_signal.=%eval(&iter.-1);
%let conv_&current_signal.=&conv.;
/* update polished variable by adding effects of all factors */
%do j=1 %to &dim_num.;
   data _null_;
   call symput('current_dim',scan("&dim.",&j.));
   run;
   data &dataset.;
   set &dataset.;
   if &current_signal._pol ne . then
      &current_signal._pol =
      &current_signal._pol +
      &current_dim._effect_&current_signal.;
   run;
/* plot effect of current factor for current variable */
%if &plot.=1 %then %do;
   title "&current_dim._effect_&current_signal. - by &current_dim.";
   footnote;
   goptions ftext=SWISS ctext=BLACK
       htext=1 cells;
   axis1 width=1 offset=(3 pct)
       label=(a=90 r=0);
   axis2 width=1 offset=(3 pct);
   symbol1 c=BLUE ci=BLUE v=dot
       height=0.5 cells interpol=NONE
       l=1 w=1;
   proc gplot data=&dataset.;
   plot &current_dim._effect_&current_signal.
   name='SCAT'
       caxis = BLACK
       ctext = BLACK
       cframe = CXF7E1C2
       hminor = 0
       vminor = 0
       vaxis = axis1
       haxis = axis2
       vref = 0
   ;
   run;
   symbol1;
   goptions ftext= ctext= htext=;
   axis1;
   axis2;
title; quit; %end;/*%if*/
proc sort data=&dataset.; by &current_dim.; run;
data effects_for_plot;
set &dataset.; by &current_dim.;
if last.&current_dim.;
length measure $ 15 variable $ 15;
measure="&current_dim.";
variable="&current_signal.";
rename
&current_dim._effect_&current_signal.=effect;
keep measure variable
&current_dim._effect_&current_signal.;
run;
data for_plot;
set for_plot effects_for_plot;
if effect= . then delete;
run;%end;/*second j*/
data resids_for_plot;
set &dataset.;
length measure $ 15 variable $ 15;
measure=" residuals";
variable="&current_signal.";
rename resid_&current_signal.=effect;
keep measure variable resid_&current_signal.;
run;
data for_plot;
set for_plot resids_for_plot;
run;%end;/*i: signal*/
/* write to log iteration and convergence records of all variables */
%do i=1 %to &signal_num.;
data _null_;
call
symput('current_signal',scan('&signal.',&i.));
run;
%put number of iterations for &current_signal.: &&&iter_&current_signal.;
%put convergence value for &current_signal.: &&&conv_&current_signal.;
%put convergence list for &current_signal.: &&conv_list_&current_signal.;
%end; /*second i*/
%mend;
A3. Complex research analysis (Chapters 7-8)

A3.1 Analysis of variance

Language:
SAS Macro

Objective:
This program applies multiple analyses of variance and produces FDR adjusted p-values.
Options other than FDR available for multiple testing adjustment: hoc, bon, sid, hom (see SAS reference).

Reference:

Input Data:
The input file should contain all variables that are in the model.

Output:
1. Files:
   - anova_p_fdr: FDR adjusted p-values
   - anova_p_fdr_[p max]: FDR adjusted p-values with max. specified p-value
   - [input file]_anova: original data with ANOVA parameters and p-values
2. Plots:
   - raw p-values
   - FDR adjusted p-values

Macro arguments:
- dataset: input file name
- dep: dependent variable
- class: list of class variables
- indep: list of independent variables
- int: list of interaction effects
- source: effect variable for which p-values should be adjusted
- byvar: multiplicity variable
- bytree: family variable for hierarchical tree testing scheme
- method: p-value adjustment method
- p_max: max. FDR adjusted p-value for selection
%macro anova_mult(dataset,dep,class,indep,int,source,byvar,bytree,method,p_max);
  %ods_close;
  %by_st(byvar.);
  proc glm data=&dataset. noprint outstat=anova_out;
  %by_st.;
  class &class.;
  model &dep.=&class. &indep. &int.;
  output out=anova_resid r=resid h=hat;
  run;
  quit;
  %ods_open;
  %create_gseg;
  %create_anova_mse(anova_out,byvar.);
  %create_anova_p(anova_out,byvar.,source.);
  %mult(anova_p,method,bytree.);
  %mult_plot(anova_p,method.,method.);
  %mult_select(anova_p,method.,method.,p_max.);
  %merge_macro(anova_resid anova_mse anova_p,method.,
               dataset..anova,byvar.);
  %delete_sets(anova_out anova_resid anova_mse anova_p);
%mend;
A3.2 Pairwise comparison

Language:
SAS Macro

Objective:
This program applies multiple analyses of variance with post-hoc pairwise comparisons and produces FDR adjusted p-values.

Reference:

Input Data:
The input file should contain all variables that are in the model.

Output:
1. File named ‘prwise_p_fdr’ containing FDR adjusted p-values for each pair comparison
2. Plots:
   - raw p-values
   - FDR adjusted p-values

Macro arguments:
- dataset: input file name
- dep: dependent variable
- class: list of class variables
- indep: list of independent variables
- int: list of interaction effects
- pair_var: variable for which pairwise comparisons should be done
- byvar: multiplicity variable
- bytree: family variable for hierarchical tree testing scheme
- method: p-value adjustment method (if option other than FDR is wanted)
- p_max: max. FDR adjusted p-value for selection
%macro pairwise_mult(dataset, dep, class, indep, int, pair_var, byvar, bytree, method, p_max);

%ods_close;
%by_st(&byvar.);
proc glm data=&dataset.;
%by_st.;
class &class.;
model &dep.=&class. &indep. &int.;
lsmeans &pair_var./pdiff;
ods output Diff=out_diff;
run;
quit;

%stack_macro(out_diff, raw_p, &byvar.);
%triangle_matrix(out_diff_stack, &byvar.);
%rename_dataset(out_diff_stack_tri, prwise_p);
%mult(prwise_p, &method., &bytree.);
%ods_open;
%mult_plot(prwise_p_&method., &method.);
%mult_select(prwise_p_&method., &method., &p_max.);
%delete_sets(out_diff_stack out_diff prwise_p
out_diff_stack_freqs);
%mend;

%mend;
A3.3 Identifying interaction

Language:
SAS Macro

Objective:
This program applies two-stage selection of interaction, where first stage filters by multiple analysis of variance and the second identifies interaction by standardized residuals. FDR control is either by subset approach, or hierarchical testing scheme approach. Options other than FDR available for multiple testing adjustment: hoc, bon, sid, hom (see SAS reference).

Reference:

Input Data:
The input file should contain all variables that are in the model.

Output:
1. Files:
   - anova_p_fdr: FDR adjusted p-values - ANOVA (1st stage)
   - anova_p_fdr_[p max]: FDR adjusted p-values - ANOVA (1st stage) with max specified p-value
   - int_p_fdr: FDR adjusted p-values - interaction (2nd stage)
   - int_p_fdr_[p max]: FDR adjusted p-values - interaction (2nd stage) with max specified p-value
   - [input file]_anova_int: original data with ANOVA parameters interaction parameters and p-values

2. Plots (for each stage):
   - raw p-values
   - FDR adjusted p-values

Macro arguments:
- dataset: input file name
- dep: dependent variable
- class: list of class variables
- indep: list of independent variables
- int: list of interaction effects
- source: effect variable for which p-values should be adjusted
- byvar: multiplicity variable
- bytree: family variable for hierarchical tree testing scheme
- method1: p-value adjustment method - ANOVA (1st stage)
- method2: p-value adjustment method - interactions (2nd stage)
- p_max1: max. FDR adjusted p-value for selection - ANOVA (1st stage)
- p_max2: max. FDR adjusted p-value for selection - interaction (2nd stage)
%macro anova_int(dataset, dep, class, indep, int, source, byvar, bytree, 
method1, method2, p_max_1, p_max_2);
    %create_gseg;
    %anova_mult(&dataset., &dep., &class., &indep., &int., &source., 
    &method1., &p_max_1.);
    %adj_name(&method1.);
    %column_act(&dataset._anova, rename, 
    'raw_p=raw_p_1 
    &adj_p._p=&adj_p._p_1');
    %compute_stat(&dataset._anova, resid, mean 
    n, &byvar. &class., 
    unique_result=no);
    data &dataset._anova;
    set &dataset._anova;
    if resid_N=1 then std=sqrt(((1-hat)*mse));
    else std=sqrt((0.5*(1-2*hat))*mse);
    test_stat=resid_mean/std;
    raw_p=2*(1-probnorm(abs(test_stat)));
    run;
    %sort_macro(&dataset._anova, &byvar. &class., int_p, 
    &adj_p._p_1 le &p_max_1. and 
    &adj_p._p_1 ne ., nodupkey);
    %column_act(int_p, keep, &byvar. &class. raw_p);
    %mult(int_p, &method2. &bytree.);
    %mult_plot(int_p &method2., &method2.);
    %mult_select(int_p &method2., &method2., &p_max_2.);
    %column_act(int_p &method2., rename, 
    'raw_p=raw_p_2 
    &adj_p=p&adj_p._p_2');
    %column_act(&dataset._anova, drop, raw_p);
    %merge_macro(&dataset._anova 
    int_p &method2., &dataset._anova_int, 
    &byvar. &class.);
    %delete_sets(int_p &dataset._anova);
%mend;
A3.4 Simple effect analysis of variance

Language:
SAS Macro

Objective:
This program applies multiple simple effect analyses of variance and produces FDR adjusted p-values. This type of ANOVA is appropriate when there is insufficient number of replication for each hypothesis. Options other than FDR available for multiple testing adjustment: hoc, bon, sid, hom (see SAS reference).

Reference:

Input Data:
The input file should contain all variables that are in the model.

Output:
1. Files:  
   - anova_p_fdr: FDR adjusted p-values
   - anova_p_fdr_[p max]: FDR adjusted p-values with max. specified p-value
   - [input file]_anova: original data with p-values
2. Plots:  
   - raw p-values
   - FDR adjusted p-values

Macro arguments:
- dataset: input file name
- dep: dependent variable
- class: list of class variables
- indep: list of independent variables
- int: list of interaction effects
- slice_var: independent variable for which categories effects should be calculated
- byvar: multiplicity variable
- bytree: family variable for hierarchical tree testing scheme
- method: p-value adjustment method
- p_max: max. FDR adjusted p-value for selection
%macro anova_mult_simple_effect(dataset,dep,class,indep,int,
     slice_var,by var,by tre e,m eth,od,p_max);

%create_gseg;
%by_st(&byvar.);
%if "&byvar." ne "" %then %do;
   %sort_macro(&dataset.,&byvar.);
%end;
%ods_open;
ods listing close;
ods html close;
proc glm data=&dataset.;
   %by_st.;
   class &class.;
   model &dep.= &indep. &int./ss3;
   lsmeans &int./slice=&slice_var.;
ods output slicedanova=anova_p;
run;
quit;
%ods_open;
data anova_p;
set anova_p(keep=&byvar. &slice_var. probf rename=(probf=raw_p));
run;
%mult(anova_p,&method,&bytree.);
%mult_plot(anova_p,&method.,&method.);
%mult_select(anova_p,&method.,&method.,&p_max.);
%merge_macro(&dataset.
   anova_p,&method.,&dataset._anova,
   &byvar. &slice_var.);
%delete_sets(anova_p);
%mend;
A3.5 Identifying correlation

Language:
SAS Macro

Objective:
This program applies correlation analysis to test correlation between a single variable and a group of other variable, and produces FDR adjusted p-values.
The single variable is assumed to be measured in replicates, where each replicate is a pool of several individuals.
Therefore, the FDR is estimated by repeated simulated analysis, where in each repetition this variable is randomly jittered according to an estimate of its variability.
The estimated variance is:
replicate variance * num. of individuals / num. of replicates

Options other than FDR available for multiple testing adjustment:
hoc, bon, sid, hom (see SAS reference).

Reference:

Input Data:
File containing all variables to be tested for correlation

Output:
1.Files:
   - cor_p_fdr: FDR adjusted p-values
   - cor_p_fdr_[max p]: FDR adjusted p-values with max. specified p-value

2.Plots (for each stage):
   - raw p-values
   - FDR adjusted p-values

Macro arguments:
- dataset: input file name
- var: jittered variable
- byvar: multiplicity variable
- bytree: family variable for hierarchical tree testing scheme
- jitter_var: name of variable after jittering
- var_est: variance estimate
- n_indiv: number of individuals
- n_rep: number of replicates
- method: p-value adjustment method
- p_max: max. FDR adjusted p-value for selection
%macro
sim_cor_jitter(dataset, var, byvar, bytree, jitter_var, var_est,
    n_indiv, n_rep, method, p_max);
    %compute_local_jitter(dataset, var, byvar, jitter_var,
        var_est, n_indiv, n_rep);
    %sim_cor(dataset, sim_num, cor_type, var, with_var,
        jitter_var, byvar, bytree, method, p_max);
%mend;
A4. Accompanying SAS macro programs (sorted alphabetically)

%macro adj_name(mult_method);
  %global adj_p;
  data _null_;  
  if substr(lowcase("&mult_method."),1,3) in ("fdr", "hoc", "bon", "sid", "hom") then do;
    call symput('adj_p',substr(lowcase("&mult_method."),1,3));
  end;
  else if lowcase("&mult_method." in ("holm", "stepbon") then do;
    call symput('adj_p','stpbon');
  end;
  else if lowcase("&mult_method." in ("stepsid") then do;
    call symput('adj_p','stpsid');
  end;
run;
%mend;

%macro by_st(byvar);
  %global by_st;
  %if "&byvar." ne "" %then %do;
    %let by_st=by &byvar.;
  %end;
  %else %do;
    %let by_st=;
  %end;
%mend;

%macro column_act(dataset,func,var_expr);
  /*if func=rename, var_expr is within ' '*/
  %symput_macro(final_var_expr,compress("&var_expr.","'"));
  data &dataset.;
  set &dataset.;
  &func.  &final_var_expr.;
run;
%mend;

%macro compute_local_jitter(dataset,var,byvar,jitter_var,var_est,n_indiv,n_rep);
  %compute_stat(&dataset.,&var.,&var_est.mean,&byvar.,,unique_result=no);
  %column_act(&dataset.,drop,&jitter_var.);
  %sort_macro(&dataset.,loc b sn,,nodupkey);
  %insert_var(&dataset.,local_jitter,&var._&var_est.*&n_indiv./&n_rep);
%mend;

%macro compute_raw_p_general(dataset,var,std);
  data &dataset.;
  set &dataset.;
  test_stat=&var./&std.;
  raw_p=2*(1-probnorm(abs(test_stat)));
run;
%mend;

%macro compute_stat(dataset,var,stat,byvar,where_condition,
  unique_result=yes,out=no);
  %words(&stat.);
  %let stat_num=&word_num.;
  %words(&var.);
  %let var_num=&word_num.;
  %local i j result_string;
  %do i=1 %to &stat_num.;
    %do j=1 %to 1;
      data _null_;  
      call symput('current_stat',upcase(scan("&stat.",&i.)))
      if 1;
      run;
      data _null_;  
      call symput('result_string',"&result_string."||"="||"&current_stat."||"=");
      run;
  %end;
%mend;
%do j=1 %to %var_num.;
data _null_; call symput('current_var',scan("&var.",&j.)); run;
%if "&byvar." ne "" and "&unique_result."="yes" %then %do;
data _null_; call symput('result_string','"&result_string."||"|compress("&current_var._")||compress(translate("&byvar._","_",""))||compress("&current_stat.")'); run;
%end;
%else %do;
data _null_; call symput('result_string','"&result_string."||"|compress("&current_var._&current_stat.")'); run;
%end;
%end;
%end;
%by_st(&byvar.);
%where_st(&where_condition.);
data _null_; call symput('outfile','"&dataset._means"'); run;
%if "&byvar." ne "" %then %do;
%sort_macro(&dataset.,&byvar.);
%end;
proc univariate data=&dataset. noprint;
%where_st.; %by_st.; var &var.; output out=&outfile. &result_string.; run;
%if "&out." ne "yes" %then %do;
%merge_macro(&dataset. &outfile.,&dataset.,&byvar.);
%delete_sets(&outfile.);
%end;
%end;
%macro cor_mult(dataset,cor_type,var,with_var,byvar,bytree ,method,p_max);
%by_st(&byvar.);
%if "&byvar." ne "" %then %do;
%sort_macro(&dataset.,&byvar.);
%end;
%ods_open;
ods listing close;
ods html close;
proc corr data=&dataset. nosimple spearman;
%by_st.; var &var.; with &with_var.; ods output kcor_type.Corr=cor_p; run;
quit;
%column_act(cor_p,rename,'p&var.=raw_p'); %mult(cor_p,&method.,&bytree.);
%mult_select(cor_p,&method.,&p_max.);
%delete_sets(cor_p);
%end;
%macro count_distinct(dataset,var);
%global value_num;
%freq_macro(&dataset.,&var.);
data _null_; set &dataset._freqs end=last; if last then call symput('value_num',_n_); run;
%macro create_anova_mse(dataset, byvar);
data anova_mse;
set &dataset.;
if compress(_type_)="ERROR";
mse=ss/df;
keep &byvar. mse;
run;
%mend;

%macro create_anova_p(dataset, byvar, source);
data anova_p;
set &dataset.;
if compress(_type_)="SS3" and _source_="&source.";
rename prob=raw_p;
keep &byvar. prob;
run;
%mend;

%macro create_value_list(dataset, var);
%global value_list;
%let value_list=;
%local i;
%count_distinct(&dataset.,&var.);
%freq_macro(&dataset.,&var.);
%do i=1 %to &value_num.;
data current_data;
set &dataset._freqs;
if _n_=&i.;
run;
data _null_;;
call symput('current_value',&var.);
run;
data _null_;;
call symput('value_list',"&value_list."||"|"||"|"&current_value."|"|");
run;
%put &value_list.;
%delete_sets(current_data);
%end;
%mend;

%macro define_gcat(cat_name);
proc catalog catalog=work.gseg;
copy out=work.&cat_name new;
run;
quit;
%delete_plots(&cat_name,_all_);
%mend;

%macro delete_plots(cat_name, entry_list);
proc greplay igout=&cat_name. nofs;
delete &entry_list;
run;
quit;
%mend;

%macro delete_sets(data_list);
proc datasets nolist nodetails;
delete &data_list.;
run;
quit;
%mend;

%macro freq_macro(dataset, var);
proc freq noprint data=&dataset. order=data;
tables &var./out=&dataset._freqs;
run;
%mend;

%macro insert_var(data_list, newvar, value, label, if_condition);
%if_st(&if_condition.);
%words(&data_list.);
%let data_num=&word_num.;
%local i current_data;
%do i=1 %to &data_num.;
data _null_;  
call symput('current_data',scan("&data_list.",&i.," "));  
run;  
data &current_data.;  
set &current_data.;  
%if_st.;  
&newvar.=&value.;  
%if "&label." ne "" %then %do;  
label &newvar.="&label.";  
%end;  
run;  
%end;  
%end;  
%end;  
%end;  

%macro merge_macro(data_list,merged_data,byvar);  
%insert_var(&data_list.,key,1);  
%sort_macro(&data_list.,&byvar. key);  
data &merged_data.;  
merge &data_list.;  
by &byvar. key;  
drop key;  
run;  
%column_act(&data_list.,drop,key);  
%end;  

%macro mult(dataset,method,byvar);  
%sort_macro(&dataset.,&byvar. raw_p);  
%by_st(&byvar.);  
odds listing close;  
proc multtest pdata=&dataset. noprint &method.  
out=&dataset._&method.;  
%by_st.;  
run;  
odds listing;  
%end;  

%macro mult_plot(dataset,method);  
%define_gcat(pv);  
%adj_name(&method.);  
%ods open;  
*** plot raw_p ***;  
%sort_macro(&dataset,raw_p,for_plot);  
data for_plot;  
set for_plot;  
if raw_p ne .;  
run;  
data for_plot;  
set for_plot end=last;  
rank=1;  
if _n_=1 then line=0;  
else if last then line=1;  
run;  
title;  
footnote;  
options ftext=SWISS ctext=BLACK htext=1 cells;  
axis1 width=1 offset=(3 pct) label=(a=90 r=0);  
axis2 width=1 offset=(3 pct);  
symbol1 c=RED ci=RED v=dot height=0.1 cells  
interpol=none l=1 w=1;  
symbol2 c=BLACK ci=BLACK v=dot height=0.1 cells  
interpol=join l=1 w=1;  
proc gplot data=for_plot gout=pv;  
plot (raw_p line)*rank/overlay  
name="raw_p"  
caxis = BLACK  
cframe = BLACK  
cf = white  
min = 0  
minor = 0  
xaxis = axis1  
haxis = axis2  
run;  
quit;  
options ftext= ctext= htext=;  
symbol1;
axis1;
*** plot adjusted_p ***;
%sort_macro(dataset.,&adj_p..p,for_plot,yes);
data for_plot;
set for_plot (keep=&adj_p..p);
rank=1;
run;
data tmp;
set for_plot;
if &adj_p..p ne 1;
run;
data tmp;
set tmp;
if _n_=1 then call symput('start_rank',rank);
run;
data _null_;
call symput('start_rank',&start_rank.-1);
run;
title;
footnote;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
axis1 width=1 offset=(3 pct) label=(a=90 r=0);
axis2 width=1 offset=(3 pct);
symbol1 c=RED ci=RED v=dot height=.1 cells
interpol=join l=1 w=1;
proc gplot data=for_plot gout=pv;
where rank ge &start_rank.;
plot &adj_p..p*rank/
  name="&adj_p..p"
  caxis = BLACK
  ctext = BLACK
  cframe = white
  hminor = 0
  vminor = 0
  vaxis = axis1
  haxis = axis2
; run;
quit;
goptions ftext= ctext= htext=;
symbol1;
axis1;
%delete_sets(for_plot tmp);
%mend;

%macro mult_select(dataset,method,p_max);
  %adj_name(method.);
  %symput_macro(p_max_char,translate("&p_max.","_","."))); 
  %subset(dataset..&p_max_char.,&dataset.,
  &adj_p..p le &p_max. and &adj_p..p ne .); 
%mend;

%macro ods_close;
  ods _all_ close;
  ods results off;
%mend;

%macro ods_open;
  filename tmp 'C:\Documents and Settings\anat\Local 
  Settings\Temp\SAS Temporary Files\';
  ods html file='odshtml-tmpbody.html' path=tmp;
  ods listing;
  ods results on;
%mend;

%macro rename_dataset(old_name,new_name);
  proc datasets nolist nodetails;
    change &old_name.=&new_name.;
  run;
%mend;

%macro sim_cor(dataset,sim_num,cor_type,var,with_var,
   jitter_var,byvar,bytree,method 
   ,p_max);
  proc printto log=dummy;
  run;

%macro ods_close;
  ods _all_ close;
  ods results off;
%mend;

%macro ods_open;
  filename tmp 'C:\Documents and Settings\anat\Local 
  Settings\Temp\SAS Temporary Files\';
  ods html file='odshtml-tmpbody.html' path=tmp;
  ods listing;
  ods results on;
%mend;

%macro rename_dataset(old_name,new_name);
  proc datasets nolist nodetails;
    change &old_name.=&new_name.;
  run;
%mend;

%macro sim_cor(dataset,sim_num,cor_type,var,with_var,
   jitter_var,byvar,bytree,method 
   ,p_max);
  proc printto log=dummy;
  run;
%let current_op=&op.;
options nomlogic nomprint nosymbolgen;
%subset(workfile,&dataset.);
%local i;
%do i=1 %to &sim_num.;
  %insert_var(workfile,&var._jitter,&var.+sqrt(&jitter_var.)*rand('NORMAL'));
  %cor_mult(workfile,&cor_type.,&var._jitter,&with_ var.,&byvar.,&bytree.,&method.,&p_max.);
  data p_new;
  set cor_p.&method.;
  test_stat&i.=quantile('NORMAL',1-(fdr_p/2));
  corr_coef&i.=&var._jitter;
  keep &byvar. variable test_stat&i. corr_coef&i.;
  run;
  %if &i. ne 1 %then %do;
    %merge_macro(p p_new,p,&byvar. variable);
  %end;
  %else %do;
    data p;
    set p_new;
    run;
  %end;
%dend;
data cor_p_final; set p;
  test_stat=mean(of test_stat1-test_stat&sim_num.);
  test_stat_std=std(of test_stat1-test_stat&sim_num.);
  fdr_p=2*(1-probnorm(abs(test_stat)));
  corr_coef=mean(of corr_coef1-corr_coef&sim_num.);
run;
%column_act(cor_p_final,keep,&byvar. variable fdr_p corr_coef);
%mult_select(cor_p_final,&method.,&p_max.);
%let op=&current_op.;
options &op.mlogic &op.mprint &op.symbolgen;
proc printto;
run;
%delete_sets(p p_new);
%mend;

%macro sort_macro(data_list,byvar,out_list,descend,where_condition,options);
%local i order;
%let out_list_final=&data_list.;
%if "&out_list." ne "" %then %do;
  data _null_;
  call symput('out_list_final',"&out_list.");
  run;
%end;
%let order=descending;
%where_st(&where_condition.);
%words(&data_list.);
%let data_num=&word_num.;
%doo i=1 %to &data_num.;
  data _null.;
  call symput('current_out',compress("out="||scan("&out__list_final.",&i.," ")));
  run;
  data _null.;
  call symput('current_data',scan("&data_list.",&i.," "));
  run;
proc sort data=&current_data. &current_out. &options.;
  by &order. &byvar. &where_st.;
run;
%mend;
%macro stack_macro(dataset,stack_var,byvar);
%local i;
data &dataset._stack;
set &dataset.;
tmp=compress("_"||rowname);
run;
%create_value_list(&dataset._stack,tmp);
data &dataset._stack;
set &dataset._stack;
keep &byvar. /*optional - by variable for the procedure that created the infile*/
_stack_ /*stacked value - new numeric variable that is now created*/
_source_ /*columns that disappear and their names are now stacked*/
rowname;
length _stack_ 8;
length _source_ $ 8;
array __ta__ {*} &value_list.;
do __ti__ = 1 to dim(__ta__);
_stack_ = __ta__{__ti__};
_source_ = scan("&value_list.",__ti__); 
output;
end;
run;
%column_act(&dataset._stack,rename,'_stack_=&stack_var._source_=colname');
%mend;
%macro subset(newset,oldset,if_condition);
%if_st(&if_condition.);
data &newset.;
set &oldset.;
%if_st.;
run;
%mend;
%macro symput_macro(var,var_def);
%global &var.;
data _null_;
call symput("&var.",&var_def.);
run;
%mend;
%macro triangle_matrix(dataset,byvar);
data &dataset._tri;
set &dataset.;
if substr(left(colname),2) ne left(rowname);
if substr(left(colname),2) lt left(rowname) then do;
small=substr(left(colname),2);
big=rowname;
end;
else do;
small=rowname;
big=substr(left(colname),2);
end;
comparision=compress(small||"_"||big);
run;
%sort_macro(&dataset._tri,&byvar. comparision);
%if "%byvar."="" %then %do;
%let condition=comparision=lag(comparision);
%end;
%else %do;
%let condition=(%byvar.=lag(%byvar.) and comparision=lag(comparision));
%end;
data &dataset._tri;
set &dataset._tri;
if %condition. then delete;
drop rowname colname small big;
run;
%mend;
%macro where_st(where_condition);
%global where_st;
%if &where_condition. ne "" then %do;
  %let where_st=where &where_condition.;
%end;
%else %do;
  %let where_st=
%end;
%mend;

%macro words(string,delim=%str( ));
  %global word_num;
  %local count word;
  %let count=1;
  %let word=%qscan(&string.,&count.,&delim.);
  %do %while(&word. ne);
    %let count=%eval(&count.+1);
    %let word=%qscan(&string.,&count.,&delim.);
  %end;
  %let word_num=%eval(&count.-1);
  %put word number is: &word_num.;
%mend;